Going viral

An integrated view on virological data analysis from basic research to clinical applications

Dissertation
zur Erlangung des Grades
"Doktor der Naturwissenschaften" (Dr. rer. nat.)
der Naturwissenschaftlich–Technischen Fakultät I
der Universität des Saarlandes
Saarbrücken
November 2013

Vorgelegt von
Sven-Eric Schelhorn

Tag des Kolloquiums 04.02.2014

Dekan der Fakultät .... Prof. Dr. Mark Groves

Vorsitzender des Prüfungsausschusses ...... Prof. Dr. Thorsten Herfet

Erstgutachter .... Prof. Dr. Thomas Lengauer

Zweitgutachter .... Prof. Dr. Hans-Peter Lenhof

Drittgutachter .... Prof. Dr. Gunnar Rätsch

Akademischer Beisitzer .... Dr. Marcel Schulz

Short abstract

Viruses are of considerable interest for several fields of life science research. The genomic richness of these entities, their environmental abundance, as well as their high adaptability and, potentially, pathogenicity make treatment of viral diseases challenging. This thesis proposes three novel contributions to antiviral research that each concern analysis procedures of high-throughput experimental genomics data. First, a sensitive approach for detecting viral genomes and transcripts in sequencing data of human cancers is presented that improves upon prior approaches by allowing detection of viral nucleotide sequences that consist of human-viral homologs or are diverged from known reference sequences. Second, a computational method for inferring physical protein contacts from experimental protein complex purification assays is put forward that allows statistically meaningful integration of multiple data sets and is able to infer protein contacts of transiently binding protein classes such as kinases and molecular chaperones. Third, an investigation of minute changes in viral genomic populations upon treatment of patients with the mutagen *ribavirin* is presented that first characterizes the mutagenic effect of this drug on the hepatitis C virus based on deep sequencing data.

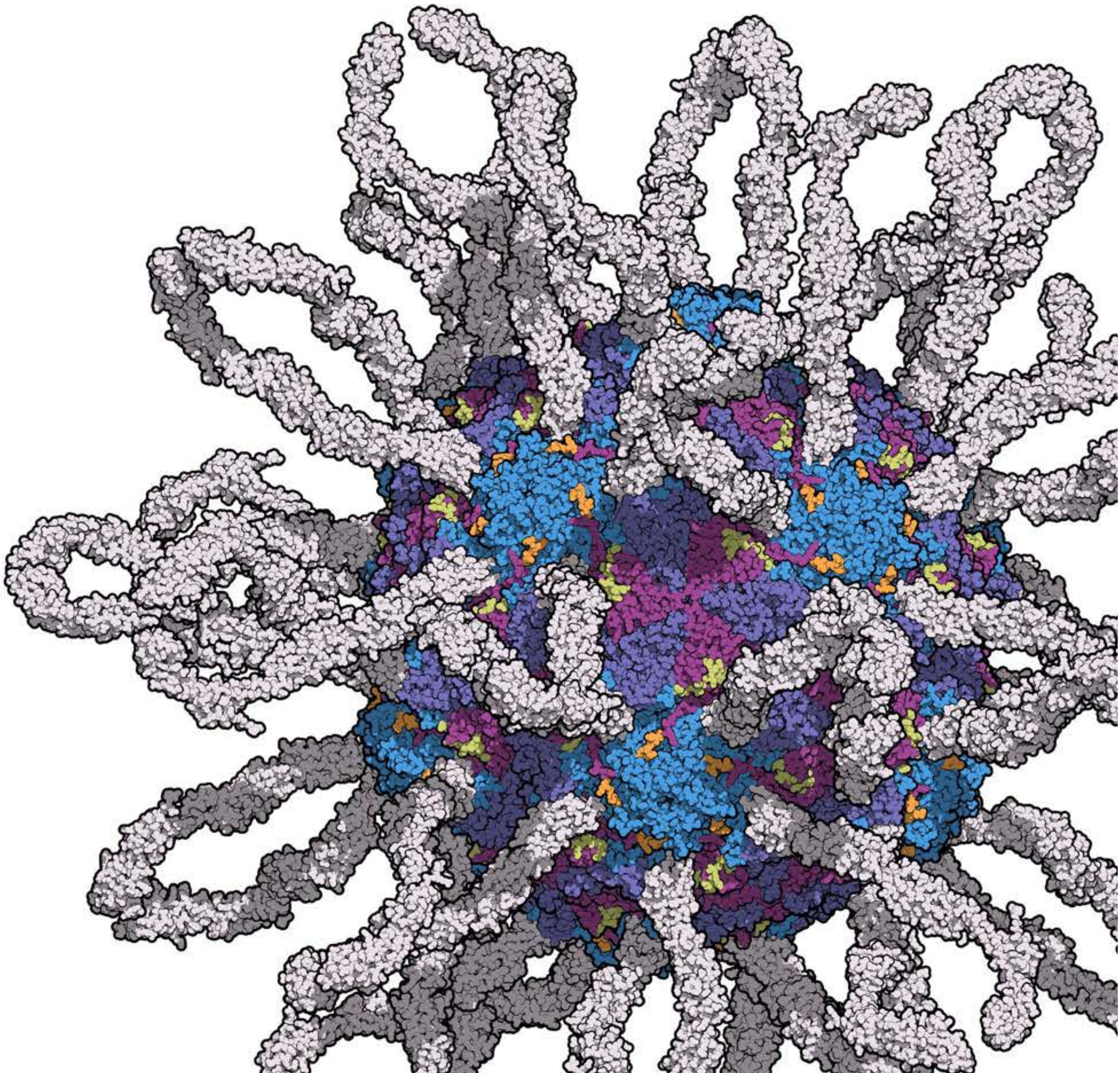
Kurzzusammenfassung

Viren sind von beträchtlichem Interesse für die biowissenschaftliche Forschung. Der genetische Reichtum, die hohe Vielfalt, wie auch die Anpassungsfähigkeit und mögliche Pathogenität dieser Organismen erschwert die Behandlung von viralen Erkrankungen. Diese Promotionsschrift enthält drei neuartige Beiträge zur antiviralen Forschung welche die Analyse von experimentellen Hochdurchsatzdaten der Genomik betreffen: erstens, ein sensitiver Ansatz zur Entdeckung viraler Genome und Transkripte in Sequenzdaten humaner Karzinome, der die Identifikation von viralen Nukleotidsequenzen ermöglicht, die von Referenzgenomen abweichen oder homolog zu humanen Faktoren sind. Zweitens, eine computergestützte Methode um physische Proteinkontakte von experimentellen Proteinkomplex-Purifikationsdaten abzuleiten welche die statistische Integration von mehreren Datensätzen erlaubt um insbesondere Proteinkontakte von flüchtig interagierenden Proteinklassen wie etwa Kinasen und Chaperonen aus den Daten ableiten zu können. Drittens, eine Untersuchung von kleinsten Änderungen viraler Genompopulationen während der Behandlung von Patienten mit dem Mutagen *ribavirin* die zum ersten Mal die mutagene Wirkung dieses Medikaments auf das Hepatitis C Virus mittels Tiefensequenzdaten nachweist.

SVEN-ERIC SCHELHORN

GOING VIRAL

DISSERTATION



AT MAX PLANCK SOCIETY AND SAARLAND UNIVERSITY

THESIS.SCHELHORN.ORG

Graphic models of virions and viral capsids displayed on the cover and at chapter headings were generated by the author based on public atomic protein structures deposited in the RCSB Protein Data Bank. Front cover, from top left to bottom right: 1EJ6, 1NOV, 1HB5, 4BX4, 1LD4, 2W0C, 2XVR, 3J0C, 3IYU, 4BTG, 3J26, 1Z7Z. Back cover: 3J27. Cover within the book proper: 1Z7Z. Chapter I: 1NOV. Chapter II: 3J0C. Chapter III: 1LD4. Chapter IV: 1EJ6. Chapter V: 3IYU. Colors within these models serve illustrative purposes only.

First printing, November 2013

難之。故終無難。
不爲大。故能成其大。夫輕諾必寡信。多易必多難。是以聖人猶
于其細。天下難事。必作于易。天下大事。必作于細。是以聖人終
爲無爲。事無事。味無味。大小多少。報怨以德。圖難于其易。爲大
第六十三章　恩始

Think of the small as large
  and the few as many.
Confront the difficult
  while it is still easy;
accomplish the great task
  by a series of small acts.

Daodejing Chapter 63 (excerpt).
Laozi, probably 6th century BC.

Abstract

Viruses are of considerable interest for several fields of bioscience research. The genomic richness of these entities, their environmental abundance, as well as their high adaptability and, potentially, pathogenicity make treatment of viral diseases challenging and antiviral research medically relevant. Especially if seen in a context of clinical reality that is dominated by chronic disease, drug resistance, and side effects, it becomes clear that an integrated view on virological data analysis spanning aspects from basic research to drug development and clinical applications is required in order to facilitate medical progress.

Such an integrated view should, at the least, encompass early detection of emergent human pathogenic viruses (*virus discovery*), support antiviral drug discovery (*drug target identification*), and aid in developing and optimizing clinical treatment of infectious diseases in order to to curb viral drug resistance and increase therapy success (*treatment optimization*). This treatise aims to provide contributions to all three of these areas.

Specifically, the thesis proposes three novel contributions to antiviral research that each concern analysis procedures of high-throughput experimental genomics data. First, a sensitive approach for detecting viral genomes and transcripts in sequencing data of human cancers is presented that improves upon prior approaches by allowing detection of viral nucleotide sequences that consist of human-viral homologs or are diverged from known reference sequences.

Second, a computational method for inferring physical protein contacts from experimental protein complex purification assays is put forward that allows statistically meaningful integration of multiple data sets and is able to infer protein contacts of transiently binding protein classes such as kinases and molecular chaperones.

Third, an investigation of minute changes in viral genomic populations upon treatment of patients with the mutagen *ribavirin* is presented that first characterizes the mutagenic effect of this drug on the hepatitis C virus based on deep sequencing data.

Acknowledgements

I would like to express my deep gratitude to Thomas Lengauer, Mario Albrecht, and Elena Zotenko for for their patient guidance, steady encouragement, and often useful critiques of the research work presented in this thesis. Thomas in particular has been a paragon for enthusiastic scholarship, superb management, and humane leadership. I hope to have learned by his example and feel privileged to have spent these years in his group.

I would like to offer my special thanks to Gunnar Rätsch and Hans-Peter Lenhof who have agreed to attend the examining board at short notice. Gratitude is also extended to the staff of the MPII, particularly Ruth, Achim, and Georg, without whom my work would have run much less smoothly than it did.

My great appreciation also goes to the external cooperation partners and scholars I had the good luck to converse with over the last years; while certainly monomaniacs, personages like Harald zur Hausen, Eugene Myers, and Oliver Smithies have made lasting impressions on me.

I am grateful for the friendship of the many graduate and undergraduate students I have worked, eaten, and discussed with in the recent years. Hagen in particular has taught me much in his seemingly reticent way; I believe I would have been another person without him. I greatly appreciate the many insightful and pleasant conversations with Jasmina, Laura, Kasia, Konstantin, Basti, Glenn, Ingolf, Andi, and Fabian – you are great. I also regret not having spent more time with the "old guard" before they left: Tobi, Andreas, and Olli. Perhaps we can rectify this over the years.

My parents Dr. Ulf Schelhorn and Silke Petersen-Schelhorn have my deep gratitude for their support and encouragement (both emotionally and financially) that allowed me to pursue my studies.

Finally to my beautiful and loving fiancé, Christine: thank you. Together we have weathered many a storm and it is to you and our wonderful daughter Marlene Sophie that I dedicate this work.

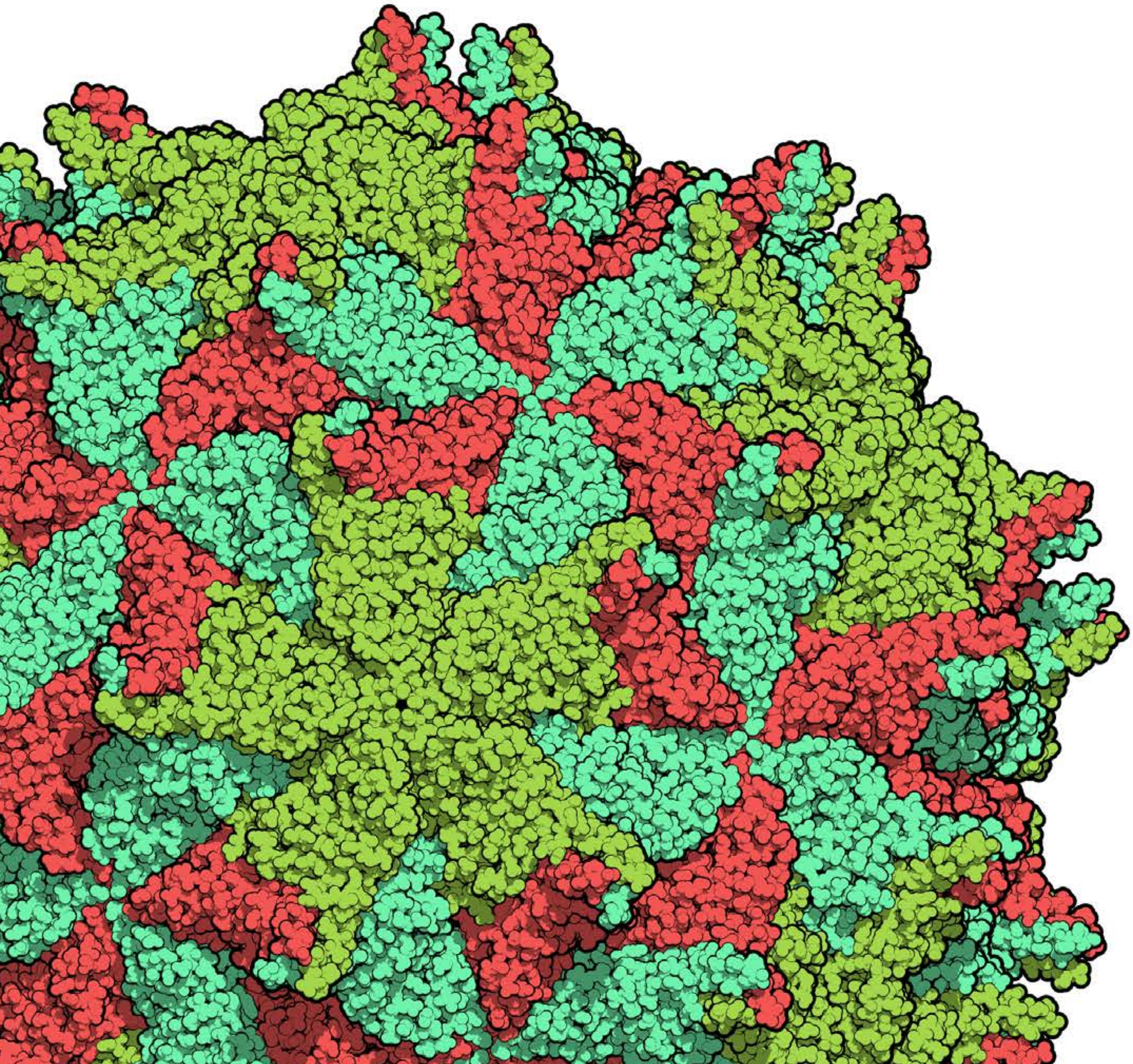
Contents

List of Figures

List of Tables

I

Introduction



1 Preface

Viruses are of considerable *interest for several fields of bioscience research such as genomics, epidemiology, medicine, bioinformatics, as well as, naturally, virology. The genomic richness of these entities, their environmental abundance, as well as their high adaptability and, potentially, pathogenicity make viral research both relevant and diverse. In addition, and unbeknownst to many, viruses are not only subjects of research but also essential tools of microbiology; indeed, viral proteins such as polymerases are now standard components in molecular laboratories and the study of viruses resulted in serendipitous moments such as the discovery of cellular oncogenes, a concept of crucial importance for oncology. In a similar manner, many viral innovations found their way into genomes of cellular organisms, including human: so are important factors of placenta formation of viral origin, and viruses may even have been instrumental to the emergence of cellular life as we know it.*

Last, and most obviously, viruses are important pathogens that, while representing the minority of microbial organisms infecting humans, are the major source of newly emerging pathogens. Both the SARS and the swine flu outbreaks may serve as current examples. The genomic diversity as well as the astonishingly high adaptability of viruses make treatment of both known and emergent viral diseases challenging. Especially if seen in a context of clinical reality that is dominated by chronic disease, drug resistance, and side effects, it becomes clear that an integrated view on virological data analysis spanning aspects from basic research to drug development and clinical applications is required in order to facilitate medical progress. It is the opinion of the author that such an integrated view should, at the least, (1) encompass early detection of emergent human pathogenic viruses, (2) support antiviral drug target discovery, and (3) aid in developing and optimizing clinical treatment of infectious diseases. This treatise aims to provide contributions to all three of these areas.

Contributions

This thesis is organized along three major themes that may aid the reader in connecting the diverse research results presented herein into a coherent whole. First among these themes is the aforementioned integrated view on virological data analysis consisting of (1) *virus discovery* to support sensitive detection of novel human pathogens, (2) *drug target identification* to facilitate development of novel antivirals, and last, (3), *treatment optimization* to curb emergence of viral resistance and increase therapy success. Besides this main theme, a second recurring *leitmotif* is the use of particularly sensitive statistical methods that aid in separating faint biological signals from technical noise; such methods are critical components of all novel approaches introduced in this thesis. Last, the third and most speculative theme of this treatise concerns the subtle similarities of viral infection disease to another class of heterogeneous and adaptive disorders: heterogeneous human cancers. While we will revisit this last theme only at the very end, the reader is encouraged to keep it in mind as an ulterior motive, a *Hintergedanken*.

Apart from these overarching themes, this thesis proposes three novel and specific contributions to antiviral research that each concern analysis procedures of high-throughput experimental genomics data.[1] First, a sensitive approach for detecting viral genomes and transcripts in mixed sequencing data of human cancers (presented at the end of Chapter II). Second, a statistical method for inferring physical protein contacts from experimental protein purification assays (presented at the end of Chapter III). Third, an investigation of minute changes in viral population structures upon treatment of patients with mutagenic drugs (presented at the end of Chapter IV). All these contributions represent published, first-author works of the author as referenced in the introduction of each manuscript as well as in the list of List of publications.

While the chapters of this thesis quite naturally fall into place when considered from the proposed, integrated viewpoint of virological data analysis, this orderly view is not how the manuscripts that compose the core of this work have been originally conceived. Instead, the chronology of these manuscripts follows the order (and, perhaps, the scientific development) of the author as he ventured from *systems biology* (Chapter III) to *clinical virology* (Chapter IV) and further towards *high-throughput genomics and oncology* (Chapter II).

It is only in hindsight that the three aforementioned themes connect these manuscripts in a favorable manner. The reader is therefore kindly asked to indulge the author at his attempts to form a coherent whole of the disparate parts. If this works seems to delude the reader into thinking that basic research were a predictable process, it is to convey the material in a more concise manner rather than to belie its less structured origins.

[1] Even the experimental protein purification data that is analyzed in the second contribution is, in its essence, of genetic origin since open reading frames of the investigated proteins form the principle input data.

Limitations

There are several topics of research that would be a natural fit to the aforementioned integrated view on virological data analysis but that are intentionally omitted from this thesis either for the sake of increased focus, or because the author does not feel sufficiently competent in discussing them here.

Among the prior group of topics are presentations of the various software systems that the author developed for conducting the analyses presented in this thesis. One such system in particular, *Virana*, has been used extensively for preparation and analysis of viral next-generation sequencing data and can be considered the major technical contribution of the author. Nevertheless, since a detailed discussion of this software system does not significantly advance the main aim of this thesis, i.e., the understanding and treatment of viral disease, it is left out of this treatise.

Similarly, prior work of the author regarding the estimation of interacting protein regions and investigation of protein interaction networks(2) is omitted here. In particular, an unpublished pilot study conducted in cooperation with Elena Zotenko that proposed a novel optimization procedure for deducing physical protein contacts based on protein interaction networks is left out: although biologically insightful and methodologically interesting, the publication at the end of Chapter III seemed to be more concise and the method less easily confounded by incomplete protein domain annotations.

Last, although Chapter IV presents approaches to modeling viral evolution and treatment response in patients *in vivo*, these results predominantly relate to basic rather than to applied research. Applications of these results, for example within rational approaches to therapy optimization as brought forward by Altmann et al. 2009, Beerenwinkel et al. 2003, Roomp et al. 2006, Sing et al. 2005 and others are not explicitly treated here. While the author gained expertise in this field during his PhD studies as demonstrated by the development of geno2pheno[hcv],(3) he felt that further consideration of this vast and interesting subject would extend this thesis even further. Instead, the interested reader is encouraged to consider rational therapy optimization as a continuation of Chapter IV of this thesis and review the aforementioned citations in order to obtain an overview of this field.

Among the second group of topics that are left out due to limited competence of the author are the large fields of target validation, combinatorial drug discovery, and rational drug design. It is quite obvious from the structure of this thesis that, while antiviral drug targets relating to both viral and host factors are among the potential results of the protein interaction analysis featured in Chapter III, these targets are neither validated nor further utilized here in order to propose novel antivirals. Instead, the existence

(2) Assenov et al. (2008), Blankenburg et al. (2009), Schelhorn et al. (2008)

(3) The web service geno2pheno[hcv] provides decision support for clinicians and virologist in the context of anti-HCV precision medicine. It has been developed by Sven-Eric Schelhorn and Michael Zeidler and is now continued by Bastian Beggel. It is widely used by the HCV community and more than 20,000 genotype samples have been analyzed by the service since May 2011. See http://hcv.bioinf.mpi-inf.mpg.de/

of safe and effective compounds with antiviral activity is already taken as a given in Chapter IV where the clinical effects of such drugs on the viral population are discussed.

Nevertheless, in order to supplement the interested reader with additional material in this regard, a short list of references may be of use. A bird's eye view on the antiviral drug development process is given in two multi-volume works edited by Thomas Lengauer.[4] For introductions to principles of drug discovery the reader may wish to review Hughes et al. 2011, Pauwels 2006. Several works discuss virtual and experimental screening approaches for lead discovery and optimization.[5] A special focus of future antivirals will likely be put on biologicals, e.g., protein-like compounds such as antibodies, vaccines, as well as on drugs targeting host factors.[6] These and related future directions of antiviral research are extensively discussed by Erik de Clercq and others.[7]

Outline

This thesis is divided into five chapters. While the the middle three chapters from the core of the treatise and discuss research of the author, the leading and closing chapters feature introductions to viruses and and concluding remarks, respectively.

The introductory Chapter I reviews general concepts of viruses such as their physical and genomic characteristics before aiming to familiarize the reader with theories on the evolutionary origin of viruses and their relation to the three domains of cellular life. The tight relation of cellular and viral life is further supported by recent results on ancient proviral integration events in the human germline. Subsequently, estimated trends of global disease burdens of viral infections will be discussed before concluding this chapter with an investigation of worldwide viral abundance and the zoonotic emergence of viral disease.

Chapter II features the first contribution of the author to the field of viral research. After an introduction to metagenomics and the burgeoning field of next-generation sequencing, aspects of computational analyses of mixed sequencing data are discussed. Subsequently, a conceptual interlude provides a review on the history, virology, and epidemiology of tumor viruses. The chapter concludes with an investigation of low-frequency viral factors in metagenomes and -transcriptomes of human tumors, a study that contributes to both viral surveillance and the identification of novel causes of cancer.

Chapter III shifts the focus from genomics to systems biology and presents the second major contribution of this thesis. The chapter sets out by introducing the concept of protein interactions and reviews experimental and computational assays to measure these interactions. Subsequently, the biology of viral host factors is discussed and approaches for their determination are explicated. The

[4] Lengauer (2007), Lengauer et al. (2002)

[5] de Clercq (2011), Hopkins and Bickerton (2010), Jhoti and Leach (2007), Kirchmair et al. (2011), Kuntz (1992), Lengauer et al. (2004), Menéndez-Arias and Gago (2012), Real et al. (2004)

[6] Burton (2002), Fox (2007), Marasco and Sui (2007), Sawyer (2000), Tan et al. (2007)

[7] de Clercq (2007), Flexner (2007), Wainberg (2008)

importance of these host factors is further highlighted by a description of antiviral drugs that are in current use or in development, as well as a first introduction to the concept of antiviral drug resistance. The chapter concludes with a contribution for statistical inference of physical protein contacts, an approach that may aid in the identification of novel antiviral drug targets.

The last core section of this thesis, Chapter IV, deals with clinical applications concerning the determination of effects of antiviral drug treatment. After introducing concepts of viral population structures and intra-host viral adaptation, current approaches for treating infections with highly divergent viruses are discussed. These concepts are applied to the therapeutic context of one particular virus, Hepatitis C (HCV), for which current as well as future treatment options are reviewed. Last, the modes of action of a class of antivirals, mutagens, are introduced before the chapter concludes with an experimental investigation of the effects of a specific mutagenic drug, ribavirin, on the HCV population structure.

Finally, in Chapter V, this thesis is brought to a conclusion in the context of the aforementioned integrated view on virological data analysis. In addition, and by way of expanding on thoughts presented earlier, a wider view on subclonal diseases is offered that encompasses heterogeneous human cancers. The thesis finishes with predictions concerning the future development of the field from which specific recommendations for further research may be derived.

The core chapters of this thesis are relatively independent from each other. While each of the research items presented can be arranged into the aforementioned integrated view on virological data analysis, these items relate to diverse fields of bioinformatics (i.e., *metagenomics*, *systems biology*, and *viral population genomics*) that have only low methodological overlap. It is due to this diversity that the thesis, while cumulative in nature, is relatively explicative in order to best introduce and relate each of the research items. The author hopes that the reader finds the breadth of the material pleasantly extensive rather than tedious.

Last, it is the modest hope of the author that this work may convey some of the fascination he feels for viruses in their twofold role as both pathogens and explanatory devices of Nature. By way of extending a related quote printed in the forematter of this thesis: *"If you cannot do great things, do small things in a great way." – Napoleon Hill.*

Despite its volume, the author aimed at making this thesis such a small thing.

2 *Viral genomes and structures*

THIS THESIS BEGINS *with a presentation of the historic and current conceptualizations of its main protagonists, viruses. After reviewing their physical characteristics, we focus on the question if viruses should be considered living entities. Based on these insights, this section presents on overview of modes of viral replication before concluding with a review of viral taxonomic classification. While this chapter may, in principle, be skipped by readers with a background in virology, the sections on viral taxonomy and modes of replications have some bearing on Chapter II.*

Physical characteristics

At the end of the 19th century, discoveries by Dmitri Ivanovsky, Martinus Beijerinck, and Friedrich Loeffler of non-bacterial pathogens heralded the beginning of a new biological science, virology.[1] Viruses, later shown to depend on living tissue for replication, were soon recognized as important complements of the germ theory of disease as promoted by Louis Pasteur and others. Historic definitions of viruses define these entities as

> *"strictly intracellular and potentially pathogenic entities with an infectious phase, and (1) possessing only one type of nucleic acid, (2) multiplying in the form of their genetic material, (3) unable to grow and to undergo binary fission, (4) devoid of a Lipmann [energy and carbon metabolic] system."*[2]

Today, viruses are predominantly seen as a class of obligate intracellular parasites, invasive biological agents that infect and reproduce inside the cells of every known organism. In addition, viruses function as important scientific discovery tools for various disciplines of biology,[3] thereby further emphasizing the ancient Latin meaning of the term *virus* that incorporates aspects of both destruction and creation.[4]

Physically, known viruses are often conceptualized as *virions* , minuscule containers with physical dimensions of 20–200nm (about 1/100th of the size of a bacterium) that consist of RNA or DNA genomic material, supplementary proteins, and a protein shell (the *capsid*) that may additionally be surrounded by a lipid membrane, the *envelope*. Capsids come in a variety of shapes that reflect the particular selective pressure the virus is subjected to[5]; apart from the usual icosahedral configurations, some of these shapes are

[1] Levine and Enquist (2007)

[2] Lwoff (1957)

[3] Rozenblatt-Rosen et al. (2012)

[4] *Virus*: the German, and therefore English, origin of the word is Latin although it may have Indo-European roots older than the Latin language. Due to the relatively high lexicographic parsimony of Latin, the word is semantically overloaded, i.e., polysemic and may have meant *slimy liquid, bitter juice, poison, snake venom, stinking odor, malignant personality in a person, therapeutic salve, magical secretion, animal sperm*, and, in later forms of Latin, *human semen*. Obviously, these meanings have not been included in the English, or for that matter, German languages where the term virus is defined as *"any of a large group of submicroscopic infective agents that are regarded either as the smallest microorganisms or extremely complex molecules and are composed typically of a protein coat surrounding an RNA or DNA core of genetic material, that are capable of growth and multiplication only in living cells, and that cause various important diseases in man, animals, or plants (...)"* from: Webster's Third New International Dictionary, Unabridged (Merriam-Webster, 2002).

[5] Bruinsma et al. (2003), Grayson et al. (2006), Zandi et al. (2004)

bewildering and resemble spindles, lemons, contain terminal hooks or claws, and are bottle-shaped or fully linear.[6]

After successful entry into a susceptible host cell (see Section 14), viruses set up complex replication machinery (the 'virus factory') that employs host components to replicate viral genomes and produce new viral progeny. Due to their obligate parasitic nature, viruses cannot procreate without the host cell. Viral replication usually terminates lytically by bursting the cell after large numbers of viral progeny have been created. This process usually kills the cell by exocytosis or generation of a large number of viral antigens that lead to activation of innate and adaptive arms of the immune system and thus induce immune-mediated cell death (see Section 14). Alternatively, infection can occur chronically by limited production and budding of viral progeny that does not lyse the cell, or by a process denoted as lysogeny or *latency* where the viral genome becomes part of the hosts genetic material by either episomal or proviral mechanisms (see Section 10). In the latter case, lytic reproduction is entered at a later time of the viral life cycle.

While traditionally the nucleotide-containing virion is considered to be the definitive viral entity, it was recently proposed that the virus is not the infective viral particle but the established virus factory within the infected cell.[7] This contested[8] definition seems to put viral behavior, arguably a more specific and comprehensive view of this entity, into the limelight while demoting structural aspects of the capsid to a mere storage form, very much like a bacterial or fungal spore. This view puts the virus on equal footing with intracellular bacterial parasites which exhibit life styles that are at least in some regards comparable with that of viruses.

However, even the more general definition of viruses as small, protein-containing obligate intracellular parasites does not seem to cover well all observed viral entities. The recent identification of giant viruses, as well as non-viral sub-cellular pathogens like viroids, RNA satellites, protein-based prions, and retro-elements, all of which exhibit virus-like behavior but are missing many viral attributes blur the lines between viral and non-viral entities.[9] While non-viral 'selfish' genetic elements have recently been denoted as 'orphan replicons' and were suggested to be part of an universal tree of life together with viruses,[10] a well-defined notion of viruses remains controversial and viruses are now considered to be an intrinsically ill-defined taxonomic group.[11]

[6] Brussow (2009)

[7] Claverie (2006)

[8] Moreira and López-García (2009)

[9] Alper et al. (1967), La Scola et al. (2008), Raoult et al. (2004), Sanger et al. (1976)

[10] Raoult and Forterre (2008)

[11] Koonin et al. (2006), Moreira and López-García (2009)

Viruses as living systems

The definition of life has long been the subject of both biological and philosophical discussions.[12] Several notions of life have been proposed taking thermodynamic, computational, evolutionary, biochemical, and network-centric views.[13] The status of viruses in this regard has been argued since their discovery; indeed, viruses were first considered living entities and present-day successors of primitive genes in the primordial world, possibly predating cellular life.[14] Viruses are near in complexity to the single gene, arguably the simplest entity that ever has been suspected to be living based on scientific arguments.[15] As a consequence of their simplicity, neither "nucleocentric" (the genetic material was first) nor "cytoplasmist" (the metabolism was first) theories on the origin of life are unequivocally applicable to viruses and a "virocentric" synthesis of these theories was therefore proposed.[16] Historically, the notion of viruses as living entities was most prominently represented by J. B. S. Haldane who besides pursuing influential medical genetic research and terming the modern word 'clone', was one of the first who stated that

> *"life may have remained in the virus stage for many millions of years before a suitable assemblage of elementary units was brought together in the first cell."*[17]

(before revoking this opinion later).[18] Later arguments state that viruses miss life-like properties such as self-organization, self-maintenance, common ancestry, energy metabolism, or evolution independent of cells; thus, viruses are now considered non-living entities by the majority of experts in the field.[19]

However, new arguments regarding the structural similarity of viral capsids, theories on viral invention of the DNA replication machinery, and the discovery of giant viruses containing cellular genes, have recently renewed the debate about the status of viruses as living beings.[20] In particular, initial arguments stating that viruses miss important features of living systems such as self-maintenance, self-replication, common ancestry (or, at least no detectable common ancestry due to extensive gene transfer), structural continuity across generations, as well as metabolic and protein-synthesis related genes have sparked lively debates.[21] Some of the opinions thus expressed judge the discussion as so useless as to

> *"propose an eleventh reason to exclude viruses from the tree of life: there is no such thing as a tree of life"*[22]

– at least if applied to species instead of single genes. Indeed, alternative notions of a forest of life connected by extensive gene transfer (cf. Section 3 of this chapter) should be well considered.[23]

In conclusion, the controversy about the definition of life may well be termed unscientific from the beginning (as quoted by K. R. Popper)[24] and thus the question of the status of viruses in this regard should perhaps be pursued with a certain lightheartedness rather than in a strictly scientific manner.

[12] Luisi (1998)

[13] "*system that can maintain itself in a state far from equilibrium, and that can grow and multiply with the help of a continual flow of energy and matter from the environment*", C. de Duve (DeDuve, 1991); "*self-reproducing automata*", J. von Neumann (von Neumann, 1966); "*a self-sustained chemical system capable of undergoing Darwinian evolution*", G. Joyce (Deamer and Fleischaker, 1994); "*the operation of proliferating, programme-controlled fluid chemical automatons*", T. Ganti, (Gánti, 2003); "*autopoietic system with a network of processes of production (synthesis and destruction) of components such that the components continuously regenerate and realize the network that produces them and constitute the system as a distinguishable unity in the domain in which they exist*", F. J. Maturana (Varela et al., 1974)

[14] Haldane (1929), Muller (1929), Simon (1923)

[15] Muller (1929)

[16] Moreira and López-García (2009), Podolsky (1996)

[17] Haldane (1929); the statement may seem paradoxical given the obligate cellular parasitism of viruses today. However, assuming the "*existence of a complex, precellular, compartmentalized but extensively mixing and recombining pool of genes*" in the primordial RNA world as it is still hypothesized today (Koonin et al., 2006), Haldane's statement seems less implausible.

[18] Oparin (1961)

[19] Moreira and López-García (2009), van Regenmortel (2000, 2008)

[20] Brussow (2009), Hegde et al. (2009), Ludmir and Enquist (2009), Moreira and López-García (2009), Raoult (2009)

[21] Hegde et al. (2009), Koonin et al. (2009), Ludmir and Enquist (2009)

[22] Raoult (2009)

[23] Dawkins (2006), Doolittle and Bapteste (2007), Koonin (2009)

[24] Popper (2002)

Viral taxonomy and classification

Viruses are commonly named in semi-structured fashion, either for the disease of which they are causative (e.g., *polio* virus), the organ or tissue they are affecting (e.g., *Hepatitis* viruses from Greek *hepar*, liver, or *Rhino* viruses, from the Greek word for nose), their physical characteristics (e.g., *rota* virus, Latin for *wheel*), the place of their first isolation (e.g., *Marburg* virus, *Coxsackie* virus), the researcher who discovered them (e.g., *Ebstein-Barr* virus), or names associated with folklore (e.g., Influenza, Dengue). As discussed previously, from a structural point of view viral particles consist of a membrane-enveloped or naked protein capsid that surrounds the viral genome and often displays a variety of protein receptors. Viral capsids have a variety of functions, most of them concerning protection of the viral genome from external influence, while others are regarding viral entry into the host cell, genome packaging, and assembly of new virions.[25] Capsids commonly consist of tens to many thousands of protein subunits that form macromolecular assemblies predominantly spherical (more specifically, icosahedral) or helical symmetry.[26]

[25] Rux and Burnett (1998)

[26] Carrillo-Tripp et al. (2009), Crick and Watson (1956)

Table 2.1: *Baltimore classification of viruses.* Classes I–VI were originally proposed by David Baltimore (Baltimore, 1971) and later extended to also include Hepadnaviruses. Examples for specific viruses were added by the author.

Group	Genome	Orientation	Viruses
I	DNA	+/- double stranded	Bacteriophages, Adenoviruses, Herpesviruses, Polyomavirus
II	DNA	+ single stranded	Parvoviruses, Torque teno virus
III	RNA	+/- double stranded, segmented	Rotavirus
IV	RNA	- single stranded	Coronavirus, Rubellavirus, Rhinovirus, Poliovirus, Hepatitic C virus
V	RNA	+ single stranded	Coronavirus, Influenzavirus A/B, Rabies virus, Ebola virus, Hanta virus
VI	RNA	+ singe stranded, RT	HIV-1/2, HTLV-1
VII	DNA	+/- partially double stranded	Hepadnavirus

After determination of the first protein structure, the capsid of the tobacco mosaic virus,[27] and followed by visualizations of viral particles by electron microscopy in the 1960s, a hierarchical system consisting of phylum, class, order, family, genus and species was proposed to classify viruses based mostly on structural aspect of their capsid. Classification was undertaken based on size, symmetry, presence of an envelope, and number of capsomers, and only secondary by the nature of the nucleic acid the capsid contains (namely, type of acid, segmentation, linearity, ploidy, strandedness, and coding orientation).[28]

However, these taxonomical classification based primarily on structural aspects of viral capsid proteins were soon shown to be flawed; as these proteins have evolved independently at multiple occasions during evolutionary history,[29] viruses may utilize very different life cycles even if sharing a capsid protein.[30] Following advances in nucleotide sequencing in 1970s it became clear that viruses are polyphyletic[31],[32] and thus cannot be taxonomically organized based on their genetic divergence from a single common

[27] Watson (1954)

[28] Lwoff et al. (1962)

[29] Bamford et al. (2005)

[30] Krupovic and Bamford (2009)

[31] *polyphyletic*: Greek for "many races". The term designates organisms that share phenotypic or genetic traits that do not stem from a common ancestor. Phyletic placement of viruses is inherently difficult due to their high genomic variability and extensive gene transfer between hosts, pehnomena which may mask patterns of common descent.

[32] Koonin et al. (2006), Rohwer and Edwards (2002)

ancestor.[33] Additionally, and in contrast to organisms within the three accepted domains of life, viruses in general do not possess a common nucleic acid such as rRNA based on which genetic divergence can be estimated. In addition, newly originating viruses are often a result of considerable horizontal gene transfer between viruses or between viruses and their hosts, thus further complicating the determination of evolutionary descent.[34]

Due to the lack of common evolutionary descent, viruses are therefore often not classified based on intrinsic attributes of their genes but rather by mechanistic properties pertaining to their mode of nucleic acid replication. Even though viruses are obligate cellular parasites and do not contain or encode for a complete replication machinery, they follow the general paradigm of microbiology (DNA makes RNA makes protein) in the sense that all viruses have to produce mRNA from their genome on order to exploit the cellular machinery for viral protein synthesis. Following the highly influential Baltimore scheme of viral classification,[35] all viruses are thus grouped into originally six (later extended to seven, designated as I-VII) classes based on the general microbiological pathway that produces mRNA from the viral genome (see Tables 2.1 and 2.2). Regardless of their exact mode of replication, critical aspects of viral life cycles become readily apparent once their mode of mRNA synthesis has been identified. As a consequence of the usefulness of this manner of classification, today both structural aspects and the Baltimore scheme are employed for assigning viruses to unified taxonomies by the International Committee on the Taxonomy of Viruses[36] and the National Center for Biotechnology Information, NCBI.[37]

Group	Replication
I	DNA-dependent DNA polymerase copy both strands to produce a dsDNA viral genome. DNA-dependent RNA polymerase copies the - strand into a + strand mRNA for translation.
II	DNA-dependent DNA polymerase copies the + strand to produce a dsDNA intermediate. DNA-dependent DNA polymerase copy the - DNA strand into ss + strand DNA genome. DNA-dependent RNA polymerase enzymes copy the - strand DNA into + strand mRNA.
III	RNA-dependent RNA polymerase copies both strands to generate a viral genome copy. RNA-dependent RNA polymerase copies the - RNA strand into + viral mRNA.
IV	RNA-dependent RNA polymerase enzymes copy the - RNA genome to a - strand RNA genome via a + strand RNA intermediary. This genome copy also functions as mRNA.
V	RNA polymerase enzymes copy the + RNA genome to a + strand RNA genome via a - strand RNA intermediary. RNA-dependent RNA polymerase enzymes copy the (-) RNA strand into (+) viral mRNA
VI	RNA-dependent DNA polymerases copies the + RNA genome into - DNA strand. DNA-dependent DNA polymerase copies the - DNA strand into a dsDNA intermediate. DNA-dependent RNA polymerase copies the - DNA strand to produce the + RNA genome. DNA-dependent RNA polymerase copies the - DNA strand into + viral mRNA.
VII	Partially double stranded DNA is completed by DNA-dependent DNA polymerase and circularized. DNA-dependent RNA polymerase produced + mRNA from the circularized DNA. The partially double stranded genome is produced by RNA-dependent DNA polymerase and DNA-dependent DNA polymerase from the + mRNA.

Table 2.2: *Viral replication strategies.* Process of viral replication depending on Baltimore groups (Roman numerals, cf. Table 2.1).

By 2012, the ICTV (International Committee on Taxonomy of Viruses) , the institution tasked to develop, refine, and maintain a universal virus taxonomy, recognized 96 viral taxonomic *families*,

[33] Only particular viral families contain highly homologus genes such as reverse transcriptase or other polymerases and can thus be phylogenetically analyzed

[34] Krupovic and Bamford (2007), Pedulla et al. (2003)

[35] Baltimore (1971)

[36] Buechen-Osmond and Dallwitz (1996)

[37] Sayers et al. (2012)

i.e., the major taxonomic category employed for differentiating and categorizing viral life cycles.[38] These families encompass 2,618 viral species which, in turn, are subdivided into additional strains and genotypes. Many of the latter are represented in the NCBI viral RefSeq sequence database which currently contains more than 3,000 entries.[39] In contrast to the three established domains of life that usually have a well-defined notion of the term ’species’, the definition of this term is not straightforward for viruses.[40] This is a result of the fact that viruses are often divergent and only few exemplars, termed ’isolates’,[41] of any closely related ensemble of viruses (as, for example, genomes of viruses infecting a single host) are usually known. Critically, isolates are not directly transferable to the species concept; rather, multiple related isolates are usually assigned to a single species.[42] The ICTV proposal of

> *"A virus species is a polythetic class of viruses that constitutes a replicating lineage and occupies a particular ecological niche"*[43]

is highly flexible and transfers the responsibility for correctly applying the term to virologists who are experts for a specific viral family.[44]

[38] King et al. (2011)

[39] Sayers et al. (2012)

[40] Fauquet and Fargette (2005)

[41] Fauquet and Stanley (2005)

[42] Van Regenmortel (2003)

[43] Van Regenmortel (1989)

[44] Van Regenmortel et al. (1996)

3 The origins of viruses

After discussing *the difficulties of defining the concept of a virus, its status as living organism, and its taxonomic classification, we next look in more detail at viral evolution in order to establish from where these difficulties originate. In particular, we will review the relationship between viruses and the three domains of cellular life and highlight the importance of viruses as gene transfer vehicles; last, a bold theory on the origin of cellular life is introduced that features viruses as protagonists.*

Viruses and the three domains of life

All cellular life present today, as categorized into the three domains Bacteria, Archaea, and Eukarya, has descended from a single ancestor (LUCA: last universal cellular ancestor),[1] a complex life form with complete, protein-based ribosomal machinery[2] that already used the universal genetic code.[3] Based on sequence similarity of ribosomal proteins and rRNA-components, it is possible to assign to each cellular organisms a position within a single tree of life, a common representation of ancestry and speciation.[4],[5]

As briefly touched upon before, families of giant viruses such as *mimiviruses*, *marseilleviruses*, and *mamaviruses*[6] were identified that possess genomes and physical dimensions rivaling that of bacteria. The genomes of these and other viral families contain genes for carbon-, energy-, and cellular metabolism[7] – functions that are more characteristic for cellular than of viral life[8]. These findings have lead to the proposition of an additional, fourth domain of life consisting of giant viruses,[9] a notion that was later found to be invalid due to erroneous phylogenetic assumptions.[10]

Since novel genes cannot appear in viral particles *de novo*, it is very likely that genes of giant viruses are homologs of cellular genes from either current or long extinct cellular lines, possibly reaching back before the LUCA. These genes may then have been manipulated by gene duplication, recombination, and frameshift mutations and thus repurposed by selection and divergent evolution while also being subject to frequent horizontal gene transfer between viruses and their hosts. The latter may have occurred to an extent resulting in more novel genes having been transferred from viruses to cellular life than vice versa[11] (a notion that is, however, contested).[12]

[1] Glansdorff et al. (2008), Ranea et al. (2006)

[2] Lecompte et al. (2002)

[3] Vetsigian et al. (2006)

[4] Woese and Fox (1977), Woese et al. (1990)

[5] Interestingly, the original depiction of the*tree of life* was the only figure in Charles Darwin's book "On the Origin of Species". The concept it depicted should prove to be visionary. It repeatedly weathered the tests of time by incoporating new discoveries Padian (2008). As we will argue in the following, this view may be challenged by horizontal gene transfer between viruses and their hosts.

[6] Boyer et al. (2009), La Scola et al. (2008), Raoult et al. (2004)

[7] Claverie and Abergel (2010), Dinsdale et al. (2008)

[8] These giant viruses are in turn preyed upon by prophages: *"a flea/ Hath smaller fleas that on him prey;/ And these have smaller fleas to bite 'em;/ And so proceed ad infinitum."* –Jonathan Swift

[9] Boyer et al. (2010), Raoult et al. (2004)

[10] Williams et al. (2010)

[11] Forterre and Prangishvili (2009a)

[12] Moreira (2000)

Established and novel theories on the origin of viruses

Several conflicting hypotheses for the origin of viruses have been proposed; these include notions that viruses are remnants of precellular life forms (*virus-first hypothesis*), that viruses are reduced forms from cellular life that either escaped the cellular environment and adapted to a parasitic style of life (*escape hypothesis*), or that viruses are parasites of evolutionary more competent rivals (one of whose descendants developed to be the LUCA instead of becoming extinct (*reduction hypothesis*). Newer investigations have re-evaluated these hypotheses based on the current data and confirm an ancient origin of viruses[13] that probably goes back to a primordial environment, termed the *RNA-world*, in which viruses either preceded cellular life[14] or existed concurrently with it.[15]

These arguments pertaining to the origin of viruses are based on the observation that viral genes originally involved in DNA and RNA replication are structurally more similar between viruses infecting all three domains of life than between viruses and cellular organisms.[16] Similarly, specific protein folds of viral capsids are also present in RNA and DNA viruses and in viruses infecting the three domains of life, thus strongly suggesting that these viruses predate the LUCA and may have significantly shaped the emergence of cellular life.[17] Further insights derived from the comparison of protein structures suggest that the world of viruses may have been split from the beginning: while RNA viruses originated by escape or reduction, DNA viruses may have evolved only later from RNA viruses, possibly due to an evolutionary arms race between RNA-based viral and RNA-based cellular life in which viruses changed the chemical implementation of their genomes in order to protect themselves against host nucleases.[18]

Virus-like gene transfer agents (GTAs) are hypothesized to be major drivers in the development of complexity in early evolution.[19] While prokaryotes are under constant selective pressure to maintain small genomes,[20] many viruses are less constrained in this regard due to the need for a certain minimal genome size in order to ensure high-pressure capsid-packing required for successful infection.[21] Horizontal gene transfer events are a common mode of genomic exchange between viruses and prokaryotes:[22] more than 60% of the sequenced bacterial genomes contain at least one integrated viral genome (a *provirus*) and up to 3% of the nucleotide content of all bacterial genomes may consist of such proviruses.[23] The widespread use of horizontal transfer may confer selective advantages to the host and thus increase metabolic resources for the virus: this case is illustrated by marine phages that infect *Prochlorococcus* and *Synechococcus*, the most abundant photosynthetic organisms in oceanic ecosystems, and supplement the host repertoire of photosynthesis genes, thereby increasing host fitness.[24]

[13] Forterre (2006a), Forterre and Prangishvili (2009b)

[14] Koonin et al. (2006)

[15] Forterre (2005)

[16] Koonin et al. (2006)

[17] Forterre and Prangishvili (2009b), Holmes (2011)

[18] Forterre (2006a), Gorbalenya et al. (1990), Hansen et al. (1997)

[19] Koonin and Dolja (2006)

[20] Mira et al. (2001)

[21] Kindt et al. (2001)

[22] Koonin et al. (2001)

[23] Casjens (2003), Edwards and Rohwer (2005)

[24] Lindell et al. (2005)

Viruses as midwifes of the tree of life

Due to the high abundance of viruses and other virus-like gene transfer agents, the virosphere forms a gigantic reservoir of genes that can be transferred to cellular organisms, in principle.[25] Interestingly, it is therefore at least conceivable that DNA replication may have originated first in the virosphere, for example as a result of gene duplication of existing RNA polymerases, followed by divergent evolution. More daring hypotheses even suggest that components of the cellular DNA replication machinery could have been transported by horizontal gene transfer from the first DNA viruses to cellular life, thereby aiding in the emergence of the first DNA cellular life forms.[26] Indeed, there is strong evidence for the fact that while DNA was invented before the LUCA,[27] it was not yet replicated by that time. Instead, data suggest that replication mechanisms were invented by Bacteria and Archaea independently *after* the divergence of these domains.[28]

This hypothesis is compatible with the proposed introduction of DNA replication systems into RNA-based cellular life by DNA viruses.[29] While certainly a provocative theory, a similar event also demonstrates large-scale intercellular innovation originating from viruses at later evolutionary times: it is now well established by comparative genomics data that eukaryotic mitochondria originated from a free-living α-Proteobacterium.[30] However, it is less commonly known that the bacterial RNA and DNA polymerases of the proto-mitochondria have been replaced by more efficient viral homologs of T-odd bacteriophages at the time of endosymbiosis, an event that is conceptually similar to the proposed introduction of DNA replication in bacteria.[31]

Interestingly, viral invention and the proposed transfer of viral DNA replication mechanisms into cellular organisms solves an apparent evolutionary paradox: DNA is more stable than RNA and protects genetic information against oxidization and cytosine-uracil mutations, thus providing a necessary precondition for the evolution towards larger genome sizes.[32] However, it is unclear what a viable RNA-DNA intermediate that is required for evolving such a transition may look like, or how the future potential for encoding a larger genome may confer an immediate selective advantage to a cell that has just transitioned to a DNA-based genome. The proposed stepwise origin of DNA replication that makes use of an RNA template[33] may explain the diversity of proteins replicating DNA in the three domains of life,[34] provides an immediate benefit to DNA viruses due to the protection of their genome from host RNA nucleases,[35] and is further supported by evidence for a potential RNA-DNA intermediary that has been identified in phages.[36]

[25] Angly et al. (2006)

[26] Forterre (2002)

[27] Mushegian and Koonin (1996)

[28] Leipe et al. (1999)

[29] Forterre and Philippe (1999)

[30] Gray and Lang (1998)

[31] Filée et al. (2003), Shutt and Gray (2006)

[32] Lazcano et al. (1988)

[33] Wintersberger and Wintersberger (1987)

[34] Forterre (2006b)

[35] Warren (1980)

[36] Takahashi and Marmur (1963)

By way of extending these provocative assumptions, an elegant theory of the origin of DNA-based life proposes that the three domains of life are the results of three independent fusion-events between viral DNA replication machinery and RNA-based cellular life.(37) Depending on the phylogenetic data used for analysis, multiple conflicting scenarios for the evolutionary relation between the three domains of life are proposed, each of which fails to explain all of the observed data(38): either two lineages diverged from the LUCA, giving rise to Bacteria and the common ancestor of Archaea and Eukarya, respectively,(39) or two primordial lineages gave rise to Eukarya and the common ancestor of Archaea and Bacteria.(40) While the identity of the LUCA is argued to be either an RNA-based cellular life form,(41) or a bacterium giving rise to the Archaea,(42) the popular chimeric theory of eukaryotic evolution suggests that eukaryotes are a product of extensive genome fusion event between archaeal and bacterial lineages.(43)

The 'three viruses, three domains' theory, in contrast, suggests that the LUCA was RNA-based cellular life that gave rise to Bacteria and the common ancestor of Archaea and Eukarya, the latter of which might have arisen by divergence or by fusion of Bacteria and Archaea,(44) resulting in the three accepted domains of life. Three independent fusions of a founder cell of each domains with a different DNA virus, respectively, may then have lead to a transformation of the RNA genomes into DNA genomes. As a result, the transformed cells and their descendants may have been able to stabilize their genomes with regard to mutations and afford genome sizes larger than the largest currently known RNA genomes.(45) Due to the higher genetic stability and functional versatility, these cells may have then been in a position to quickly outcompete other RNA-based life forms, which were consequently lost from evolutionary records.

This theory, while not falsifiable by available data, has considerable explanatory power. First, it explains why there are only three domains of life: viral fusion is a rare event and RNA-based cells were outcompeted and removed from the biosphere, thus eliminating the basis for further fusion events. The three domains of life that originated before loss of RNA-based life specialized in different life styles (fast replicating Bacteria versus predatory Eukarya) or invaded environmental niches (Archaea).(46) Second, the theory justifies why there are three different canonical versions of ribosomal proteins, as well as extensive differences between components of the DNA replication machinery in the three domains of life.(47) Third, the hypothesis provides arguments for mutation rates in the time period between the LUCA and the origin of the three domains of life being in accordance with RNA-based life, while later evolutionary rates are suggestive of DNA-based life.(48) In summary, these hypotheses suggest that viruses may play an influential role in the evolution of the three kingdoms of life. Whether viruses should be considered to be part of the tree of life or only act as mediators between its branches is open to debate.

(37) Forterre (2005)

(38) Forterre (2006b)

(39) Woese et al. (1990)

(40) Forterre and Philippe (1999)

(41) Woese (1987)

(42) Gupta (2000)

(43) López-Garćia and Moreira (1999)

(44) Forterre (2011)

(45) Which are of viral origin (*Coronaviridae*) and are about 32 kbp long; cf. Lauber et al. (2013) for a recent discussion on the relation of replication fidelity, genome size, and genetic complexity of RNA genomes in general and *Coronaviridae* in particular.

(46) Forterre (2006a)

(47) Forterre (2006a)

(48) Woese (2000)

4 The virus within

As briefly discussed in the previous sections *viruses do not only exists as infectious virions that produce acute or chronic infections, but may also integrate their genomes directly into the host genome, thus effectively becoming part of the host and, in specific cases, of the host germline. As this section will discuss, such proviral integration events litter the genomes of most or all vertebrates, including human, and the study of these integrated genomes by sequencing offers new insights into the evolution of viruses and their hosts. This integration proves to be a two-edged sword: at least some human diseases are expected to be related to ancient integration events, human physiology critically depends on factors of viral origin. While this section will offer some information in this regard, a detailed discussion of the pathogenic effects of proviral integration is deferred until later in the next chapter.*

Proviral integration

Viruses are environmentally ubiquitous and able to infect all three domains of life[1] as well as other viruses.[2] While human pathogenic viruses dominate our understanding of and the research on viruses, most of these entities have a host range not overlapping with primates, are highly adapted to the human host, or have achieved germ-line symbiosis with it. The latter viruses are achieve achieve vertical transmission in the offspring of the host by a process termed *proviral integration* that inserts viral genomes either by a complex, *integrase*-mediated process that protects the integrity of the insert (*Retroviridae*), or by homologous recombination/non-homologous end joining[3] that often produces defective (i.e., replication incompetent) inserts (some *Adenoviridae* and *Herpesviridae* as well as viral families listed in Table 4.1). Such integration events may be permanently fixated into a species if the integration occurs in germ-line cells and is either evolutionary neutral or confers selective advantages to the host.

The human genome, for example, contains about half a million fixations of endogenous viral elements (EVEs),[4] originating from both historical reinfections and intracellular transpositions of retrotransposons, mobile genetic elements that can move within genomes via an RNA intermediate and who may have been the

[1] Rice et al. (2004)

[2] Levin and Moran (2011)

[3] cf. Vasileva and Jessberger (2005) for a recent overview on the latter two processes.

[4] Feschotte and Gilbert (2012)

ancestors of modern retroviruses.[5] While integration of EVEs has long been supposed to be an exclusive property if retroviruses, recent studies have demonstrated that all seven major groups of eukaryotic viruses have the propensity for subgenomic integration in a wide range of hosts, sometimes even crossing biological kingdoms.[6] Insertion site duplications and poly-A tails have indicated that retro-transmission by retrotransposons or retroviruses as well as by non-homologous end joining may be instrumental in formation of most non-retroviral EVEs.[7] Several of these non-retroviral ERVs demonstrate host ranges of previously unknown extent (see Figure 4.1), thereby suggesting that these viral families may form viral reservoirs in animal hosts and may thus provide the basis for potential zoonotic pandemic events affecting human society (cf. Section 6 of this chapter).[8]

Ancestral viruses in the human genome

Evolutionary dating of current viral strains is complicated due to the frequent incidence of mutations and recombinations among viruses at rates several orders of magnitude higher than those of their host.[9] These circumstances have considerably confounded previous phylogenetic analyses of viruses. The study of EVEs fixated in host genomes gave rise to the field of *paleovirology*, i.e., the analysis of integrated proviruses, and firstly enabled the investigation of viral evolution across hundreds of million of years. These analyses recently revealed that long term viral substitution rates are significantly slower than expected from the short-term data, a fact relevant for current research on human pathogenic viruses.[10] At the same time, some viral families appear to be much older than originally expected and infected mammals more than a hundred million years ago, thus highlighting aspects of virus-host coevolution and potentially informing theories on viral emergence.[11]

Several mammals such as mice and Koalas currently express infectious ERVs and are under constant epidemic pressure by these elements. These organisms are heavily influenced by EVE activity, both genotypically and phenotypically[12] More dramatically, recombination of EVEs may results in fully replication-competent viruses that cause spontaneous lymphomas and mammary tumors in mice.[13] Notably, these mechanisms are reminiscent of the origin of XMRV, a putative murine retrovirus that has falsely been associated with human prostate cancer.[14]

No fully infectious endogenous element seems to remain in the human germline as of today. This is most probably due to losses of genetic material on the one hand and purifying selection on the other hand. In evolutionary terms, however, the last infectious activity of such elements in human is still recent: approximately 150,000 years ago, well after the emergence of the modern human.[15] As a result of this relatively recent activity, about 10^5 fragments of vial origin remain in the human germline, constituting over 8% of the human genome.

Viral family	Host
Baculovirus	Insects
Herpesviridae	Humans
Nudivirus	Parasitic wasps
Phycodnaviridae	Brown algae
Circoviridae	Mammals
Geminiviridae	Tomentosae (tobacco and three other species)
Parvoviridae	Mammals; shrimp
Partitiviridae	Plants; arthropods; Protozoa
Reovirus	Aedes spp. mosquitoes
Totiviridae	Fungi; plants; ticks
Dicistroviridae	Honeybees
Flaviviridae	Medaka fish; mosquitoes
Potyviridae	Grapes
Bornaviridae	Vertebrates
Bunyaviridae	Ticks
Filoviridae	Mammals
Nyavirus	Zebrafish
Orthomyxoviridae	Ticks
Rhabdoviridae	Insects
Retroviridae	Vertebrates
Hepadnavirus	Passerine birds, Humans
Pararetrovirus	Plants

Table 4.1: *Retroviruses and non-retroviruses that integrate into eukaryotic genomes.* See Feschotte and Gilbert (2012) for a list of references supporting these findings.

[5] Jern and Coffin (2008), Koonin and Dolja (2013), Xiong and Eickbush (1990)

[6] Cui and Holmes (2012)

[7] Bill and Summers (2004), Geuking et al. (2009), Gilbert and Feschotte (2010), Horie et al. (2010)

[8] Belyi et al. (2010)

[9] Awadalla (2003), Drake et al. (1998), Duffy et al. (2008)

[10] Gilbert and Feschotte (2010)

[11] Katzourakis et al. (2009)

[12] Maksakova et al. (2006), Yalcin et al. (2011)

[13] Coffin et al. (1997a), Stoye et al. (1991)

[14] Kearney et al. (2011)

[15] Jha et al. (2011)

Gene transfer and pathogenicity

In contrast to the more common transposable elements in eukaryotic genomes that increase species diversity and have dominantly deleterious impact,[16] EVEs have introduced aspects of biological innovations advantageous to their hosts. These innovations primarily consist of domesticated viral factors that were originally introduced to subvert the host immune system,[17] or that add new protein-coding genes to the host repertoire.[18] Such beneficial gene transfers seem to have occurred multiple times in invertebrates at formative moments of their evolution. For instance, totipotent stem cells during the first cell divisions of the ovum are controlled by a master switch promoter of viral origin.[19] This promoter is repeatedly silenced and reactivated during embryonal development, resulting in cycles of pluripotency and totipotency of the affected cells that serve important regulatory purposes.

These findings are in line with earlier research that identified the viral origin of the *synctin* gene family – essential factors for both placenta-uteral membrane fusion during placenta formation and for modulation of the maternal immune system during pregnancy that prevents abortion of the foreign embryonal tissue.[20] Interestingly, synctins modulate trophoblast cell fusion through mechanisms similar to *env*-mediated retroviral entry and have been incorporated in the mammal germ-line independently on at least six occasions in rodents, primates, rabbits, and other species, notably excluding pig and horse which have other, and potentially less efficient, mechanisms of placenta formation.[21] This domesticated viral DNA has been established in the population by positive selection and may have been crucial for the rise of placental mammals a hundred million years ago.

There is currently no strong evidence for *de novo* EVE integration associated with any human disease and known human EVEs seem to lack the ability for autonomous retro-transposition and construction of infectious particles either due genomic losses or epigenetic deactivation.[22] However, loss of replication competence does not exclude the possibility of genomic rearrangements based on non-allelic homologous recombination. Such recombination events are likely to have significantly shaped human genomic architecture across evolutionary time[23] and are currently implicated in diseases such as male infertility and malign neoplasms.[24] In combination with the ability of EVEs to disseminate transcription factor binding sites and control p53 binding motifs,[25] these facts highlight the potential impact of EVEs on tumor development. Indeed, breakdown of epigenetic maintenance of EVE integration sites has the potential of malignantly affecting gene expression at up to 7% of all transcription start sites in the human genome.[26] This finding is implicated as contributing factor not only to cancer but also to autoimmune and neurological diseases, such as multiple sclerosis, type I diabetes, rheumatoid arthritis, and schizophrenia.[27]

[16] Kidwell and Lisch (2001)

[17] Arnaud et al. (2007), Coffin (1992)

[18] Mi et al. (2000)

[19] Macfarlan et al. (2012)

[20] Mangeney et al. (2007)

[21] Blaise et al. (2003), Heidmann et al. (2008), Mallet et al. (2004), Mi et al. (2000)

[22] Maksakova et al. (2008)

[23] Hughes and Coffin (2001)

[24] Sun et al. (2000), Tomlins et al. (2007), Wang-Johanning et al. (2003)

[25] Levin and Moran (2011)

[26] Conley et al. (2008), Kurth and Bannert (2010)

[27] Gifford and Tristem (2003), Jern and Coffin (2008), Perron and Lang (2010)

5 Global burden of viral disease

By way of continuing the theme *of the role of viral infections in human disease, the following section aims to summarize existing statistics on the influence of different public health factors on human well-being. These data are based on WHO studies that aim to quantify relative risks originating from infectious and other diseases and predict their development into the near future. For viral infections, these predictions may serve as rough landmarks for the relative importance of future treatment and preventive efforts.*

Quantifying human suffering

A large scale study by the World Health Organization (WHO) termed *Global Burden of Disease*[1] (GBD) undertook the task of globally estimating the influence of diseases on health. In this first-of-a-kind study on quantification of human suffering, the *disability-adjusted life year* (DALY) is used as an epidemiological unit of disease burden and firstly quantified the global impact non-fatal diseases in a transparent and reproducible manner. DALYs are defined as the absolute number of human life years lost due to premature mortality and years lived with disability, and thus serve as a time-based health-outcome measures, similar to quality-adjusted life years (QALY). In contrast to the latter, DALYs allow quantifying premature mortality with non-fatal health outcomes. DALYs are modified based both on the severity of the disability and on the period of life being lost to the individual, with suffering and disability at younger ages being weighted higher than disease at old age. Although later studies criticized the GBD approach based on unquantified uncertainties in low-income regions[2] and missing robustness of parameter choices for computing DALYs.[3] Regardless of these confounds, the GBD remains to the most comprehensive series of studies on the subject.

In the original WHO study, DALYs for three causal groups (group I–III, see Table 5.1) were computed by regression models based on death registry data and supplemental epidemiological information from 1950 to 1990. The data contained 107 causes of death tabulated by age, sex, and geographic region. In addition, incidence rates prevalence, duration, and case-fatality rates for 483

[1] Murray and Lopez (1997a,b,c)

[2] Mathers and Loncar (2006)

[3] Arnesen and Kapiriri (2004)

disabling *sequelae*[4] of the causes of death were considered. The original study by Murray *et al.*[5] was later significantly refined by Mathers *et al.* to correct for additional risk factors such as the previously underestimated severity of the AIDS pandemic, increasing tobacco use, as well as by adapting the estimates to projections for the increase of world population.[6] These refinements were applied on WHO GBD 2004 data and published in 2008, yielding updated estimates from 2008 to 2030.

Estimating global mortality rates

In general, the 2006 study predicts large declines in mortality in the next decades for all principal Group I diseases, including HIV/AIDS, tuberculosis, and malaria. In particular, global HIV/AIDS deaths are predicted to decline after the year 2012 and result in 1.2 million deaths in 2030 (compared to 1.7 million deaths in 2011).[7] While age-specific death probabilities for most Group II conditions are also estimated to decline, the increasingly aging population of most countries will likely result in significantly more deaths due to Group II conditions in general, culminating in 70% of all deaths by 2030. In particular, global cancer deaths and global cardiovascular deaths are projected based on the 2006 estimates to increase by 61% and 72% in 2030, respectively. Last, group III deaths are estimated to rise by 28% by 2030, predominantly due almost twice the number of road traffic incidents in 2030 than in 2004 as a result of increasing motorization of low-income countries.

If ranked by single causes of death, ischemic heart disease (atherosclerosis of the coronary arteries, mostly due to lifestyle choices such as smoking), cerebrovascular disease (stroke), chronic obstructive pulmonary disease (chronic bronchitis due to smoking, infections, and environmental toxins), and lower respiratory infections (pneumonia) are predicted to be the leading causes of death by 2030. Cumulatively, tobacco-related deaths, including lung cancer, will represent near 10% of global mortality by 2030. While HIV/AIDS deaths are projected to decrease, they will remain a major factor of mortality for the next decades.

Expected development of disease burden

In contrast to death rates, DALYs also incorporate people living with disabilities and are projected to decline by 10% by 2030 in comparison to 2004. These estimates already take into account the estimated population increase of 25% in the same period, thereby highlighting the significant reduction in relative disease burden of about 30% per capita. Driven by economic growth, global DALYs are predicted to decrease at a rate faster than the overall death rate due to the aging population that will encounter death at a later point in life. In particular, Group I causes of DALYs are predicted to halve by 2030 – then accounting for only 20% of all global DALYs

Groups	Disorders
I	Communicable diseases, maternal, perinatal, and nutritional disorders
II	Non-communicable diseases such as malignant neoplasms and cardiovascular and neuropsychiatric disorders
III	Intentional or unintentional injuries, including accidents and war-like conflict

Table 5.1: *Burden of disease incident groups.* The ratio of group II deaths to group I deaths aids as a crude indicator for the epidemiological transition of low-income countries dominated by preventable diseases to higher-income countries dominated by an aging population. Source: WHO burden of disease report.

[4] a pathological condition resulting from a disease

[5] Murray and Lopez (1997a,b,c)

[6] Mathers and Loncar (2006)

[7] the latter numbers are based on the UNAIDS World AIDS Day Report, http://www.unaids.org/en/resources/publications/2012/name,76120,en.asp

– while the Group II disease burden is projected to increase to two thirds of all global DALYs. Interestingly, cancers with viral risk factors are not considered within Group I DALYs in these projections. Although cancer incidents with viral cofactors such as liver cancer (HBV and HCV) and cervical cancer (HPV) will continue to increase due to long time periods between initial infection and development of cancer, these DALYs do not contribute to the Group I burden in 2030.

Leading causes of DALYs in 2030 are predicted to be unipolar depressive disorders, ischemic heart disease, and road traffic accidents. In particular lower respiratory infections and HIV/AIDS, the dominant infectious causes of DALYs in previous decades, are projected to decline significantly. This development is predominantly a result of great successes in antiviral research, such as antiretroviral drugs effective against HIV and the wide use vaccines, as for example against polio. On the other hand, diseases associated with old age or increasing wealth such as diabetes mellitus, road traffic accidents, hearing loss and tobacco-associated diseases are predicted to increase substantially, resulting from increases in income and longevity especially in low-income countries.

In general, the next years will probably show a shift in the distributions of deaths and DALYs from younger to older ages and from Group I causes (communicable diseases such as viral infections) to Group II causes (non-communicable diseases such as most cancers). This phenomenon likely is an indicator for the epidemiological transition from low-income countries dominated by preventable diseases to higher-income countries dominated by an aging population. While epidemiological changes resulting from better clinical care will particularly mitigate the negative effects of non-communicable as well as cardiovascular and infectious diseases, such factors may be overpowered by population growth and aging. The latter will especially amplify non-communicable disease such cancers as well as maladies associated with tobacco smoking as, for example, cardiovascular and chronic obstructive pulmonary diseases.

While therefore infectious diseases will probably take a backseat in the next decades compared to other diseases due to the successes of antiviral research and increasing standard of living in low-income countries, the latter phenomenon also has detrimental consequences. As the next section will demonstrate, sequela of rising incomes such as increased urbanization and air travel may facilitate the emergence of novel viral diseases as for example swine flu and SARS, in principle. Antiviral research will therefore be of continuing importance especially with respect to worldwide surveillance of viral disease.

2004 · Incident	DALYs
Lower respiratory infections	6.2
Diarrhoeal diseases	4.8
Unipolar depressive disorders	4.3
Ischaemic heart disease	4.1
HIV/AIDS	3.8
Cerebrovascular disease	3.1
Prematurity and low birth weight	2.9
Birth asphyxia and birth trauma	2.7
Road accidents	2.7
Neonatal infections	2.7
2030 (projected) · Incident	DALYs
Unipolar depressive disorders	6.2
Ischaemic heart disease	5.5
Road accidents	4.9
Cerebrovascular disease	4.3
COPD	3.8
Lower respiratory infections	3.2
Hearing loss, adult onset	2.9
Refractive errors	2.7
HIV/AIDS	2.5
Diabetes mellitus	2.3

Table 5.2: *Relative DALYs 2004 and 2030 (projected).* DALYs are expressed as percent of total DALYs. Source: WHO burden of disease report.

6 *Viral abundance and emergence*

Viruses are highly abundant and diverse *entities that permeate both the external world as well as our bodies and germlines. The present section introduces recent estimates of the copiousness and global environmental effects of viral particles and argues for a great genetic richness of viral genomes and proteomes. Apart from their significance as a reservoir for genetic innovations, the diversity of viruses also has implications for public health: since viruses are able to switch host species, in principle, many or most new human pathogenic viruses emerge from a background of zoonosis, i.e., transmission from animal life to human. This fact has been exemplified several times in the last years and surveillance of viral factors is therefore of growing importance for public health.*

Environmental abundance of viruses

Prokaryotes are highly numerous entities that represent 90% of the ocean biomass, feature abundances in excess of 40 million entities in a gram of soil, and differ drastically in their composition between geographical locations.[1] Overall, there are an estimated 5 $\times$ 10^{30} prokaryotes present in the worldwide ecosystem, the most abundant single species of which being photosynthetic marine cyanobacteria which consists of approximately 10^{27} individuals.[2] For comparison, the number of cyanobacteria therefore is about a million-fold higher than the number of stars in the universe, estimated as 5×9^{21} entities based on stellar density and the observable volume of the universe[3].

In addition, the combined mass of prokaryotic biomass[4] of 350,000–550,000 megatons (Mt) equals or exceeds the total biomass of plant life.[5] In contrast, the biomass of high-mass animal species like human (an estimated 350 Mt given seven billion individuals with an average weight of 50kg) or antarctic krill (500 Mt) is easily surpassed by cyanobacteria alone (1,000 Mt).[6]

[1] Suttle (2007), Whitman et al. (1998), Zinger et al. (2011)

[2] Whitman et al. (1998)

[3] `http://www.grc.nasa.gov/WWW/K-12/Numbers/Math/documents/ON_the_EXPANSION_of_the_UNIVERSE.pdf`

[4] This is considering the whole mass or *fresh biomass* of an organism and includes intracellular water. In contrast, ecological biomass may also be quantified by the mass of all organically bound carbon contained in an organism, which usually is 30% of the fresh weight and sometimes considerably lower.

[5] Whitman et al. (1998)

[6] Atkinson et al. (2009), Garcia-Pichel et al. (2003)

Sequence-independent experimental methods such as direct-count epifluorescence and transmission electron microscopy were first to indicate that, similar to Bacteria and Archaea, viruses and virus-like particles are also highly abundant in the environment. Given the copiousness of bacteria within both seawater and soil, it is not surprising that phages, i.e., bacteria-infecting viruses, are at least as bountiful. Indeed, since viral abundance seems to be highly correlated with the occurrence of prokaryotes, the prior may exceed the latter by one to two orders of magnitude,[7] a number that is also supported by experimentally measured phage *burst sizes*[8] Indeed, the resulting estimates for the number of virus-like particles in the biosphere are ranging from 10^{31} to 10^{33}, depending on the biome under investigation.[9],[10] The approximately 10^{13} human cells are outnumbered 10-fold by prokaryotes and 100-fold by viruses and up to 94% of all existing nucleic acid containing particles may be of viral origin.[11]

While many viral strains pathogenic for human are known today, it is estimated that less than 1% of the global viral diversity has been sampled scientifically, leading to the designation of the *virome* as "dark matter" of ecology.[12] However, pathogenic effects of viruses are not limited to single hosts: although constituting only 5% of the oceans' biomass, viruses are generating approximately 10^{23} infections per second, thereby lysing in excess of 18% of the oceans' biomass per day and mediating 25% of the oceans' primary production of nutrients.[13] The dissolved organic matter set free by this process of "viral priming" constitutes a major geochemical forces in the oceans, heavily influencing large-scale carbon cycles and green-house gas emissions.[14] In addition, these processes also substantially influence the global food web by controlling the amount of iron available for photosynthesis and limiting the propagation of dominant prokaryotic species such as phytoplankton populations during bloom, thus increasing biological variety in the oceans.[15]

Based on techniques such as micropore filtration that allow investigation of the nucleotide content of viral particles, studies were conducted on viral abundance in external environments such as fresh and salt water as well as within terranean samples[16] (see Fancello et al. (2012) for a review). These investigations further demonstrated the vast diversity of the virome, especially regarding bacteriophages.[17] Indeed, recent results indicate that between 500 and 129,000 viral genotypes are present in a liter of seawater (the term *genotype* is roughly equivalent to a species in the context of metagenomics but is usually employed to denote viral subspecies in clinical virology). These numbers can be extrapolated to 10^{30} unique viral genotypes present in the oceans, indicating that the marine virome may be the most diverse ecological community on earth.[18] Given this high diversity, it is not surprising that the great majority of genotypes found by recent studies in oceanic samples do not have identifiable homology to known genes.[19]

[7] Bergh et al. (1989), Rowe et al. (2012)

[8] Average number of phages that are produced per infected bacterium; this quantity can be experimentally determined by single-step growth experiments; can range between a few hundred particles for DNA phages and up to 10^4 particles for RNA phages.

[9] Whitman et al. (1998)

[10] For comparison, this number of particles stretches longer than the nearest 60 galaxies if aligned end to end.

[11] Mokili et al. (2012), Suttle (2007)

[12] Rohwer and Youle (2011); The term *virome* denotes the collection of all viruses in a given habitat. If not habitat is specified, the term applies to all viruses on earth.

[13] Suttle (1994, 2007), Wilhelm (1999)

[14] Evans et al. (2007), Suttle (2005)

[15] Bratbak et al. (1993), Murray (1992), Poorvin et al. (2004)

[16] Angly et al. (2006), Breitbart et al. (2004, 2002), Desnues et al. (2008), Dinsdale et al. (2008), Rice et al. (2001), Williamson et al. (2008)

[17] Culley et al. (2006), Pride et al. (2012), Reyes et al. (2010), Suttle (2005), Willner et al. (2009)

[18] Angly et al. (2006), Breitbart and Rohwer (2005), Rohwer and Thurber (2009), Rosario and Breitbart (2011)

[19] Kristensen et al. (2010), Rohwer and Thurber (2009)

Extrapolating these numbers to global estimates is problematic[20] and extensive gene transfer between viruses may mask the underlying diversity of unique genetic and proteomic components. However, the proposed vast richness of the virome is also supported by proteomics. Indeed, the total viral proteome space is estimated to encompass up to a billion virally encoded ORFs.[21] Even though more comprehensive and recent studies based on a wider range of data and using more stringent analysis approaches[22] drastically reduced these estimates to between 3.9 million protein clusters derived from metagenomics studies and at least 0.6 billion proteins clusters based on known phage-host systems,[23] viruses may still represent the largest genetic reservoir on the planet.

These studies are paralleled by related results stating that most viral proteins have no homologs in modern cells, a fact seemingly in contradiction with the traditional concept of viruses as "gene robbers", i.e., recruiters of and shuttle vectors for genes of cellular origin.[24] It has recently become clear that this abundance of viral genes without cellular homologs is not a consequence of the large number of yet unknown cellular genes or an artifact resulting from distant divergence of the homologous sequences. Instead, it is now accepted that at least 63 protein domain superfamilies (of the about 2,000 superfamilies known to SCOP[25]) have no structural relatives among cellular life and thus represent uniquely viral innovations.[26]

Due to the large abundance of viruses, the fact that all known forms of cellular life are subject to infection and consequently, in principle, to horizontal gene transfer by viruses,[27] and the high rates viral mutation of up to 5–6 orders of magnitude higher than the mutation rates of their hosts,[28] the uniquely viral superfamilies may be considered a core mechanism of protein evolution. Effectively, the virome may thus act as a continuous source of novel, stable protein folds that supply cellular organisms with structural innovations. If stable cellular protein conformations represent islands in theoretical structure space that are embedded in a 'ocean' of structurally unstable conformations, these stable but yet-unobserved viral structures may act as land bridges between cellular protein folds.[29] These metaphorical bridges may thus amplify development of new protein functions by pre-selecting and importing stable protein folds into cellular organisms, thus potentially facilitating evolution of cellular organisms[30]

[20] Duhaime and Sullivan (2012)

[21] Rohwer (2003)

[22] Hurwitz et al. (2013), Hurwitz and Sullivan (2012), Roux et al. (2011)

[23] Cesar Ignacio-Espinoza et al. (2013)

[24] Moreira and López-García (2009)

[25] Andreeva et al. (2004)

[26] Abroi and Gough (2011)

[27] Sano et al. (2004)

[28] Sanjuán et al. (2010)

[29] Taylor et al. (2009)

[30] Abroi and Gough (2011)

Emerging human pathogenic viruses

The majority of newly identified or *emerging* human viral diseases such as severe acute respiratory syndrome (SARS) and the Nipah virus arise from a background of zoonosis, i.e., originate from non-human hosts, most of them mammal wildlife.[31] Although the trajectory of viral emergence is currently contested (both increasing and mildly decreasing rates are reported in the literature),[32] these numbers are still sufficiently high to cause alarm: indeed, while viruses make up a comparably small part of known human pathogenic vectors, the majority of all newly identified pathogens are viruses.[33] Recent estimates based on only one species of bats, a notoriously super-infected mammal here employed as a representative for the approximately 5,500 mammals known, revealed more than 50 previously unknown viral species in only nine viral families assayed.[34] Based on statistical estimates about the completeness of sampling and the number of known mammals, the total richness of unknown mammalian viruses may thus exceed 300,000 species within the nine assayed viral families alone.[35]

Viruses have evolved towards broad host ranges that may even encompass multiple domains of life: some *Nodaviridae*, for instance, infect insects, plants, as well as fungi.[36] Interestingly, host change may evolve quickly especially in highly divergent viruses such as avian influenza that may require only few mutations in order to acquire airborne transmission capabilities between mammals.[37] At the same time, host specificity in itself is very diverse and phylogenetic analyses have determined that even closely related viral species may have distinct host ranges.[38]

It is therefore not surprising that up to a third of the viral families known to infect eukaryotes are also infecting humans, and more than two thirds of viruses infecting humans also infect other vertebrates.[39] Most of the hosts susceptible and permissive to infection by human viruses are mammals, in particular rodents, hoofed animals, primates, and bats; less than 20% of these hosts are birds. Emerging from this reservoir of zoonotic viruses, new human viruses are discovered at rates of about three a year.[40] As a consequence, over two thirds of all new human pathogens consist of viruses.[41]

Thankfully, although an estimated one third of the viruses infecting livestock may also infect humans, transmission of these viruses between humans is limited: only about half of all viruses infecting humans can also be transmitted within the human species, and only 25% are considered sufficiently infective to spread efficiently in a susceptible population.[42] Still, examples for past zoonotic viruses originating from mammals and birds that have established themselves in the human population exist, as for example IV-1 and HIV-2 (chimpanzees and mangabeys, respectively), SARS (bats), HTLV, dengue, and yellow fever (primates), measles and smallpox (livestock), as well as Influenza A (wild birds).[43]

[31] Morse (1995), Morse et al. (2012), Woolhouse and Gowtage-Sequeria (2005)

[32] Jones et al. (2008), Woolhouse et al. (2012)

[33] Woolhouse and Gaunt (2006)

[34] Anthony et al. (2013)

[35] Anthony et al. (2013)

[36] Price (1996), Selling et al. (1990), Woolhouse et al. (2005)

[37] Herfst et al. (2012), Imai et al. (2012), Russell et al. (2012)

[38] Woolhouse (2001)

[39] King et al. (2011), Taylor et al. (2001), Woolhouse and Gowtage-Sequeria (2005)

[40] Based on estimates by Woolhouse et al. (2012) which are derived from historical data since 1954; the rate of emergence is consistent with a Poisson process with an expected value of 3.37 species a year.

[41] Woolhouse and Gaunt (2006)

[42] Taylor et al. (2001), Woolhouse et al. (2012)

[43] Woolhouse et al. (2012)

Viral epidemiology and surveillance

As a result of significant increases in global trade and air travel that accompany globalization, travelers infected with human pathogenic agents are now more prone to transmitting pathogens to new hosts around the globe in a fashion that is both rapid and unforeseeable. Mass migrations of human populations in general create ideal conditions for the transmission of infectious diseases[44] and air travel in particular has been associated with the global spread of the initial waves of infections with HIV, SARS coronavirus, West Nile virus, and tuberculosis.[45] The increased global transportation of food produce occasionally contaminated with pathogenic animal faeces[46] and intensive farming accompanied with heavy use of antibiotics[47] provide ideal conditions for the evolution of viruses with broad host ranges. These animal pathogens are considered important sources for emerging human diseases, a fact that is illustrated by the recent transmissions of Influenza H7N7 from poultry and H1N1 from pigs.[48]

In addition to agricultural sources of novel pathogens, climate change modifies the geographical distribution of viral hosts, resulting in the occurrence of tropical infectious diseases in new locations as exemplified by the recent outbreak of West Nile encephalitis in the United States.[49] Deforestation of high-diversity biotopes, and illegal trafficking of animal wildlife allows pathogens to reach new environments, thereby increasing the likelihood of zoonotic events originating from invaded biotopes[50] that have already been implicated in the origin of HIV, Nipah virus, and filoviruses and may be responsible for up to 75% of all emerging infectious diseases in humans.[51]

Last, both idiopathic pathologies suspected to involve unknown viruses and medical conditions such as organ transplantation, HIV-mediated T-cell depletion, and some forms of cancers that involve host immunosuppression, favor emergence of human pathogenic viruses from the background of usually non-pathogenic latent infections with viral commensals. Indeed, several such diseases as for example respiratory diseases, diarrhea, multiple sclerosis, Kawasaki disease, as well as rheumatoid arthritis and inflammatory bowel disease have been associated with known and unknown viral factors, some of which may have zoonotic origins (see Delwart (2013), Wylie et al. (2013) for a list of current publications concerning these associations).

Similarly, highly prevalent commensals and chronically infecting human viruses such as TTV, one of the newest emergent viruses for which there is currently no clear disease association, may reveal pathogenic potential in contexts of immunodeficiency or within specific patient populations.[52] Other zoonotic commensals exist which can cause human diseases in rare cases: for example the lymphocytic choriomeningitis virus, a zoonotic pathogen originating from mice and hamsters for which up to 3% of the German popu-

[44] Jones et al. (2008), Morse (2004)

[45] Dowdall et al. (2010), Hosseini et al. (2010), Mangili and Gendreau (2005)

[46] Harris et al. (2010), Hutson (2007), Maki (2006)

[47] Garcia-Alvarez et al. (2012), McEwen and Fedorka-Cray (2002)

[48] Cleaveland et al. (2001), Koopmans et al. (2004), Peiris et al. (2009)

[49] Roehr (2012), Shuman (2010)

[50] Delwart (2007), Karesh et al. (2012), Smith et al. (2012), Wang (2011)

[51] Chua et al. (1999), Pulliam et al. (2011), Taylor et al. (2001)

[52] Lipkin (2010), Virgin et al. (2009)

lation carries antibodies and that may cause deadly meningitis in about 1% of the infected patients.[53]

In combination with studies demonstrating that divergent viruses may quickly adapt to new hosts (see arguments in the previous section), these results suggests that additional human diseases with yet unknown etiology may be caused by emergent viral agents, thus highlighting the need for ongoing global pathogen surveillance.

Global response to viral epidemics is heavily depending on early and fast identification of the causative pathogens. For instance, the 1918 pandemic of H1H1 Influenza that infected approximately half a billion people is estimated to have killed 3% of the global population.[54] However, the causative factor of the disease was only later shown to be a virus rather than, as first suspected, the bacterium *Haemophilus influenza*, both of which have significantly different disease mechanisms, epidemiological behavior, and, at least nowadays, treatment options. In contrast, the 2002 pandemic associated with the severe respiratory syndrome (SARS) was linked to a novel coronavirus by Sanger (dideoxy) sequencing within weeks of the initial outbreak.[55] Both increased global communications and better understanding of disease mechanisms enabled dramatically improved medical and preventive response in the 2002 pandemic, resulting in significantly lower numbers of infected people (~8000) and fatalities despite comparable transmissibility (R_0=2–5) and case fatality rates (~10%) of the H1H1 and SARS infections.[56]

In order to detect emergent human pathogens as early as possible, basic research initiatives investigating the environmental virome are currently supplemented with applied studies on human-proximal viral communities. These studies focus on the development and spread of human pathogenic viruses in the context of preemptive epidemiology and public health and employ systematic screening in order to support rapid detection of known and novel viral pathogens.

The speed of identification of causative viral agents is currently further accelerating due to advances in sample preparation and next-generation sequencing techniques that allow identification of uncultured pathogens in clinical samples with unprecedented velocity. Recent examples employing this technology that indicate the future of rapid-response epidemiology include the identification and genomic characterization of the *Lujo* virus in South Africa in 2008[57], the cattle Schmallenberg virus in Germany and the Netherlands in 2011,[58] and the toxic *Escheria coli* bacterium O102:H4 in Hamburg in 2011 within days of receiving the first clinical sample.[59]

In addition, there is rising public awareness of early detection of novel and known human pathogens as exemplified by scenarios of general public health and surveillance,[60] and discussions about the possible use of weaponized organisms employed by terrorist organizations.[61]

[53] Lehmann-Grube et al. (1979), Schelhorn (1993)

[54] Durand (1977), Taubenberger and Morens (2006)

[55] Marra (2003), Rota (2003)

[56] Durand (1977), Lipsitch (2003), Mills et al. (2004), Riley (2003), Taubenberger and Morens (2006), WHO (2003)

[57] Briese et al. (2009)

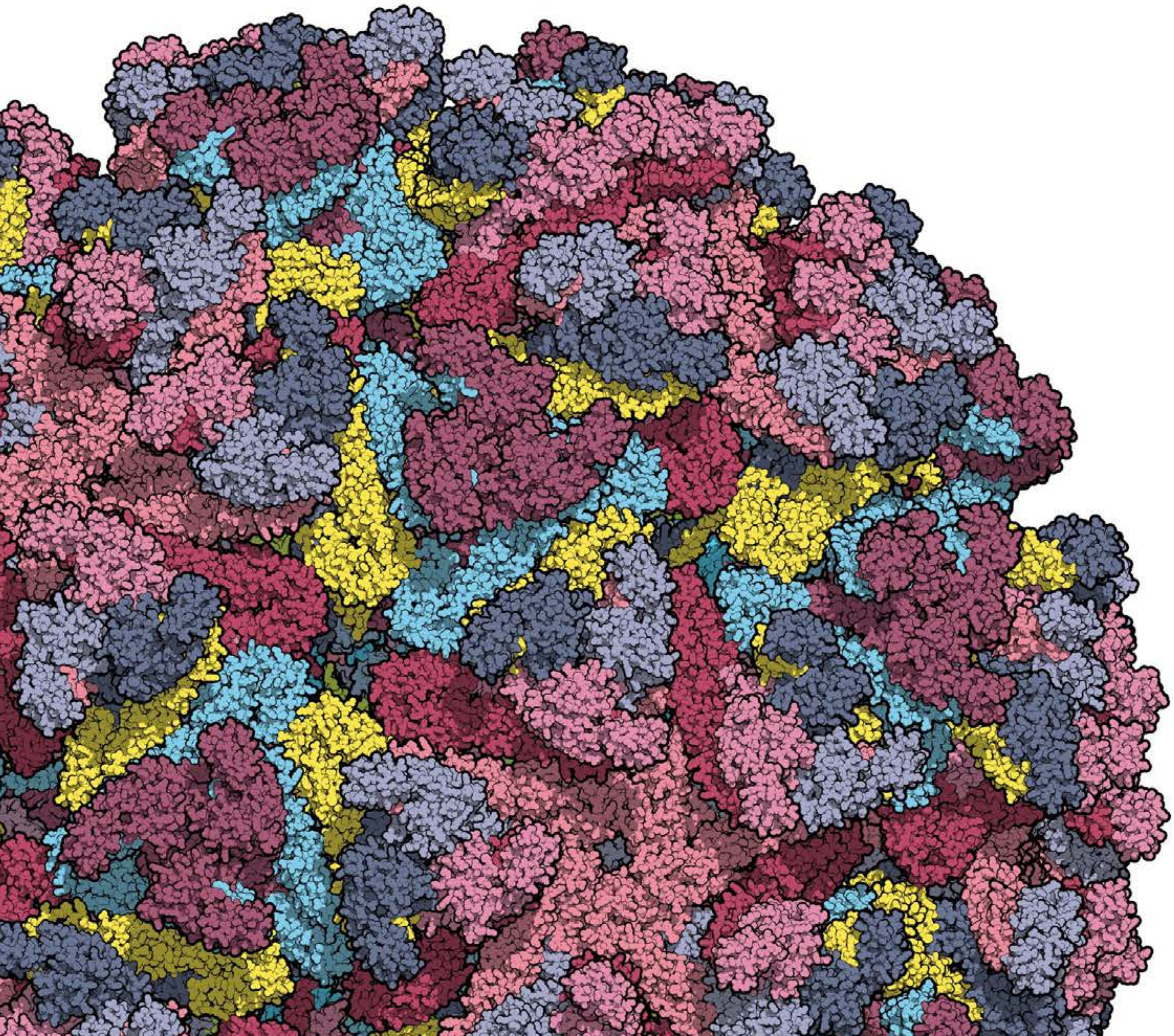
[58] Hoffmann et al. (2012)

[59] Frank et al. (2011), Karch et al. (2012), Mellmann et al. (2011)

[60] Mykhalovskiy and Weir (2006)

[61] cf. the heated discussion about the Herfst et al. (2012) publication in the news media

II
Viral metagenomics



THE PRESENT CHAPTER *forms the vanguard of the three core chapters of this thesis. It discusses the concept of viral metagenomics, i.e., the systematic search for viral nucleotide signatures in mixed samples, in the context of both surveillance of emerging pathogens and the search for previously unknown causes of human cancers.*

After briefly introducing the concept of metagenomics, this chapter presents an introduction to sequencing approaches in virology in order to provide sufficient background for later sections. Subsequently, an outline is presented of both founding studies and recent developments in metagenomics, with a special focus on the detection and characterization of viral pathogens. In a conceptual interlude, the biomedical notion of oncogenic viruses and their multifaceted modes of interaction with cellular systems are surveyed. The bioinformatics paradigm of metagenomics and the virological paradigm of oncogenic viruses are then synthesized into a computational method for detecting low-abundance transcripts of known and novel viruses in human tissues. The chapter concludes with an application and validation of the aforementioned method on deep sequencing data from neuroblastoma, a human pediatric tumor.

7 Introduction to metagenomics

Due to their high genetic diversity, *their possible role as gene transfer vehicles between species, and the selective pressure they exert on many cellular populations, viruses are critical regulators of global biodiversity.*[1] *Investigation of viral environmental diversity is now considered to be a fruitful area for basic research initiatives that aim to classify novel functional genetic elements in the virome, gain insight into the origin and evolution of viruses, or attempt to understand patterns of interactions between viral, bacterial, and eukaryotic components of marine and terranean ecosystems. While often applied to a wide range of bioinformatics approaches and technologies, the term* metagenomics *foremost denotes the application of sequencing methodologies for the investigation of an entire, uncultured community of microbial*[2] *organisms directly sampled from their natural environment.*[3]*Although first based on clonal sequencing data,*[4] *subsequent systematic metagenomics studies employed high sensitivity deep sequencing technology in order to systematically analyze microbial communities. Since metagenomics is highly dependent on nucleotide sequencing, the following section will provide a brief introduction to this bourgeoning subfield of genomics.*

Viral next-generation sequencing

Besides cell cultures, polymerase chain reaction (PCR), and electron microscopy, sequencing has become one of the major experimental technologies of virology. As of September 2013, GenBank contained 1,619,495 sequence entries assigned to viral taxonomies (NCBI taxon identifier 10239), the majority of which originate from HIV-1, Influenza A, and hepatitis viruses. Based on historical data, these entries represent an annual growth of more than 20%[5] and almost entirely reflect sequences of high clinical significance: indeed, 18 of the 20 most frequently sequenced viruses are pathogenic in humans. Based on these sequences, more than 4,000 full viral genomes have been determined as of today. Scientific insights derived from these genomes had large impact on various subfields of virology and sequencing of viral factors is now considered one of the most important experimental technologies employed in virology.

Nucleotide sequencing is based on two breakthrough microbiological technologies, namely DNA amplification by PCR[6] and

[1] Futse et al. (2008), Suttle (2007)

[2] The term *microbe* is here used in its most general sense, i.e., it encompasses all microorganisms including bacteria and archaea, fungi, protozoa, and viruses. While the Greek roots of the constituting terms *mikro* and *bios* refer to living organisms, viruses are included here in spite of their somewhat ill defined role as life-like but not fully living entities.

[3] Handelsman et al. (1998), Petrosino et al. (2009)

[4] Breitbart et al. (2003, 2002)

[5] Benson et al. (2013)

[6] Mullis et al. (1986)

the first generation of DNA sequencers using chain-terminating nucleotides.[7] While these technologies provide an affordable way to sequence templates up to 1000 bp, their low scalability as well as their reliance on prior template amplification using specific primers restrict their use on samples containing large or highly diverse sequence content. In addition, PCR amplification followed by direct Sanger sequencing exhibits only limited sensitivity at resolving minority variants, i.e., low-frequency sequence fragments in mixed samples; indeed, these variants can only be detected by Sanger sequencing if represented at more than 10-20% frequency.[8] While this sensitivity can be increased by sequencing multiple individual DNA molecules from each sample, a process termed *limiting-dilution PCR*, this procedure is costly and displays only low scalability.[9] As a consequence of these confounds, Sanger sequencing is not suited for investigating metagenomics samples that contain samples with large volumes of unknown species with highly varying abundances and that are expected to include large fractions of minority variants.

Next-generation sequencing. By building on prior breakthroughs in semiconductor technology and microfluidics, novel sequencing technologies have been developed that afford both dramatically increased throughput and high sensitivity. These technologies, commonly termed next-generation sequencing[10] firstly provided researchers with the ability to generate millions or billions of sequence bases over night, thus empowering even single labs to sequence a human genome or hundreds of viral genomes in a short time frame and at acceptable costs. In contrast to Sanger technology, NGS platforms can detect minority variants with high sensitivity even at low frequencies of 1-2%.[11]

As a result of the high acceptance of next-generation sequencing by academic and industrial research centers and the ever increasing capacities of sequencing platforms, the available sequencing capacity is now significantly higher than abilities for analyzing the generated sequencing data. The resulting large volumes of sequencing data, also denoted as 'data deluge', are exemplified by the large numbers of sequences deposited in specialized data archives such as ENA and SRA,[12] and pose significant challenges concerning storage, transfer, and analysis of these data.

While differing in their exact implementation, current *second generation* sequencing technologies have in common that they operate on pre-amplified DNA pools of large ensembles of cells that are processed in complex library preparation steps, sequenced in a massively parallel fashion, and read out by high-throughput imaging technology that deduce the actual base content (*base calling*) of the sample. While deep sequencing allows for determination of millions or billions of sequence fragments within a single run of the sequencing instrument, the length of each sequence fragment (*read*) is limited, depending on platform specifics, to about 150–800

[7] Sanger et al. (1977)

[8] Varghese et al. (2009), Wang et al. (2007)

[9] Palmer et al. (2004)

[10] also known as new generation sequencing, second generation sequencing, or deep sequencing; the exact designation varies between scientific fields and reflects the scientific application as well as the level of demarkation with respect to related technologies

[11] Margeridon-Thermet et al. (2009), Varghese et al. (2009)

[12] Cochrane et al. (2013), Kodama et al. (2012)

bp due to biochemical and biophysical constrains of the sequencing process. As a consequence, these technologies are also often termed 'short read' technologies and are contrasted with third generation ('long read') sequencing technologies that allow for dramatically longer reads lengths in the kilobase range, albeit at lower overall throughput and, currently, accuracy.[13]

Specific formulations for the library preparation process exist that influence the sequencing process; for example, reads can be sequenced from the opposite ends of a single DNA molecule, a process that provides additional linkage information and is denoted as *paired end* or *mate pair* sequencing. In addition, sequence fragments can be annealed to additional adapters that allow for mixing of sequencing libraries, a process termed *multiplexing*. In addition, a wealth of library formulations exist that aim to measure different biological entities or relations using sequencing; an incomplete list of these approaches is provided in Table 7.1.

ALEXA-Seq
Apopto-Seq
AutoMeDip-Seq
Bind-n-Seq
Bisulfite-Seq
ChIA-PET
ChIP-Seq
ChIRP-Seq
ChiRP-Seq
CIIP-Seq
CNV-Seq
Degradome-Seq
DGE-Seq
DNA-Seq
DNase-Seq
dRNA-Seq
F-Seq
FAIRE-Seq
FRT-Seq
Frag-Seq
HITS-CLIP
Immune-Seq
indel-Seq
MBD-Seq
MeDIP-Seq
MethylCap-Seq
microRNA-Seq
mRNA-Seq
NA-Seq
NET-Seq
NOMe-Seq
NSR-Seq
PARE
PAS-Seq
Peak-Seq
PhIP-Seq
Protein-Seq
ReChIP-Seq
RIP-Seq
RIT-Seq
RNA-Seq
RNASeq
rSW-Seq
SAGE-Seq
Shape-Seq
Seq-Array
Sono-Seq
Sort-Seq
Tn-Seq

Table 7.1: *Library formulations for deep sequencing.* Incomplete list. From: `http://core-genomics.blogspot.de/2011/09/next-generation-sequencing-acronyms.html`

Sequencing platforms. In the context of this work, two second-generation sequencing platforms are of prime importance: the first technology, Roche/454 pyrosequencing, relies on internal amplification of the pre-amplified library product using micro-droplet emulsion PCR. Each droplet contains a single bead, i.e. a nanotechnological entity that fixates a single DNA molecule. This molecule is amplified to several thousand copies that remain attached to the bead. Beads are located on a picotiter plate consisting of millions of wells. These wells are organized into flowcells that allow for adding of nucleotides and washing reagents in all wells in parallel. DNA molecules on loaded beads within the flowcell are simultaneously sequenced by sequentially adding one of four deoxy-nucleotides (dNTPs) nucleotides to the flowcell. These nucleotides are polymerized by enzymatic synthesis of complementary DNA strands within each well (thus the term *sequencing by synthesis*).

Apart from employing emulsion PCR for amplification, a notable technical innovation of the Roche/454 platform is the pyrosequencing detection process by which each nucleotide incorporation results in the release of pyrophosphate. Pyrohosphate is converted by the ATP sulfurylase into ATP, which in turn activates luciferase and results in emission of a light signal whose intensensity is proportional to the number of bases incorporated. Since nucleotides are added to the flowcells with precise timing at each sequencing cycle, peaks of photonic signals can be translated into base calls for each well, which are concatenated into sequencing reads.

In contrast to Roche/454 technology, the Illumina HiSeq and MiSeq platforms employ different approaches that allow for significantly higher throughput, although at lower read length. Illumina instruments rely on high-density amplification of clonal colonies of DNA strand that are fixed onto a glass surface. Amplification occurs on this surface in a process termed *bridge* amplification and results in clusters of clonaly amplified molecules. Sequencing is

[13] Flaherty et al. (2012), Hou et al. (2012), Lien et al. (2010), Navin et al. (2011), Nora Dickson et al. (2011), Xia et al. (2011), Xu et al. (2012b)

performed by simultaneously flushing all four nucleotides over all colonies of a sequencing lane, a specified compartment of the flowcell. Since nucleotides are fluorescently labeled with distinct fluorophores, and nucleotide synthesis is self-terminating (*reversible chain termination*), laser illumination of the sequencing lane can be employed in order to detect the last incorporated nucleotides in all colonies using an imaging instrument. After enzymatic cleaving of the fluorescent dye and 3′ terminators of the last nucleotide, the procedure can be repeated for a limited number of cycles.

By employing pyrophosphate-based chemistry, the 454 platform avoids using reversible chain termination chemistries used by other sequencing platforms that afford only shorter read lengths. However, the Roche/454 platform gains this advantage at the cost of increased difficulties in separating homopolymer runs which tend to produce a continuous light signal across multiple base incorporations, thus increasing the error rate of the 454 sequencing process to about 1% (compared to about 0.1% for Illumina platform).[14] A second, and perhaps more important advantage of the Illumina technology is sequencing throughput: since sequence clusters amplified by Illumina bridge amplification can be attached to the flowcell at higher densities compared to wells employed by the 454 technology, Illumina devices typically afford significantly higher numbers of sequenced bases per instrument run.[15]

Additional sequencing platforms of note are "2.5"th generation and third generation technologies such as the Ion Torrent Personal Genome Machine and the Pacific Biosciences RS, respectively. While the Ion Torrent can be viewed as a low-cost relative to 454 machines, affording significantly simpler library preparation and cheaper machines by using pH change upon base incorporation as signal and electrical signals for detection, the PacBio platform employs dramatically different technologies that do not require amplification of molecules. Instead, the RS employs *single-molecule sequencing* by observing synthesis activity of individual polymerases trapped in one of 150,000 wells that use phospho-linked fluorescent nucleotides as substrates. While featuring only low signal-to-noise ratios due to its reliance on non-amplified molecules, this technology affords very long read lengths of more than 3 kbp and avoids biases associated with DNA amplification.

Sequencing errors. For the second generation sequencing process, image analysis results in reconstructed sequencing reads that represent noisy measurements of large numbers of overlapping fragments that originate from an ensemble of genomes. Due to oversampling (i.e., each genome position is sequenced multiple times, typically 30–200-fold for DNA sequencing), and the usage of pre-amplified DNA, each sequencing experiment is able to generate more base pairs than the original sample contained. This high *average* coverage is generally required in order to ascertain sufficient *minimal* coverage of high-fidelity reads at each genome position.

[14] Error rates are reported as percentage of errors per base within individual reads of the maximum length Glenn (2011); however, as the author notes, error rates among platforms are not exactly comparable

[15] Typically, an Illumina HiSeq 2000 run yields on the order of 600 Gbp of sequence data per run. In contrast, a 454 FLX+ instrument is limited to about 0.6 Gbp per run. While the 454 FLX+ instrument affords longer read lengths of 650-800 bp (compared to 2 × 150 bp generated by current-generation Illumina instruments), this is achieved at significantly higher costs per Mbp sequenced ($7.00 per Mbp for 454 versus $\geq$ $0.04 per Mbp for Illumina HiSeq 2000). Source: http://www.molecularecologist.com/next-gen-fieldguide-2013/

In order to compare the accuracies of sequencing platforms, error rates of the sequencing process can be determined by amplifying clonal samples (which have been obtained by limited dilution, for example) within bacterial plasmids that display significantly reduced error rates compared to PCR amplification.[16] Variation from the consensus sequence of the sequenced reads can thus be associated with the sequencing process rather than with amplification. These analyses have resulted in specific error profiles for the 454 and Illumina sequencing platforms that predominantly consist of erroneous base substitutions that depend on the sequence context[17] as well of base under-/overcalls in homopolymeric regions that result in artifactual indels.[18]

Other technical errors are frequently introduced prior to sequencing by the library preparation process: as already touched upon, current second-generation sequencing platforms require sufficient DNA material in order to operate at high accuracy. Therefore, the sample DNA (or reverse-transcribed RNA) has to be amplified prior to sequencing using PCR technologies. Amplification can be confounded by several types of technical errors, such as primer mismatches and selective amplification which may increase the frequency of certain DNA molecules. In addition, nucleotide substitutions due to errors of the DNA polymerase, as well as artifactual recombinations resulting from template switching and unspecific hybridizations may occur.[19]

Finally, sequencing of RNA poses additional sources for errors. RNA or transcriptome sequencing requires additional steps prior to sequencing; first, RNA has to be isolated from genomic DNA in the sample. Subsequently, contaminating rRNA is removed by antisense oligos or selection of polyadenylated RNA. Finally, complementary DNA synthesis by reverse transcriptase is primed by oligo(dT) or random hexamer primers. These primers as well as gene length, GC content and dinucleotide frequencies may introduce additional biases.[20] Since RNA has to be reverse-transcribed for sequencing by the reverse transcriptase, an enzyme that usually lacks proof-reading ability, additional nucleotide substitutions may be introduced at this level.[21]

Reads generated by sequencing platforms are commonly annotated with quality scores that quantify the uncertainty of each particular base call.[22] However, the interpretation and calibration of these scores is often difficult and may not sufficiently reflect biases resulting from the sequence content.[23] In addition, quality scores cannot capture well errors associated with the placement of deletions read alignments commonly have to be post-processed.[24] In spite of these shortcomings, quality values are regularly employed in a first step in order to trim or remove unreliable sequence reads, thereby significantly reducing error rates of downstream analyses.[25] More involved quality control procedures are available, as for example removal of rare (and thus more likely to be erroneous) sequence fragments based on k-mer analysis.[26]

[16] Margeridon-Thermet et al. (2009), Mitsuya et al. (2008), Solmone et al. (2009)

[17] Archer et al. (2012), Nakamura et al. (2011)

[18] Balzer et al. (2011), Margulies et al. (2005)

[19] Eckert and Kunkel (1991), Kanagawa (2002), Liu et al. (1996)

[20] cf. Zheng et al. (2011) for an overview of the primary literature that discusses these biases

[21] Ji and Loeb (1994), Roberts et al. (1988)

[22] Ewing and Green (1998)

[23] Brockman et al. (2008), Dohm et al. (2008), Rieber et al. (2013), Ross et al. (2013)

[24] DePristo et al. (2011)

[25] Reumers et al. (2012)

[26] Schröder et al. (2009), Yang et al. (2010a)

Read analysis. Subsequent to error correction, quality-controlled reads are either aligned (*mapped*) to a reference genome or assembled *de novo* if no such reference is available. For each of these purposes, a variety of specialized software implementations are available that consider various phenomena such as transcriptome splicing, viral diversity, or different library types.[27] Based on comparing several alignments or assemblies as well as by incorporation of reference sequence information, genomic and transcriptomic variants such as SNVs, short indels, splice variants, and larger structural variations can be extracted from the alignment, in principle. These data can be further functionally annotated using protein structures and disease gene and pathway information in public databases, thus yielding clinically actionable results.[28]

Determination of genomic and transcriptomic variants poses unique challenges, many of them facilitated by mapping ambiguities, highly diverse genomes, or inaccuracies of reference sequences. With regard to viruses, identification of SNVs at individual positions (*SNV calling*) or multiple correlated positions on the same genome (*haplotype estimation*) are of foremost importance due to their clinical relevance for viral phenotypic traits. SNVs are usually identified based on the sequence alignment by counting base calls of multiple overlapping reads at a individual genome position. However, due to high viral diversity and considerable error rates of the sequencing process, true minority variants are often difficult to discern from technical artefacts, even at high sequence coverage. This difficulty has lead to the development of statistical models by the virology and oncology communities, both of which regularly deal with low-frequency SNVs and heterogeneous samples (cf. Chapter IV of this thesis and Beerenwinkel et al. 2012 for an overview of such models that are frequently used in virology).

Deep sequencing approaches in virology

Next-generation sequencing has been employed in various applications of basic and clinical virology, such as determining population diversity of viral species,[29] viral transmission of resistance variants,[30] and whole-genome sequencing of medically relevant viruses.[31] Chapter IV of this thesis is dedicated to exploring some of these medical applications in greater detail.

In addition to sequencing viral DNA genomes, gene expression patterns of viral RNA genomes, viral mRNA, and of viral microRNA are often determined to provide a more complete picture of the virome. In particular, large DNA viruses with many genes employ a complex array of coding and non-coding transcripts that are of medical interest.[32] Complementary to determining the genome and transcriptome of viruses, analyses of genomes and transcriptomes of infected hosts are currently also being pursued using deep sequencing technologies.[33]

As touched upon previously in Chapter I, the discovery of novel

[27] Baker (2012), Bao et al. (2011), Finotello et al. (2012), Hatem et al. (2013), Li and Homer (2010), Li et al. (2012), Nagarajan and Pop (2013), Oliver (2012)

[28] Pabinger et al. (2013)

[29] Eriksson et al. (2008), Wang et al. (2007)

[30] Varghese et al. (2009)

[31] Bimber et al. (2010)

[32] Ertl et al. (2011), Riley et al. (2012), Yang et al. (2010b, 2011b)

[33] Ge et al. (2009), Khalid et al. (2011), Woodhouse et al. (2010)

emergent viruses by metagenomics approaches plays a key role in preventive surveillance and timely reaction to viral outbreaks. For the latter, deep sequencing in addition to electron microscopy and non-specific PCR are especially versatile tools since they do not rely on prior knowledge about the pathogen being sought. As mentioned in Chapter I, recent outbreaks of haemorrhagic fever and epidemic infection of cattle with the Schmallenberg virus have been traced to their causative factors within days, thus allowing for rapid responses to the epidemic.[34] In a related manner, as will be expanded upon later in this chapter, analysis of host-viral metagenomes and metatranscriptomes comprise a current field of research that aims to identify clinically relevant viruses within patients with unknown disease etiology.[35]

In accordance with the specific aim of a given metagenomic study, two classes of metagenomics sequencing approaches are often distinguished: targeted, or *amplicon, sequencing* and *shotgun sequencing*.[36]

Metagenomics sequencing approaches. *Targeted* sequencing is based on selected amplification of known, evolutionarily conserved genetic marker sequences in a metagenomic sample. In order to afford a systematic and comprehensive view of a microbial community, this approach requires that each member species of the community is covered by at least one known genetic marker sequence known to the researcher. Individual marker genes conserved across all three domains of life such as 16S and 18S rRNA genes and factors required for nucleotide metabolism and protein processing are widely used for identification and taxonomic characterization of prokaryotic lineages.[37]

However, as discussed earlier, marker gene approaches are not amenable to characterizing viral populations which lack common genetic markers due to their polyphyletic evolutionary origin (cf. Chapter I). In addition, targeted approaches are limited to characterizing very limited sequence regions by design, confounding hindering functional investigation of complete microbial genomes. Therefore, while being relatively cost effective due to both the lower amount of genetic material required and the limited amount of involved computational processing, reliance on marker genes is prone to underestimating microbial diversity.[38] Consequently, targeted approaches are currently used for determining the taxonomic composition of metagenomic samples or for specialized clinical applications but not for more comprehensive functional studies of microbial communities.

As an alternative to targeted approaches, current metagenomics investigations usually employ *shotgun sequencing* that are not limited to specific marker regions but sequence the full nucleotide content of a given sample. While being more cost intensive than targeted approaches, shotgun sequencing is more generally applicable to investigating a broad range of taxonomic, genomic, and func-

[34] Briese et al. (2009), Hoffmann et al. (2012)

[35] Arron et al. (2011), Feldhahn et al. (2011)

[36] Dröge and McHardy (2012)

[37] Eisen (2007), Huse et al. (2008)

[38] Medini et al. (2005)

tional properties of microbial communities. Additionally, it is also amenable to the reconstruction of entire microbial (draft) genomes by *de novo* metagenomic sequence assembly. Due to the increased availability of deep sequencing and bioinformatics methodologies as well as resulting from the shifting aim of metagenomics approaches toward more detailed functional investigations, shotgun sequencing has become the method of choice of current investigations.

Human and viral metagenomics

The main advantage of the metagenomics approach over traditional mechanisms of culturing and analyzing individual organisms is the access to the vast majority of microbes that cannot be cultured given the current biological knowledge;[39] indeed, only about half of the known phyla of bacteria have at least one member species that can be cultured in the laboratory[40] and an estimated 99% of all microbial species cannot be grown *in-vitro* at all.[41] In addition to providing a more complete picture of the microbial world, metagenomics also allows for the investigation of both the taxonomic structure and mutual functional interdependence of whole microbial communities. As a consequence, metagenomics offers a relatively unbiased[42] view on the microbial world that is both wider and deeper than was previously possible.

Early metagenomics studies have focused on characterizing uncultured microorganisms in environmental samples in a context of basic research independently of the pathogenicity of these agents.[43] Metagenomics has since then been recognized as a critical component of the applied human health sciences, in particular epidemiology and medicine. Availability of deep sequencing methods has enabled targeted investigations of microbial diversity associated with human health[44] that revealed a large abundance of microorganisms as well as an intricate interplay between bacterial communities and their human host[45] (see Table 7.2 for a chronology of viral metagenomics studies).

While these interactions are usually non-pathogenic for individuals with an intact immune response,[46] the characterization of microbial communities of healthy humans is now actively pursued in order to define baseline communities whose deviations may be indicative of a variety of pathological disorders.[47] Such baselines have been measured within several human biomes[48] and first insights derived from such studies are currently being employed for designing human fecal transplants that are able to restore productive bacterial populations in diseased patients.[49]

Viral metagenomics. Throughout the history of medicine, the discovery of microbial agents, in general, and viral agents, in particular, followed (rather than preceded) the identification of the pathogenic condition itself. AIDS, poliomyelitis, liver cancer, and

Year	Discovery
2002	Marine viral metagenome
2003	Gut microbiome
2004	Marine sediment metagenome
2005	Novel human viruses detected in blood, plasma, and nasopharyngeal aspirates
2006	RNA viruses detected in seawater and human feces
2007	New polyomavirus discovered. First metagenome of virioplankton and soil published. First use of next-generation sequencing for detecting viruses in honey bees.
2008	Several metagenomes of fecal and water biomes, widespread use of next-generation sequencing
2009	Numerous viruses in potable water, human liver, pants, wild animals, and insects detected
2010	Metagenomes of chimpanzee and mosquito
2011	Viral metagenomes of pigs, simian monkeys, and turkeys

Table 7.2: *Major milestones in viral metagenomics.* Major milestones in environmental and animal viral metagenomics. From: Mokili et al. (2012)

[39] Amann et al. (1995), Hugenholtz (2002)

[40] Rappé and Giovannoni (2003)

[41] Pace (1997)

[42] While being less biased than traditional methods of molecular biology employed for characterizing microbial genes like PCR, it may be noted that metagenomics approaches still suffer from subtle biases (Morgan et al., 2010, Thomas et al., 2012) originating from sample preparation (Thurber et al., 2009) or amplification (Kim and Bae, 2011).

[43] Breitbart et al. (2002), Handelsman et al. (1998)

[44] Lipkin (2010), Minot et al. (2011), Ravel et al. (2011), Sullivan et al. (2011), Turnbaugh et al. (2007)

[45] Claesson et al. (2012), Faust and Raes (2012), Sears (2005)

[46] Breitbart et al. (2003)

[47] Fancello et al. (2012)

[48] Breitbart et al. (2008, 2003), Kim and Bae (2011), Minot et al. (2012, 2011), Reyes et al. (2010)

[49] van Nood et al. (2013)

cervical cancers were all described in the literature prior to detecting the causative viral agent. Traditionally, virologists depended on methods like inoculation (*anno* 1796), filtration and reinfection (1892), cell cultures (1909), and electron microscopy (1933) to infer the presence of pathogens, all of which exhibit considerable experimental biases. These traditional approaches were later succeeded by sequence-dependent methods such as PCR and microarrays[50] that relied on prior knowledge of a specific marker sequence of the pathogen and resulted in the discovery of several new HIV genotypes and the SARS coronavirus, respectively.

Epidemiological studies support the conjecture that many human pathogenic viruses remain to be discovered[51] and viral communities have received renewed attention as cofactors of cancer, as well as of autoimmune and degenerative disorders,[52] and causes of idiopathic diseases such as infantile diarrhea, influenza-like illnesses, and chronic fatigue syndrome.[53] Similar to prior studies of the human bacterial baseline microbiome, investigations aiming to characterize the human baseline virome of diverse human tissue samples as for example originating from blood, oral cavity, sputum, gut, and fecal matter were undertaken.[54] These investigations resulted in large proportions of 66% of sequence reads that had not been characterized before in public nucleotide archives.[55]

Disease	Nucleotide	Virus discovered	Technology	Reference
Lower respiratory tract infection	Both	Parvovirus, coronavirus	Sanger	Allander et al. (2005)
Human merkel cell carcinoma	RNA	Polyomavirus	454	Feng et al. (2008)
Diarrhea	RNA	Astrovirus, torque teno virus, norovirus, picobirnavirus, enterovirus, nodavirus	Sanger	Finkbeiner et al. (2008)
Fatal transplant-associated disease	RNA	Arenavirus	454	Palacios et al. (2008)
Hemorragic fever	RNA	Arenavirus	454	Briese et al. (2009)
Acute flaccid paralysis	DNA	Bocavirus, picornaviruses, circovirus, nodavirus, dicistroviruses	454	Victoria et al. (2009)
Encephalitis	RNA	Astrovirus	454	Quan et al. (2010)
Lower respiratory tract infections	RNA	Rhinovirus C	454	Lysholm et al. (2012)
Tropical febrile illness	DNA	Circovirus	Illumina	Yozwiak et al. (2012)

Table 7.3: *Recent literature in clinical metagenomics.* Selection of recent literature in clinical metagenomics that lead to discovery of a candidate pathogenic viruses. Derived from Fancello et al. (2012).

As already discussed in Chapter I, viruses are of polyphyletic descent and do not possess common genetic marker genes. Therefore, sequence-dependent approaches that rely on hybridization of a known marker sequence (e.g., using a primer or hybridization probe) to DNA or RNA contained in the sample are of only limited use for detecting novel viruses. While individual viral families can be identified by amplifying viral hallmark genes (VHG) that encode conserved viral proteins such as viral capsids or polymerases by intentionally unspecific methods as, for example, degenerative PCR,[56] such approaches are not applicable for sizable fractions of to the virome.[57] Additionally, the use of VHGs is confounded by horizontal gene transfer and viral variability that may delete or modify the selected VHGs and thus hindering detection.[58]

[50] Mokili et al. (2002), Wang et al. (2003)

[51] Woolhouse et al. (2008)

[52] Dalton-Griffin and Kellam (2009), Peterson et al. (2008), Relman (1999b)

[53] Mokili et al. (2012)

[54] Breitbart et al. (2008, 2003), Breitbart and Rohwer (2005), Willner et al. (2009, 2011)

[55] Minot et al. (2011), Reyes et al. (2010)

[56] Rose et al. (1998)

[57] Rohwer and Edwards (2002)

[58] Lawrence et al. (2002)

These restrictions on the use of early sequence-dependent methods for detecting novel viruses have led to the development of non-targeted methods[59] that do not presuppose knowledge about the pathogen being sought. These methods rely on extensions or innovations of existing microbiological protocols and are either suited for amplifying purified viral DNA without the need for specific primers (sequence-independent single-primer amplification and random PCR) or applicable to differentially analyzing the sequence content of two samples by comparative hybridization (suppression subtractive hybridization and representational difference analysis).[60] While applications of these methods have lead to the discovery of HTLV-1, Torque Teno virus, parvoviruses, coronaviruses, and polyomaviruses in clinical samples,[61] they have since then been replaced by shotgun metagenomics approaches featuring higher sensitivities and a more comprehensive view on the microbiome.[62]

Several of these shotgun metagenomics approaches were recently employed for identification of novel viruses in patients suffering from a range of adverse conditions. The human biomes under investigation included human feces, blood, and brain[63] and yielded several new, potentially pathogenic viruses (cf. Table 7.3 and Fancello et al. 2012). As a consequence of these successes, clinical metagenomics is now considered a premier tool for detecting novel human pathogens in clinical as well as in public health settings.[64]

[59] often also denoted as *sequence independent* methods; however, the term is misleading since these methods still rely on nucleotide sequences for detection.

[60] Ambrose and Clewley (2006), Chang et al. (1994), Froussard (1992), Reyes and Kim (1991)

[61] Allander et al. (2007, 2005), Chang et al. (1994), Fouchier et al. (2004), Nishizawa et al. (1997)

[62] Nakamura et al. (2009)

[63] Briese et al. (2009), Finkbeiner et al. (2008), McMullan et al. (2012), Palacios et al. (2008), Quan et al. (2007), Sullivan et al. (2011), Victoria et al. (2009)

[64] Anderson et al. (2003), Greninger et al. (2010), Rose et al. (1998), Staheli et al. (2011)

8 Computational metagenomics

Metagenomic approaches *generate large amounts of sequencing data that have to be processed by computational methods in order to come to conclusions about the distribution of microorganisms within the sample under investigation. As detailed earlier, the analysis of viral sequence data is complicated by high viral mutation rates and resulting divergence of viral metagenomic sequences from the known viral reference sequences. In addition, viruses are subject to considerable gene transfer and do not possess as common marker gene such as the 16S/18S rRNA genes. In combination, these confounds make the wealth of computational tools that have been developed for the analysis of bacterial metagenomics data less suitable for analyzing viral samples.*[1]

Read mapping and assembly

The analysis of shotgun metagenomics data falls under the domain of *sequence-based bioinformatics* and commonly includes a range of methods that are organized into computational pipelines. As indicated earlier, samples undergo purification and library preparation. The resulting library is then sequenced by one of the deep sequencing platforms discussed earlier.

The processing of the resulting sequencing reads is highly dependent on the specific aim of the metagenomics study. Human clinical metagenomics approaches usually remove known bacterial contaminants and and reads homologous to the human reference genome or transcriptome by means of read mapping or sequence clustering.[2] In particular, since many human pathogens are already sequenced, read mapping against these references often serves as the initial step for analyzing clinical metagenomics data.

In contrast, environmental and human baseline metagenomics approaches usually refrain from read mapping due to the sparseness of suitable reference sequence for most of the microbiota under investigation. These approaches instead often rely on *taxonomic binning* or on *de novo* sequence assembly in order to combine the short reads of current sequencing technologies in into longer sequence contigs[3] that allow for more specific taxonomic annotation. Contigs can then be further processed in order to construct longer sequence scaffolds that serve as draft genomes for species within

[1] Kunin et al. (2008), Raes et al. (2007)

[2] Bhaduri et al. (2012), Kostic et al. (2011)

[3] Contigs: overlapping nucleotide segments that represent a consensus region of the genome. The set of all contigs is the primary output of a sequence assembly.

the metagenomic sample and enable functional analyses and comparisons.

Sequence assembly is frequently considered the computationally hardest problem of sequence-based bioinformatics. It is influenced by a variety of biases and technical considerations too vast to be presented here but extensively reviewed in Baker 2012, Myers 2005, Paszkiewicz and Studholme 2010. In simple terms, assembly is the problem of reconstructing the whole genome (or transcriptome) of an organism from sequenced nucleotide fragments. Due to a variety of biological factors such as large genome size, short read lengths, genomic repeats, low complexity regions, contaminant organisms, sequencing errors, heterozygosity, ploidy, and uneven sequence coverage, an unique or even continuous assembly is currently only possible for small genomes. These confounding factors of sequence assembly are further complicated in metagenomic analyses where complex samples exceed available sequencing capacities, orthologue regions of different species merge into chimeric contigs, and uneven sequence coverage is the norm rather than the exception.

In contrast to assembly of metagenomics samples consisting mainly of bacteria, however, less ambiguous assemblies are produced from viral genomes due to the compactness and low number of repeats in the latter. In addition, high viral replication rates at least in clinical settings allow for sufficient amounts of sample, thus providing high sequencing coverage and contiguous assemblies of viral genomes.[4]

Most established assemblers developed for deep sequencing data are relying on varying formulations of Overlap-Layout-Consensus graphs and related string graphs on the one hand[5] or k-mer graphs (i.e., de Bruijn graphs) for assembly of DNA genomes and RNA transcriptomes,[6] the later of which is subject to additional confounds such as dynamic intron/exon structure as well as differential expression of transcripts. While the string graph can utilize reads of arbitrary length and is less vulnerable to sequencing errors than de Bruijn graphs, in principle, the reliance of string graph approaches on sequence alignments makes them inherently slower compared to de Bruijn graph approaches if no algorithmic preprocessing is undertaken.[7] In contrast, the de Bruijn graph data structure is not depending on sequence alignments and allows for the computationally efficient computation of sequence assemblies; in return, post-processing (*scaffolding*) and the use of multiple sequencing libraries with varying characteristics is required to make use of longer read lengths and to allow for panning of repeat regions[8] However, the memory requirements of de Bruijn graph approaches are high (up to 512 GB for eukaryotes with large genomes) since the size of the graph depends on the number of distinct k-mers in the genome, and read data generally contain sequencing errors that induce additional, artifactual k-mers. As a consequence, de Bruijn graph memory requirements tend to increase linearly with the number of sequence reads rather than with

[4] Yang et al. (2012)

[5] Myers (2005), Myers et al. (2000), Simpson and Durbin (2012)

[6] Butler et al. (2008), Grabherr et al. (2011), Li et al. (2010), Schulz et al. (2012), Simpson et al. (2009), Zerbino and Birney (2008)

[7] Simpson and Durbin (2012)

[8] Zerbino (2009)

the size of the genome if no error correction is applied, thus making high sequencing coverages as required for assembly of eukaryotic especially problematic.[9]

The memory problem is especially grievous for metagenomic assembly, as sufficiently capturing the genomic richness of complex samples requires extremely high sequencing depth. In order to accommodate confounds of metagenomic assemblies such as uneven coverage and cross-species chimeras, specialized algorithms have been developed that explicitly consider species distributions and uneven coverage[10] or implement novel data structures to decrease memory consumption.[11] Based on finished assemblies, absolute and relative species abundances as well as community diversity and structure of metagenomic samples can be estimated (cf. Fancello et al. 2012) while the taxonomic origin of each contig can be identified via taxonomic annotation methods.

Taxonomic annotation

Viral metagenomics firstly provided the biomedical sciences with an opportunity to systematically investigate both environmental and human biomes in order to identify emerging and established human pathogenic viruses. However, the identification of such novel pathogens is significantly impeded by our lack of knowledge about the virome. The majority of metagenomic sequences with any homology to GenBank records are assigned to 60 viral families (of the 96 viral families known according to ICTV), representing about equal numbers of RNA and DNA viruses.[12] However, these sequence hits exhibit only low sequence similarity (<50% amino acid similarity) and comprise genes associated with central metabolic functions rather than specifically viral genes.[13]

More critically, 51%–98% of the sequences produced by viral metagenomics studies that sampled the human biome did not have apparent homology to any entries in public nucleotide archives.[14] While a lesser part of these unknown sequences may be part of a *pangenome* of poorly conserved genes that occur in both cellular and viral genomes, many of these sequences are likely to represent signatures of unknown viruses[15] Interestingly, similar studies on bacterial habitats found only 10% unknown sequences, indicating that the extent of diversity and the amount of genes that cannot be classified might be significantly higher for viruses than for bacteria.[16]

Likewise, analyses on the level of reading frames (ORFs) showed that 30% of ORFs in the sampled viral genomes are not homologous to known eukaryotic or prokaryotic factors, thrice as many as in bacterial metagenomes and cultured bacteria.[17] Cumulatively, the 24 metagenomic studies published by 2011 have yielded 0.8 Gbp (70%) of ORFs with no homology to GenBank records.[18]

[9] Brown et al. (2012)

[10] Koren et al. (2011), Lai et al. (2012), Laserson et al. (2011), Namiki et al. (2012), Peng et al. (2012), Ye and Tang (2009)

[11] Bankevich et al. (2012), Boisvert et al. (2010, 2012), Chikhi and Rizk (2012), Conway et al. (2012), Simpson and Durbin (2012), Ye et al. (2011)

[12] Rosario and Breitbart (2011)

[13] Angly et al. (2006), Edwards and Rohwer (2005)

[14] Benson et al. (2006), Breitbart et al. (2003), Edwards and Rohwer (2005), Huson et al. (2009), Zhang et al. (2006)

[15] Breitbart and Rohwer (2005), Finkbeiner et al. (2008), Koonin and Wolf (2008), Kristensen et al. (2010), Lapierre and Gogarten (2009), Paul et al. (1993)

[16] Edwards and Rohwer (2005)

[17] Daubin and Ochman (2004), Yin and Fischer (2008)

[18] Rosario and Breitbart (2011)

Means of taxonomic annotation. A prerequisite for the identification and detailed functional and metabolic investigation of novel microbial organisms in metagenomic sequence data is a taxonomic annotation, or taxonomic *binning*, of sequence fragments.[19] Binning is commonly based on local sequence similarity of reads to entries in protein or nucleotide reference sequence databases such as NCBI RefSeq, nr, or nt[20] that have a known taxonomic identity. Additionally, sequence data under investigation are often correlated with phylogenetic information in order to derive the lowest common ancestor of taxa corresponding to several related taxonomic bins.[21]

Technically, the search for locally similar sequences in reference databases is conducted by either employing hash-based seed-and-extend methods such as BLAST[22] or methods derived therefrom and adapted to metagenomics data.[23] Alternatively, abstracted models of sequence similarity implemented via Hidden Markov Models that can detect specific marker genes or motifs with high computational efficiency are employed.[24]

As indicated before, databases of reference sequences capture only a small fraction of the microbial sequence space sampled by current metagenomics studies. As a consequence, many sequence reads cannot be taxonomically categorized using *similarity*-based methods that rely on the existence of similar sequences in nucleotide archives. In order to address this confound, *composition*-methods (also termed *alignment-free* methods) have been developed that categorize sequence reads not by similarity but by more approximate nucleotide signatures such as GC content and the over-representation of specific k-mers. These nucleotide signatures are shaped by biological factors as for example translational codon selection, context-dependent mutation pressures, and polymerase nucleotide incorporation biases that are hypothesized to be specific for microbial species or taxonomic clades.[25]

Composition-based methods for taxonomic annotation are computationally less demanding than similarity-based approaches and are additionally able to identify highly divergent species as well as novel clades. On the other hand, composition-based methods display reduced accuracy and, due their reliance on nucleotide count statistics that require sufficient read lengths to achieve high specificity, suffer more from the short read length of current sequencing technologies than similarity-based approaches. The major subclass of composition based methods employ supervised learning approaches that rely on comparing sequence reads to previously known genomic signatures; these methods require sufficient amounts of known reference sequences as training material which are not always available.[26] Consequently, both similarity-based and supervised composition-based approaches are affected by the low coverage of public reference archives.[27]

In practice, metagenomic binning, phylogenetic placement, as well as in-depth functional and pathway annotations of sequence

[19] Kunin et al. (2008), Simon and Daniel (2011)

[20] Sayers et al. (2012)

[21] Chatterji et al. (2008), Matsen et al. (2010), Mirarab (2012), Patil et al. (2011)

[22] Altschul et al. (1990)

[23] Ghosh et al. (2010), Huson et al. (2011), Monzoorul Haque et al. (2009), Xie et al. (2010)

[24] Finn et al. (2011)

[25] Deschavanne et al. (1999), Karlin et al. (1994, 1997), Perry and Beiko (2010)

[26] Patil et al. (2011)

[27] McHardy and Rigoutsos (2007)

reads are conducted in an integrated fashion by computational annotation frameworks. These frameworks internally employ local similarity-based methods (such as BLAST[28] and HMMs[29] or rely on genomic signatures of nucleotide composition using either supervised or unsupervised[30] machine learning methods (cf. Mande et al. 2012 for an exhaustive list of annotation frameworks).

Viral sequence annotation

Among the vast range of computational annotation and binning methods, only few are directly tailored to viral sequences. Viruses pose a special problem for taxonomic classification due to their polyphyletic origin, absence of common markers genes, high sequence divergence, extensive gene transfer, short genome, and the potentially high similarity between host and viral genomes in terms of both homologous genes and codon usage patterns.

Most similarity-based approaches to taxonomic annotation rely on hash-based local gapped alignments using variants of BLAST,[31] including comparisons of query and reference sequences in nucleotide space (BLASTn), protein space (BLASTp), query nucleotides translated to protein space (BLASTx), and reference and query nucleotides both translated to protein space (tBLASTx). Public BLAST reference databases offered by NCBI and EBI such as nt (nucleotide database) or nr (protein database) also include secondary reference sources as, for example, GenBank/RefSeq entries and Swissprot/PDB entries, respectively. These databases were recently supplemented by environmental sequence tags (env) resulting from metagenomics studies, making these databases the most comprehensive source for biological sequence information currently available.

Due to the ability of BLASTx and tBLASTx methods to bypass synonymous mutations during codon translation and thus identify functionally conserved homologs, these approaches are recommended for discovering remote similarities and are particularly well suited for identification of divergent viral species.[32] Bacterial metagenomics studies commonly post-process BLASTx results in order to better reflect taxonomic relations.[33] However, such approaches, while represented in the literature,[34] are of only limited applicability to viral data where taxonomic relations do not follow a Linnaean-like taxonomic hierarchy. In addition to directly comparing query and reference on the sequence level, more remote homologies or functional annotations of sequence regions can often be inferred by using abstractions of sequence motifs in terms of Hidden-Markov Models or position-specific scoring matrices that have been trained on reference protein sequences. However, although facilitating functional characterization of highly divergent query sequences, their reliance on the availability of protein reference sequences makes HMM models by Pfam, InterPro, PHI-Blast/Psi-Blast, or, more recently, HMM-FRAME[35] less well suited

[28] Meyer et al. (2008), Rosen et al. (2011), Su et al. (2011), Sun et al. (2011)

[29] Gerlach et al. (2009), Gerlach and Stoye (2011), Stark et al. (2010), Wu and Scott (2012)

[30] Brady and Salzberg (2009), Chatterji et al. (2008), McHardy et al. (2007), Teeling et al. (2004)

[31] Altschul et al. (1990)

[32] Kunin et al. (2008)

[33] Huson et al. (2007, 2009)

[34] Ghosh et al. (2011), Hanekamp et al. (2007), Kim et al. (2011), Yang et al. (2011a)

[35] Altschul et al. (1997), Bateman et al. (2003), Mulder et al. (2007), Zhang and Sun (2011)

for viral taxonomic annotation where protein references may be not available.

Several similarity-based approaches aim to address some of the aforementioned confounds of viral taxonomic annotation; however, these approaches presently are either limited to BLASTN/BLASTP similarity searches that are unsuited for detecting similarities between highly diverged viral genomes and their references,(36) or they employ only a limited range of known reference sequences without making use of the wealth of the full NCBI nr and nt nucleotide archives,(37) thereby potentially missing known viral genes. In a similar fashion, composition-based taxonomic classification methods that are currently in use by the metagenomics community(38) may be confounded by viral species that emulate the codon usage patterns of their hosts(39) or are not specifically trained or validated on viral genomes.(40)

In summary, therefore, there currently is no golden bullet for viral taxonomic annotation and the reliance on imperfect reference archives and incomplete viral taxonomies remains a hindering factor of viral metagenomics.

(36) Lorenzi et al. (2011), Wommack et al. (2012)

(37) Angly et al. (2005, 2009), Meyer et al. (2008)

(38) Bazinet and Cummings (2012), Dröge and McHardy (2012), McHardy and Rigoutsos (2007)

(39) Karlin and Burge (1995), Lucks et al. (2008)

(40) Trifonov and Rabadan (2010), Williamson et al. (2012), Willner et al. (2009)

9 Introduction to tumor viruses

The following conceptual interlude *changes the focus of this chapter from a bioinformatics topic, metagenomics, to a central theme of microbiology: oncogenic or tumor viruses.*[1] *As will be discussed in the next sections, the origins of the two scientific disciplines* virology *and* oncology *as well as many methods of microbiology are highly related to investigations of animal tumor viruses. Detection and characterization of oncogenic pathogens has important practical consequences for medicine and epidemiology. In addition, basic researchers are interested in tumor viruses due to the highly effective genetic control elements that these pathogens employ to modulate core cellular pathways such as replication, apoptosis, and immune response. These cellular pathways are not only utilized by viruses but are also frequently aberrant in malignant human cells, thus providing a conceptual bridge between virology and oncology.*

Retroviruses and discovery of cellular oncogenes

Soon after characterization of the first plant viruses, two landmark discoveries marked the beginning of both animal and tumor virology: first the result that avian leukemia, at that time not considered to be a cancer, can be caused by an unknown *filterable agent* of sub-bacterial size[2] that later was identified as the first *retrovirus*. Second, the evidence for infectiousness of another cancer, avian sarcoma, by inoculation of cell-free tumor extracts in healthy animals with the Rous sarcoma virus (RSV).[3] Subsequent discoveries revealed infectious oncogenic viruses in malignancies of several animals such as rabbits and mice[4] and indicated that tumor viruses encompass both RNA and DNA viruses of several viral families.

Fundamental to further research on oncogenic viruses was the development of quantitative transformation assays[5] which first facilitated a detailed investigation of several different RSV variants. These investigations demonstrated the distinctness of cellular transformation and viral replication, thereby suggesting that RSV variants contain an oncogenic gene dispensable for viral replication.[6] This gene, *src*, was later shown to stably integrate into the host genome and cause cellular transformation; as such, it constitutes the first identified *oncogene*. These results gave rise to the *provirus* theory of retroviral persistence that was later experimentally proven

[1] The terms *oncogenic* virus, *cancer* virus, and *tumor* virus are here used interchangeably. While the concept of *oncovirus* has roots in the study of cancer-causing RNA retroviruses, the term *tumor virus* has been predominantly employed by virologists targeting DNA cancer viruses. As both DNA and RNA viruses have later been shown to promote cancer by manipulating similar cellular pathways (although usually by different means), the semantic distinction implied by the different lexemes is void. An exception is the word *transforming virus* as it explicitly specifies pathogens that change the phenotype of a cell in an *in-vitro* environment rather than in an *in-vivo* environment which is required for true oncogenesis.

[2] Ellerman (1908)

[3] Rous (1973)

[4] Gross (1951), Rous and Beard (1935)

[5] which earned Rous the 1966 Nobel Prize for Physiology or Medicine

[6] Martin (2004), Temin and Rubin (1958)

in two landmark studies by David Baltimore and Howard Temin.[7]

While oncogenes were first believed to be entities of exclusively viral origin that are activated by somatic mutations or external carcinogens, it was soon shown that retroviruses such as RSV are able to *capture* (i.e., copy) proto-oncogenes from the host cell. These capturing retroviruses probably originated from non-transforming retroviruses, for example the Avian Leukemia virus that incorporated a cellular proto-oncogene into their genome during proviral integration and subsequent read-through transcription. Upon subsequent infections, the captured proto-oncogene may be integrated into the chromosomes of the same or other host cells, often either in a mutationally activated form or integrated into a cell type lacking transcriptional control for this oncogene. As a consequence, the integrated oncogene is either permanently activated or its products are phenotypically inhibiting its cellular homologues,[8] thus predisposing the cell for oncogenesis.[9]

Due to their capability to rapidly and efficiently cause cellular transformation *in vitro* as well as *in vivo*,[10] viruses containing cellular oncogenes such as RSV, Murine Leukemia virus and Feline Leukemia virus[11] were later termed *acutely transforming*. Importantly, capture of cellular oncogenes usually leads to deletions of essential genes within the viral genome, rendering the capturing virus replication incompetent and thus only able to initiate a single round of infection[12].

Historically, the investigation of captured cellular genes in transforming viruses was critical for understanding the molecular mechanisms of oncogenesis by activation of oncogenes.[13] As mentioned previously, not only was Rous Sarcoma virus the first discovered tumor virus but its transformative gene, *src*, was also the first oncogene to be identified.[14] While later studies on virally activated cellular proto-oncogenes uncovered many central elements of the cellular regulatory machinery involved in oncogenesis (see Table 9.1), it was shown that RSV is among the few acutely transforming viruses that retain replication competence after oncogene capture. Thankfully, however, such replication-competent acutely transforming viruses are extremely rare and no such retrovirus is known to infect humans.

Subsequent studies on non-viral activation of oncogenes in a range of tumors demonstrated that activating mutations both in oncogenes captured by viruses and in somatically mutated cellular oncogenes often occurred at identical sites, thus indicating common molecular mechanisms of oncogenesis.[15] Investigations of insertional mutagenesis of non-acutely transforming retroviruses affecting cellular oncogenes later substantiated this *cellular oncogene theory of cancer*, thus providing the first comprehensive theory on cancer genesis by combining etiologies associated with viral cofactors, with chemical and environmental carcinogens, as well as with somatic mutations[16] (cf. Table 9.2).

Class	*Proto-oncogenes*
Growth factor	sis
Kinases	erbB, fms, kit, abl, src, raf, akt
G proteins	H-ras, K-ras
Transcription factors	erbA, ets, myc, rel

Table 9.1: *Proto-oncogenes associated with retroviruses.* Proto-oncogenes discovered by studying retroviruses. From: Butel (2000), Javier and Butel (2008)

[7] Baltimore (1970), Duesberg and Vogt (1970), Temin (1964), Temin and Mizutani (1970)

[8] for instance, by being incorporated in cellular protein complexes and there having a dominant-negative effect on the functioning of the whole complex

[9] Huebner and Todaro (1969), Stehelin et al. (1976)

[10] Coffin (1997)

[11] Nevins and Vogt (1996)

[12] Interestingly, although replication-incompetent, acutely transforming tumor viruses may still retain continued invectiveness if the infected cell is co-infected with a related and replication competent virus, such as the Friend virus or wild type Avian Leukemia virus. These *helper viruses* provide critical services such as capsid assembly and genome packaging to the damaged tumor virus. Thus, the presence of helper viruses, while not having an oncogenic effect itself, may mask the activity of low-abundance tumor viruses.

[13] Bishop (1991)

[14] Stehelin et al. (1976)

[15] Der (1987), Der et al. (1982), Parada et al. (1982), Sukumar et al. (1983)

[16] Butel (2000)

DNA tumor viruses and discovery of tumor suppressors

After the discovery of viral capture mechanisms of cellular oncogenes and the rise of the oncogene theory of cancer, the study of human tumor viruses rapidly expanded to include DNA viruses such as *polyomaviridae*, *adenoviridae*, and *papillomaviridae*. It became soon clear that, in contrast to both rapidly and non-rapidly transforming retroviruses, oncogenes within transforming DNA viruses had a genuinely viral origin (*v-onc*), had multiple functional roles within viral infections, and were essential for viral replication. In particular, these viral oncogenes affected a newly discovered class of cellular proteins such p53 and Rb that modulate cellular transcription factor binding to control cell cycle progression, senescence, apoptosis, and DNA repair.[17]

p53 and Rb possess regulatory and signaling functions that have suppressive effects on oncogenesis and were later found to be mutated in over half of all cancers.[18] In contrast to cellular oncogenes that dominantly promote activation of the mitogenic pathways, tumor suppressor proteins have a recessive effect, inhibit cell cycle progression, and are activated by a range of conditions such as cellular stress (DNA damage), transformation (as indicated by aberrant proliferative signaling), or viral replication (as indicated by activation of innate immune system). DNA tumor viruses as, for example, adenoviruses, polyomaviruses, and herpes- and papillomaviruses counte the cellular antiviral response in a variety of ways by inhibiting tumor suppressor pathways with viral oncogenes, thereby increasing the permissiveness of the cell for viral replication and also predisposing it to oncogenesis.

1. Cellular origins of oncogenes: acutely transforming retroviruses carry oncogenes that are derived from cellular genes
2. Genetic basis of cancer: cancer may arise through activated cellular oncogenes that positively affect the mitogenic cycle
3. Multistep oncogenesis: multiple genomic events are necessary for cancer development and progression
4. Negative regulators of cancer: cancer may rise through deactivated cellular tumor suppressors that negatively affect the mitogenic cycle and may initiate apoptosis
5. Unification of etiologies: viruses, chemical carcinogens, radiation, and somatic mutations may cause cancer by affecting the same mitogenic pathway

Table 9.2: *Oncological discoveries based on virology.* Oncological discoveries based on virological experiments.

Direct and indirect mechanisms of oncogenesis

The currently known repertoire of human oncogenic viruses includes both large and small DNA viruses, as well as retroviral and non-retroviral RNA viruses that have either acutely or non-acutely transforming properties. Besides grouping viral pathogens by nucleic-acid type (RNA or DNA), tumor viruses are often classified by their *direct* or *indirect* oncogenic mechanisms. As will be discussed later in more detail, tumor viruses may induce cancer by one or several such oncogenic mechanisms either in parallel or in succession.

Many human oncogenic viruses such as HPV, MCV, EBV and KSHV contain a virus-derived oncogene, termed *v-onc*.[19] *v-onc*s are usually activators of cellular proto-oncogenes or deactivators of cellular tumor suppressors and constant *v-onc* expression is a necessary condition for maintaining the transformed cell state. While the targeted oncogene or tumor suppressor often requires additional mutations to become *activated*,[20] *v-onc* transcripts are expected to be present in each cancer cell and directly predispose the cell to oncogenesis. Consequently, viruses encoding a *v-onc* are denoted as *directly* oncogenic.

[17] Braithwaite and Prives (2006), Knudson (1971), Lane and Crawford (1979)

[18] Finlay et al. (1989), Greenblatt et al. (1994), Harris (1996), Vogelstein et al. (1989)

[19] Usually denoted *v-onc* in contrast to an activated cellular oncogene, *c-onc*

[20] Activated: here, this denotes the change from a proto-oncogenic to an oncogenic state

On the other hand, predominantly indirectly acting oncogenic viruses such as HBV, HCV, and retroviruses achieve transformation by more distal mechanisms such as insertional mutagenesis, prevention of apoptosis of pre-cancerous cells, opportunistic infections with other viruses by virus-induced immunosuppression, or oxidizing DNA damage by chronic inflammation. Indirect tumor viruses are not depending on the continued expression of a viral oncogene but influence the viral context in a more circumstantial manner.

While the distinction of directly and indirectly acting tumor viruses may serve as an useful organizing principle in tumor virology, several viruses such as human polyomaviruses, HBV, HCV, and HTLV-1 promote both direct and indirect mechanisms,[21] thereby indicating a continuum rather than distinct classes of viral oncogenic mechanisms.

A third order of oncogenic viral cofactors, here loosely denoted as *hit-and-run* viruses, also employs indirect oncogenic mechanisms whose exact definition is currently contested.[22] Hit-and-run viruses are hypothesized to predispose the cell to chromosomal instability by an array of molecular mechanisms that result in aneuploidy[23], large-scale chromosomal rearrangements, and genomic losses that may promote oncogenesis.[24] In the hit-and-run model of oncogenesis, oncoviral factors are essential for the initial transformation but may be lost from progressed tumors due to activation of additional cellular oncogenes and selection of growth-autonomous cells that make the original viral factor dispensable.

The following section will expand on these topics and investigate in more detail the exact molecular mechanisms of tumor virus transformation.

[21] Jeang et al. (2004), Seeger and Mason (2000), Tsai and Chung (2010), Yasunaga and Matsuoka (2007)

[22] Gallagher et al. (2003)

[23] *Aneiploidy*: Abnormal number of chromosomes within a cell.

[24] Duensing et al. (2000), Hein et al. (2009)

10 Molecular mechanisms of tumor virus transformation

Both cancer cells and oncogenic viruses *modulate mitogenic, apoptotic, and immune pathways in order to influence the cell cycle and promote rapid replication. Importantly, however, oncogenesis is not an integral part of the life cycle of tumor viruses. Instead, it is to be considered as an accidental side effect of long term molecular interactions of a persistent virus with limited or lost replication competence with the host cell machinery. As briefly discussed in previous sections, known human tumor viruses predispose cells to oncogenesis by manipulating host cellular pathways and the cellular environment by a broad range of mechanisms. These mechanisms can be summarized in two broad classes of molecular events: (1) by direct means and mostly enacted by oncogenes of viral origin that modulate the host cellular machinery in order to promote replication and lessen immune surveillance*[1] *(see Table 10.1), or (2) by indirect means such as insertional mutagenesis at host genomic loci near cellular oncogenes, modulation of the cellular inflammatory context, epigenetic effects, and inducement of chromosomal instabilities.*[2] *The oncogenic etiology of tumor viruses is strikingly different between these two means of oncogenesis as well as between simple retroviruses, complex retroviruses, and DNA viruses, as will be illustrated in the following section.*

1. Promotion of genetic instability and DNA damage, for instance in HPV (Duensing and Munger, 2003) and EBV (Liu et al., 2004)
2. Cell immortalization and telomere lengthening, for example in EBV (Sugimoto et al., 2004) and KSHV (Verma et al., 2004)
3. Control of mitogenic and tumor suppression pathways, as in HPV (Scheffner and Whitaker, 2003) and KSHV (Friborg et al., 1999)
4. Inhibition of intrinsic and extrinsic apoptosis pathways, in particular caspase activation (EBV Grimm et al. (2005), KSHV Matta et al. (2003))
5. Modulation of the cell micro-environment by cytokine and growth factor production causing inflammation (EBV Maggio et al. (2002))
6. Evasion of immunological surveillance by episomal and proviral latency or inhibition of host immunoregulatory molecules such as interferones, interleukins, and MHCs Coscoy and Ganem (2000)

Table 10.1: *Molecular pathways modulated by tumor viruses.* Molecular pathways modulated by tumor viruses that may lead to oncogenesis.

Tumor viruses modulate mitogenic and immune pathways

Cellular proto-oncogenes code for gene products that regulate cell growth and differentiation by modulating mitogenic pathways. Known classes of oncogenes include growth factors, tyrosine/serine/threonine kinases, GTPases, and transcription factors.[3] Due to a variety of mutational events such as single nucleotide polymorphisms or indels within the coding sequence, insertional misregulation in coding or promoter sequences, genomic aberrations such as translocations, copy number variations, and gene fusions, as well as modifications of regulatory transcription factors and miRNAs,[4] these genes can be *activated* and become tumor inducing agents. In contrast to the strict definition of cellular oncogenes as positive regulators of the mitogenic pathway that may induce cancer if mutated or misexpressed, the concept of virally derived oncogenes is used in a considerably broader sense and

[1] Matsuoka and Jeang (2007)

[2] Mikkers and Berns (2003)

[3] Coffin (1997)

[4] Esquela-Kerscher and Slack (2006), Negrini et al. (2007)

includes all genes that code for protein factors that modulate mitogenic, apoptotic or or immune pathways in ways that predispose the cell to malignant transformation.

Proto-oncogenes are positive regulators of the cell cycle and their activation usually has a dominant oncogenic effect (that is, *in vitro* transformation or *in vivo* oncogenesis occurs if at least one allele of the proto-oncogene is activated).[5] In contrast, tumor suppressors such as *p53* or *rb* that are engaged DNA repair, repression of cell devision, and triggering of apoptosis are negative regulators of the cell cycle and their deactivation by means similar to oncogene activation has a recessive effect, requiring at least two transformative mutations or "hits" in order to trigger oncogenesis.[6] The realization that multiple genetic events are necessary for cancerous transformation gave rise the theory of multi-step oncogenesis and it is now accepted that discrete genetic events cooperatively and successively confer survival advantages and thus clinical malignancy to the cancerous cell.[7]

As mentioned in previous sections, advances in the technology of culturing cells that can be infected with viruses lead to the possibility of employing viruses as biological model systems for oncology.[8] Based on the identification of the first cellularly derived oncogene in retroviruses[9] and followed by the realization that mutations in cellular proto-oncogenes are causative for most somatic cancers,[10] experimental investigations resulted in the discovery of more then 70 cellular proto-oncogenes[11] (see Table 10.2 for a list of selected examples). In addition, these studies revealed one of the major innovations of the viral world: the reverse transcriptase.[12]

Many viral factors modulate core cellular regulatory and defense mechanisms in order to increase the viral replicative time window, for example by inhibiting or modulating tumor suppressor proteins such as p53 and Rb to fixate the cell cycle and prevent apoptosis, or by inhibiting antiviral defense mechanisms such as cellular endonucleases.[13] While all these events predispose the cell to oncogenesis by enhancing proliferation of aberrant cells, these events are not sufficient for oncogenesis and further conditions such as somatic mutations, environmental risk factors, and immunosuppression are generally considered to be necessary for cancer formation.[14]

Viral genomes frequently contain cellularly or virally derived genes that facilitate persistence of the virus by decreasing recognition of viral factors by components of the adaptive and innate immune system, such as chemokines, cellular proliferation factors, and inhibitors of apoptosis (cf. Chapter III for a discussion of viral strategies of immune evasion in a context of viral host factors).[15] Especially larger DNA viruses, as for example KSHV and Poxvirus, employ several passive and active strategies to evade the host immune system (see Table 10.3). These strategies induce prolonged immunosuppression of infected cells and result in cellular stress as well as in unhindered modulation of mitogenic and apoptotic pathways, thereby creating an ideal background from which cancer

Virus	Viral oncogene	Cellular factor
HBV	HBx	p53, CBP/p300, PI3K
HCV	Core, NS3, NS5A	p53, pRb, TNFR, TBP, PI3K
EBV	LMP-1	TRAFs, TRADD, RIP, vimentin, JAK3
HPV	E5-E7	p53, TNFR, pRb, p21, CycA, CycE
HTLV-1	Tax	pRb, CBP/p300, p21, CycA, CycE
HIV	Tat	CycT1, TFIIH, P-TEFb, PKR, pCAF, CBP/p300, TAFII250
KSVH	vIRF, vGCR, Kaposin	CBP/p300

Table 10.2: *Frequent cellular targets of viral oncogenes.* Investigation of frequent targets of viral oncogenes lead to the identification of cellular proto-oncogenes. Only selected cellular factors are shown. From Boccardo and Villa (2007).

[5] Stehelin et al. (1976)

[6] Knudson (1971)

[7] Bishop (1991), Weinberg (1989)

[8] Butel (2000), Temin and Rubin (1958)

[9] Duesberg and Vogt (1970), Martin (2004)

[10] Der et al. (1982), Parada et al. (1982)

[11] Butel (2000)

[12] Baltimore (1970), Temin and Mizutani (1970)

[13] Javier and Butel (2008), McLaughlin-Drubin and Munger (2008), Teodoro and Branton (1997)

[14] McLaughlin-Drubin and Munger (2008)

[15] Smith (1994)

can arise.[16]

Similarly, the replication of some oncogenic viruses is controlled in healthy patients but increases to pathogenic levels in contexts of immunosuppression as, for example, induced by HIV infections or medication after organ transplants. While an intact immune system can remove rare neoplastic cells, immunosuppressed and persistently infected patients may harbor a critical mass of cells with virus-induced decreased immune surveillance, thus making successful proliferation and immune escape of cancerous cells possible. In addition, primary infections that favor survival of altered cells and promote opportunistic infections may act cooperatively with a subsequent infection of an oncogenic driver virus, as for example shown for HIV/KSHV and EBV/KSHV.[17] KSHV probably infects the whole genus *homo*; however, the cases of Kaposi's sarcoma historically had only low incidence rates until the onset of the AIDS epidemic, after which the number of KSHV-associated cancer cases in immunosuppressed patients rose by several thousand percent[18]

Oncoviral persistence and latency

Most human pathogens trade off between two strategies of replication: high rates of reproduction and rapid viremia,[19] or long term persistence in the host and control of the host immune system.

Viral strategies of replication and persistence While most acute human pathogens such as Influenza, Noroviruses, and Rhinoviruses rapidly induce viremia and achieve high replication rates in order to infect other hosts, viruses establishing persistent infections in a host population have a higher chance of vertical transmission and require molecular mechanisms in order to constantly evade host immune clearance. Bacterial and protozoan organisms as well as many DNA viruses utilize the large protein repertoire of their comparatively large and complex genomes in order to achieve continued immune evasion by exploiting regulatory (e.g. by inhibiting host factors) or immuno-combinatorial (e.g. by changing surface receptors) mechanisms.[20] In contrast, RNA viruses such as HCV and HIV have compact genomes that allow for rapid replication. In combination with high mutation rates and large viral burst sizes that rapidly create novel genetic variants, these pathogens are able to evade immune recognition and establish chronic infections.[21]

Finally, both DNA viruses (EBV) and RNA-viruses (HIV, HTLV-1) often support lysogenic as well as lytic replication cycles and may employ episomal[22] and proviral[23] latency mechanisms as part of the lysogenic cycle. These latency mechanisms, augmented by epigenetic modifications of the viral genomes, may limit viral transcription or remove viral genomes from the cytoplasmic context in order to lessen immune surveillance and allow for long-term (or even permanent) viral persistence.

1. Restriction of viral transcription by proviral, episomal, or epigenetic latency mechanisms
2. Infection of cellular compartments that are under lessened immune surveillance such as brain or kidney
3. Variation of the viral genome by error-prone viral polymerases to decrease antibody and T-cell recognition of viral antigens
4. restriction of host major histocompatibility complex expression by viral inhibitors
5. Inhibition of antigen processing and presentation
6. Infections of immune cells

Table 10.3: *Viral modulation of the host immune system.* Viral strategies to modulate the host immune system in order to support viral persistence. Derived from Butel (2000).

[17] Sun et al. (2005), Trivedi et al. (2004)

[18] Engels et al. (2006), Miller et al. (1996)

[19] *Virimia*: Medical condition where viral particles have access to the bloodstream.

[20] Centurion-Lara et al. (2004)

[21] Johnson and Desrosiers (2013)

[22] An *episome* is a viral genome or parts thereof persisting as linear or lariat (lasso-shaped) structures in the cellular cytoplasm or nucleus independently and physically separated from the host genetic material.

[23] Retroviruses insert their genome into the host genome as an obligatory part of their life cycle. After viral entry, the viral RNA genome is reverse-transcribed into a double-stranded DNA molecule by a viral enzyme, *reverse transcriptase*. The double-stranded viral DNA is integrated into the host genome by the viral *integrase* enzyme. The integrated copy of the viral genome is termed a *provirus* and would resemble cellular genes if it were not under transcriptional control of a specific viral promoter element, the long terminal repeat (LTR). The integration site of the integrated genome, the provirus, is essentially random, although preferences for specific sequence motifs or chromatin states have been reported (Deichmann et al., 2011).

Latency-associated mechanisms of oncogenesis In contrast to animal retroviruses as for example Rous sarcoma virus, Harvey's murine sarcoma virus, and simian sarcoma virus that induce cancer by integrating cellularly derived oncogenes such as *src*, *H-ras*, and *sis* into the cellular context,[24] transforming retroviruses that lack cellularly derived oncogenes are typically replication competent and undergo polyclonal integration into the host genome as part of the viral life cycle.

Occasionally, proviral integration may occur in or near a cellular proto-oncogene or tumor suppressor, thus modulating expression of the cellular factor by the integrated viral promoter and enhancer elements, presumably even over large distances as a consequence of chromatin looping.[25] Alternatively, viral integration may occur within a cellular factor, thus creating an activated fusion transcript.[26]

In either case, integration can disrupt the genomic organization and regulation of the proto-oncogene or tumor suppressor locus, thus predisposing the host cell to oncogenesis. Such mechanisms of oncogenic activation that are a direct consequence of proviral integration are termed *cis* activation of the cellular oncogene. They are contrasted with oncogenic mechanisms depending not on the integration event itself (although integration may be part of the viral life cycle) but on expressed viral factors that modulate cellular pathways in a manner similar to viral oncogenes of DNA tumor viruses. The latter mechanism is termed *trans* activation and is predominantly employed by complex retroviruses such as HIV and HTLV-1.[27]

Oncogenesis is circumstantial. The ability for prolonged persistence, either by active immune evasion, chronic infection, or viral latency is a critical characteristic of all human tumor viruses as it provides a necessary precondition for inducing an oncogenic state in the host cell.[28] Viral persistence may be further enforced by infections of cell types not permissive to viral replication, cross-species infections (zoonosis), viral genomic deletions (acutely transforming retroviruses), and infections of immunocompromised hosts, all of which may prevent cell lysis and thus successful production of new viral particles.

This self-limiting replicative behavior of tumor viruses is one of the hallmarks of tumor virology and strongly suggests that transformative behavior is not part of the normal viral life cycle. Instead, inducing malignant transformation is a detrimental biological accident for tumor viruses, either as a side effect of viral manipulation of mitogenic and apoptotic pathways aimed at facilitating viral replication,[29] or as a consequence of long-term proviral, episomal, or chronic persistence.[30]

[24] Der et al. (1982), Devare et al. (1983), Stehelin et al. (1976)

[25] Maeda et al. (2008), Mikkers and Berns (2003), Moloney (1960), Neel et al. (1981)

[26] Hayward et al. (1981)

[27] Matsuoka and Jeang (2007)

[28] zur Hausen (2009b)

[29] Levine (2009)

[30] Moore and Chang (2003)

Tumor viruses promote oncogenesis by indirect mechanism

In addition to directly oncogenic mechanisms enacted by viral oncogenes and indirect activation of cellular oncogenic factors by proviral integration, many tumor viruses also influence the cellular context, DNA damage response, or epigenetic pathways and thus predispose the cell to cancer in a more distal manner. The indirectness of these molecular mechanisms combined with the temporal delay and stochastic nature of oncogenesis results in considerable uncertainties regarding their characterization. Lytic viral activity is generally not observed in progressed cancer cells due to changes in replicative permissiveness or damages to the viral genome.[31] Instead, viral genomes either continue to latently persist in the transformed cell, thus further evading immune detection and replicating whenever the cancer cell divides[32] or are fully lost from progressed tumor cells. The latter event, termed ’hit-and-run’ signature of viral activity,[33] is conceptually related to virally induced oncogenesis that does not require continued maintenance of nucleotide signatures in the transformed cell and thus is especially hard to trace by molecular methods.

DNA damage and genomic instability. Most tumor viruses such as adenoviruses, EBV, Hepatitis B and C viruses, herpesviruses, papillomaviruses, HTLV-1, KSHV, and polyomaviruses manipulate cellular DNA damage response pathways in a variety of ways in order to increase replicative capacities, support viral latency, or protect viral genomes from degradation (reviewed inWeitzman et al. (2010)). As a side effect, these manipulations may result in chromosomal instability, aneuploidy, large-scale chromosomal rearrangements, and genomic losses that promote oncogenesis.[34] Additionally, oncogenic viruses such as HBV and HCV that cause chronic infection implicitly modulate the cellular context by virus-specific T cells mediated inflammation or chronic immune response, resulting in mitochondrial damage and DNA double strand breaks due to production of oxidative chemical agents,[35] for instance as a result of viral inhibition or metabolic depletion of cellular antioxidants such as glutathione that also exhibit antiviral activity.[36] Last, retroviruses may induce mobilization of endogenous transposons or produce genomic breakpoints by polyclonal viral integration, both of which may also result in chromosomal instabilities.[37]

Epigenetic modifications. In additional to promoting genomic instabilities, virally mediated epigenetic modifications of the viral and human genomes may also directly promote oncogenesis: several epigenetic markers such as DNA methylation and acytelation, histone modification, and miRNA signatures are important factors of oncogenesis.[38] DNA methylation of CpG islands, for instance, is a common cellular mechanism for regulation of expression and is employed in genomic imprinting, X-chromosome inactivation, and

[31] Small et al. (1982)

[32] Frenkel and Roizman (1972), Lilley et al. (2007), Moore and Chang (2003)

[33] McLaughlin-Drubin and Munger (2008), Si and Robertson (2006)

[34] Duensing et al. (2000), Duensing and Munger (2002), Elgui de Oliveira (2007), Gruhne et al. (2009), Hein et al. (2009), Kamranvar et al. (2007)

[35] Gallagher et al. (2003), Levrero (2006), Machida et al. (2006)

[36] ; and Staal, F. J. F., Ela, S. W. S., Roederer, M. M., Anderson, M. T. M., Herzenberg, L. A. L., and Herzenberg, L. A. L. (1992). Glutathione deficiency and human immunodeficiency virus infection. *Lancet*, 339(8798):909–912

[37] Nabirochkin et al. (1998), Sung et al. (2012)

[38] Esteller (2008), Jones and Baylin (2002)

silencing of foreign genomic elements.[39] As discussed previously, human tumor viruses frequently establish latent infections and thus have to evade immune surveillance. DNA methylation has been proposed as a viral regulatory mechanisms that restricts expression of viral elements associated with an immune response[40] (see Table 10.4 for a list of known and suspected tumor viruses that show evidence for employing epigenetic regulation).

Epigenetic mechanisms for silencing viral genomes are also employed by the host cell in order to limit expression of endogenous virus-like elements such as retrotransposons as well as for constraining proviruses that may negatively impact genomic stability.[41] Several viruses are able to specifically recruit or inhibit these cellular epigenetic mechanisms for their own purposes.[42] Tumor viruses such as HSV1, EBV, KSHV and adenoviruses, for instance, employ epigenetic mechanisms to regulate expression of viral latency proteins that are recognized by cytotoxic T cells[43] or more generally influence histone acetyltransferase activity,[44] while herpesviruses such as KSHV specifically methylate promoters of host genes involved in immune surveillance.[45]

As a consequence of these early insights in viral epigenetic processes, antiviral therapy that nonspecifically removes genome methylation in order to unmask HIV proviruses in latent reservoirs has been proposed.[46]

Name	Genome	Family	kbp	Methylation	Viral oncogenes	Known or suspected cancers
EBV	dsDNA	Herpesviridae	172	Partial	LMP1, BZLF1, EBNA2, EBNA3, BRLF1	Burkitt's lymphoma, nasopharyngeal primary carcinoma, Hodgkin disease, gastric carcinoma
KSHV	dsDNA	Herpesviridae	138	Partial	LANA, vIRFs	Kaposi's sarcoma, primary effusion lymphoma, multicentric Castleman's disease
HPV	dsDNA	Papillomaviridae	8	Complete	E2, E6, E7	Cancer of cervix, vulva, vagina, penis, anus, orial cavity, oropharynx, and tonsil
HBV	partial dsDNA	Hepadnaviridae	3	Complete	HBx	hepatocellular carcinoma
HCV	ssRNA	Flaviviridae	10	None	NS4B	hepatocellular carcinoma, non-Hodgkin lymphoma
HTLV-1	ssRNA (RT)	Retroviridae	9	Partial	Tax	adult T-cell leukemia, lymphoma
HIV	ssRNA (RT)	Retroviridae	9	partial	indirect	indirect
HCMV	dsDNA	Herpesviridae	230	partial	IE1-72, IE2-86	colorectal cancer, glioma, prostate
SV40	dsDNA	Polyomaviridae	5	None	T-Ag	Osteosarcoma, mesothelioma, brain
JCV	dsDNA	Polyomaviridae	5	None	T-Ag	Brain, colorectal, glioma, medulloblastoma
BKV	dsDNA	Polyomaviridae	5	None	T-Ag	Prostate, brain
MCV	dsDNA	Polyomaviridae	5	None	?	Merkel cell carcinoma

Table 10.4: *Known and suspected cancer viruses.* Known and suspected cancer viruses and their methylation status. Derived from Fernandez and Esteller (2010) and Sarid and Gao (2011).

[39] Doerfler (1991), Feinberg et al. (2002), Herman and Baylin (2003), Payer and Lee (2008)

[40] Butel (2000), Fernandez et al. (2009), Verma et al. (2007)

[41] Colot and Rossignol (1999), Doerfler (1991), Yoder et al. (1997)

[42] Ferrari et al. (2009), Flanagan (2007), Javier and Butel (2008)

[43] Fernandez et al. (2009), Knipe and Cliffe (2008), Pantry and Medveczky (2009), Robertson and Ambinder (1997)

[44] Ferrari et al. (2008), Nyborg et al. (2010)

[45] Di Bartolo et al. (2008), Shamay et al. (2006)

[46] Richman et al. (2009)

11 Epidemiology of tumor viruses

Today, several infectious agents *have been identified as either cause or contributing factor of human cancers. The four major cancer viruses HBV, HCV, HPV, and EBV alone are causative for approximately 12% all human cancer cases,*[1] *with an estimated 26% of all cancer cases in developing countries being caused by an infectious agent.*[2] *The known human oncogenic viruses are distributed throughout complex retroviruses,* Flaviviridae, *as well as all known DNA viral families, with exception of* Parvoviridae *that may have a too limited genomic architecture to influence the cell cycle.*[3] *While epidemiology is frequently employed to detect novel tumor viruses, the stochastic nature of oncogenesis, temporal delays between oncoviral infection and the diagnosis of cancer, as well as the commonness of many oncoviral infections (such as with HPV and Herpesviruses) make establishing a causal relation between cancer incidence and viruses challenging. This section will discuss the epidemiological state of known oncogenic viruses as well as problems with establishing causal relations based on epidemiological data. In this manner this section sets a stage for the next section that will introduce molecular methods for discovery of new tumor viruses.*

Known and novel human tumor viruses

The concept of tumor viruses received renewed attention in the 1960s when the first human oncogenic agent, Epstein–Barr virus (EBV), was discovered by electron microscopy (EM) within Burkitt lymphomas (BL).[4] Since then, the carcinogenic potential of viruses from the five taxonomic families *Herpesviridae*, *Retroviridae*, *Hepadnaviridae*, *Flaviviridae*, and *Papillomaviridae* is recognized as sufficiently evident by the International Agency for Research in Cancer (IARC) to designate these viruses as group I carcinogens. In addition, HIV-2, MCV, and several additional human papillomaviruses have been classified as possibly (IARC group 2A/B) carcinogenic to humans and both adenoviruses and simple retroviruses have been suspected of potential oncogenic activity (see Table 11.1). Infections by all these viral vectors are accountable for approximately 15% of all human cancer cases and constitute the second major preventable cancer risk factor after tobacco use[5] (see Table 11.2 for relative cancer incidence rates of cancers with known oncoviral cofactors).

Viral family	Oncogenic	Model
Adenoviridae	potential	yes
Flaviviridae	confirmed	no
Hepadnaviridae	confirmed	yes
Herpesviridae	confirmed	yes
Papillomaviridae	confirmed	yes
Polyomaviridae	confirmed	yes
Retroviridae (simple)	potential	yes
Retroviridae (complex)	confirmed	yes

Table 11.1: *Viral families associated with oncoviral activity.* Viral families associated with oncoviral activity and presence of an animal model system for the association. From Butel (2000), Javier and Butel (2008)

.

[1] Parkin et al. (2005), zur Hausen (2006)

[2] Parkin (2006)

[3] Butel (2000)

[4] Epstein et al. (1964)

[5] Boccardo and Villa (2007)

Virus	Fractions	Number of cases (in thousand)
HPV	3% (Mouth)-100% (Cervix)	6.3 (Oropharynx) - 492.8 (Cervix)
HBV	Liver 54%	340
HCV	Liver 31%	195
EBV	46% (Hodgkin's lymphoma) - 98% (Nasopharyngeal carcinoma)	6.7 (Burkitt's lymphoma) - 78.1 (Nasopharyngeal carcinoma)
KSHV	Kaposi's sarcoma 100%	66.2
HTLV	Adult T cell leukaemia/lymphoma 2%	3.3

Table 11.2: *Prevalence and number of cases of virally induced cancers.* Prevalence and number of cases of virally induced cancers. From Schiller and Lowy (2010)

Importantly, the recent detection of several polyomaviruses with suspected oncogenic potential indicates that there still are human cancer-causing viruses to be identified.[6] Notably, lung cancers in never smokers, neuroblastoma, diffuse large B-cell lymphoma, childhood acute lymphocytic leukemias, as well as some instances of breast cancer and genital cancers are suggestive of viral cofactors.[7] Additionally, viruses such as the Mouse Mammary Tumor Virus (MMTV), Polyomaviruses, Adenoviruses, and the Torque-Teno virus (TTV) are currently being investigated for their potential of inducing human oncogenesis (see Table 11.3). The following paragraphs will discuss some of these potentially oncogenic viral families in more detail.

Virus	Cancers	Reference
JSRV	lung cancer	Felini et al. (2012), Sun et al. (2007)
TTV	Colon cancer, pancreatic cancer, liver cancer, lung cancer, cervical cancer	zur Hausen (2012), zur Hausen and de Villiers (2009)
MMTV	Breast cancer	
JCV	Colon cancer, anal cancer	

Table 11.3: *Suspected cancer-virus associations currently being investigated.* Suspected cancer-virus associations currently being investigated. From Sarid and Gao (2011)

MMTV. MMTV is a retrovirus that induces breast cancer by insertional mutagenesis in mice and is often transmitted to the offspring through breast milk.[8] Viral particles similar to MMTV identified in high-risk human breast cancer patients were shown to exhibit reverse transcriptase activity and MMTV sequences could be identified in over half of breast cancer biopsies according to several studies.[9] However, the exact identity of the possible MMTV homologue is still debated and it is presently unclear if it constitutes evidence for a zoonotic event, a human MMTV homologue, or lab contamination.[10]

Polyomaviruses. Human polyomaviruses are particular interest for studies that aim to identify possible new oncogenic cofactors. It has been suggested that replication-competent polyomaviruses that procreate at high multiplicities and are transmitted during consumption of red beef produce replication-incompetent, carcinogenic variants at a low rates that may be a cofactor of childhood leukemias and colon cancer.[11] Similar to EBV, the human polyomaviruses Jamestown Canyon virus (JCV) and BK virus (BKV) are very common in humans and may infect up to 90% of the adult human population, thus constituting a viral commensal.[12] While infections with polyomaviruses usually remain asymptomatic, both JCV and BKV contain an oncogenic T (tumor) antigen and have been associated with triggering disease in immunocompromised patients,[13] the latter fact being a general indicator for potential oncogenic activity of viruses. In spite of ongoing research on more specific associations of these viruses with human tumors, only JCV

[6] zur Hausen (2009b)

[7] zur Hausen (2009a,b)

[8] Callahan (1996)

[9] Etkind et al. (2000), Ford et al. (2004a,b), Moore et al. (1969), Schlom et al. (1971)

[10] Lawson et al. (2001), Mant and Cason (2004, 2005), Mant et al. (2004a,b), Stewart et al. (2000), Szabo et al. (2005)

[11] zur Hausen (2009b)

[12] Gardner et al. (1971), Padgett et al. (1971), Shah et al. (1973), Weber and Major (1997)

[13] Eash et al. (2006)

could yet be loosely associated with brain and gastric tumors.[14]

A third human polyomavirus, simian virus 40 (SV40), infects up to 20% of the human population and encodes a known viral oncogene, the large T antigen that initiates viral DNA synthesis and inactivates the tumor suppressor proteins p53 and pRb in order to promote DNA synthesis.[15] Most human tissues are permissive to SV40 replication and undergo lysis, thereby preempting transformation. An exception is presented by human mesothelial cells that may be non-permissive to efficient viral replication and lysis and are therefore prone to virally induced oncogenesis as a result of chronic, low-level infection.[16] However, while SV40 is indeed found in many malignant mesothelioma tumors and *in-vitro* models, the evidence for the oncogenic activity of the virus is conflicting[17] and the role of SV40 as a human carcinogen remains controversial.

Adenoviruses and HERV-Ks. Human adenoviruses are viral commensals that may cause respiratory illnesses in children as well as conjunctivitis and gastroenteritis. While these viruses harbor two known viral oncogenes and are associated with malignant tumors in rodents, no human cancers are currently known to be caused by adenoviruses.[18] Notably, also endogenous retroviral elements of the HERV-K family have characteristics that are suggestive of oncoviral activity. These viruses are still able to code for complete if non-infectious viral particles that can rendered infectious by removal of stop-codons.[19] Increased HERV-K expression was observed in breast cancer,[20] HIV-associated lymphomas, non-HIV-associated lymphomas, and HIV-associated Hodgkin's lymphomas,[21] possibly indicating a role for these viruses in oncogenesis.

Role of epidemiology in detecting tumor viruses

The search for novel cancer viruses is primarily guided by prioritizing relatives to known tumor viruses or cancers with certain epidemiological characteristics for detailed investigation. In particular, cancers that are related to immunosuppression or that are localized in specific geographic regions as in the case of HBV may point to an infectious carcinogen.[22] These kinds of investigations have characterized the epidemiological profile of both KSHV and MCV ahead of the detection of the infectious agent and are currently persued to identify novel cancer viruses in immunosuppression-related tumors and cancers related to red meat consumption.[23]

Identification of tumor viruses by means of epidemiology is confounded by the fact that interactions between oncoviral and host factors can result in a variety of outcomes, from no apparent effects to viral lysis, apoptosis, or cancer.[24] In addition to this stochastic nature of oncogenesis, several factors significantly hamper the identification by epidemiological means; among these factors are the

[14] Enam et al. (2002), Imperiale (2001), Krynska et al. (1999)

[15] Butel and Lednicky (2000), Vilchez and Butel (2004)

[16] Gazdar et al. (2002)

[17] Rizzo et al. (1999), Shah (2007), Shivapurkar et al. (1999)

[18] Berk (2005), Elgui de Oliveira (2007)

[19] Dewannieux et al. (2006), Ruprecht et al. (2008)

[20] Wang-Johanning et al. (2008)

[21] Contreras-Galindo et al. (2008)

[22] Chen et al. (1991), Grulich et al. (2007)

[23] Beral et al. (1990), Engels et al. (2002), Schulz (2009), Vajdic and van Leeuwen (2009), zur Hausen (2012)

[24] Roulston et al. (1999)

long clinical latency of cancer, low viral replication in progressed tumors, the ubiquity of infections with viral commensals, the requirement for further mutations in the host and viral genomes, as well as the influence of other external mutagens that may act synergistically with infections to cause cancer.(25)

Due to these difficulties, epidemiological insights are primarily used for prioritizing suspected tumor viruses for experimental research. In this manner, several oncogenic viruses have been identified by observing epidemiological indicators such as geographic coincidence of viral infections and specific cancers (HBV), regional clustering of cancer cases that may suggest an infectious cause (EBV), dependency on sexual contacts or blood transfusion (HPV, HCV), or cancers arising under immunosuppression (EBV, KSHV, HPV).(26)

Viral commensals. The high abundance of viruses in the human microbiome suggests that, contrary to both the common concept of viral infections and the focus of medical research on pathogens, most viruses inhabiting humans are not pathogenic. Indeed, many viruses initially suspected to be transfusion contaminants such as Torque Teno virus (TTV) and Hepatitis G virus (HGV) were later shown to be highly abundant in several human organs and are now considered to be viral commensals that may provide advantages to the host by restricting prokaryotic replication.(27) While uncommon viruses such as KSHV and HTLV-1 are specifically present in a particular cancer and are thus comparatively simple to identify as oncogenic cofactors,(28) viral commensals are non-specifically present in afflicted as well as healthy populations and are therefore notoriously difficult to causally relate to any disease state.

Stochasticity and clinical latency of oncoviral transformation. As discussed in previous sections, oncogenic retroviruses are classified into acutely and non-acutely transforming viruses. Acutely transforming retroviruses contain an oncogene derived from a captured cellular proto-oncogene, are typically replication defective, and rapidly induce tumors. Non-acutely transforming retroviruses, on the other hand, do not encode a cellular-derived oncogene but may activate cellular proto-oncogenes through insertional mutagenesis, inducing tumors with long clinical latency due to the inherent stochasticity of the integration process. RNA and DNA viruses that lead to cancer due to chronic infection are also associated with similarly long clinical latency until cancer onset.

As touched upon earlier in this chapter, the long clinical latency and the low rate of transformation of virally induced human cancers indicate that virally caused oncogenesis in human is not part of the replication strategy of tumor viruses. Instead, viral oncogenesis is a consequence of either rare events (such as insertional mutagenesis) or long-term exposure to weakly viral oncogenic factors, both of which result in comparatively slow carcinogenic kinetics.(29)

(25) zur Hausen (2009b)

(26) zur Hausen (2009b)

(27) Bernardin et al. (2010), Willner et al. (2009)

(28) Sarid et al. (1999), Yasunaga and Matsuoka (2007)

(29) Nevins and Vogt (1996)

In addition, oncogenesis is a stochastic multi-step process that requires additional somatic mutations or genomic aberrations for the cell to transform and evade immune control; while viral infection may be essential for the first steps of oncogenesis mainly by providing the cell with growth advantages, immortality, or lessened immune surveillance, these viral factors are not sufficient for fully malignant transformation. However, the initial advantages conferred by the tumor viruses may provide the basis for additional transforming events which cumulatively and auto-catalytically enhance oncogenesis.

Given the raw number of commensal viruses and the number of infections they cause in large portions of the population, persistent viral commensals are interesting candidates for triggering such low-probability transforming events. Similarly, viruses that arise in a context of immunosuppression may transform cells that are not under immune surveillance, thereby increasing the chance of oncogenesis even at low rates of infection. Indeed, all known human cancer viruses are either ubiquitous viral commensals or are targeting immunocompromised subgroups such as patients suffering from AIDS or having received organ transplants.

Tumor viruses and causal inference. In addition to the ubiquitousness of viral commensals, the stochastic nature of oncogenesis, and the long clinical latency of cancer, causal reasoning regarding viral-cancer associations is confounded by three further factors that generally apply to all pathogens. First, any causal inference pertaining to disease-pathogen relationships is usually based on inductive reasoning, i.e., generalizations based on particular observables that might be misleading or incomplete in specific cases. Second, human diseases often are multi-factorial and include many variables such as length and type of pathogenic exposure, state of the host immune system, infected cell type, and particularities of the genetic variation in host and pathogen.[30] The interactions of these variables are difficult to assess in any scheme of causal reasoning. Finally, prospective studies based on pathogen exposure are often impossible due to ethical or practical limitations, thereby hampering systematic approaches of causal inference (i.e., experimental testing).

Traditionally, the causal relationship between a pathogen and a disease is considered proven beyond reasonable doubt by adhering to Koch's postulates[31]. For instance, the causal links between several viruses and human pathogenic conditions such as HCV with hepatitis,[32] and HTLV-1 with Kaposi's sarcoma[33] have been successfully established using this scheme. More recently, advances in microbiological techniques, the realization that many pathogens cannot be cultured[34] or may be present at lower abundances in healthy subjects, and the fact that pathogens may cause the pathogenic condition after considerable temporal delay led to several reformulations of Koch's postulates that emphasize

[30] Relman (1999a)

[31] *"(i) The parasite occurs in every case of the disease in question and under circumstances which can account for the pathological changes and clinical course of the disease. (ii) The parasite occurs in no other disease as a fortuitous and non-pathogenic parasite. (iii) After being fully isolated from the body and repeatedly grown in pure culture, the parasite can induce the disease anew."*, original in Rivers (1937), quotation from Fredericks and Relman (1996)

[32] Kuo et al. (1989)

[33] Moore and Chang (1995)

[34] Amann et al. (1995)

biomolecular methodology and relaxed criteria for the consistency of pathogen-disease association.[35] The reformulations presently culminated in the *Metagenomic Koch's postulates*.[36]

Disproving virus-cancer associations. In spite of these reformulations of Koch's postulates, causally relating tumor viruses to a cancer remains a task that is both conceptually and methodologically challenging.[37] The causal relation between infection and disease state is especially hard to prove in cases where several viral commensals are infecting the same disease tissue, or the virus is only a contributing factor for the disease, as it is the case for many human tumor viruses.[38] Although many novel viruses have been identified in animals or humans afflicted with a specific disease, most of these viruses were not causally related to the disease but constituted viral commensals or chance associations.[39] For instance, although several viruses were associated with multiple sclerosis, none of these infections could be conclusively proven to be causally related to the disease.[40] In a similar manner, XMRV was long thought to be causative for chronic fatigue syndrome and some instances of prostrate cancer until this association was shown to be based on experimental confounds.[41] It is therefore notable that falsifying spurious and false associations of diseases with specific pathogens can be significantly more difficult than proving true associations.[42] This is especially the case in contexts of limited treatment options where possible viral infections constitute a straw of hope for patients that they cannot let go of; this may be exemplified by amyotrophic lateral sclerosis that has wrongly been linked to enteroviruses, or the temporal pattern of age of onset (or diagnosis) of autism that coincides with measles, mumps and rubella (MMR) vaccine administration, events that have been falsely associated and later economically exploited by certain individuals.[43]

Treatment of oncoviral infections

Global prophylactic vaccination against viral infections such as polio, measles, and smallpox are among the most successful medical endeavors ever undertaken. Currently, infections of only two human tumor viruses, HPV and HBV, can effectively be prevented by vaccination. These vaccinations lead to long-term reductions in cancer incidents in the immunized populations that constitute conclusive proof for the oncogenic effect of these viruses.[44]

As discussed previously, virus-mediated oncogenesis is a rare event and most cancer viruses act as co-carcinogens rather than deterministically triggering cancer at each infection; indeed, only KSHV and HPV are now considered to be necessary causes for Kaposi's sarcoma and cervical cancer, respectively, while other oncoviral infections mostly remain asymptomatic and contribute only slightly to the total cancer risk.[45] In the vast majority of cases, tumor viruses within cancer cells have neither increased transmis-

[35] Evans (1976), Falkow (1988), Fredericks and Relman (1996), Hill (1965)

[36] (i) The diseased metagenome is required to be significantly different from the metagenome constructed with the same sample type obtained from a healthy matched control subject. (ii) The suspected metagenomic traits must be present and more abundant in the diseased subject compared to matched control. (iii) Inoculating a healthy individual with a sample from a diseased subject must result in disease state. (iv) Differential metagenomic traits in step (i) recovered in the newly induced diseased subject may be the biomarker of the candidate etiological agent. (v) Selective inoculation of samples from the disease subject (ii) must induce disease in another healthy control subject if the metagenomic contains the trait associated with the etiological agent of the disease, or phenotype under investigation. Adapted from Mokili et al. (2012)

[37] Pagano et al. (2004), zur Hausen (2002)

[38] This epidemiological challenge has been encountered before when smoking was investigated as a possible risk factor for lung cancer. Since smoking was widespread in the population at the time of the study and lung cancer is commonly diagnosed well after exposure to smoking, large cohorts and prospective studies where required to obtain the necessary statistical power for proving increased cancer risk in smokers (Doll and Hill, 1956)

[39] Bexfield and Kellam (2011)

[40] Münz et al. (2009)

[41] Paprotka et al. (2011)

[42] cf. Lipkin (2013); this parallels the well known mathematical fact that existence proofs are much easier to provide than non-existence proofs.

[43] Whose scientific work or names will not be mentioned here so as to not perpetuate the ghost of false association and scientific wrongdoing.

[44] Kao and Chen (2002), Taira et al. (2004)

[45] Moore and Chang (2010)

sibility nor higher replication fitness than viruses infecting healthy tissues. Instead, cancer is a ’dead end’ for cancer-causing viruses that are often replication incompetent due to mutations in their genomes. As a result of this propensity of tumor viruses against producing large numbers of new viral particles, long-term viral persistence is often asymptomatic, i.e., it elicits only transient or limited immune response and very limited viremia. It is primarily due to this asymptomatic behavior that oncoviral infections are difficult to diagnose and patients are often clinically presented after onset of cancer. Also, as a consequence of their resilient nature and limited interaction with the host, persistent viral infections are inherently hard to treat with antivirals; consequently, medical treatment is often limited to preventive measures in susceptible populations.(46)

Sadly, identifying new cancer viruses does not necessarily lead to increased prevention of oncoviral infection or to vaccine development: in spite of the past successes of HBV vaccination, the development of the recently approved HPV vaccines was hindered by significant political and economical pressures. Due to limited profitability, vaccines against the first known human cancer virus, EBV, and against the leading cofactor to adult cancer in Africa, KSHV, are not expected to be developed anytime soon. Since vaccination is by far the best means to fight infectious diseases and chemo- and radiation therapies are not available in many regions of the world, oncogenic viruses will prove to be significant burdens to public health for years to come at least in low-income nations.

(46) Cannon et al. (1999), Little and Yarchoan (2003)

12 *Tumor viruses and viral metagenomics*

As discussed previously, *identification of tumor viruses and their causal relation to cancers faces several methodological and conceptual difficulties. This section reviews metagenomics approaches for identification of these elusive pathogens in deep sequencing data and provides the background for the last section of this chapter that introduces a novel approach for detecting signatures of tumor viruses in human cancers.*

Pathogen-focused viral metagenomics

At the same time as the first studies on environmental viral metagenomics[1] that characterized viral diversity of ecological domains such as sea water, soil,[2] and a variety of human environments[3] were carried out, (see earlier sections of this chapter) investigations of animal and plant metagenomes were undertaken with the expressed aim to systematically identify known and novel viral pathogens (see Bexfield and Kellam 2011 or Table 12.1 for an overview).

One of the first of these sequence-based studies[4] introduced the *subtractive approach*[5] to pathogens discovery in human EST data by aligning sequence fragments to reference entries in public databases.[6] Several other subtractive approaches expanded on this strategy and used suffix-tree (or related Burrows-Wheeler transform) short-read mapping technologies,[7] or hash-based read mappers to sequentially align query reads to human and microbial reference sequences in order to detect signatures of known pathogens.[8]

Assembly-based apporaches. An interesting alternative approach to sequence-based detection of viral factors employs small interfering RNA (siRNA),[9] a species of host RNA that is used as part of the host innate immune system to post-transcriptionally silence genes of intracellular pathogens. Since by virtue of their mode of molecular action siRNA are highly enriched in sequences homologous to microbial genomes, the sequencing of siRNA provides a high signal-to-noise ratio for identifying new viruses. Using either mapping-based approaches to detect previously known pathogens or assembly-based approaches to reconstruct novel viruses from the

[1] Breitbart et al. (2002)

[2] Angly et al. (2006), Breitbart et al. (2004, 2002), Desnues et al. (2008), Dinsdale et al. (2008), Rice et al. (2001), Williamson et al. (2008)

[3] Briese et al. (2009), Finkbeiner et al. (2008), McMullan et al. (2012), Quan et al. (2007), Sullivan et al. (2011), Victoria et al. (2009)

[4] Weber et al. (2002)

[5] *Subtractive approaches* to sequence-based pathogen discovery identify and discard human sequence homologs from the sequenced data and consider the remaining transcripts as potential viral signatures. They stand in contrast to *de novo sequence assembly* approaches that aim to reconstruct whole viral genomes from overlapping reads

[6] Zhang et al. (2000)

[7] Li and Durbin (2009), Li et al. (2008)

[8] Kostic et al. (2011), Moore et al. (2011)

[9] Hamilton (1999)

Library	Sequencing	Strategy	Reference database	Reads	Mapping	Annotation	Assembly	Tissue	Reference
RNA Tag	Sanger	Subtraction	Various human	7×10^3	BLAST	BLAST	NA	HeLa	Weber et al. (2002)
RNA	Sanger	Subtraction	Various human & microbial	2.7×10^4	BLAST	BLAST	NA	PTLD	Xu et al. (2003)
RNA	Sanger	Subtraction	Refseq DNA & Transcriptome	2.6×10^5	BLAST	BLAST	NA	sCCC	Feng et al. (2007)
RNA	Roche-454	Assembly	Genbank	5×10^4	NA	BLAST	CAP3	Bee guts	Cox-Foster et al. (2007)
RNA	Roche-454	Subtraction	Human Refseq genome and transcriptome	3.8×10^5	BLAST	BLAST	NA	Merkel Cell Cancer	Feng et al. (2008)
RNA	Roche-454	Assembly	Genbank	1×10^5	NA	BLAST	CAP3	Various tissues	Palacios et al. (2008)
DNA Tag	Illumina	Subtraction	Virtual tag library from microbial DNA	8.9×10^6	Exact match	BLAST	NA	Various cancers	Duncan et al. (2009)
siRNA	Illumina	Assembly	Genbank	3×10^6	NA	BLAST	Several DNA	Infected plants	Kreuze et al. (2009)
siRNA	Illumina	Assembly	NCBI NR	6.5×10^6	NA	BLAST	Velvet	Fruitfly, Worm	Wu et al. (2010)
RNA	Illumina	Subtraction	Human genome fusion transcription	1×10^7	BLAT	BLAST	NA	cuSCC	Arron et al. (2011)
siRNA	Illumina	Assembly	NCBI NT	1.7×10^7	NA	BLAST	CLC Bio	Mosquito	Ma et al. (2011)
RNA	Illumina	Subtraction	Human genome and transcriptome viral and microbial refseq	6×10^6	MAQ	BLAST	Velvet	Various human tumors and simulated	Kostic et al. (2011)
siRNA	Illumina	Subtraction	Human genome	1×10^7	BWA	BLAST	Velvet	HIV-positive controls	Isakov et al. (2011)
RNA	Illumina	Subtraction	Human genome and NCBI references	4.4×10^7	BWA	Novoalign	NA	Simulated reads and one positive-control virus	Moore et al. (2011)
RNA, DNA	Illumina	Subtraction	Human genome transcriptome NR	1.8×10^8	BLAT	BLAST	NA	Dengue serum samples and positive controls	Yozwiak et al. (2012)
RNA	Illumina	Homology	Human genome NT NR	2.6×10^8	GSNAP	BLAST	Trinity	Neuro-blastoma	Schelhorn et al. (2013)

Table 12.1: *Approaches to systematic pathogen detection.* Approaches to systematic pathogen detection in animal and plant species. The column "Reads" designates the number of reads sequenced in the study.

sequenced siRNA, studies successfully recovered RNA and DNA viruses from disease vectors such as mosquito, as well as from several model organisms, honey bees, and crop plants.[10] Based on these successes, more detailed investigations of RNA samples of patient populations with idiopathic syndromes of acute viral infections[11] as well as of immunosuppressed patients suffering from AIDS and organ transplants[12] were undertaken, the latter of which resulted in the identification of a novel human pathogenic arenavirus. However, the use of siRNA-based approaches requires specialized sequencing libraries and is thus not amenable to analyses based on general RNA-Seq libraries.

Confounds of oncoviral metagenomics

The search for human tumor viruses poses particular challenges to sequence-based viral discovery pipelines. As discussed in previous sections, virological confounds such as latent viral replication strategies, viral replication incompetence, and loss of viral genetic material from the cell often result in limited or selective transcription of oncoviral genomes.[13] Low concentration and extratumoral location of viral producer cells[14] or selection of growth-autonomous cells in progressed tumors[15] can significantly dilute the number of viral transcripts in a sample. Additionally, transcription of human oncogenic factors modulated by viral[16] or endogenous[17] retroviral promoters as well as 'hit-and-run' mechanisms of viral oncogenesis that imply loss of viral material[18] may predispose cells to transformation without requiring maintenance of viral transcripts. Known tumor viruses such as high-risk HPV strains, EBV, and MCPyV, which selectively transcribe their genome during viral latency[19] (HPV: E6/7, EBV: EBNA1/2, MCPyV: large T antigen) and generate low abundances of tens (MCPyV) to hundreds (KSHV, EBV)[20] of transcripts per cell that require especially sensitive methods of detection.

In addition to virological confounds affecting detection sensitivity, sequence-based approaches to pathogen detection are also hampered by varying degrees of sequence similarity between sequence reads and human and viral references sequences. Specifically, viral oncogenes homologous to human factors and chimeric transcripts originating from proviral insertion sites may share significant sequence similarity with human transcripts,[21] thus making unequivocal identification of viral factors difficult. In addition, high rates of viral sequence divergence from 10^{-5} to 10^{-8} (dsDNA viruses) up to 10^{-4} (ssRNA viruses) substitutions per site and year[22] may mask true similarities between sequence reads and viral references, thereby further hindering recognition of known viruses.

[10] Cox-Foster et al. (2007), Kreuze et al. (2009), Ma et al. (2011), Wu et al. (2010)

[11] Patowary et al. (2012), Yozwiak et al. (2012)

[12] Isakov et al. (2011), Palacios et al. (2008), Simbiri et al. (2010)

[13] zur Hausen (2006)

[14] zur Hausen (2009b)

[15] Voisset et al. (2008)

[16] Coffin et al. (1997b)

[17] Ono et al. (1986), Tomlins et al. (2007)

[18] McLaughlin-Drubin and Munger (2008), Si and Robertson (2006)

[19] Dyson et al. (1989), Feng et al. (2008), Houben et al. (2010), Kelly et al. (2009), Klein et al. (2007), Scheffner et al. (1990), Young and Rickinson (2004)

[20] Cornelissen et al. (2003), Feng et al. (2008), Metzenberg (1990)

[21] Butel (2000)

[22] Duffy et al. (2008), Firth et al. (2010)

Systematic approaches to oncoviral metagenomics

In the last years, several systematic viral metagenomics studies have investigated human cancers or cancer-related diseases in order to detect putative viral cofactors. These systematic approaches are primarily based on shotgun deep sequencing technologies(23) and identified several cancer-virus associations. Among the viruses thus identified are MCPyV, a human polyomavirus, as a cofactor of Merkel cell carcinoma.(24) In addition, analysis of transcriptomes of Lymphoproliferative disorders(25) resulted in identification of EBV, expressed sequence tag (EST) libraries of body cavity-based lymphoma cells yielded the detection of KSHV fragments,(26) and comparison of human-derived EST tag sequences with samples obtained from several human tumors found Human Herpesvirus 6 signatures.(27)

Finally, investigations of several idiopathic human cancers that were likely to involve a viral cofactor but did not result in identification of any human pathogens provided important negative proofs of cancer-virus associations. Among these studies are the the analysis of cutaneous squamous cell carcinoma and metastatic melanoma that were based on physiological or epidemiological evidence for a viral cofactor but yielded negative results(28) as well as high-throughput investigations of thousands of tumors (not including human neuroblastoma) from the Cancer Genome Atlas that involved near a trillion RNA-Seq reads and also did not result in novel cancer-virus associations.(29)

Currently, the three published pipelines PathSeq, RINS, and CaPSID(30) undertake to automatically identify known and, to a limited extend, novel human pathogens in human deep-sequencing samples.

(23) Bexfield and Kellam (2011)

(24) Feng et al. (2008)

(25) Xu et al. (2003)

(26) Feng et al. (2007)

(27) Duncan et al. (2009)

(28) Arron et al. (2011), Feldhahn et al. (2011)

(29) Khoury et al. (2013), Tang et al. (2013)

(30) Bhaduri et al. (2012), Borozan et al. (2012), Kostic et al. (2011)

Preliminary conclusion

In conclusion, this chapter has presented a novel field of research, metagenomics, and a class of experimental technologies, deep sequencing, the combination of which allows for the first time to elucidate the complete genomic content of natural habitats. One of these habitats, the human body, is of particular importance for identifying a highly elusive type of pathogens: human tumor viruses. These entities have a unique biology and their manifold interactions with human factors as well as their potential to explain the genesis of several types of cancers with unknown etiology make these entities highly interesting for biomedical research.

13 Detecting tumor viruses in human cancers

AS DISCUSSED AT LENGTH *in previous sections, detection of tumor viruses in cancer tissues is challenging, both due to the high amount of sequencing data, the low expected abundance of tumor viruses within the data, and the unknown identity of these viruses. This section presents work that aims to extend the aforementioned approaches to systematic viral metagenomics by addressing several perceived shortcomings of the established methods and to achieve best-in-class sensitivity at detection of divergent viruses and taxonomic annotation while at the same time improving on processing speed and the detection of human-viral homologs such as cellularly derived viral oncogenes.*

Apart from formatting changes and few stylistic corrections such as shifting citation marks to the end of sentences for better readability, the manuscript presented here is identical to the published version.[1] It may be mentioned that since publication of the manuscript, the analysis of neuroblastoma data has been further extended by the author to more than 200 transcriptome samples and 30 whole-genome samples. The propositions made within the manuscript also hold given the new data.

Acknowledgements. The study was supported by a DFG grant (Klinische Forschergruppe, KFO 129) and was the result of a collaboration between Laura Toloşi and Thomas Lengauer (MPI Informatics, Saarbrücken), Matthias Fischer and Frank Berthold (University of Cologne Department of Pediatric Oncology and Hematology), Janine Altmüller and Peter Nürnberg (Cologne Center for Genomics), as well as Herbert Pfister (Institute of Virology, University of Cologne). Within the context of the investigations presented here, Thomas Lengauer was further partially supported by grant No. 01GS08103 (Oncogene) of the German Federal Science Ministry. The Children's Cancer Research Fund Neuroblastoma (Fördergesellschaft Kinderkrebs-Neuroblastom-Forschung) supported the deep sequencing analysis (awarded to Matthias Fischer and Frank Berthold). Writing of the manuscript as well as all implementations and analyses were performed by Sven-Eric Schelhorn. An initial prototype of the study was conducted by Sven-Eric Schelhorn and Laura Toloşi.

[1] Schelhorn et al. (2013)

Introduction

Abstract. In excess of 12% of human cancer incidents have a viral cofactor. Epidemiological studies of idiopathic human cancers indicate that additional tumor viruses remain to be discovered. Recent advances in sequencing technology have enabled systematic screenings of human tumor transcriptomes for viral transcripts. However, technical problems such as low abundances of viral transcripts in large volumes of sequencing data, viral sequence divergence, and homology between viral and human factors significantly confound identification of tumor viruses.

We have developed a novel computational approach for detecting viral transcripts in human cancers that takes the aforementioned confounding factors into account and is applicable to a wide variety of viruses and tumors. We apply the approach to conducting the first systematic search for viruses in neuroblastoma, the most common cancer in infancy. The diverse clinical progression of this disease as well as related epidemiological and virological findings are highly suggestive of a pathogenic cofactor. However, a viral etiology of neuroblastoma is currently contested.

We mapped 14 transcriptomes of neuroblastoma as well as positive and negative controls to the human and all known viral genomes in order to detect both known and unknown viruses. Analysis of controls, comparisons with related methods, and statistical estimates demonstrate the high sensitivity of our approach. Detailed investigation of putative viral transcripts within neuroblastoma samples did not provide evidence for the existence of any known human viruses. Likewise, *de novo* assembly and analysis of chimeric transcripts did not result in expression signatures associated with novel human pathogens. While confounding factors such as sample dilution or viral clearance in progressed tumors may mask viral cofactors in the data, in principle, this is rendered less likely by the high sensitivity of our approach and the number of biological replicates analyzed. Therefore, our results suggest that frequent viral cofactors of metastatic neuroblastoma are unlikely.

Introduction. To date, pathogenic agents are known to be causally related to 20% of human cancer cases and significantly affect the global health burden of this disease.[2] The majority of these agents comprise oncogenic viruses such as human papilloma virus (HPV), Epstein-Barr virus (EBV), hepatitis B virus (HBV), and hepatitis C virus (HCV).[3] Characterizing the oncogenic potential of viral pathogens has important consequences for prevention, diagnosis, and treatment of malignant neoplasms.[4] Tumor viruses in particular have received renewed attention in the context of recent global efforts to characterize the etiology of cancer.[5] Consequently, viral cofactors for several idiopathic cancers are currently investigated and epidemiological indicators suggest that additional human tumor viruses remain to be discovered.[6]

[2] Moore and Chang (2010), Parkin (2006)

[3] Sarid and Gao (2011)

[4] Schiller and Lowy (2010), zur Hausen (2006)

[5] International Cancer Genome Consortium et al. (2010), zur Hausen (2012)

[6] Javier and Butel (2008), zur Hausen (2009b)

Neuroblastoma is a heterogeneous embryonal tumor[7] that is accountable for 15% of deaths caused by malignant conditions in children.[8] The disease is associated with an exceptionally low median age of presentation of 17 months[9] and is often diagnosed *in utero*. Metastatic neuroblastoma has two biologically divergent subtypes. Stage 4S is characterized by an age of presentation between *in utero* and 18 months, metastases confined to liver, skin, lymph nodes and bone marrow, and its ability to regress spontaneously; in contrast, stage 4 tumors are presented at any age, demonstrate high infiltration rates in bone marrow and bone, and are most often progressive.[10] While genes related to neuronal differentiation have been described to be upregulated in stage 4S in comparison to stage 4 neuroblastoma, thereby indicating distinct levels of neuronal differentiation,[11] little is currently known about the differences between molecular etiologies of stage 4 and stage 4S neuroblastoma.

The variation of clinical outcomes between neuroblastoma subtypes indicates distinct genetic and environmental factors affecting the development of this malignancy. Interestingly, the early onset of the disease overlaps with periods of high susceptibility to viral infections and is reminiscent of acute lymphoblastic leukemia – another pediatric tumor with uncertain etiology for which an infective cofactor has long been suspected.[12] Furthermore, epidemiological studies have associated reduced neuroblastoma risk with immunologic indicators such as previous childhood infections, day care attendance, and breast feeding that are suggestive of an infective cofactor.[13] While transforming polyomaviruses such as JCV and BKV were previously identified within neuroblastoma samples and other pediatric embryonal tumors, newer studies seem to render these associations inconclusive.[14] Therefore, the role of pathogenic cofactors of neuroblastoma oncogenesis remains unresolved.

In general, the search for suspected viral cofactors of idiopathic diseases requires systematic screening of human tissues for viral biomarkers such as virus-derived nucleotide sequences. Unfortunately, viruses are of polyphyletic origin and thus lack common universal marker genes as they are frequently exploited in metagenomics studies targeting cellular microorganisms. Consequently, it is not currently possible to specifically PCR-amplify viral nucleotide sequences within a given tissue without prior information about the infective agent being sought.[15] As a result, several systematic assays for pathogen detection have been developed that do not rely on targeted PCR-amplification of viral factors[16] and were employed to identify Kaposi's sarcoma-associated herpes virus (KSHV) as a human tumor virus.[17] These systematic approaches were recently supplemented by sensitive deep sequencing technologies and applied to exclude several cancer-virus associations based on negative evidence and aided in the identification of MCPyV, a human polyomavirus, as a cofactor of Merkel cell carcinoma.[18]

[7] Brodeur (2003), Maris et al. (2007)

[8] Janoueix-Lerosey et al. (2010)

[9] Kaatsch (2010)

[10] Brodeur (2003), D'Angio et al. (1971), Janoueix-Lerosey et al. (2009), Shuangshoti et al. (2012)

[11] Fischer et al. (2006)

[12] Roman et al. (2007)

[13] Heck et al. (2009), Menegaux et al. (2004), zur Hausen (2009a)

[14] Flaegstad et al. (1999), Jørgensen et al. (2000), Krynska et al. (1999), Stolt et al. (2005)

[15] Rohwer and Edwards (2002)

[16] Bexfield and Kellam (2011)

[17] Chang et al. (1994)

[18] Arron et al. (2011), Bexfield and Kellam (2011), Feldhahn et al. (2011), Feng et al. (2008)

Deep sequencing technologies have enabled detection of both known and novel viruses with unprecedented sensitivity.[19] However, the large numbers of sequence fragments ("reads") generated by these methods necessitate data reduction approaches for filtering and condensing the list of putative viral transcripts. Two such approaches are currently represented in the literature: *digital transcript subtraction* that discards human sequence homologs from the sequence data and considers the remaining transcripts as potential viral signatures,[20] and *de novo sequence assembly* that aims to reconstruct whole viral genomes from overlapping reads.[21] Recently, variants of these of two approaches have been implemented in several computational pipelines such as PathSeq, RINS, and CaPSID.[22]

Identification of tumor viruses in particular poses several important challenges to existing computational pipelines. Confounding factors such as loss of viral genetic material from progressed tumors as well as limited replication competence or latent replication strategies often result in low or selective transcription of tumor viruses.[23] In addition, viral oncogenes homologous to human factors and chimeric transcripts originating from proviral insertion sites may share significant sequence similarity with human transcripts,[24] thus making unequivocal identification of viral factors difficult. Finally, high rates of viral sequence divergence from 10^{-5}-10^{-8} (dsDNA viruses) up to 10^{-4} (ssRNA viruses) substitutions per site and year[25] hinder recognition of known viruses based on known reference sequences.

We have developed Virana, a novel computational approach specifically tailored to detecting low-abundance transcripts that diverge from known viral reference sequences or share significant sequence homology with human factors. In particular, our method maps sequence reads to a combined reference database comprising the human genome and all known viral reference sequences. The approach is configured to allow for high mismatch rates and mappings to multiple reference sequences (*'multimaps'*). By using this combined and sensitive mapping strategy, our approach is especially well suited for detecting human-viral chimeric transcripts and viruses diverging from known references. In contrast to existing subtractive approaches for viral transcript discovery, our method abstains from discarding reads homologous to the human genome from further analysis. Instead, Virana exploits multimaps to assign sequence reads to a homologous context comprising human reference transcripts and viral reference genomes. These homologous regions retain the full, unfiltered information contained in the raw sequence data while also being amenable to further analyses by multiple sequence alignments, human-viral phylogenies, and orthogonal taxonomic annotations, thus greatly aiding in the interpretation of the results.

We applied our novel approach on an overall number of 14 deep sequencing transcriptomes of stage 4 and stage 4S metastatic neuroblastoma in order to identify putative viral cofactors associated with this idiopathic disease.

[19] Lipkin (2010)

[20] Arron et al. (2011), Duncan et al. (2009), Feng et al. (2008, 2007), Isakov et al. (2011), Moore et al. (2011), Patowary et al. (2012), Weber et al. (2002), Xu et al. (2003)

[21] Kreuze et al. (2009), Ma et al. (2011), Palacios et al. (2008), Wu et al. (2010)

[22] Bhaduri et al. (2012), Borozan et al. (2012), Kostic et al. (2011)

[23] zur Hausen (2006)

[24] Butel (2000)

[25] Duffy et al. (2008), Firth et al. (2010)

Materials and methods

Clinical samples and experimental deep sequencing data. Primary neuroblastoma samples from stage 4 (progressive) patients (n=7) and stage 4S (regressive) patients (n=7) were obtained prior to treatment from the central neuroblastoma tumor bank at the University Hospital of Cologne, Germany. None of the tumors harbored amplification of the MYCN proto-oncogene as determined by two independent laboratories for each case by fluorescence in situ hybridization (FISH) and Southern blot.[26] Only neuroblastoma samples with a tumor cell content of above 60% as assessed by a pathologist were selected for deep sequencing. Integrity of RNA was evaluated using the Bioanalyzer 2100 (Agilent Technologies) and only samples with an RNA integrity number of at least 7.5 were considered for further processing. Quality of all neuroblastoma samples and related deep sequencing data was additionally confirmed by an orthogonal computational analysis focusing on human gene expression in the context of differential splicing.[27]

All patients were enrolled in the German Neuroblastoma trials with informed consent. In order to validate our approach we additionally employed a positive control panel consisting of tumors with known viral cofactors. An EBV-positive B-cell-lymphoma (BCL) was received from the Pediatric Oncology and Hematology Department of the Hannover Medical School. Deep-sequencing reads obtained from full transcriptome libraries of two HPV18-positive HeLa samples (HeLa) and a HPV16-positive primary cervical squamous cell carcinoma (ceSCC) were downloaded from the Short Read Archive (SRA) and preprocessed as specified in the original publication.[28] Transcriptome data of a HBV-positive hepatocellular carcinoma (HCC) HKCI-5α cell line with confirmed HBV integration events was downloaded from the SRA based on information in the original publication.[29] A negative control panel consisting of a normal brain transcriptome generated as part of the Illumina BodyMap 2.0 project was obtained from the SRA at run accession number ERR030882.

Library preparation and sequencing. mRNA libraries of the EBV-positive B-cell lymphoma and 14 neuroblastomas were prepared following the Illumina RNA Sample Preparation Kit and Guide (Part #1004898 Rev. A). For each sample, 5 μg high-quality total RNA was processed for mRNA purification, chemical fragmentation, first strand synthesis, second strand synthesis, end repair, 3′-end adenylation, adapter ligation, and PCR amplification. Validated libraries underwent gel size selection and final paired-end sequencing with an effective read length of 2 × 36 bp on the Illumina Genome Analyzer IIx following Illumina standard protocols. Additionally, libraries for two of the 14 neuroblastoma samples were generated using the same protocols and sequenced with an effective paired-end read length of 2 × 95 bp on an Illumina HiSeq 2000. All

[26] Spitz et al. (2003)

[27] Aschoff et al. (2013)

[28] Arron et al. (2011)

[29] Li et al. (2013)

libraries had insert size distributions approximating $\mu = 150$ bp, $\sigma = 50$ bp as later confirmed by read mapping. The data were filtered according to signal purity by the Illumina Realtime Analysis (RTA) software.

Simulated sequencing data. In this study we employ simulated sequencing data from three viral genomes that are homologous to human factors. Reads originating from the ABL1-homologue of the Abelson murine leukemia virus (A-MuLV, GI:9626953, positions 1326-2605), from the the *gag* region of HERVK22I (obtained from Repbase,[30] positions 1-1452), and from Bo17, a GCNT3-homolog of the bovine herpesvirus 4 (BoHV-4, GI:13095578, positions 107098-108748) were generated *in silico* by dwgsim, a read simulator based on wgsim.[31] In addition, we produced simulated chimeric transcripts by fusing each of the aforementioned sequence regions to the human TP53 gene, a known proto-oncogene (UCSC build hg19, GRCh37, chr17, positions 7572926-7579569). These artificial fusion transcripts were generated using Fusim[32] based on TP53 exon models obtained from the UCSC refGene database.[33] Fusion transcripts were then used as templates for generating simulated data sets with dwgsim. In all cases, dwgsim was applied using the default empirical error model. Paired-end read lengths and insert size distributions were chosen according to the neuroblastoma sequencing data (see above). Additional simulated sequencing data generated by a related publication were analyzed as described in Section "Estimation of read mapping sensitivity".

Sample data notation. Sample panels containing neuroblastoma transcriptomes sequenced at 2 × 36 bp and 2 × 95 bp effective read lengths are denoted as NB1 and NB2, respectively. While the NB1 panel contains seven transcriptomes of neuroblastoma stages 4 and 4S each, the NB2 panel contains one sample of stages 4 and 4S each (see Table 13.1). Positive control panels of human cancer transcriptomes with known viral cofactors (BCL, HeLa, ceSCC, and HCC) are denoted as POS. The negative control panel consisting of a normal human brain transcriptome is denoted as NEG.

Reference genomes. The current assembly of the human reference genome (UCSC build hg19, GRCh37) as well as corresponding refGene splice-site annotations were obtained from UCSC. Splice variant annotations and cDNA sequences for the human genome were downloaded from Ensembl.[34] A set of all 4,680 available complete viral reference genomes and their taxonomic lineages were obtained from NCBI via the E-utilities web service[35] and the database query: "Viruses[Organism] AND srcdb_refseq[PROP] NOT cellular organisms [ORGN]". In addition, we obtained consensus reference sequences for all human endogenous retroviruses (HERV-K/HML-2) represented in Repbase (Primate HERV, HERVK11DI, HERVK11I, HERVK13I, HERVK22I, HERVK3I, HERVK9I, HERVKC4)).[36] All

[30] Jurka et al. (2005)

[31] Li et al. (2009a)

[32] Bruno et al. (2013)

[33] Pruitt et al. (2009)

[34] Flicek et al. (2012)

[35] Maglott et al. (2011)

[36] Jurka et al. (2005)

Panel	Source	Sample ID	Read length	Depth (Gbp)	Read pairs (M)
POS	HeLa	15	2 × 54 bp	0.076	0.737
POS	ceSCC	16	2 × 54 bp	0.157	1.527
POS	ceSCC	17	2 × 54 bp	0.041	0.400
POS	BCL	18	2 × 36 bp	3.134	43.527
POS	HCC	19	2 × 100 bp	11.22	55.547
NEG	Brain	20	2 × 50 bp	7.351	73.513
NB1	4	1	2 × 36 bp	1.184	16.439
NB1	4	2	2 × 36 bp	0.770	10.695
NB1	4	3	2 × 36 bp	0.881	12.236
NB1	4	4	2 × 36 bp	0.744	10.345
NB1	4	5	2 × 36 bp	1.207	16.759
NB1	4	6	2 × 36 bp	1.050	14.581
NB1	4	7	2 × 36 bp	0.829	11.527
NB1	4S	8	2 × 36 bp	1.031	14.317
NB1	4S	9	2 × 36 bp	1.172	16.282
NB1	4S	10	2 × 36 bp	0.868	12.065
NB1	4S	11	2 × 36 bp	0.890	12.368
NB1	4S	12	2 × 36 bp	0.845	11.737
NB1	4S	13	2 × 36 bp	1.174	16.300
NB1	4S	14	2 × 36 bp	0.847	11.772
NB2	4	7	2 × 95 bp	9.284	48.863
NB2	4S	13	2 × 95 bp	8.748	46.041

Table 13.1: *Sequencing characteristics of data sets.* Sequencing characteristics of neuroblastoma (NB), positive control (POS), and negative control (NEG) panels. Depths are reported in basepairs (bp).

reference genomes were combined into a single human-viral reference database for Virana. Since RINS and CaPSID cannot use such a combined database, human and viral reference sequences were collected within two separate databases for these approaches.

Quality control, mapping, and assembly. Paired-end reads from the neuroblastoma panels and positive control panels were quality-controlled with an in-house sequence analysis framework in order to identify sample contamination, adapter contamination, and batch effects. After quality control, the sequence data consisted of 13.494 Gbp (NB1), 18.032 Gbp (NB2), 14.63 Gbp (POS), and 7.351 Gbp (NEG) of sequence reads, respectively (see Table 13.1).

All data were mapped against a combined human-viral reference database with the splicing-aware and gapped read mapper STAR[37] in paired-end mode. While Virana considers the read mapper to be a replaceable component, in principle, we decided to employ STAR due to its mapping speed, high sensitivity settings, and its consideration of putative chimeric transcripts. We configured the mapper for high sensitivity by following recommendations of the author of STAR (personal communication). In particular, we set the rate of acceptable mismatches to 0.3 times the length of each read and the *seedSearchStartLmax* and *winAnchorMultimapNmax* parameters to 12 and 50, respectively. The minimum length of chimeric segments (*chimSegmentMin*) was reduced to 15 in order to detect fusion transcripts at short read lengths. Known splice sites from splice annotations of the human reference genome as well as canonical splice sites were considered in the mapping. For each read, multiple mapping locations with alignment score distances of up to 10 ranks relative to the best score were permitted

[37] Dobin et al. (2013)

('multimaps'). Read alignments were stored in standardized BAM files. STAR supports detection of chimeric transcripts by reporting discordant read pairs whose ends map to different chromosomes. These discordant read pairs were employed in further analyses as detailed in the next section.

In order to identify putative new viral transcripts, read pairs with at least one unmapped read end were extracted from BAM files by the Samtools suite[38] and assembled into longer contigs by the *de novo* transcriptome assemblers Trinity and Oases[39] using default parameters. Oases was configured for using different k-mer values in order to facilitate reconstruction of low-abundance viral transcripts. Contigs of length less than 300 bp were considered to be spurious assemblies and excluded from further processing.

Detection of chimeric transcripts. Virana supports detection of human-viral chimeric transcripts in two different manners. First, the read mapper employed in our study is able to partially align reads that contain a human-viral chimeric breakpoint to multiple reference sequences. Consequently, these partially aligned reads can be detected by Virana within the generic analysis of homologous regions (see below). The second, more sensitive approach to detecting chimeric transcripts is based on paired-end read information. Since the STAR mapper assigns reads to a combined reference database comprising both human and viral reference sequences, ends of paired-end reads whose inserts span the breakpoint of a chimeric transcript will be aligned to different reference sequences. These discordant read pairs are reported by STAR during read mapping (see above) and can further be filtered by mismatch score or sequence complexity in order to yield a high-confidence list of chimeric transcripts.

Generation of homologous regions. A distinguishing feature of Virana is its ability to automatically reconstruct the homologous context of reads that map to both viral and human reference sequences. This homologous context is constructed in four steps:

(1) First, reads that map to at least one viral reference are extracted from the mapping together with their primary (highest alignment score) and secondary (up to ten ranks of alignment scores below the highest score) mapping positions (see Figure 13.1). Since viruses of the same taxonomic family often exhibit significant sequence similarity, reads that map to one family member often also map to related family members as well as to homologous loci in the human reference. Based on these primary and secondary mapping locations, Virana obtains overlapping human reference transcripts, viral genomic references, and viral taxonomic information pertaining to the location. For each sequence read, information obtained in this manner is collected in a data structure denoted as HIT. HITs originating from the same analysis panel are pooled for further analysis.

[38] Li et al. (2009a)

[39] Grabherr et al. (2011), Schulz et al. (2012)

(2) Second, pooled HITs originating from the same analysis panel are assigned to viral taxonomic families based on the viral genomic references they refer to. Sets of HITs assigned to the same viral taxonomic family are denoted as the *homologous group* (HOG) of that family. The same HIT may, in principle, be assigned to several HOGs.

(3) Third, since reads and references generally share local rather than global sequence similarity, sequences in HOGs cannot conveniently be aligned in a multiple sequence alignment. This circumstance considerably complicates interpretation of homologous relationships between multiple reads and references. Virana therefore applies a three-step greedy clustering approach to split HOGs into manageable and alignable clusters denoted as homologous regions.

(3a) The set of all reads within a HOG is re-aligned to the set of all references (human reference transcripts and viral reference genomes) within the HOG using a highly sensitive BLASTN[40] alignment (word size 7). Since all possible mapping locations are required for further processing, BLAST is configured for high permissiveness (E-value 10).

(3b) Each HIT is assigned to a singleton cluster. Clusters containing reads that map to the same reference are merged if their reference mapping locations (as determined by BLASTN) are less or equal than L=25 basepairs apart (L-gaps). Optimal values for L are determined empirically, see Section "Estimation of required sequencing coverage for detection of a homologous region" for a robustness analysis. Merging continues until the number of clusters converges. Subsequently, all clusters with fewer than an empirically chosen cutoff of t=5 reads are discarded in order to remove spurious hits. After filtering, each remaining cluster represents a candidate HOR. Since cluster membership is defined by reads mapping to common references, each pair of references within the candidate HOR shares one or more regions of high local sequence similarity (e.g., the loci the read mapped to) connected by L-gaps.

(3c) For each HOR, parts of reference sequences that are neither covered by a read mapping location nor by an L-gap between read mapping locations are trimmed.

(4) Last, due to the high mutual similarity of sequences within trimmed HORs, sequences within each HOR are now amenable to sequence alignment against the longest reference sequence within that HOR using LASTZ, the successor of BLASTZ.[41] The resulting star-shaped multiple sequence alignment is then used for construction of per-sample (for reads) and per-gene (for human reference transcripts) consensus sequences. Aligned consensus sequences retain information on non-consensus nucleotides due to the usage of IUPAC ambiguous nucleotide codes. Consensus sequences can then be manually inspected in order to determine single nucleotide permutations and indels up to length L that distinguish sequence reads, viral references, and human reference transcripts.

[40] Altschul et al. (1990)

[41] Schwartz et al. (2003)

Consensus sequences can be further processed by phylogenetic analyses. For generating phylogenies, Virana employs the software PhyML[42] following the maximum likelihood approach and using default parameters recommended by the HIV sequence database (http://hiv.lanl.gov, GTR model of nucleotide substitution, transition/transversion ratio: 4, gamma shape parameter: 1, number of substation rate categories: 4, approximate Likelihood Ratio Test (aLRT) using SH-like supports where applicable). We note that the topology of the phylogenetic trees constructed in this manner is stable with regard to the model choice; while more complex model parameters may yield better likelihoods in some instances, these differences do not influence interpretation of our results.

[42] Guindon et al. (2010)

Taxonomic annotation. In this study, we additionally compare consensus sequences of aligned HOGs as well as *de novo* assembled sequence contigs to nucleotide (NCBI NT) and protein (NCBI NR) reference archives in order to assign transcripts to a taxonomic origin. To this end, we employ several BLAST[43] search strategies (BLASTN, BLASTX, and TBLASTX) with sensitive word sizes (4, 3, and 3, respectively). TBLASTX bypasses synonymous mutations during similarity search and is particularly suited for detecting functionally conserved homologs. This approach is therefore recommended for discovering remote similarities[44] and is widely used in environmental metagenomics.[45] A permissive E-value threshold of 0.1 is used for all comparisons in order to reduce the possibility of missing true viral hits. For each query transcript and search strategy, the three highest-scoring reference sequences are extracted from the BLAST results. Subsequently, descriptions, taxonomic information, and available gene annotations for high-scoring reference hits are pooled and query transcripts are assigned a putative viral, human, or ambiguous origin based on the pooled information. In order to limit the search space of the computationally intensive TBLASTX procedure, we constrain the allowed taxonomic origin of reference sequences to only viral (NCBI taxon ID 10239) or human (NCBI taxon ID 9606) hits while excluding artificial sequences (NCBI taxon ID 81077) using the NCBI database query "(((txid10239 [ORGN]) OR (txid9606 [ORGN]) OR (human [ORGN])) NOT (txid81077 [ORGN]))".

[43] Maglott et al. (2011)

[44] Kunin et al. (2008)

[45] Mokili et al. (2012)

Estimation of read mapping sensitivity. We quantify the ability of our novel method Virana and the related methods RINS[46] and CaPSID[47] at detecting diverged viral transcripts among human sequence data by employing a recently published validation data set.[48] This data set consists of a negative control background set of reads simulated from the human reference genome that is spiked with four sets of 10,000 reads simulated from 10 viral reference genomes. Nucleotide positions within reads of each of the four viral spike-in data sets are mutated randomly independently and uniformly with a set-specific probability $\theta \in \{0, 0.05, 0.1, 0.25\}$ be-

[46] Bhaduri et al. (2012)

[47] Borozan et al. (2012)

[48] Borozan et al. (2012)

fore being merged with the background data set. The set of viral reference sequences represents 10 different viral families that infect plants (Cherry green ring mottle virus, Cestrum yellow leaf curling virus, Elm mottle virus, East African cassava mosaic virus), birds (Gallid herpesvirus 1), insects (Cotesia congregata bracovirus), bacteria (Guinea pig Chlamydia phage), amphibians (Frog adenovirus 1), and mammals (Rat coronavirus Parker, Banna virus).

All five data sets (non-spiked human negative control and four human-viral spike-in sets) are analyzed by Virana, RINS, and CaPSID using identical reference sequences as described in Section "Reference genomes". Sensitivity (fraction of correctly identified viral reads among all viral reads) and specificity (1 - fraction of falsely identified human reads among all human reads) of viral read detection are determined for each method and data set. Analyses are performed with either default parameters (Virana), parameters published in the original validation data set (CaPSID), or settings adapted by us in order to maximize sensitivity (RINS: minimal contig length decreased to 100, read lengths and insert size distributions according to input data).

Since all methods map to the same complete viral reference set, reads from a particular viral genome of the validation data set may be distributed across several closely related reference genomes, all of which may be considered valid mappings. For this reason, we added post-processing steps to CaPSID and RINS and performed this validation on the level of viral taxonomic families rather than on the level of single viral species. We note, however, that results of all tested methods including Virana retain information on single viral species throughout the analysis. In particular, sensitivity and specificity of the methods change only minimally if data is analyzed on the single species level.

Analysis of human-viral homologous and chimeric transcripts. Analysis of the human-viral homologous regions and chimeric transcripts based on simulated read data (see Section "Simulated sequencing data") was conducted by configuring CaPSID, RINS, and Virana analogous to the previous section. For the validation of fusion transcript detection, the number of true positives is set to the number of all reads originating from the human-viral fusion transcript. Since all detection methods in this validation are configured to only report reads mapping to the viral part of the fusion transcript, sensitivity estimates are scaled down equally for all methods in this particular validation. Analysis of discordant read ends in order to detect the origins of chimeric transcripts was performed as described before (see Section "Detection of chimeric transcripts").

Estimation of required sequencing depth. Expanding on related work,[49] we quantify the theoretical sensitivity of Virana by estimating the number of viral transcripts per cell that are required for achieving a certain minimal sequencing coverage at a probabil-

[49] Feng et al. (2007), Moore et al. (2011)

ity of at least 95%. Based on human genome annotations obtained from UCSC, we determined an average length of human coding sequences (CDS) of l = 1,634 bp. By conservatively assuming that an idealized cell contains 200,000 mRNAs(50) of average length l fragmented at f = 500 bp as a result of library preparation, an expected number of m = 653,600 cDNA fragments are generated per cell. For a given viral transcript of length r and a viral transcript abundance x per cell, we expect a number of $v = x\, r / f$ viral transcript fragments. Assuming a theoretical, unbiased sequencing process, the probability of sequencing a viral transcript fragment among the overall m transcript fragments is $p_{viral} = v/m$. Given a single-end read length of j, a number $k = rc/(2j)$ reads are required to achieve a sequence coverage c of that viral transcript. The probability p^k_{viral} of observing at least k reads during sequencing with a sequencing depth n is specified by the cumulative binomial distribution function with parameters k, n and p_{viral}. Due to numerical instabilities of computing the cumulative binomial distribution for large values n, we exploit the Central Limit Theorem and estimate p_{viral} by the Camp-Paulson normal approximation to the binomial distribution. This approach has a negligible approximation error of $< 0.007/\sqrt{n\ p_{viral}\ q}$, where $q = 1 - p_{viral}$.(51)

Our approach further depends on successfully reconstructed homologous regions, each requiring an empirically determined minimum number of t = 5 transcripts separated by no more than L = 25 base pairs. Although the probability p_{region} of a homologous region being successfully constructed from viral transcripts at a given sequence coverage can be derived analytically for a special case,(52) this solution neither considers edge effects occurring for small transcripts nor takes into account the distribution of insert sizes of paired-end reads. We therefore approach the problem empirically by *in silico* simulation of paired-end reads that are assigned randomly independently and uniformly to transcripts of different lengths and at varying coverages. This simulation process addresses the aforementioned confounding factors by considering transcript boundaries and sampling insert sizes from a normal distribution parametrized according to neuroblastoma sequence data employed in this study (see Section "Library preparation and sequencing"). A mean estimator for p_{region} and its standard error $SE_{p_{region}}$ were derived by averaging the success rates of homologous region constructions across 1,000 simulations for each transcript length, read length, region linkage, and read coverage.

Availability. All sequence data generated in this study are publicly available in the European Nucleotide Archive (ENA) at study accession number PRJEB4441. Software implementations of our method and all validation procedures are available from `http://mpii.de/~sven/virana` and/or from the authors upon request.

(50) Feng et al. (2007)

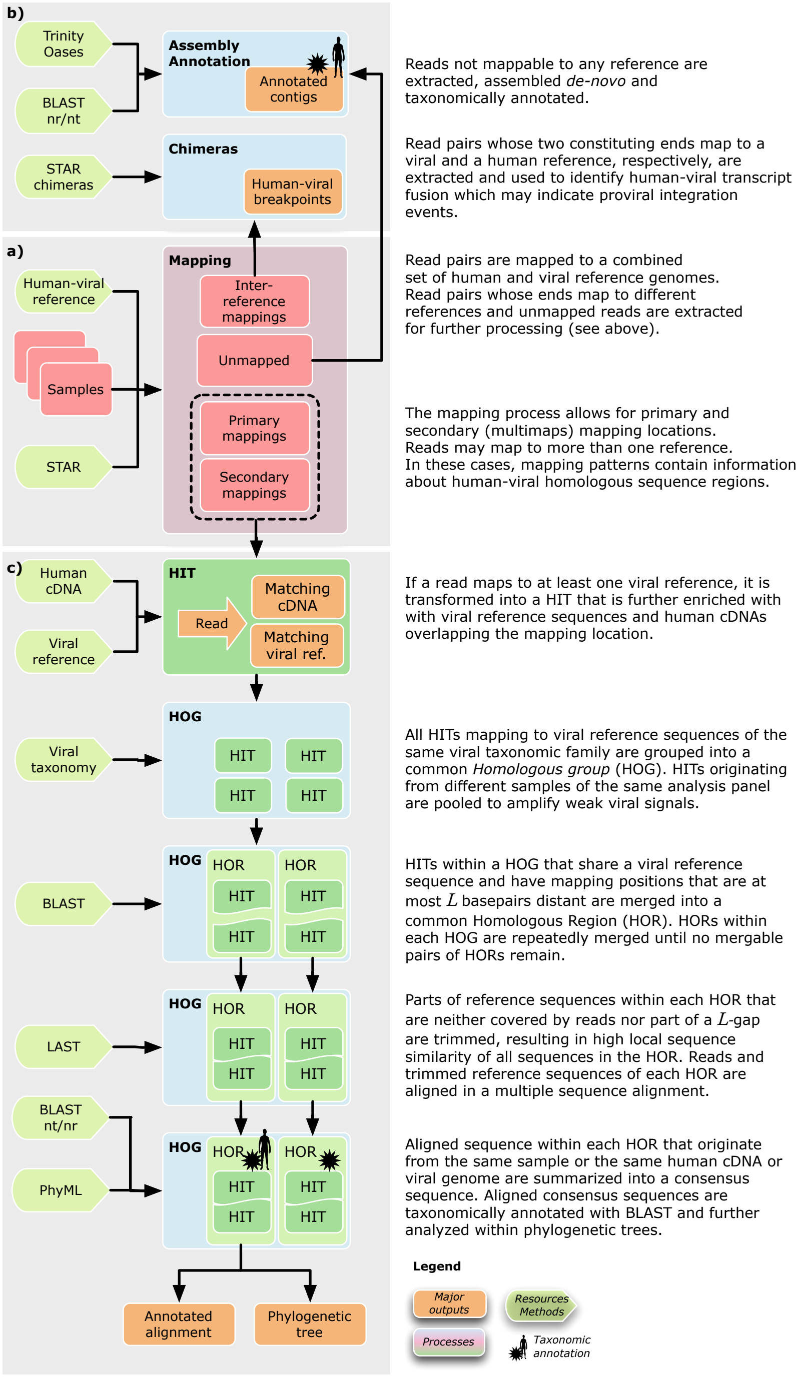
(51) Lesch and Jeske (2009)

(52) Breitbart et al. (2002)

Results and discussion

This study presents a novel approach to detecting viral transcripts in human tumor transcriptomes. In contrast to related approaches such as RINS and CaPSID that rely on subtracting reads homologous to human transcripts from the analysis, our novel method Virana assigns sequence reads to a combined human-viral reference database without discarding homology information (see Figure 13.1). By employing a particularly fast and sensitive read mapper, Virana gains sensitivity at discovering highly divergent and chimeric viral transcripts. In addition, this configuration allows for exploitation of multimaps (e.g., sequence reads mapping to several reference genomes with varying mismatch rates) to discover the homologous context of sequence reads with regard to viral and human reference sequences. Last, Virana employs chimeric alignments as well as *de novo* assembly of unmapped sequence reads followed by taxonomic annotation in order to discover proviral integration events and novel viruses, respectively.

Detection of divergent viruses. In order to compare Virana and the two subtractive approaches CaPSID and RINS in a controlled environment we rely on a previously published simulated data set consisting of a negative control data set free of viral reads, here denoted as background set. The background set is used to construct four additional validation data sets spiked with viral reads at increasing rates of sequence divergence (0%, 5%, 10%, 25%, see Materials and Methods). Performance is quantified in terms of sensitivity and specificity (see Materials and Methods). Applying all three viral detection methods on the validation data sets reveals comparatively high rates of correctly detected viral reads for CaPSID and RINS at low sequence divergences between 0% and 5%. Specifically, the two subtractive methods achieve 0.99-1.13 fold higher sensitivities compared to Virana (sensitivities of 0.835-1.0 versus 0.844-0.882 for subtractive approaches and Virana, respectively, see Figure 13.2). In contrast, Virana substantially surpasses subtractive approaches at higher rates of viral sequence divergence (10-25%), offering comparatively stable sensitivities up to 7-fold and 182-fold higher than Capsid and RINS, respectively (sensitivities of 0.0008-0.6578 versus 0.1456-0.7880 for subtractive approaches and Virana, respectively, see Figure 13.2, left panel). Notably, while subtractive approaches fail to identify 20-90% of viruses in settings of high sequence divergence, Virana is the only approach able to reliably detect the full set of viruses in all validation scenarios (see Figure 13.2, right panel). As a result of Virana's ability to detect human-viral transcript homologs, reads originating from several human endogenous retroviruses (HERVs) that are part of the human reference genome but technically also belong to the viral family *Retroviridae* are detected in validation data at all levels of sequence divergence. Since the detected HERV reads originate from the human rather than from the viral part of the validation data,



Reads not mappable to any reference are extracted, assembled *de-novo* and taxonomically annotated.

Read pairs whose two constituting ends map to a viral and a human reference, respectively, are extracted and used to identify human-viral transcript fusion which may indicate proviral integration events.

Read pairs are mapped to a combined set of human and viral reference genomes. Read pairs whose ends map to different references and unmapped reads are extracted for further processing (see above).

The mapping process allows for primary and secondary (multimaps) mapping locations. Reads may map to more than one reference. In these cases, mapping patterns contain information about human-viral homologous sequence regions.

If a read maps to at least one viral reference, it is transformed into a HIT that is further enriched with with viral reference sequences and human cDNAs overlapping the mapping location.

All HITs mapping to viral reference sequences of the same viral taxonomic family are grouped into a common *Homologous group* (HOG). HITs originating from different samples of the same analysis panel are pooled to amplify weak viral signals.

HITs within a HOG that share a viral reference sequence and have mapping positions that are at most L basepairs distant are merged into a common Homologous Region (HOR). HORs within each HOG are repeatedly merged until no mergable pairs of HORs remain.

Parts of reference sequences within each HOR that are neither covered by reads nor part of a L-gap are trimmed, resulting in high local sequence similarity of all sequences in the HOR. Reads and trimmed reference sequences of each HOR are aligned in a multiple sequence alignment.

Aligned sequence within each HOR that originate from the same sample or the same human cDNA or viral genome are summarized into a consensus sequence. Aligned consensus sequences are taxonomically annotated with BLAST and further analyzed within phylogenetic trees.

Figure 13.1: *Virana's approach to identifying viral transcripts in human tumors.* a) Transcriptome sequence samples are first mapped to a combined set of human and viral reference sequences in a splicing-aware fashion. b) Unmapped or discordantly mapped read pairs are further processed by assembly methods to detect novel viruses or transcript chimeras that may indicate proviral integration events. c) Reads mapping to one or more viral genomes (HITs) are analyzed in an integrated fashion by considering human homologous mapping locations and viral taxonomies. This process results in a number of homologous regions (HOR) for each viral family. HORs are represented as multiple sequence alignments incorporating a wealth of sequence information. Alignments are further enriched by taxonomic annotations and phylogenetic analyses.

these reads classified as false positive (FP) hits for the purpose of this validation. As a result of this artifact, Virana exhibits a slightly lowered specificity compared to subtractive approaches (0.99985 versus 1.0 for Virana and CaPSID/RINS, respectively). However, we note that HERV reads are correctly classified by Virana during homologous region construction and by optional BLAST-based taxonomic annotation. These reads can therefore be safely and automatically ignored in subsequent analyses if HERV expression is of no interest to the researcher.

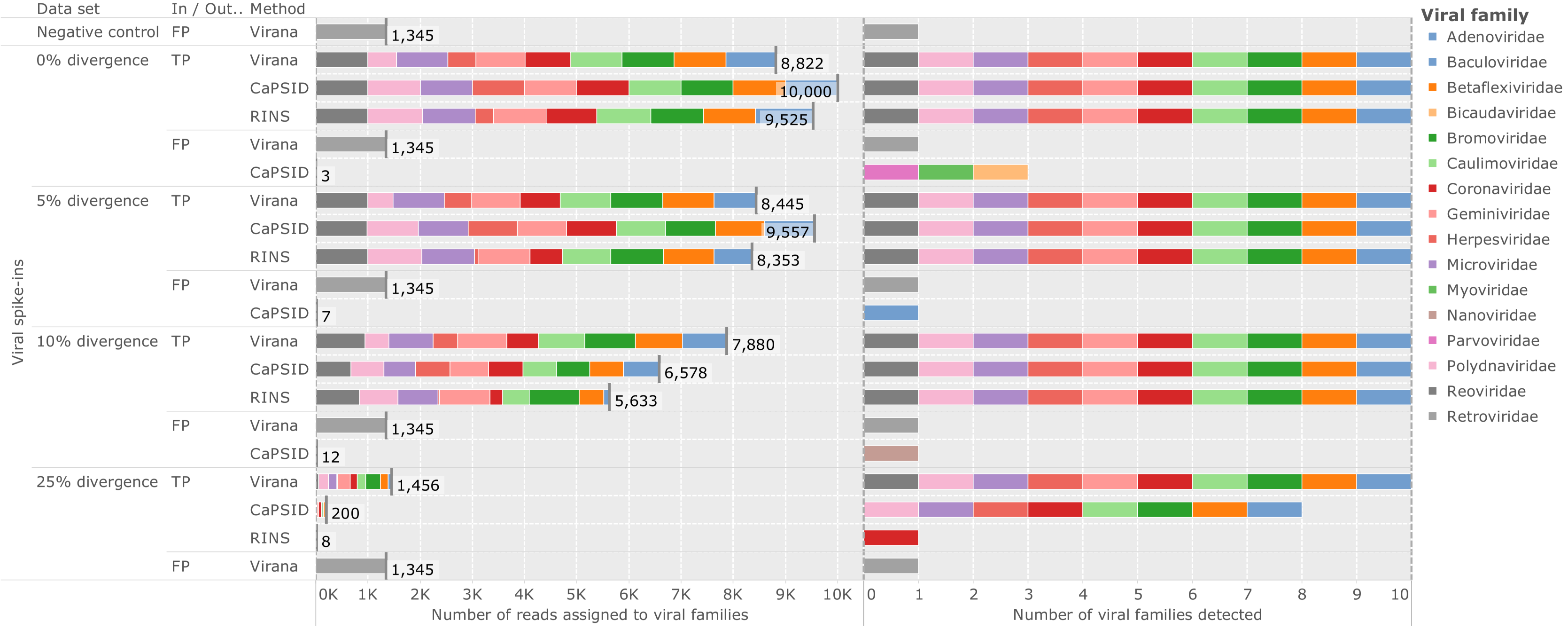


Figure 13.2: *Detection of divergent viruses.* Performance comparison of Virana, CaPSID, and RINS at detecting viral reads at different rates of simulated sequence divergence among a background set comprising human genomic reads. The background set without any spike-ins of viral reads serves as negative control. Left panel: stacked bars represent absolute numbers of detected reads grouped by sequence divergence, correctness of classification (TP: true positive, FP: false positive), and detection method. Falsely classified reads not assigned to any of the viral families present in the validation are labeled as false positives (FP). Colored segments indicate to which viral families the reads were assigned. Each condition allowed for the correct detection of up to 10,000 reads. Right panel: color coded markers for each condition and detection method indicating which viral families were identified. A maximum number of 10 viral families could be correctly identified in each condition.

In spite of the involved construction process of homologous regions, Virana is fastest among the three viral detection approaches, requiring only about half an hour per sample analyzed. In contrast, RINS and CaPSID require two to 17 times longer per sample, respectively (see Figure 13.3). Interestingly, the majority of time spend by CaPSID is lost on subtraction, indicating that this step is a limiting factor of subtractive approaches. We note than reported times are based on analyses using a single compute core. Since all evaluated methods benefit from multithreading, dedicating additional compute cores to the analysis allows for further reduction in processing time.

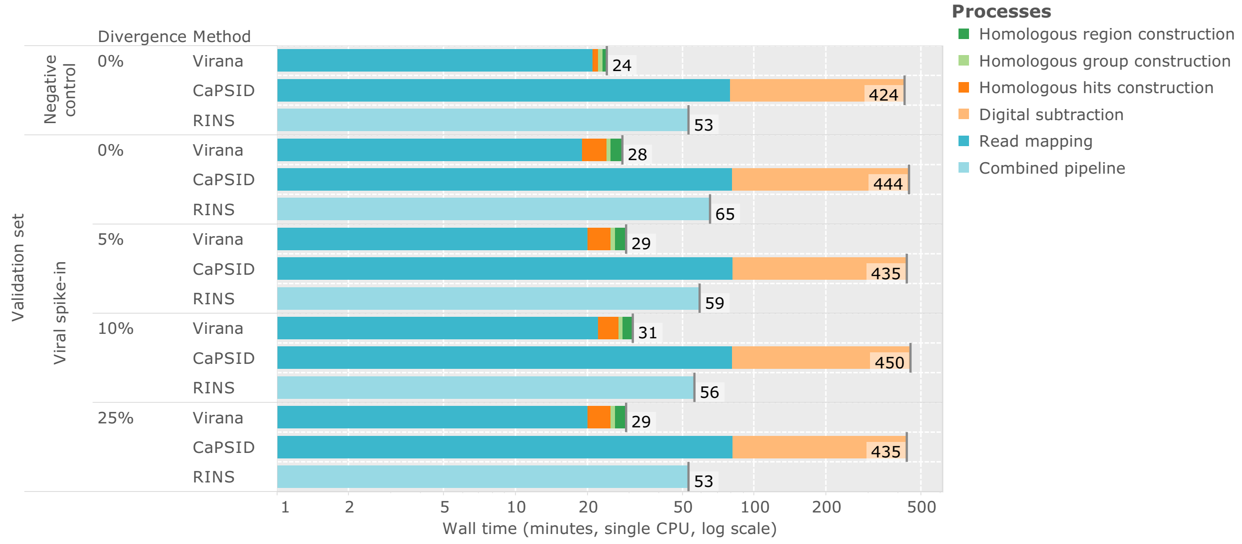


Figure 13.3: *Time required for data analysis.* Cumulative time in minutes required for analysis of the divergence validation set. Times are reported for the negative control without viral spike-ins as well as for four mixed data sets consisting of negative control background set with viral spike-ins at different divergence rates. Segments within bar plots represent different analysis processes employed by the three viral detection methods Virana, CaPSID, and RINS. All measurements are based on a single CPU Intel(R) Xeon(R) E5-4640 clocked at 2.40 GHz.

Detection of low-coverage, homologous, and chimeric viral transcripts. Having established Virana's ability to detect reads sampled at comparatively high coverage from viral genomes with low or no human-viral sequence similarity, we next test the sensitivity of the viral detection methods in a more challenging scenario involving gene regions of animal viruses that have close human homologs and are sampled at low sequencing coverages. Three such human-viral homologs are used in the analysis: V-ABL of the acutely transforming retrovirus A-MuLV, Bo17 of herpesvirus BoHV-4 (a model virus for oncogenic gammaherpesviruses such as EBV and KSHV and implied in several animal cancers)[53] and *gag* of HERV-K(HML2)22I, a class of human endogenous retroviruses associated with some forms of breast cancer).[54] Validation is based on simulated sequencing data and split into two scenarios (see Materials and Methods for details). Within the first scenario, simulated se-

[53] Thiry et al. (1992)

[54] Wang-Johanning et al. (2003)

quencing reads are sampled directly from human-viral homologs while in the second scenario reads are generated from artificial fusion transcripts that each involve one of the three homologs fused to the human TP53 proto-oncogene. The resulting human-viral fusion transcripts mimic transcriptional signals indicating retroviral integration or homologous recombination of viral DNA next to a human gene which may result in activation of the latter by insertional mutagenesis.

We apply the viral detection methods Virana, CaPSID, and RINS on these two validation data sets in order to evaluate sensitivity at detecting viral genes that are similar to human factors either due to natural sequence homology or due to gene fusions. Performance is quantified by detection sensitivity, specificity, as well as by the absolute number of reads correctly detected. While all methods performed at a perfect specificity of 1.0, only Virana detects viral transcripts at all coverages and with two to three-fold higher sensitivities compared to competing methods (Figure 13.4). In particular, sequence reads originating from endogenous retroviruses were almost always subtracted from the analysis by RINS and CaPSID. In addition, RINS seemed to be confounded by low sequencing coverage, a fact most probably resulting from its heavy reliance on *de novo* transcript assembly. Subsequent analysis of discordantly mapped read pairs by Virana (see Materials and Methods) correctly identified the TP53 gene as fusion partner of both V-ABL and B017, indicating that detection of human-viral chimeras is reliable even at low twofold coverage. Due to the repeat nature of the HERV-K sequence in the human genome and the resulting re-occurrence of HERV-K homologs at multiple loci in the human reference it was not possible to unambiguously identify the fusion partner of the HERV-K *gag* gene.

Estimation of optimal sequencing depth. Due to a variety of factors (see Discussion) human tumor viruses often replicate at very low levels within the infected cell. Determining the required sequencing depth for detecting viral transcripts present at specific cellular abundances is therefore crucial for planning transcriptome experiments designed to identify tumor viruses. Based on statistical arguments and average mRNA sizes (see Materials and Methods), we inferred the minimal abundances of viral transcripts required in an average cell required for detection depending (1) on the length of the transcript being sought and (2) on the sequencing depth employed in the experiment. Here we report results for an average viral cDNA-transcript (795 bp), an average viral transcript region analyzed in the validation of human-viral homologs (B017 and vABL, 1,465 bp, see previous section), an average length human CDS (1,634 bp), and the genome size of a small tumor virus (A-MuLV, 5,896 bp). Based on these estimates and given an average sequencing depth as employed in the NB1 analysis panel, Virana requires a minimum twofold sequence coverage of an aver-

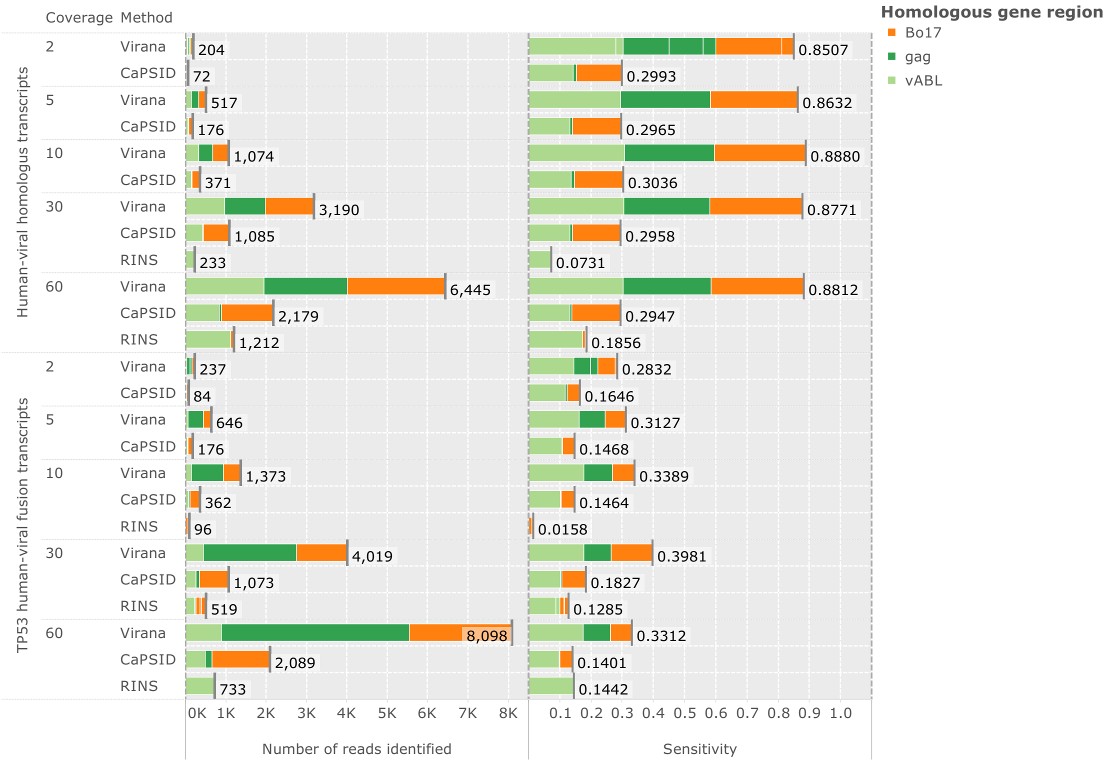


Figure 13.4: *Detection of low-coverage, homologous, and chimeric viral transcripts.* Displayed are performances of Virana, CaPSID, and RINS at detecting the three human-viral homologous gene regions Bo17, gag, and vABL. Performance is quantified in terms of sensitivity (right panel) and absolute number of reads correctly identified (left panel) at differing sequencing coverages (2-60 fold). Methods are validated at detecting both isolated gene regions (upper part) as well as at detecting human-viral fusion transcripts involving each of the three gene regions fused to the human TP53 proto-oncogene (lower part). Specificity of detection is 1.0 (100%) for all detection methods (not displayed).

age viral cDNA transcript in order to detect the transcript within a homologous region with 99.9% probability (Figure 13.5, upper left quadrant, dashed blue vertical line). This sequence coverage is produced with 95% probability if at least one viral transcript is present per cell, on average (Figure 13.6, upper left quadrant, dashed blue vertical line). The number of viral transcripts per cell required for detection is inversely related to transcript length and sequencing depth, in principle: at a transcript length corresponding to a small viral genome (5,896 bp) and a per-sample sequencing depth of 1% of the sequencing depth generated in the NB1 panel, a transcript coverage of 0.6 and at least 55 viral transcripts per cell are required for reliable detection (Figure 13.6, upper right panel, dotted black vertical line).

Analysis of positive and negative experimental controls. In order to evaluate Virana on experimental data we conducted an analysis of several positive and negative control samples with a cumulative size of 21.982 Gbp. The negative control sequencing data originates from a normal brain transcriptome that is suitable as a control for neuroblastoma data. Positive controls span a range of cancer transcriptomes that are associated with several viral cofactors such as a hepatocellular carcinoma (HCC) cell line with proviral integration

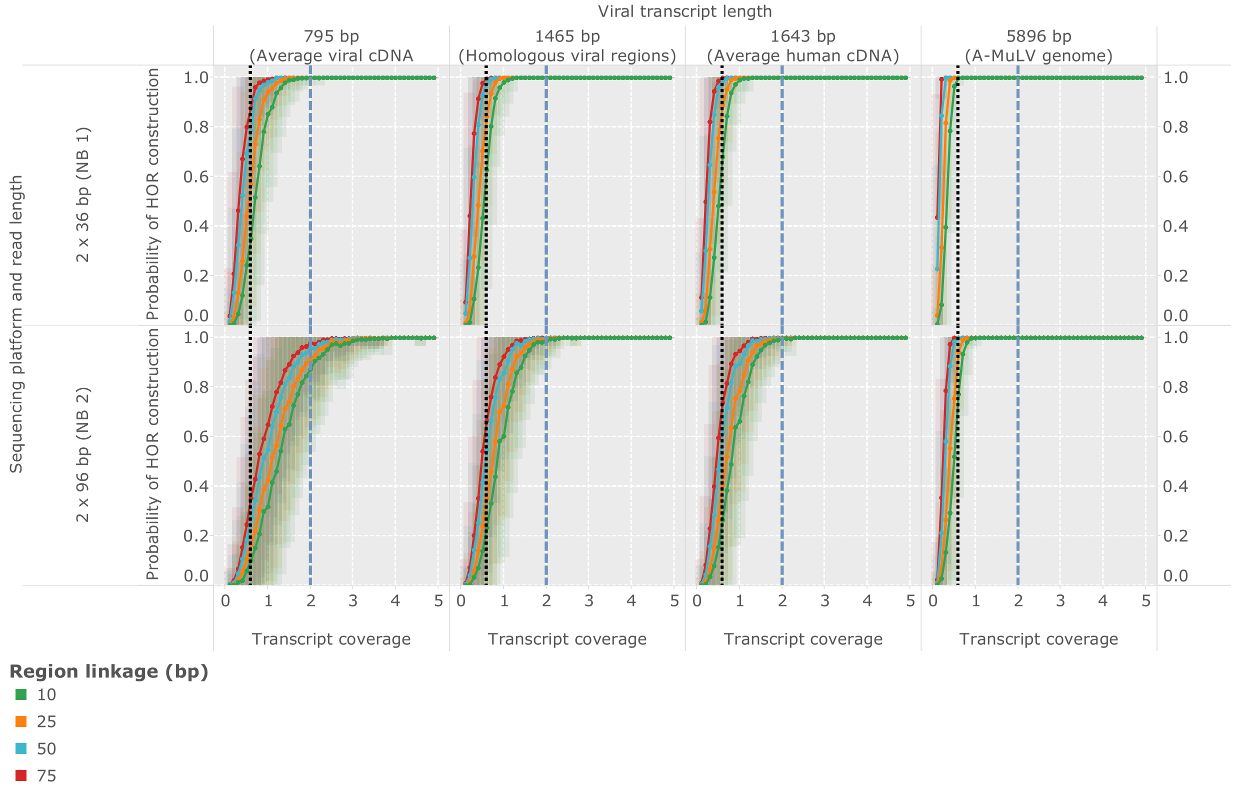


Figure 13.5: *Estimation of required sequencing coverage for detection of a homologous region.* Probability of successful region construction by Virana depending on the lengths of the transcripts being sought, the region linkage parameter L, as well as characteristics of the sequencing platform employed. Colored areas represent overlapping standard error bands of the mean, denoting the uncertainties of the estimations. The probability of Virana to detect a homologous region depends on the length of the viral transcript being sought, the linkage parameter L of the homologous region, as well as the transcript coverage and read length of the sequencing platform employed. Given characteristics of the sequencing process applied for NB1 sample panel, an average viral cDNA of length 795 bp requires a minimal transcript coverage of 2 in order to be reliably detected using a linkage parameter of $L = 25$ as employed in this study (upper left quadrant, dashed blue vertical line). Technologies affording longer read length as used for the NB2 panel typically also afford higher sequencing depths. However, at a fixed coverage these technologies generate a more highly fragmented region linkage due to a smaller number of longer reads, resulting in lower probability of generating contiguous homologous regions (lower left quadrant). Lower transcript coverage is sufficient for longer transcripts transcribed from a complete A-MuLV genome (upper right panel, dotted black vertical line) or smaller values of the region linkage parameter L.

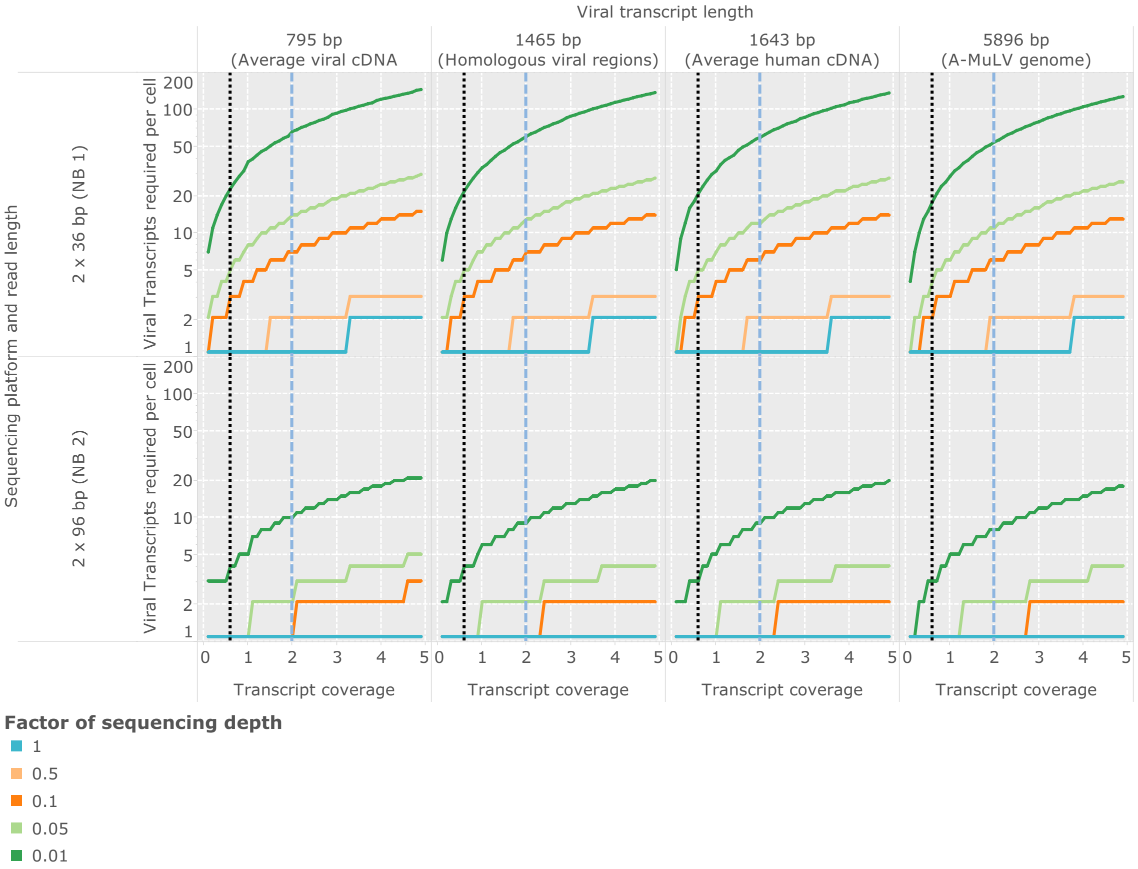


Figure 13.6: *Estimation of required cellular transcript abundances for achieving a given transcript coverage.* Sequencing coverage of viral transcripts is depending on the average number of transcript copies per cell in the sequenced sample, on the length of the viral transcript being sought, and on characteristics of the sequencing process. In order to better visualize the optimal sequencing depth required for detection of viral factors, we estimated the required number of transcript copies per cell for different sequencing depths. These sequencing depths are expressed as factors relative to the depths employed for the NB1/NB2 panel generated in this study (which are here reported as a relative sequencing depth of = 1).

of Hepatitis B virus, a cervical squamous cell carcinoma (ceSCC) and two HeLa cell line samples with associated human papillomavirus (HPV), and an Ebstein-Barr virus (EBV) positive B-cell lymphoma (BCL).

As displayed in Figure 13.7 (upper part), analysis of the brain negative control sample demonstrates that viral transcription is ubiquitous even in normal (non-cancerous) samples. Specifically, several bacteriophages of the taxonomic families *Microvirodae*, *Myoviridae*, *Podoviridae*, and *Siphovoridae* indicate sample contamination with bacteria as well as technical spike-ins.[55] Remarkably, the Coliphage phi-X174 genome of the family *Microvirodae* could be fully assembled by Virana's homologous region construction, yielding a single fragment of 99% sequence identity and 100% coverage compared to the phi-x174 reference genome. In addition, several retroviral and flaviviral hits at low abundances of 1-28 reads per million reads mapped (RPMM) highlight human factors such as HERV-Ks (endogenous retroviruses) as well as human proto-oncogenes SRC /ABL and DNAJC14/RP11 that have close homologs in the viral families *Retroviridae* and *Flaviviridae*, respectively. The taxonomic ambiguity of these regions is automatically identified during Virana's homologous region construction and confirmed by optional BLAST-based annotation compared to NCBI nt and nr databases (as indicated by thinner bars in Figure 13.7).

Analysis of positive control samples resulted in 41 homologous regions (HORs)spanning five viral families (see Figure 13.7, lower part). Viral cofactors associated with each of the cancer samples are correctly recovered at a high dynamic range of read abundances between 3 RPMM (HCC with integrated HBV provirus) and 1,628 RPMM (HeLa cell line associated with HPV18). In addition, several viral fragments were successfully reconstructed within HORs of the positive control samples, such as a 9,550 bp long EBV segment containing latency-associated factors EBNA 3b, 3c, and 4a (80% sequence identity with the wild type genome) as well as a 1,693 bp long HBV fragment containing the oncogenic HBV-X gene (98% sequence identity compared with Hepatitis B virus isolate HK1476). Similar to results on the negative control brain sample, several HORs with lower abundances assigned to the taxonomic families *Retroviridae* and *Flaviviridae* represent human-viral sequence homologies that are automatically flagged to be of ambiguous taxonomic status by Virana.

Interestingly, the HCC sample was also investigated in recent work focusing on detecting viral integration events.[56] In this recent study, the authors confirmed one integration event by Sanger sequencing while alluding to two additional events still awaiting experimental validation. By analyzing discordantly mapped read ends, Virana could correctly identify all three HBV fusion events involving human genes TRRAP (11 read pairs), ZNF48 (11 read pairs), and PLB1 (6 read pairs) as part of the primary mapping procedure.

[55] `http://res.illumina.com/documents/products/technotes/technote_phixcontrolv3.pdf`

[56] Li et al. (2013)

Panel	Source	ID	Pairs mapped	Both ends	Uniquely	Depth (Gbp)
POS	HeLa	15	94.900%	94.900%	68.422%	0.127
POS	ceSCC	16	90.803%	90.803%	69.561%	0.264
POS	ceSCC	17	96.629%	96.629%	73.921%	0.075
POS	BCL	18	91.612%	91.612%	63.528%	6.424
POS	HCC	19	94.693%	94.693%	73.500%	14.924
NEG	Brain	20	95.481%	95.481%	72.515%	11.234
NB1	4	1	95.878%	95.878%	69.422%	2.275
NB1	4	2	96.062%	96.062%	74.342%	1.43
NB1	4	3	96.385%	96.385%	75.938%	1.641
NB1	4	4	95.749%	95.749%	71.012%	1.503
NB1	4	5	95.057%	95.057%	69.203%	2.652
NB1	4	6	94.819%	94.819%	69.856%	2.39
NB1	4	7	96.597%	96.597%	72.107%	1.635
NB1	4S	8	95.952%	95.952%	70.681%	2.093
NB1	4S	9	95.242%	95.242%	74.009%	2.223
NB1	4S	10	96.854%	96.854%	74.756%	1.651
NB1	4S	11	96.819%	96.819%	75.256%	1.668
NB1	4S	12	96.710%	96.710%	74.899%	1.539
NB1	4S	13	95.344%	95.344%	72.326%	2.35
NB1	4S	14	97.110%	97.110%	74.829%	1.65
NB2	4	7	86.225%	86.225%	69.552%	12.243
NB2	4S	13	86.280%	86.280%	72.538%	11.517

Table 13.2: *Mapping rates.* Mapping ratios and depths of neuroblastoma (NB), positive control (POS), and negative control (NEG) panels. Mapped reads are relative to the number of sequenced read pairs that have passed quality control. Depths are reported in basepairs (bp) and include reads with multiple mapping locations ('multimaps').

Analysis of neuroblastoma samples. Deep-sequencing of 14 neuroblastoma samples on two sequencing platforms yielded 26.700 Gbp (NB1) and 23.760 Gbp (NB2) of mapped read pairs (including multimaps), respectively (see Table 13.2). While samples were sequenced independently and marked with unique identifiers to allow for sample tracking at each step of the analysis, reads from each sample panel and each tumor stage (4 or 4S) were pooled for analysis. Processing the pooled sample panels with Virana resulted in 46 homologous regions representing four viral families (see Figure 13.8). All HORs were associated with low relative read abundances of 1-67 RPMM compared to confirmed viral signatures of experimental positive controls (3-1,628 RPMM, see Figure 13.7). Several homologous regions assigned to bacteriophage viral families *Baculoviridae* and *Myoviridae* are attributable to sample contamination.

Reads assigned to viral families *Retroviridae* and *Flaviviridae* were determined to originate from either endogenous elements (HERVs) or from human proto-oncogenes that have close homologs in pestiviruses and acutely transforming retroviruses. HORs associated with these viral families were automatically assigned human or ambiguous taxonomic origin by Virana, as indicated by narrower bars in Figure 13.8. We undertook manual investigation of homologous relationships within each ambiguous HOR by analyzing multiple sequence alignments and phylogenetic trees of the respective regions. These analyses revealed unambiguous clusterings of neuroblastoma sequence reads near human or endogenous factors in all cases (see Figure 13.9 for an example phylogeny).

No significant differences in viral expression signatures between neuroblastoma 4 and 4S stages could be detected except for HERV-K endogenous retroviruses which display 36-86% higher abundances in stage 4S (NB1: 56 RPMM, NB2: 28 RPMM) than in stage 4 (NB1: 41 RPMM, NB2: 15 RPMM) neuroblastomas. All reads assigned to homologous regions were further analyzed for evidence of chimeric transcription (see Materials and Methods). While several read pairs with putative chimeric mappings could be identified, all viral chimeric read ends were clustered within low-complexity regions of the viral genomes. Analyses revealed that these putative chimeric mappings represent sequencing errors and low-complexity templates that non-specifically attracted reads of similarly low sequence complexity. No cluster of chimeric reads located at a specifically viral genome location and representing a human-viral breakpoint could be identified.

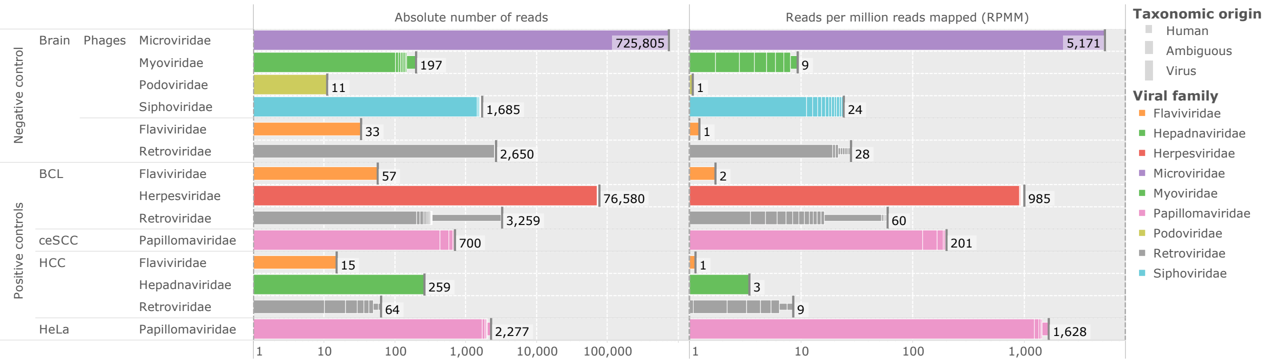


Figure 13.7: *Overview of identified homologous regions in positive and negative experimental controls.* Left panel: cumulative numbers of reads assigned to viral taxonomic families (log-scale). Each bar represents a homologous group (HOG) colored according to viral taxonomic family. Bars comprise several segments, each representing a homologous region (HOR). Heights of segments indicate the putative origin of reads assigned to this region (human, viral, or ambiguous). Viral families of bacteriophages are marked accordingly. Right panel: Analogous to left panel, but the lengths of bars represent relative rather than absolute abundances quantified in cumulative reads per million reads mapped (RPMM).

Reconstruction of novel transcripts by de novo *assembly.* In order to identify transcripts of novel viruses that do not map to known references, we generated *de novo* transcriptome assemblies of all unmapped reads. We applied the two *de Bruijn* graph based assembly methods Oases and Trinity that demonstrated best-in-class performance in recent evaluations[57] on sequencing data of the NB2 panel. This sequencing data is especially amenable to assembly due to its long read length (see Table 13.1). Assembly resulted in 14,077 and 21,510 reconstructed neuroblastoma 4S contigs for Oases, and Trinity, respectively (see Figure 13.10). Assembly of the neuroblastoma 4 sample yielded 11,828 and 12,341 contigs from the same methods. Results of Oases and Trinity assemblies are comparable in terms of contig length. All contigs were subjected to taxonomic annotation using high-sensitivity TBLASTX annotation based on

[57] Grabherr et al. (2011), Schulz et al. (2012), Zhao et al. (2011a)

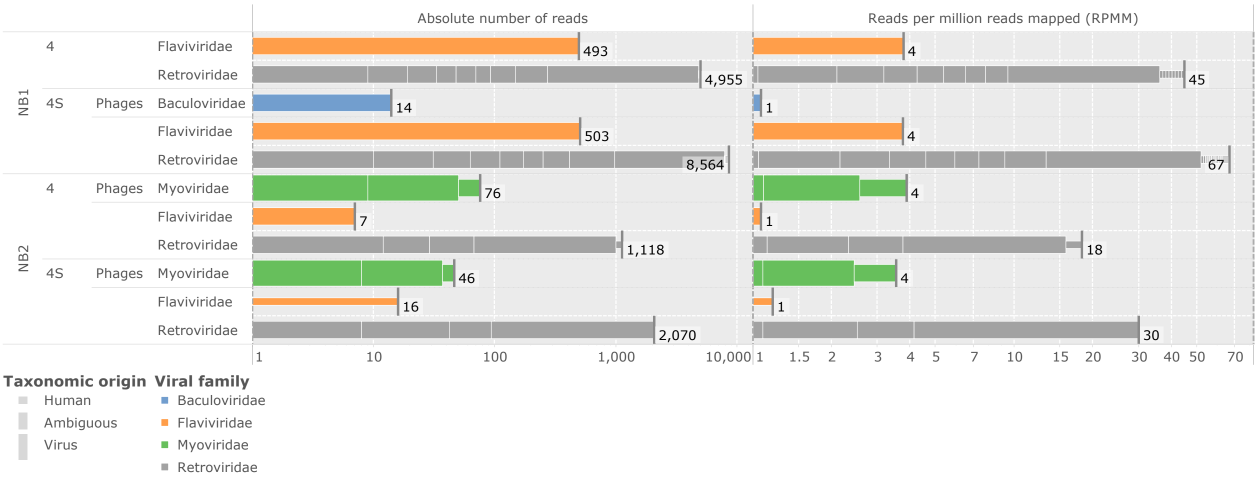


Figure 13.8: *Overview of identified homologous regions in neuroblastoma samples.* Left panel: cumulative numbers of reads assigned to viral taxonomic families (log-scale). Each bar represents a homologous group (HOG) colored according to viral taxonomic family. Bars comprise several segments, each representing a homologous region (HOR). Heights of segments indicate the putative origin of reads assigned to this region (human, viral, or ambiguous). Viral families of bacteriophages are marked accordingly. Right panel: Analogous to left panel, but the lengths of bars represent relative rather than absolute abundances quantified in cumulative reads per million mapped (RPMM).

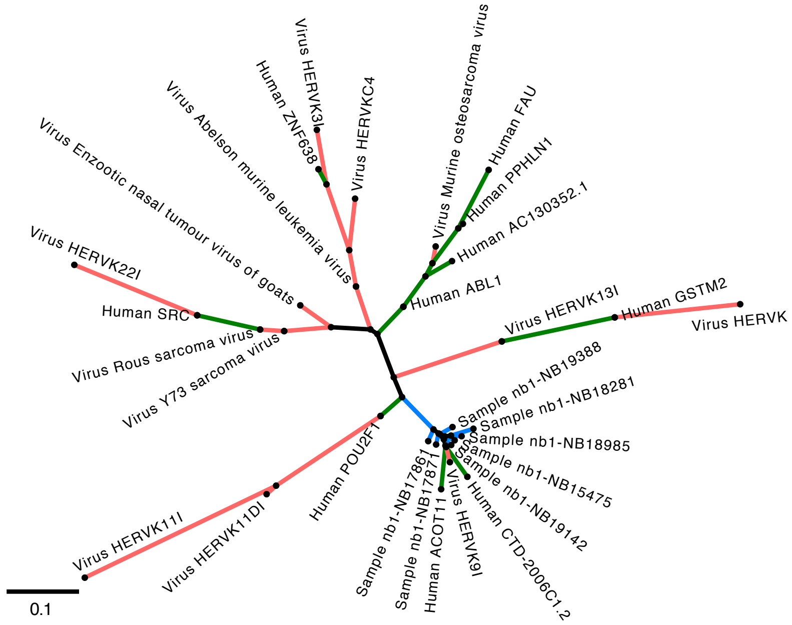


Figure 13.9: *PHuman-viral phylogeny based on a HOR.* Phylogenetic tree of HOR #16 of the NB1 stage 4 panel. Viral reference sequences are indicated with red branches and associated tip labels ('Virus') while human factors are labeled with green branches. Blue branches represent consensus sequences of neuroblastoma reads ('Sample'). The tree was generated by the maximum likelihood approach PhyML using the multiple sequence alignment of the HOR as input (see Materials and Methods). Distances between nodes are quantified as substitutions per site. As can be derived from the tree, neuroblastoma consensus sequences are tightly clustered in close proximity to the endogenous retrovirus HERVK9I and two human factors, thereby unambiguously indicating the human origin of these neuroblastoma reads. Clusters of other sequences represent well known sequence homologies, as for example between human ABL1/SRC genes and acutely transforming retroviruses.

human and viral content of the NCBI nt and nr databases (see Materials and Methods). Overall, 72 contigs (0.1-0.16% of contigs of any specific assembly) were identified to be of putative viral origin. 26 contigs were assigned to bacteriophage references and excluded from further analysis. Based on searches against the full NCBI nr

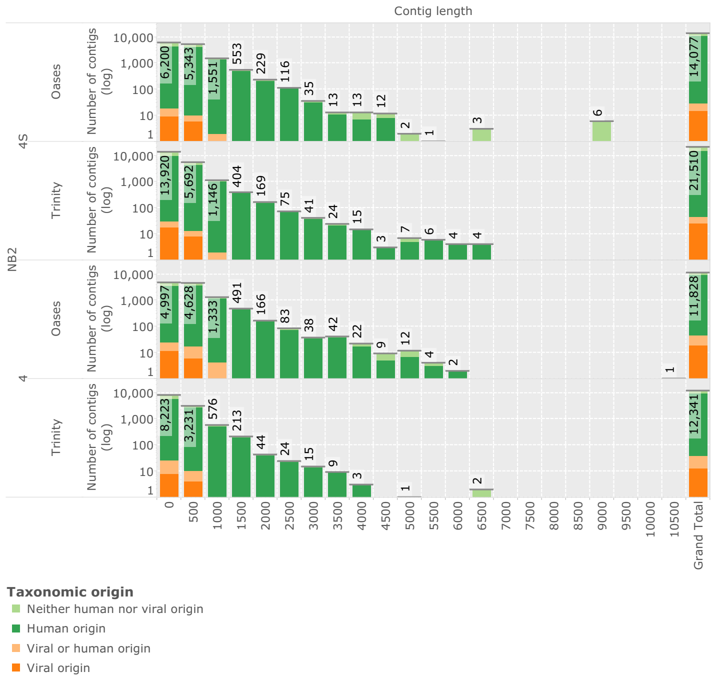


Figure 13.10: *Reconstruction of novel transcripts by* de novo *assembly.* Histograms display lengths of reconstructed sequence contigs assembled from unmapped reads of NB2 stage 4 and stage 4S samples (y-axis in log-space). Two independent assembly methods, Trinity and Oases, were used in the reconstruction. The grand total number of contigs reconstructed within each assembly is displayed in the rightmost column. Reconstructed contigs are annotated with their putative taxonomic origin as inferred by comparison with NCBI nucleotide (nt) and protein (nr) archives using TBLASTX database searches.

and nt databases followed by manual inspection, all remaining 46 contigs were determined to display higher similarities to bacterial or human sequences than to any viral reference.

Neuroblastoma is a pediatric tumor of the sympathetic nervous system that represents the most common form of cancer in infancy. It is characterized by a striking diversity in biology and clinical behavior of its subtypes. This heterogeneity as well as supporting epidemiological findings are highly suggestive of infectious cofactors involved in genesis and maintenance of the disease.[58] While several studies utilizing technologies with lower sensitivity compared to our approach have identified human polyomaviruses in neuroblastoma and pediatric embryonal tumors,[59] newer investigations seem to render these associations inconclusive.[60] However, viral commensals of the families *polyomaviridae* and *adenoviridae* are indeed suspected to acquire rare transforming properties as a consequence of viral latency or defective replication[61] and to encode oncogenes[62] whose carcinogenic potential in human is currently investigated.[63] We undertook the first systematic search for known and unknown viruses in transcriptomes of metastatic neuroblastoma by analyzing deep sequencing RNA-Seq data of 14 metastatic neuroblastomas from two tumor stages as well as positive and negative experimental controls.

Several high-throughput methods for detecting viral sequence reads among human RNA-Seq data have been developed. Among these methods, PathSeq, CaPSID and RINS are most prominent due

[58] Heck et al. (2009), Menegaux et al. (2004)

[59] Flaegstad et al. (1999), Jørgensen et al. (2000), Krynska et al. (1999)

[60] Stolt et al. (2005)

[61] zur Hausen (2001)

[62] Berk (2005), Eash et al. (2006)

[63] Elgui de Oliveira (2007), zur Hausen (2009b)

to their design as reusable computational pipelines. In this study we selected CaPSID and RINS due to their high performance and public availability and compared their detection performance with that of our novel method Virana. Both CaPSID and RINS follow a subtractive approach, e.g. they separately map input data to viral and human reference sequences and subtract viral read mappings that are similar to the human genome from the analysis. While CaPSID is conceptualized as a generalized framework that supports the subtraction process by means of a database and a web server, RINS features an integrated pipeline that splits input reads into shorter fragments in order to increase mapping sensitivity, followed by transcriptome assembly of putative viral reads into full length transcripts.

Both RNA and DNA viruses may share considerable sequence homology to human factors due to reasons such as lateral gene transfer, oncogene capture, ancestral endogenization, or insertional mutagenesis leading to chimeric transcripts.(64) Such homologous transcripts may display human-viral sequence similarities of 86% (Bovine Herpes virus) and up to 92% (acutely transforming retroviruses). Subtractive approaches silently discard these transcript from the analysis due to their similarity to the human reference genome. In contrast, our novel method Virana follows a radically different approach. Instead of separate mapping to viral and human reference database followed by digital subtraction, Virana undertakes a particularly sensitive read mapping to a combined set of human and viral references. By allowing for multimaps, this mapping strategy facilitates discovery of viral transcripts regardless of their similarity to human factors. Apart from being conceptually simpler by relying on only one mapping step and discarding the subtraction procedure that is both possibly erroneous and computationally costly, this approach empowers the mapper to make informed decisions about relative alignment quality by weighing different human and viral reference positions against each other. As a direct consequence of this increased mapping quality, paired-end reads can be mapped across human and viral references, allowing for detection of human-viral chimeric transcription and proviral integration events.

We quantitatively validated Virana's approach both in settings involving simulated reads as well as in real-world scenarios involving experimental positive and negative controls. In these validations, Virana displays significantly higher detection sensitivities than competing approaches especially at high rates of viral sequence divergence exceeding 5% that are common for tumor viruses.(65) As a consequence, Virana was the only method able to detect all viral families independent of sequence divergence in the validation data set. In spite of the additional processing undertaken by our method, Virana features between and two and three times faster execution speeds compared to related methods.

Interestingly, viral reads analyzed in the sequence divergence

(64) Butel (2000)

(65) de Villiers et al. (2004), Karlin et al. (1990), Simmonds et al. (2005)

validation originate from a broad array of viral species, only two of which infect mammalian hosts and none of which display significant human-viral sequence homology. As a consequence, this validation favors subtractive approaches by reducing the danger of erroneous subtraction of viral reads that are similar to the human genome. In addition, the sequence divergence validation contained reads sampled at high coverage. However, transcripts of tumor viruses are often expressed at only low cellular abundances and are thus expected to have low sequence coverage. We therefore next validated the ability of viral detection approaches to detect viral transcripts homologous to human factors at varying levels of sequence coverage. Virana, by virtue of not relying on digital subtraction, demonstrated superior sensitivity at this validation both in settings of natural sequence homology as well as in cases of human-viral chimeric transcription. Specifically, Virana was the only method able to detect evidence for all viruses even at low twofold coverages. We observed that both RINS and CaPSID discarded a substantial amount of human-viral homologous transcripts due to their high similarity to the human reference genome, a fact that explains the lower performance of these methods in this validation scenario.

Analysis of positive and negative experimental controls further reveals that Virana is able to detect viral transcripts associated with four types of cancer at a high dynamic range of relative abundances. While Virana displays a slightly reduced specificity in simulated and experimental evaluations, these false positive hits are limited to only two viral families (*Flaviviridae* and *Retroviridae*) that display high sequence similarity to human factors. These hits are additionally annotated with an ambiguous taxonomic origin by Virana. In addition, Virana provides extensive support for investigating such ambiguous viral hits by analyzing the homologous context of putative viral reads in a context of multiple sequence alignments and phylogenies.

In principle, several biological confounding factors may hinder detection of viral transcripts by any sequence-based method. Low concentration and extratumoral location of viral producer cells[66] or selection of growth-autonomous cells in progressed tumors[67] can significantly dilute the number of viral transcripts in a sample. Additionally, known tumor viruses such as high-risk HPV strains, EBV, and MCPyV selectively transcribe their genome during viral latency (HPV: E6/7, EBV: EBNA1/2, MCPyV: large T antigen),[68] thus generating only low abundances of tens (MCPyV) to hundreds (KSHV, EBV) of transcripts per cell.[69] Last, transcription of human oncogenic factors modulated by viral or endogenous retroviral promoters as well as 'hit-and-run' mechanisms of viral oncogenesis that imply loss of viral material may predispose cells to transformation without requiring maintenance of viral transcripts.[70]

Our approach aims to counteract these confounding factors by two strategies: first by sequencing neuroblastoma transcriptomes

[66] zur Hausen (2009b)

[67] Voisset et al. (2008)

[68] Dyson et al. (1989), Feng et al. (2008), Houben et al. (2010), Kelly et al. (2009), Klein et al. (2007), Scheffner et al. (1990), Young and Rickinson (2004)

[69] Cornelissen et al. (2003), Feng et al. (2008), Metzenberg (1990)

[70] Coffin et al. (1997b), McLaughlin-Drubin and Munger (2008), Ono et al. (1986), Si and Robertson (2006), Tomlins et al. (2007)

at comparatively high depth in order to detect rare transcripts and second by using several biological replicates at different tumor stages, thus reducing the probability of total loss of viral material from all analyzed samples. Based on statistical estimations concerning Virana's homologous region construction process and the sequencing depth of our experimental data, we can conclude that our approach requires minimal abundances of only two average-length viral transcripts per cell even under adverse conditions such as high viral divergence or extensive human-viral sequence homology. While representing a theoretical sensitivity that may be altered by sequencing biases,(71) these copy numbers compare very favorably with related estimates reporting minimal abundances of one to several complete viral genomes per cell.(72)

After applying Virana to several positive control panels of human cancers with known viral cofactors and accurately reconstructing large fragments of viruses that are causally related to the respective tumors, we analyzed neuroblastoma transcriptomes at high sequencing depth and using two different sequencing platforms. Analyses of neuroblastoma transcriptomes resulted in the detection of putative viral transcripts with high local sequence similarity to several viral families. However, automatic taxonomic annotation as well as detailed manual inspection of homologous regions pertaining to these families revealed the human or bacteriophage origin of all transcripts. While we could find differences in the abundance of HERV-K transcripts between neuroblastoma stages 4 and 4S, the causative role of HERV transcription with regard to oncogenesis is currently unclear(73) and, as to our knowledge, only tentative associations with specific cancers have been made as to date.(74) Apart from these tentative differences in HERV-K abundances, no quantitative difference between neuroblastoma stages 4 and 4S could be identified with regard to viral transcription.

In conclusion, our observations provide negative evidence regarding the contested question of putative viral cofactors of metastatic neuroblastoma by suggesting that viruses are unlikely to be frequent cofactors in the maintenance of metastatic neuroblastoma.

(71) Fang and Cui (2011)

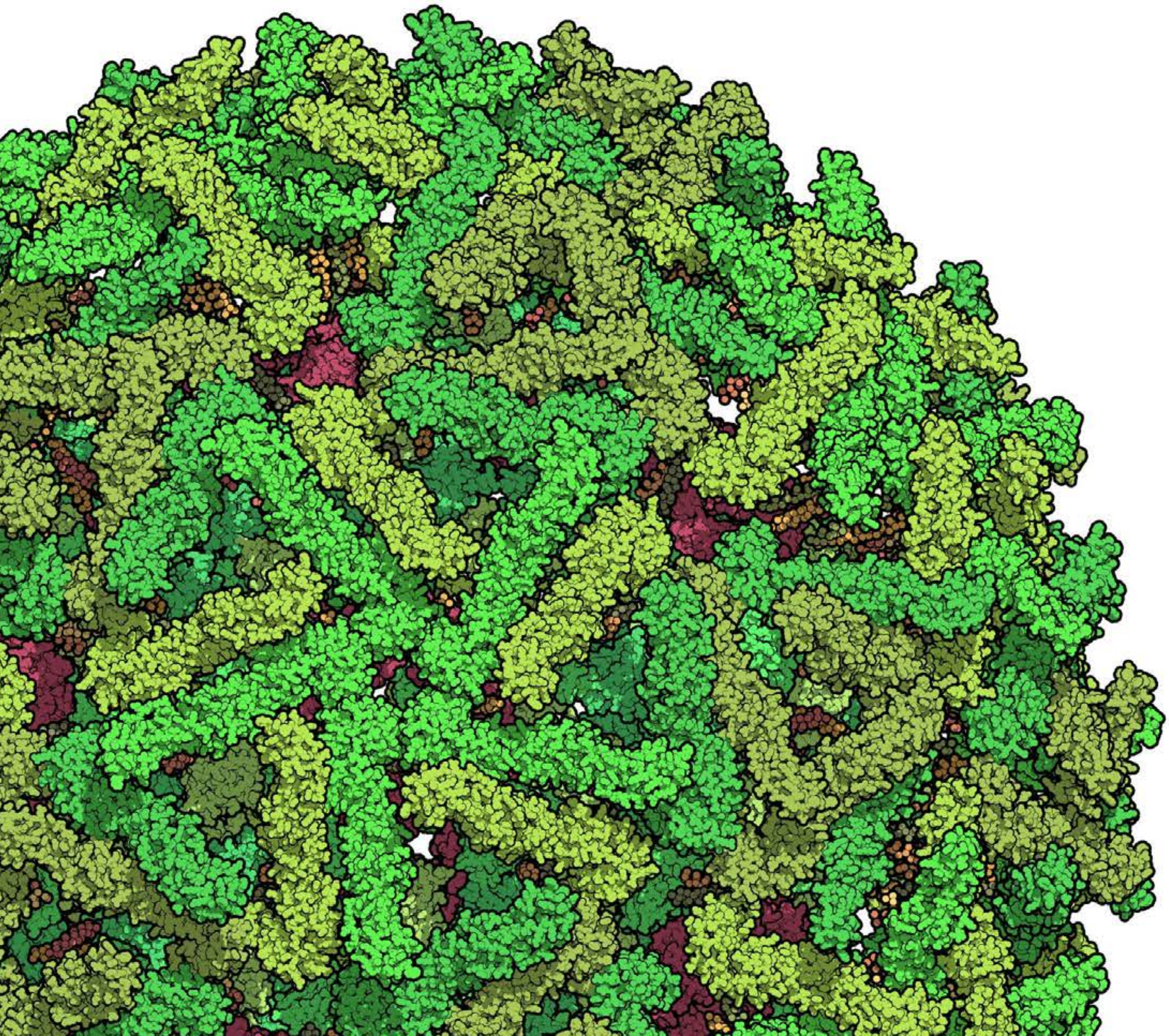
(72) Bexfield and Kellam (2011), Feng et al. (2007), Moore et al. (2011)

(73) Bannert and Kurth (2004)

(74) Wang-Johanning et al. (2003)

III

Host-pathogen protein interactions



This third chapter outlines the current state of the art in detecting protein interactions in a wider context of inferring host-pathogen interactions, i.e., protein interactions between two proteins of pathogenic and human origin, respectively. These protein interactions are especially relevant in light of antiviral therapy since drugs targeting viral host factors are promising candidates for future generations of highly effective antivirals. Apart from presenting the current state of virology with respect to viral-host interactions, this chapter additionally focuses on methods for detecting and interpreting physical protein contacts from affinity purifications, a specific class of experimental protein interaction assays.

At its beginning in Section 14, the chapter introduces the concept of viral host factors, i.e., proteins that are essential for viral infection, replication, and persistence. After detailing viral use of such factors for several purposes, the current state of development of antiviral agents such as vaccines and specifically targeted drugs is presented in Section 15 and medicines targeting or employing human host factors are highlighted. Subsequently in Section 16, an introduction to experimental and computational methods for measuring protein interactions is presented and further illustrated with current approaches for detecting viral host factors in Section 17. This chapter closes with Section 18 where a study is presented that proposes a new statistical method for inferring protein interactions from protein purification data and an application of this method on host-pathogen (HP) interactions of HIV is presented.

14 *Virally targeted host factors*

While viruses may, *in principle, exhibit highly virulent and thus pathogenic phenotypes that kill the host, such a strategy is not in the evolutionary interest of the virus since it requires a host reservoir for prolonged survival. Viral strains that co-exist with rather than overwhelm the host are therefore more likely to have epidemiologically and thus evolutionarily favorable characteristics. If regarded from an evolutionary viewpoint, this long-term coexistence of viruses and their hosts displays a "precarious balance"[1] and persistent viruses are required to develop intricate interactions with their host. Among these factors are entry factors, a range of macromolecular entities such as protein receptors, lipids, saccharides, as well as their combinations that are located in the cell membrane and promote viral entry.[2]*

Due to their large genomes, correspondingly slow rate of sequence variation, and large coding potential, DNA viruses exhibit strategies for immune evasion that are distinct from strategies of divergent RNA viruses; while the latter predominantly rely on antigenic variation of envelope proteins in order to evade immune detection, DNA viruses utilize both passive as well as active means of escaping the host immune system. In particular, poxviruses, herpesviruses, and adenoviruses utilize their large protein repertoire to actively suppress the host immune system, disable apoptosis, and thus persist in the host indefinitely.

This section introduces crucial host factors that either are utilized by viruses to gain access to the cell (entry factors) or that represent components of the host immune system which are subverted by viral factors in order to achieve persistence. These interactions are almost exclusively mediated by proteins, and the resulting virus-host protein interactions serve as important discovery tools for basic research aiming to highlight key host immune pathways. In addition, and more relevant for this thesis, such protein interactions are interesting antiviral drug targets, a topic that is explored in more detail in subsequent sections of this chapter.

Viral entry factors and coreceptors

The use of and adaption to cellular entry factors is a major determinant for infectiousness, cellular tropism, and host range of viruses.[3] Host range in itself is a highly variable characteristic

[1] Tortorella et al. (2000)

[2] Baranowski (2001), Marsh and Helenius (2006), Schneider-Schaulies (2000)

[3] Boyd et al. (1993); the term *tropism* here denotes the the specificity of a virus for either a particular host cell type (cellular tropism), a host tissue (tissue tropism), or a host species (host tropism, often also denoted as host range), cf. McFadden et al. (2009). Viral tropism is determined by both susceptibility (allowing viral entry) and permisseness (allowing viral replication) of the infected cells, tissues, or species.

among viruses (cf. Chapter I). Often, multiple entry factors are required for efficient entry and secondary factors (often termed *coreceptors*) are employed by viruses in order to increase the efficiency of the process. Of the about 90 human viruses with known cellular entry factors (representing 21 of 23 known viral families that infect humans), approximately 75% use protein receptors while the remainder specializes on linear polysaccharides for entry.[4]

The choice of host factors has important consequences for host specificity: this is demonstrated in the case of HIV-1 where small changes to viral proteins are sufficient to result in functional amino acid changes that determine CCR5 or CXCR4 coreceptor usage, thus altering cellular tropism of the virus. This propensity to adapt to host factors is varying between viruses and may result in cross-species infectivity. Pathogens as for instance HBV or Mumps virus (MuV) are *viral specialists* that probably have co-evolved with the human species and exclusively infect humans; on the other hand, viruses such as rabies virus are able to infect a wide range of hosts by utilizing cellular receptors that are highly conserved (more than 90% amino acid identity) between host species.[5]

As will be discussed later in Chapter IV, divergent viral species that contain highly variable minority variants rapidly evolve to novel phenotypes if selection pressure is sufficiently strong; these phenotypes frequently also include novel binding interfaces that promote both host tropism as well as immune escape by antigenic variation. Importantly, sufficient changes in cellular tropism may also result in modified host ranges (host tropism). This phenomenon of a pathogen adapting to a novel species is closely associated with disease emergence and viral zoonosis (cf. Chapter I) and may be a contributing factor to the current AIDS pandemic as well as for up to one novel human disease each year, many of whose are originating from bats, primates, and rodents.[6]

Due to the particular structural (and therefore, by extension, also sequence-based) characteristics that determine tropism of a virus, interfaces between viral and host proteins are often associated with antigenicity, i.e. host immune response. Indeed, receptor and antibody binding sites (epitopes) of viral proteins have been shown to frequently overlap.[7] As a consequence, immune evasion and host cell tropism seem to be intricately linked and changes of host cell tropism in order to evade immune recognition may thus be an important part of the virus-host evolutionary arms race.[8]

[4] Woolhouse et al. (2012)

[5] Simmonds (2001), Woolhouse et al. (2012), Woolhouse (2001)

[6] Antia et al. (2003), Bae and Son (2010), Haagmans et al. (2009), Parrish and Kawaoka (2005)

[7] Baranowski (2001), Iorio et al. (1989), Zhang et al. (2007)

[8] Woolhouse et al. (2002)

Viral subversion of the innate immune system

Due to their high abundance, diversity, and adaptivity, viruses pose particular challenges to the host immune system. The mammalian immune system consists of two large components. The *innate immune system* acts as a first line of defense against viral infection, has only limited specificity, and does not provide an immunological memory. The *adaptive* or *acquired* immune system, on the other hand, is mostly relevant in later phases of infection. It is driven by lymphocytes that evolve to high specificity by gene rearrangement and clonal selection and are retained for years after initial infection, thereby enabling quicker responses to subsequent infections with the same pathogen.

In the following, we will first discuss viral strategies for modulation and subversion of the host innate immune system in order to motivate and inform the search for viral host factors involved in these pathways. Subsequent sections will introduce similar interactions of viruses with the adaptive immune system. In addition, knowledge of these core immune components is relevant for the next chapter of this thesis where mechanisms of viral immune escape and the immunomodulatory activity of certain drugs will be discussed (cf. Chapter IV).

The majority of the human population is chronically or latently infected with one or several viruses such as herpes simplex viruses types 1/2 (HSV-1/2), human cytomegalovirus (HCMV), varicella-zoster virus (VZV), or human herpesviruses 6 and 7 (HHV-6 and HHV-7).(9) Indeed, there is paleovirological evidence that herpesviruses co-evolved with the human species for at least 400 million years.(10) While these viruses are usually causing asymptomatic infections, their long-time persistence within the host and within the human population as a whole well illustrates the ability of these pathogens to control the human immune response indefinitely. These viruses employ their long genomes and correspondingly large reservoir of proteins in order to inhibit multiple components of the innate immune system such as cytokine and chemokine signaling,(11) as well as pathways of the adaptive immune system such as the complement cascade, antibody-mediated effector mechanisms, and MHC I/II antigen processing and presentation.(12)

Recognition of viral factors by the host as well as proper activation of innate host defense pathways such as the *interferon* system are essential for mounting an effective immune response. Congenial failure in these components may often result in higher susceptibility to viral diseases in general; this is particularly true for highly aggressive viruses such as *Lassa* that may not leave the host sufficient time for mounting an adaptive immune response.(13)

(9) Davison et al. (2009)

(10) McGeoch and Gatherer (2005)

(11) Alcami (2003), Lanier (2008)

(12) Lambris et al. (2008), Lubinski et al. (1998), Wiertz et al. (2007)

(13) Sadler and Williams (2008), Zampieri et al. (2007)

Since successful activation of innate immune pathways may hinder initial infection of the cell, these pathways are regularly antagonized by both acutely and latently infecting viruses. In the following, we will concentrate on the two major components of the innate immune response: pattern-recognition receptors and inflammatory cytokines.

Pattern-recognition receptors. Pattern-recognition receptors (PRRs) are of particular importance in the context of virus-host protein interactions. These host proteins detect molecular patterns of pathogenic infection and can elicit signaling cascades that activate other actors of the innate immune system such as the type I interferons IFNα/IFNβ and pro-inflammatory cytokines which in turn direct T helper cells and cytotoxic T cells of the adaptive immune system.[14] In particular, PRR response activates transcription factors interferon (IFN)-regulatory factor 3 (IRF3), IRF7 and/or nuclear factor-κB (NFκB), resulting in the expression of IFNβ and, subsequently, of IFN-stimulated genes (ISGs) that collectively avert or limit viral infection.[15]

Of the known families of PRRs, the Toll-like receptors (TLRs) and the retinoic-acid-inducible gene I (RIG-I)-like receptors (here denoted as RLRs and comprising retinoic-acid-inducible gene I (RIG-I) and the melanoma differentiation-associated gene 5 (MDA5)) have been investigated in detail.[16] Both of these protein families detect pathogen-associated molecular patterns (PAMPs), in particular viral RNA and DNA, but also lipids, lipopolysaccharides, as well as heat shock proteins and glycoproteins that are specific for broad classes of pathogens.[17] TLRs are expressed both intra- and extracellularly by antigen-presenting cells (APCs) such as macrophages which ingest and degrade pathogens.[18]

In general, PRRs are specialized towards detecting viruses with specific genomic features; while double-stranded viral DNA (dsDNA) genomes are detected by Toll-like receptor 9 (TLR9), viruses with dsRNA genomes are detected by TLR3 and MDA5, the latter of which can also detect positive-sense single-stranded RNA genomes. Other single-stranded RNA genomes with retroviral and negative-sense characteristics are detected by TLR7/TLR8 and RIG-I, respectively.[19] Viral factors inhibiting PRRs often manipulate Toll/IL-1 receptor (TIR) domains that are highly conserved across TLRs and have been shown to be essential for signaling.[20]

By interacting with these cellular components, for example by protease-mediated cleavage (HCV), transcriptional down-regulation (HTLV-1) or antagonistic binding (Influenza A), viruses may inhibit downstream activation of NFκB and IRFs.[21] By contrast, and as an important feature of cellular self-recognition, endogenous host nucleotides do not activate PRRs due to their single-stranded nature (in case of RNA), reduction of immuno-stimulatory CpG-DNA (in case of DNA), extensive base modifications (pseudouridine), as well as the 5′ capping with methylguanosine and monophosphates.[22]

[14] Bowie and Unterholzner (2008)

[15] O'Neill (2008), O'Neill and Bowie (2007)

[16] O'Neill (2008), Takeuchi and Akira (2008)

[17] Jin and Lee (2008), Saito and Gale (2008)

[18] Akira et al. (2006)

[19] Bowie and Unterholzner (2008)

[20] O'Neill and Bowie (2007)

[21] Datta et al. (2006), Li et al. (2005), Pichlmair et al. (2006), Stack et al. (2005)

[22] Akira et al. (2006), Hornung et al. (2006)

These characteristics of host nucleotides are often mimicked by viruses in order to evade immune detection; in particular, viruses may employ either viral or host factors to cap viral mRNAs or shield their genomes from host proteins.[23] Conversely, viruses may also trigger PRRs to modulate host pathways for their own purposes and identical PRRs may be triggered by different viruses with either negative or positive results on virulence and disease progression.[24]

In addition, and similar to the modulation of TLR signaling pathways, viruses also directly inhibit RLR signaling by binding and cleavage of TLRs and downstream effectors.[25] Many viruses target later stages of innate immune signaling; indeed it has recently been shown that most TLR receptor signals converge at TRADD (tumor-necrosis factor receptor (TNFR)-associated via death domain)[26], an adaptor molecule that can shuttle between cytoplasm and nucleus and activate apoptosis. Similarly, all known PRRs signaling converge at the IKK family of proteins. Such bottlenecks within the human innate immune system are attractive targets for viral inhibition and subversion (see later Sections of this chapter for a detailed investigation of the importance of bottleneck proteins). Indeed, IKK proteins and the transcription factors they activate are targeted by viral factors of HCV, hantaviruses, and coronavirus.[27] Also, both HCV and HIV employ viral factors to subvert the human DEAD-box protein 3 (DDX3), a protein co-complexing with IKK factors, for replication and transcript transport, respectively.[28]

Inflammatory cytokines. Natural killer (NK) cells are important components of the innate immune system that produce inflammatory cytokines. Cytokines are polypeptides secreted by NK cells and cytolytic T cells that are essential for organizing a wide array of cellular processes such as inflammation, proliferation, and differentiation and may induce cell death or directly mediate cytotoxicity. It is assumed that activation of cytokines and the associated immune responses regularly clears less severe infections from the host, often in pre-symptomatic stages of viral disease. Interestingly, it is the effects of cytokines and not viral activity that produces the flu-like symptoms of many viral diseases such as fever, headache, and drowsiness.[29]

Due to their various protective effects on the host cell, several classes of cytokines such as interleukins, interferons, tumor necrosis factors, and chemokines (pro-inflammatory cytokines that direct other immune cells to infected tissues by chemical gradients) are prime targets for viral subversion. Many viral strategies for inhibiting cytokines focus on blocking interferon and chemokine production by blocking transcription factors,[30] by inhibiting cytokine maturation,[31] or by down-regulation through activation of cellular receptors.[32]

Another ingenious viral strategy for cytokine subversion consists of antagonistic binding (or "scavenging") of cytokines and

[23] Flanegan et al. (1977), Haasnoot et al. (2007), Habjan et al. (2008), Plotch et al. (1981)

[24] Bowie and Unterholzner (2008)

[25] Barral et al. (2007), Brand et al. (1997), Mibayashi et al. (2007)

[26] Ermolaeva et al. (2008), Michallet et al. (2008), Pobezinskaya et al. (2008)

[27] Alff et al. (2008), Devaraj et al. (2007), Otsuka et al. (2005)

[28] Ariumi et al. (2007), Yedavalli et al. (2004)

[29] Tortorella et al. (2000)

[30] Hirsch and Shenk (1998), Juang et al. (1998)

[31] Messud-Petit et al. (1998)

[32] Karp (1999)

chemokines by soluble viral receptors. These bait proteins are homologous to cellular cytokine receptors and are expressed mainly by large DNA viruses such as herpesviruses and poxviruses in order to interfere with TNF and IFN signaling.[33]Other herpesviruses are known to express chemokine receptors that continuously activate signaling cascades, thus modulating the cell cycle and contributing to tumor formation.[34] Similarly, KSHV and MCV, a poxvirus, are associated with tumor formation and secrete broad-spectrum chemokine receptor antagonists.[35]

Finally, cytokine system subversion by viruses can also serve more nefarious purposes. By expressing viral cytokine receptor homologs on the cell surface that have subtly altered signaling functions, HCMV, for instance, selectively depletes chemokines in the cellular environment and thus inhibits leukocyte activation. Other viruses employ the chemokine system for their own use and specifically activate chemokine receptors in order to attract immune cells for infection[36] or for guiding Th1/Th2 T-cell differentiation by selectively activating chemokine receptors in a manner beneficial to viral immune escape.[37]

Viral subversion of the adaptive immune system

Following the initial innate immune response to viral infection, a second layer of host defense, the adaptive immune system, is activated. In contrast to the innate immune response, the adaptive immune system coordinates more specific and longer lasting antiviral activities. The adaptive immune system consists of two large components, termed *cellular* and *humoral* arms that provide immunity for intracellular and extracellular spaces, respectively. Both arms are triggered by cytokine and chemokine signaling of the innate immune system and natural killer cells and are interconnected and coordinated by interleukins and other messenger molecules.

The humoral arm consists mainly of B cells that produce antibodies, highly diverse protein complexes that bind to antigens on viral particles or infected cells. Antibody binding marks these antigen-bearing entities for destruction by natural killer cells, components of the complement cascade, and cytotoxic T lymphocytes (CTLs or T-cells). T-cells are the most important actors of the cellular arm of adaptive immunity. These cells are able to detect foreign peptides displayed by the major histocompatibility complex (MHC) on the surface of antigen-presenting cells. Both arms of the adaptive immune system are prime targets for viral subversion as the following paragraphs will illustrate.

Major Histocompatibility Complexes Host cells present peptides that have been expressed within the cell and subsequently degraded by the proteasome in specialized protein receptors located on the cell surface. These receptors, most prominently the *Major*

[33] Lalani et al. (1997), Novick et al. (1992), Spriggs (1995)

[34] Kledal et al. (1997)

[35] Damon et al. (1998)

[36] Zou et al. (1999)

[37] Endres et al. (1999)

Histocompatibility Complexes (MHC), enable lymphocytes of the immune system such as cytotoxic T Lymphocytes to detect invading pathogens by binding to characteristic antigenic peptides, termed *epitopes*. Cytotoxic T Lymphocytes implement the cellular arms of the adaptive immune system while also providing important signaling services for humoral, or antibody-mediated, immunity.

MHC complexes play an essential role in the detection of foreign peptides and come in two varieties: MHC-I molecules are expressed by most vertebrate cells and display epitopes that originate from the cytosolic compartment and have been degraded by the proteasome. Upon binding to T-cell receptors and CD8 coreceptors on CD8$^+$ T-cells, CTLs may activate immune pathways that allow for a targeted immune responses and eliminate virus-infected cells.[38]

In contrast to MHC-I molecules that primarily display peptides of intracellular origin to CD8$^+$ T cells, MHC-II molecules present antigenic peptides also from the extracellular space that underwent phagocytosis or endocytosis to T-cells employing the CD4 coreceptor, thus termed CD4$^+$ T cells (*T-helper cells*).[39] MHC-II molecules are usually expressed on "professional" antigen-presenting cells (professional APCs or PAPCs) such as B cells, dendritic cells, and macrophages. Upon recognition of antigenic epitopes, CD4$^+$ T cells also induce antiviral response as well as T cell proliferation.[40]

Human MHC-I isoforms are predominantly constituted of products of human leukocyte antigens (HLA), a genetically highly variable family of genes distributed over at least 9 loci in the human genome that encode several thousand known alleles in the human population[41],[42] These HLA genotypes display varying MHC binding properties to peptides and are major determinants of successful human immune response to foreign as well as to autoimmune factors.[43]

Peptide transport and MHC maturation MHC-I molecules in particular undergo a process of maturation that is completed by peptide loading and translocation of the loaded complex onto the cell surface. Mature MHC-I complexes exist as a trimer comprising of a heavy chain (43-kDa membrane glycoprotein featuring a peptide groove), a light chain (12-kDa soluble protein β-microglobulin), and the ligand of the complex, a 8-10 residues peptide fragment encoding the antigenic epitope.

Antigenic peptides are generated by digestion of polyubiquitylated proteins by the proteasome, a large intracellular protein complex that functions as a multicatalytic protease. The digested peptide fragments are transported from the cytoplasm to the endoplasmic reticulum (ER) by the hetero-dimeric *Transporter complex associated with Antigen Processing* (TAP).[44] where MHC-I complexes mature and are loaded with peptides. Specifically, preparation and peptide loading of MHC-I molecules requires a macromolecular assembly termed the *peptide-loading complex* (PLC). This complex transiently interact with the MHC-I complex and are required for

[38] Braciale et al. (1987), Brodsky et al. (1999), Guidotti and Chisari (1996)

[39] Pieters (1997)

[40] Jonjić et al. (1989)

[41] Marsh (2013)

[42] The remaining component not encoded by HLAs is a protein of comparably low molecular weight, $\beta 2$ microglobulin

[43] In addition, they may be an important factor influencing human mate selection, a process that may be evolutionarily selected to increase immune diversity in the human population Brennan and Kendrick (2006), Parham and Ohta (1996).

[44] Horst et al. (2011)

stable expression of the receptor on the cell surface.[45]

Since MHC-I molecules are constructed from a highly variable set of HLA gene products that display large differences in assembly kinetics, MHC-I complexes are not fully stable after initial assembly. In addition, peptides may display varying binding efficiency and peptide loading may be deferred. For these reasons, chaperone supervision is required to facilitate correct assembly[46] One of these chaperones is *tapasin*, a transmembrane glycoprotein that is essential for MHC-I maturation by stabilizing the groove of the MHC complex until peptide loading has occurred.[47] As part of its stabilizing function, tapasin selectively chooses peptides for loading that kinetically stabilize the MHC complex, a process that is denoted as peptide editing. In addition, tapasin also performs recruitment functions of other proteins such as TAP to the peptide loading complex.[48]

Viral subversion of MHC pathways Viruses have evolved a range of strategies to subverting host MHC pathways. Among these strategies are inhibition of degradation of viral proteins by the proteasome, modulation of MHC-I transport by TAP, hindrance of tapasin-mediated localization of MHC-I complexes on the cell surface, and retainment or degradation of matured MHC molecules. Herpesviruses in particular inhibit each known step of the MHC-I presentation pathway by interfering with the degradation of viral proteins, by blocking synthesis of MHC-I molecules, by specifically inhibiting components of the MHC-I complex, or by promoting selective degradation or intracellular detention of these molecules.[49]

Degradation of ubiquitylated proteins into short peptides by the proteasome is blocked by several herpesviruses. Among the targeted host systems are the ubiquitin system that is crucial for degradation of viral proteins into epitopes,[50] as well the proteasomal degradation machinery that is inhibited by glycine-alanine and serine-proline repeat repeat motifs encoded in viral genomes, thus preventing processing of viral proteins.[51] Interestingly, these repeat motifs probably evolved independently[52] and are hypothesized to interfere with protein unfolding or recognition by the proteasome 19S subunit.[53]

Other cellular subsystems are also specifically targeted by herpesviral factors. At least eight different herpesviruses are known to degrade, induce confirmation alterations of, or interfere with peptide binding or ATP binding of the TAP complex in ways that are surprisingly unrelated in their mechanisms of action.[54] Herpesviruses such as HSV, EBV, and HCMV are known to block TAP-transport by different mechanisms such as antagonistic binding to peptide or ATP binding sites, by inducing conformational changes, or by tagging of TAP for proteasomal degradation[55]

[45] Peaper and Cresswell (2008), Purcell and Elliott (2008), Sadegh-Nasseri et al. (2008)

[46] Beck et al. (1986)

[47] Chen and Bouvier (2007), Ortmann et al. (1997), Schoenhals et al. (1999)

[48] Sadasivan et al. (1996)

[49] Hansen and Bouvier (2009), Powers and Früh (2008), Rowe et al. (2007), Wiertz et al. (2007)

[50] Isaacson and Ploegh (2009)

[51] Bennett et al. (2005), Levitskaya et al. (1995), Zaldumbide et al. (2007)

[52] Kwun et al. (2007)

[53] Daskalogianni et al. (2008), Masucci (2004)

[54] Ressing et al. (2013)

[55] Hansen and Bouvier (2009)

Several strategies of MHC retention are directly implemented by adenoviruses and poxviruses that block transport of MHC-I out of the ER by physical association.[56] Misfolded or incompletely assembled MHC molecules are recognized within the ER and degraded by the ubiquitin-proteasome pathway through a process denoted as ER-associated degradation (ERAD).[57] Herpesviruses subvert these degradation mechanisms by encoding viral proteins that bind directly to components of the MHC-I complex or by enforcing MHC ubiquitylation, thus inducing dislocation and degradation of these molecules by ERAD.[58]

Although generally less well studied compared to MHC-I molecules, MHC-II pathways are known to be also targeted by viral factors; herpesviruses, for instance, target both MHC-I and MHC-II molecules by degradation and redirection.[59] Similarly, HPV as well as HIV encode proteins that are assumed to inhibit endocytosis and thus modulate trafficking of antigenic peptides.[60]

Presentation of MHC molecules is also a prime target for viruses. Natural killer cells, usually considered to be part of the innate immune system, recognize the absence of MHC-I molecules on cell surfaces and mark these cells for destruction ('missing self hypothesis')[61] Since many pathogens subvert MHC expression by various mechanisms, NK cells thus detect the pathogen by the absence of immunocomplexes. Interestingly, viruses have evolved mechanisms to also counter this second line of defense: poxviruses as well as herpesviruses are known to encode nonfunctional MHC-I homologs that may mask reduced cellular MHC-I surface expression.[62] Similarly, other viruses such as HIV are suspected to actively regulate cellular expression of specific HLA loci (the constituent parts of MHC-I receptors) in order to increase the ratio of NK-protective alleles while down-regulating immunogenic HLA variants.[63]

Viral use of molecular chaperones. Due to its important role in MHC-I stabilization, tapasin is targeted by herpesvirus and adenovirus factors that inhibit peptide editing and thus facilitate retainment of unstable MHC molecules in the ER[64] or that blocking recruitment of other components of the peptide loading complex by tapasin.[65] Indeed, tapasin is not the only chaperone that is subject to viral inhibition; instead chaperones are believed to also buffer pathogens against mutational effects[66] and are a common target for a variety of viruses that have evolved mechanisms to utilize or subvert the host protein quality control machinery at multiple steps of the viral infection cycle, including endocytosis, early replication, and assembly[67]

The dependencies of viruses on chaperones opens attractive therapeutic possibilities for the development of broad-spectrum antivirals which are currently pursued.[68]

[56] Andersson et al. (1985), Burgert and Kvist (1985), Dasgupta et al. (2007)

[57] Vembar and Brodsky (2008)

[58] Boname and Stevenson (2001), Gewurz et al. (2001)

[59] Lewandowski et al. (1993)

[60] Lu et al. (1998), Straight et al. (1993)

[61] Ljunggren and Kärre (1990)

[62] Beck and Barrell (1988), Senkevich et al. (1996)

[63] Collins et al. (1998)

[64] Lee et al. (2000)

[65] Bennett et al. (1999), Park et al. (2004)

[66] Fares et al. (2002)

[67] Jockusch et al. (2001), Mayer (2004), Sullivan and Pipas (2001), Xiao et al. (2010)

[68] Geller et al. (2012)

Viral modulation of apoptosis Cytotoxic cytokines secreted by NK cells and cytolytic T cells as well cellularly delivered components of the complement system are important mechanisms of apoptosis, i.e., premature cell death of virally infected cells.[69] Since apoptosis severely limits the replication window of the infecting virus and would furthermore make viral latency impracticable, viruses have evolved several ways to block initiation and execution of apoptotic processes (cf. Tortorella et al. (2000) for a review).

A particularly attractive target for viral modulation in this context are TNF receptors, a family of death receptors that initiate apoptosis if activated by a suitable ligand such as tumor necrosis factor (TNF), a cytotoxic cytokine.[70] Poxviruses and adenoviruses disable TNF and related Fas receptors by direct binding to the receptor, antagonistic binding to the cytokine, or endocytosis.[71]

Similarly, many viruses also aim to gain control over components of the NFκB pathways. These pathways may hinder apoptosis and thus allow for prolonged viral replication. However, NFκB also increases production of IFNβ and chemokines which is detrimental to long-term viral infection; therefore, precise timing of the viral subversion processes is applied by viruses in order to maximize the benefit of their manipulations.[72]

Antibody-mediated humoral immunity The humoral arms of the immune response plays an important in preventing infections especially by pathogens that have been previously encountered by the immune system. *Antibodies*, also denoted as immunoglobulins, are host glycoproteins that bind to antigenic surface structures on viral particles. By thus binding to the virus, these *neutralizing* antibodies cause functional inactivation of the pathogen, for example by inhibiting viral interactions with cellular receptors. In addition, even non-neutralizing antibody binding can mark viral particles for later ingestion by phagocytes or lysing by NK cells. Finally, bound antibodies may coordinate a direct attack on antigenic membranes via the complement-mediated membrane attack complex.

In order to elicit strong and long-lasting immunization, pathogens have to sufficiently activate the immune system by presenting danger signals to pathogen recognition patterns, thus inducing the innate immune system produce chemokine and inflammatory signals. These signals attract dendritic cells, leukocytes, and monocytes to the site of injection, increasing antigen uptake and MHC-based presentation by antigen processing cells, expression of activation signatures, and cytokine production. Subsequently, antigen-processing cells with MHC-I or MHC-II molecules loaded with antigenic epitopes migrate towards the lymph nodes where both T-cell (cellular immunity by differentiation of antigen-specific CD4$^+$ T-cells into effector cells and activation of CD8$^+$ T-cells) and B-cell (humoral immunity by CD4$^+$-mediated differentiation of B-cells into antibody-secreting plasma cells and long-lasting memory cells) responses are activated.

[69] Shresta et al. (1998)

[70] Ashkenazi and Dixit (1998), Nagata (1997)

[71] Schreiber et al. (1997), Shisler et al. (1997)

[72] Hiscott et al. (2006)

Viral subversion of humoral immunity An essential component of the humoral immune system is presented by receptors located on phagocytic cells, B-cells or other leukocytes as well as free protein complexes of the complement system that bind to the invariable *Fc* locus of bound antibodies. Upon binding, these receptors induce effector immune responses, such as activation of complement pathways, phagocytosis of opsonized antigens, or B-cell activation resulting in virus-specific antibody production.[73]

Viruses commonly subvert antibody responses by either modulating Fc binding to effector complexes through endocytosis of the bound complex, or by viral Fc receptor homologs that functionally inactivate antibodies that are bound to viral antigens. The latter strategy is particularly used by herpesviruses such as HSV-1; this pathogen expresses glycoproteins that bind to antibodies, thus masking virions and infected cells from immune recognition.[74] This mechanism is of particular interest to herpesviruses since this family of pathogens persistently infects its host by a strategy of alternating latent and acute infections. Since this life cycle implies previous immunization of the host, the virus requires antibody masking to evade immune response at later stages of viral activations.[75]

Similarly to viral subversion of the chemokine system that is employed to promote viral transmission (see above), viruses such as HIV employ antibodies to facilitate uptake (but not ingestion) of virions by dendritic cells. These dendritic cells then shuttle the virus to new CD4$^+$ T cells for infection, thus increasing transmissibility of the pathogen.[76] In a related manner, viruses also employ host regulatory pathways that prevent the complement system from unrestrained over-activation. These regulators of complement activation (RCA) are readily targeted by poxviruses, herpesviruses, and retroviruses which encode RCA homologs or alter cellular expression of these factors, thus facilitating decay of the complement system (cf. Tortorella et al. (2000)).

[73] Anderson et al. (1997), Dempsey et al. (1996)

[74] Dubin et al. (1991)

[75] Tortorella et al. (2000)

[76] Embretson et al. (1993), Heath et al. (1995)

15 *An introduction to antivirals*

As reviewed previously in chapter I *of this thesis, infectious diseases such as AIDS, malaria, and lower respiratory tract infections are responsible for a significant portion of the global health burden.*[1] *In 2009, more than 30 million people were HIV-positive and an estimated 1.8 million deaths occurred as a result of this infection. Hepatitis viruses may infect more than half a billion people worldwide and other highly prevalent diseases such as influenza yearly kill on the order of 300,000 individuals, comparable to the inhabitants of Germany's former capital Bonn.*[2] *In addition, the increasing occurrence rates of zoonotic events and the emergence of multi-drug resistant pathogens further aggravate the situation. New medicines for combating viral diseases are urgently needed, especially if faced with highly adaptive viral species such as HCV, HIV, or influenza.*[3] *This section reviews the history of antivirals and presents several drugs that are in common use today. In particular, four broad classes of antiviral medicines and their relation to drug resistance are discussed. It is the last of these classes, antivirals that target host factors that is of special interest in the broader context of this chapter.*

Drugs targeting viral factors

In general, antiviral medicines fall into four classes. Specific inhibitors of viral targets such as small-molecule, high-affinity antiviral drugs and RNAis are only efficacious against one or a few closely related viral species, are vulnerable to the emergence of drug resistance, and may exhibit off-target effects. Second, specific inhibitors of host factors that are essential for viral replication may be efficacious against broader classes of viruses that rely on the same host pathway and are hypothesized to be less prone to the emergence of viral drug resistance.[4] The third class of medicines, vaccines, require specific development for each viral species or viral genotype, are less efficacious against highly adaptive viral species such as pandemic RNA viruses, must be administered preventively before infection in most cases, and are nearly always *biologicals*[5] and may therefore be difficult to mass-produce and to transport. Finally, unspecific drugs such as ribavirin and interferon that modulate the host antiviral defense system, affect viral replication by different means, or alter the host cell metabolism and are modestly effective against a range of viruses at the cost of relatively high side effects.

[1] Fauci (2001), Forst (2006)

[2] Friedel and Haas (2011)

[3] Fauci (2001)

[4] See later sections for a more detailled discussion and criticizm of these views.

[5] *Biological*, from *biologic medical product*: a therapeutic created by biotechnological rather than purely chemical processes. This class of medicines includes proteins, nucleic-acids, as well as tissue transplants.

Small-molecule inhibitors. Small molecule bacterial antibiotics, commonly also denoted as *antibiotics*, belong to the most beneficial medical compounds besides vaccines, antiseptics and anesthetics and are currently available in broad-spectrum forms for all bacterial diseases excluding multiple-drug-resistant strains. In contrast, the high dependency of viruses on cellular host factors, their intracellular nature, high replication and mutation rates, and the absence of a common drug target across viral species (as a result of the polyphyletic origin of viruses) require specialized antiviral compounds for each viral species. The considerable research investment required for developing safe medicines for each these viral targets as well the fact that drugs developed for one viral species are usually not applicable to related species result in considerable fewer medicines available for treating viral infections than for treating bacterial infections. Indeed, more than twenty classes of bacterial antibiotics covering some 160 antibiotic compounds are known today, most of which are applicable to several bacterial species.[6] In contrast, only about 50 approved small-molecule antivirals are currently available, most of which are applicable to only one viral species.[7]

History of antivirals. Interestingly, it is this status of viruses as obligate intracellular parasites that are depending on a tight interplay with host function that raised skepticism with regard to the feasibility of developing specifically targeted antiviral compounds in general. In particular, before the first antiviral was developed it was unclear if components of the viral machinery could be targeted at all by compounds without inhibiting essential host factors, thereby eliciting considerable drug side effects. This skepticism was only dissipated by the medical success of the first antiviral aciclovir, an inhibitor of herpes simplex DNA replication.[8]

Due to pressing medical needs, subsequently developments concentrated on finding effective anti-HIV compounds and devising therapy optimization techniques such as combination therapy (highly active antiretroviral therapy, HAART) that aimed to increase therapy response against drug-resistant viral strains.[9] Today, 51 FDA-approved small-molecule antivirals are in broad clinical use and while most of these compound (29) are used in anti-HIV therapy, targeted antiviral therapies against respiratory syncytial virus (RSV), Influenza A/B, various Herpes viruses, HBV and HCV are available.[10] These compounds are generally targeted against five viral processes: *viral fusion, entry and uncoating* (entry and uncoating inhibitors, 5 compounds), *viral nucleotide replication* (DNA and RNA chain terminators, mutagens, and polymerase inhibitors, 25 compounds), *viral integration* (integration inhibitors, one compound), *viral polypeptide processing* (protease inhibitors, 12 compounds), and *viral capsid construction and release* (capsid and release inhibitors, two compounds).[11] In addition, drugs against other viral diseases with major pathogenic impact (poxvirus, the hemorrhagic fever viruses

[6] Davies and Davies (2010), Lewis (2013), `http://en.wikipedia.org/wiki/List_of_antibiotics`

[7] Antonelli and Turriziani (2012)

[8] Elion et al. (1977)

[9] Palella et al. (1998), Vella (1994)

[10] Antonelli and Turriziani (2012)

[11] Antonelli and Turriziani (2012)

Ebola and Lassa, and several enteroviruses such as Coxsackie) are currently under development. Due to the specific nature of viral infections, no true broad-spectrum antivirals are currently available; however, a small class of antiviral compounds that do not target specific viral factors but activate the host immune system or indirectly hamper viral replication such as interferon and ribavirin, respectively, are in clinical use.[12]

Vaccines and antibody therapy

Vaccination has been of critical importance for efforts that aim to significantly reduce or even eliminate the disease burden of several infectious diseases such as diphtheria, pertussis, poliomyelitis, measles, smallpox, tetanus, and yellow fever. In addition, vaccines that are effective against meningitis, pneumonia, and some forms of viral hepatitis are increasingly used and preventive medicines against dengue fever, respiratory infections, and diarrhoeal diseases are currently in development. Today, vaccines are available as cheap, long-lasting, and specific protection against many viral infections. While different kinds of vaccines exist (see below), these medicines have in common that they stimulate the host adaptive immune system in order to confer humoral (antibody-mediated) or cellular (T-cell mediated) immunity against three classes of pathogenic antigens: components of the whole pathogen (such as proteins or polysaccharides located on the viral capsid), intracellularly expressed viral factors whose antigenic epitopes are presented by infected cells, and pathogenic endo- and exotoxins.

In contrast to antiviral drugs that are prescribed to patients in order to counteract ongoing viral infections[13], immunization by vaccines are today in broad use as preventive measures against viral infections. These preventive measures are especially relevant in the light of populations of low-income countries that lack medical infrastructure and funding for regularly providing antiviral drugs against chronic infections. Offering immunization to these populations enables a logistically and commercially viable solution for fighting infectious diseases while also removing a long-lasting reservoir of (potentially evolving) pathogens that may quickly spread to developed countries due to increased urbanization and air traffic.

Due to safety considerations, modern vaccines less often rely on whole live or attenuated pathogens but increasingly use *subunit vaccines*[14] or recombinant vaccines consisting of pathogens rendered non-pathogenic by genetic manipulation. Due to the lower immunogenicity of these new compounds, the provided antigens often do not induce strong neutralizing antibodies that are required for a memory response. Consequently, immunologic *adjuvants* such as potassium alum[15] or cytokines are often employed in order to

[12] McCormick et al. (1986)

[13] Exceptions exist; so for example the use of antivirals for susceptible populations that cannot be immunized due to missing vaccines (HIV) or contraindications (Influenza A)

[14] *Subunit vaccines*: components of the pathogens such as proteins, toxins, or polysaccharides with stable antigenic epitopes.

[15] Also widely used as underarm deodorant.

stimulate immune response and illicit broader, cross-reactive antibody neutralization.[16] These adjuvants are currently optimized in order to further direct $CD8^+$ and $CD4^+$ T-cell response towards increased specificity and differentiation of memory cells, respectively.[17]

While the first vaccines targeted bacteria (pertussis and tuberculosis) or bacterial toxins (tetanus and diphtheria toxins) with limited genetic diversity and practically invariant antigens, subsequently developed antiviral vaccines had to account for antigenic variation in order to immunize against several (multivalent) viral variants within a single dose.[18] Similarly, some viral pathogens such as influenza A utilize several layers of antigenic variation and recombination, thus requiring constant adaption of the corresponding vaccine to the few viral variants that are in global circulation each year.[19]

Today, several viruses exist that defy vaccination either due to antigenic diversity (HIV, HCV, as well as parasites of the genera Plasmodium, Trypanosoma, and Leishmania) or by resourceful (if such an anthropomorphic term may be used) circumvention of the adaptive immune response (in particular by large DNA viruses such as the *Herpesviridae*). All these pathogens usually cause persistent infections and offer significant challenges for vaccine design. Consequently, no comprehensive and durable preventive medicine is available for any of these pathogens as to date.[20]

In particular, pandemic HIV-1 is a highly relevant challenge for vaccine development due to its high antigenic variation within patients and also across different geographic locations that hinder identification of viral epitopes that induce cross-reactive and protective antibodies.[21] These viral epitopes can undergo viral mutations, or be affected by masking or structural burying of conserved enzymatic sites.[22] While considerable scepticism remains in the HIV community regarding the development of a vaccine anytime soon, a certain cross-reactivity can remain even in modified epitopes and broadly neutralizing antibodies are believed to exist, in principle.[23] This assumption has recently been supported by trials of bioinformatically designed 'mosaic' antigens that were pieced together from multiple HIV epitopes and resulted in broad protective efficacy in rhesus monkeys.[24]

Neutralizing antibodies are not only a product of the host immune system but also experience increased use as antiviral drug that are produced by recombinant techniques and administered to the patient. However, only one antiviral antibody (Palivizumab, a compound effective against the respiratory syncytial virus RSV) is currently approved due to the high costs associated with the synthesis of these biologicals.[25]

[16] Khurana et al. (2010)

[17] Baumgartner and Malherbe (2010), Palmer and Restifo (2009)

[18] Sabin et al. (1954)

[19] Schulman and Kilbourne (1969)

[20] Craig and Scherf (2003)

[21] Johnston and Fauci (2008), Virgin and Walker (2010), Walker et al. (2009)

[22] Colman et al. (1987, 1983), Knossow et al. (1984), Kwong (2005), Varghese et al. (1988)

[23] Burton et al. (2004), Tulip et al. (1992)

[24] Barouch et al. (2013)

[25] Russell and Cohn (2012)

Viral resistance and drugs targeting host factors

The emergence of resistant viral variants especially in high-progeny, low-fidelity RNA viruses such as HIV and HCV have resulted in the growing importance of resistance management in antiviral therapy. Also, viral latency in cellular comportments or tissues not reachable by antiviral drugs may lead to chronic infections and residual viremia even if the virus is present at undetectably low viral plasma loads.[26]

Three prominent strategies have been proposed for counteracting the emergence of viral drug resistance: first, drug combination therapies that increase the genetic barrier of the therapy against upcoming drug resistance by requiring the virus to develop escape mutations against two or more drugs with different mechanisms of action.[27] Second, methodologies of personalized medicine that afford insight in particularities of the viral and host genomes, respectively, in order to select drugs that maximize therapy success at any point in therapy.[28] Third, targeting host factors that are essential for viral infection or replication and may be associated with reduced emergence of resistance. The latter strategy will be discussed here in detail while the other two approaches, combination therapies and personalized therapies, will be discussed at some greater detail in Chapter IV.

Advantages of drugs targeting host factors. Basic research on essential host factors that are targeted by viral pathogens has lead to the development of several new compounds especially against highly diverse pathogens such as RNA viruses that rapidly acquire drug resistance. In contrast to therapeutic compounds that target viral factors and that may lose efficacy due to the emergence of resistance mutations in the viral genome, host factors are typically well conserved and can not be mutated by viral evolutionary processes. In addition, targeting host factors significantly increases the viable antiviral drug targets from about 10 proteins encoded in RNA-viral proteins to the larger number of virally accessed host factors, some of which may be relevant for other diseases and may thus have known inhibitors.

An especially interesting class of host proteins in this regard are *bridge* proteins, i.e., proteins that have a low number of binding partners but act as a common component for many host pathways. Such bridge proteins are selectively attacked by several classes of viruses and may therefore constitute factors essential for viral replication.[29] However, since these bridge proteins are also likely to be important factors for the host cell, their inhibition may cause unwanted side effects that may or may not surpass side effects of existing antivirals that target viral factors.

[26] Eigen (1993b), Palmer et al. (2008)

[27] Coiras et al. (2009)

[28] Beerenwinkel (2003), Lengauer (2011), Lengauer and Sing (2006)

[29] de Chassey et al. (2012b), Navratil et al. (2011)

The search for druggable host factors has identified several suitable targets with respect to HIV, HCV, and influenza infection.[30]Interestingly, up to 20% of all experimentally identified viral host factors across experimental studies belong to a common set of human pathways, a fact that may firstly allow for the development of broad spectrum antivirals.[31] Only 10 of the more than 50 FDA-approved antiviral compounds target host factors and none of these compounds has been especially developed for broad efficacy;[32] however, some of the most successful FDA approved inhibitors such as the HIV antiviral Maraviroc exhibit favourable pharmacokinetic and safety profiles, indicating that host factor targeting drugs are not necessarily paralleled by strong side effects.[33] In addition, supplementary approaches for using host factors to increase therapeutic efficacy exist; as touched upon previously, at least two indirect acting antiviral compounds, interferon and ribavirin, increase the natural antiviral host response.

Limitations of host factor antivirals. Although host factor targeting medicines have been shown to be effective in many different settings,[34] they may not be the golden bullet either. While viruses cannot directly prevent antivirals from binding to host factors, adaption of viral binding interfaces by mutation and selection may result in differential usage of host factors and consequent loss of drug efficacy. This is exemplified by recent results concerning HCV that have demonstrated how alternative use of host factors may result in resistance to the cyclophilin inhibitor SCY-635.[35]

Similarly, HIV-1 is also able to circumvent drugged host factors: immunodeficiency viruses employ host chemokine receptors such as as CXCR4 or CCR5 as cofactors for viral entry that are targeted by drugs in highly active antiretroviral therapy (HAART).[36] Treatment with these drugs may result in mutations of viral glycoproteins that are associated with a shift in viral coreceptor usage or by viral adaption to the inhibitor bound coreceptor, both of which result in drug resistance.[37]

While this switch in coreceptor usage may partly reflect the normal evolution of HIV-1,[38] and the reported levels of drug resistance are comparably low and affect only a minority of patients,[39] HIV resistance to host factor drugs is still considered an important contributing factor to therapy failure.[40] As a result, determination of viral host tropism profiles based on sequencing has significant clinical consequences and is commonly used to inform therapy decisions.[41]

[30] Bushman et al. (2009), König et al. (2010, 2008), Li et al. (2009b), Murali et al. (2011)

[31] de Chassey et al. (2012a), Meyniel-Schicklin et al. (2012)

[32] de Chassey et al. (2012b)

[33] Fätkenheuer et al. (2008)

[34] Garbelli et al. (2011), Geller et al. (2007), Hopkins et al. (2010), Kumar et al. (2011b)

[35] Chatterji et al. (2010), Delang et al. (2011), Kwong et al. (2011)

[36] Dragic et al. (2000), Edinger et al. (1997), Gorry and Ancuta (2011), Jones et al. (1998)

[37] Ogert et al. (2010), Pugach et al. (2007)

[38] Delobel et al. (2005), Regoes and Bonhoeffer (2005)

[39] Fätkenheuer et al. (2008), Parra et al. (2010)

[40] Fätkenheuer et al. (2008), Mosier (2009)

[41] Hung et al. (1999), Jensen and van 't Wout (2002), Schuitemaker et al. (2010), Thielen et al. (2010)

Future antivirals. Based on the successes of broadly acting drugs, the drug-induced activation of host enzymes associated with the innate immune systems such as APOBEC, TRIM, and toll-like receptors (TLRs) are considered to be an interesting approach to broadly and indirectly acting antivirals.[42] Also, novel compounds that do not target host factors but represent orthogonal approaches to attacking viral factors are under active development. These approaches include broadly acting antibodies and RNAi technologies whose therapeutic potential against HCV, HBV, HIV, and highly aggressive viruses such as Marburg virus and Ebola virus is currently explored.[43]

Current antiviral therapy is mainly aimed at lowering viral load in order to limit symptoms of the disease and facilitate immune clearance. In addition, low viral loads implicitly reduce the abundance and thus the diversity of the viral infections, thus reducing the viral adaptability to future interventions such as antiviral therapy. As will be discussed in more detail in Chapter IV, combination therapy is successful in increasing the overall genetic barrier of the treatment especially if the drugs employed have orthogonal modes of action and, as a consequence, feature differing pathways to resistance. However, ongoing low-level replication, mutation, and selection within highly adaptive viral species will commonly result in development of multi-drug resistant variants and thus therapy failure.

If antiviral compounds currently under development are a sign of things to come, then the antiviral drugs of the next two decades may increasingly focus on host factors instead of on viral factors as drug targets. While the control of side effects of these drugs remains an important goal, the potential for repurposing existing drugs that already target host factors, the promise of reduced drug resistance, and the potential of host factor targeting drugs to act as broad-spectrum antivirals makes the development of these compounds attractive endeavours.[44]

In addition, and in accordance with general trends in the pharmaceutical industry, the importance of biologicals such as antibodies, RNAi, vaccines, and synthetic interferons is also likely to increase.[45] Although significantly more expensive to produce than small molecule inhibitors and facing additional pharmacokinetic constrains, biologicals often have proven effectiveness in the host as confirmed by basic research, have high specificity towards one or multiple closely related pathways, and have important competitive advantages such as being harder to reproduce as generics. Examples for this development are early-stage broad-spectrum antivirals that employ biologicals to emulate intracellular apoptotic control circuits or inhibit viral maturation by blocking capsid assembly on a broad range of viruses.[46]

[42] Fox (2007), Huthoff and Towers (2008), Miller et al. (2008)

[43] Chen and Dimitrov (2012), Hu and Robinson (2010), Zeller and Kumar (2011)

[44] Fox (2007), Holt et al. (2010), Huthoff and Towers (2008), Lim and Murphy (2011), Miller et al. (2008), Reeves et al. (2005)

[45] Amara et al. (2001), Bacon et al. (2009), Chen and Dimitrov (2012), Ivacik et al. (2011)

[46] Lingappa et al. (2012), Rider et al. (2011)

16 Measuring protein interactions

As discussed in the last section, *virally targeted host factors are of high medical relevance and their determination is crucial for the development of novel antiviral drugs. Interactions between viral and host proteins, arguably the most promising form of molecular virus-host interaction from the standpoint of drug development, can be inferred by high-throughput experimental and computational protein interaction assays. The measured or predicted interaction patterns can subsequently be visualized in protein interaction networks (PINs). Within such networks, nodes represent proteins and edges denote interactions that vary in directness and directionality according to the network type. In contrast to detailed analyses of single protein interactions, PINs are usually analyzed using tools of interdisplinary disciplines such as network science or systems biology. While PINs serve as useful abstraction of the protein interactome (i.e., the set of all protein interactions in a organism or host-pathogen system) and are frequently used in systems biology, they also have been criticized as hard to interpret and validate in a genome-wide manner, lacking structural annotations, and not sufficiently discerning between succinct kinds of interactions and their associated experimental confidences.*[1] *Still, these networks are widely employed as discovery tools for basic research as well as for interrogating the molecular basis of disease.*[2] *In this section, we will provide an introduction to protein interactions as well as the experimental and computational methods that are commonly employed in their investigation. In addition, we will also discuss protein purifications, a specific form of protein interaction data that poses particular challenges in term of interpretation and analysis.*

Classes of protein interactions

Proteins are structurally and functionally diverse molecular entities that are responsible for the vast majority of mechanistic (i.e., ordered and reproducible as far as conceivable within a highly dynamic and stochastic environment) molecular manipulations within the cell. As such, a protein specifies a discrete component of cellular function that is embedded in a dynamic network of other proteins with which it physically interacts.[3] Proteins perform their functions by way of physical interactions with other proteins as well as with other cellular components such as metabolites, nu-

[1] Aloy and Russell (2002), Gingras and Raught (2012), Gonzalez and Kann (2012)

[2] Ideker and Sharan (2008), Kann (2007)

[3] Eisenberg et al. (2000)

cleotides, and membranes. While some proteins form only transient interactions (for example, most classes of enzymes and kinases), many proteins co-complex and form longer lasting assemblies that perform a concerted function and differ in size, heterogeneity, and temporal dynamics.

Proteins may form both transient binary interactions form homogeneous or heterogeneous protein complexes which are longer lasting (although not always permanent) and are also denoted as protein *assemblies*. Examples for such assemblies range from dimers consisting of only two polymers, over larger homo-multimers, up to multi-component conglomerates such as ribosomal subunits, the proteasome, and the gigantic nuclear pore complex.[4] Interactions within these assemblies are inherently governed by both thermodynamic principles,[5] as well as on the chemical and biological characteristics of cell physiology, competitive binding to other factors, and protein abundance.[6]

In the context of this chapter, experimentally derived protein interactions can conceptually be separated by at least two dimensions; one such dimension is the distinction between *direct physical*, *indirect physical*, and *non-physical associative* interactions. Direct physical protein interactions are interactions between peptides that are not obligate (i.e., the peptides are modular entities that are not required to bind and may exists separately) but either permanently or transiently share a common binding interface where residues of both peptides are in direct and non-covalent contact with each other. Often, direct physical protein interactions are also shorthandedly denoted as *binary* protein interactions, although these terms should not be considered to be strictly identical.

Indirect physical protein interactions involve two peptides that co-complex (i.e., are belonging to the same assembly of non-covalently bound protein modules) but do not share a common interface. Rather, these interactions are mediated by additional direct physical protein interactions that may or may not be observed in a given experimental procedure. Finally, non-physical associative interactions denote pairs of proteins that co-occur in a common pathway or sub-cellular compartment or are experimentally associated (for example, by gene co-expression or synthetic lethality) without any additional evidence of direct or indirect physical interactions between the subunits.[7] As such, these associations often describe metabolic or genetic correlations rather than co-localizations in time and space.

The other classification of interest here is the distinction between endogenous (i.e., interactions between proteins of the same species) versus exogenous (interactions between proteins originating from different species) protein interactions. While endogenous interactions usually act cooperatively to implement a specific cellular function, exogenous protein interactions are driven by a wider array of mechanisms that range from from parasitic to symbiotic.[8]

[4] Alber et al. (2007), Alberts (1998), Ciehanover et al. (1978), Lee et al. (2002), Pu et al. (2009)

[5] i.e., whether two given protein subunits bind depends on the differential free energy of binding between bound and unbound states of the proteins

[6] Carbonell et al. (2009), Johnson and Hummer (2011), Przytycka et al. (2010)

[7] De Las Rivas and de Luis (2004)

[8] Dethlefsen et al. (2007)

Experimental protein interaction assays

In order to investigate the precise function of a protein assembly, it is crucial to infer the set of binding patterns between all its members in a spatially and temporally highly resolved manner. Several experimental protocols and technologies, in the following denoted as *assays* or *systems*, have been developed in order to measure the transient and permanent features of protein interactions.

Today, protein interactions can be measured by several biophysical and genetic systems[9]; while historically small-scale interaction assays focused on structural methods such as X-ray crystallography and NMR spectroscopy, modern genetic approaches neglect structural aspects of the measured interactions in favor of increased experimental throughput. In particular, today binary interactions assays and co-complex detecting methods such as Y2H and (T)AP/MS, respectively, are considered to be premier approaches for orthogonaly and complementarily masuring the interactome.[10]

Yeast Two-hybrid. Yeast Two-hybrid (Y2H) is an experimental protein interaction assay that measures direct protein interactions by means of transcriptional activity.[11] Yeast transcription factors such as Gal4 that contain an activating (AD) and DNA binding (BD) domain are fused to two open reading frames (ORFs), respectively, that encode the pair of proteins under investigation, functionally denoted as *bait* and *prey*. The fusion constructs are transfected into yeast. Upon physical interaction of bird and prey the transcription factor domains are brought into close proximity and induce expression of a reporter gene that has phenotypic and observable effects on the yeast cell. Due to the granularity and sensitivity of this approach, large scale studies aided by robotics are able to interrogate tens of thousands of individual interactions in parallel.[12] However, the false positive rate of the Y2H system has been reported to be high, mostly due to biases in the experimental quantification of the interactions, artificial interactions between protein partners not co-located in the same sub-cellular compartments under native conditions, and occasional indirect physical interactions.[13] However, recent advancements of the original Y2H system concerning the activation of reporter factors and the use of statistical scoring schemes have been employed to significantly lower the error rate and Y2H is now the most widely used system to measure protein interactions.[14]

Affinity purification. Affinity purification and mass spectrometry (AP/MS) has been established as a more recent physical protein interaction assay that is designed to detect co-complexing proteins and thus measures both direct and indirect physical protein interactions in a single assay.

[9] Gonzalez and Kann (2012)

[10] Braun (2012)

[11] Chien et al. (1991)

[12] Fields et al. (2000), Guruharsha et al. (2011), Ito et al. (2001)

[13] Deane et al. (2002), Fields (2005), Rajagopala et al. (2009), Sprinzak et al. (2003)

[14] Barrios-Rodiles et al. (2005), Braun et al. (2009), Eyckerman et al. (2001), Nyfeler et al. (2005), Ramachandran et al. (2008), Tavernier et al. (2002)

Similar to Y2H, affinity purification follows a bait/prey approach that requires genetic tagging of the bait ORF with a known epitope such as GFO, StrepII, or FLAG prior to transfection into a yeast strain. The expressed protein complex is subsequently purified from the cellular medium using antibody affinity to the genetic tag. The components of the purified complex are then identified using liquid chromatography and tandem mass spectrometry (LC-MS).

A prominent variant implementing this interaction detection system is tandem affinity purification (TAP/MS) that allows for several sequential purifications in order to reduce contaminating proteins prior to identification.(15) In contrast to Y2H, affinity purification systems are biased towards more stable interactions that survive purification and are thus less likely to detect transient protein interactions. Affinity purification systems have been employed in large-scale studies interrogating the protein interactome of several model organisms such as yeast, worm, fly, and also human.(16)

Due to their ability to measure larger protein assemblies that may involve dynamic, i.e., functionally optional or temporally variable interaction partners, (T)AP/MS approaches are also suited for measuring protein complex dynamics using experimental perturbation techniques where the same complex is assayed multiple times in different states.(17)

Affinity purification exhibits significant false positive experimental errors due to its requirement to over-express bait proteins, as well as due to off-target effects or reduced binding of antibodies, nonspecific binding events, and missing cellular compartmentalization of the tested proteins as a result of cell lyses.(18)

Refinements of the genetic protocols have been proposed that allow for measurements in slightly more physiological environments but are still biased towards detecting highly stable interactions.(19) Similarly, statistical approaches that better quantify the uncertainty of MS protein identification have been utilized to reduce error rates.(20) For instance, while the raw false positive rate of affinity purification systems is broadly comparable to that of Y2H assays,(21) methodologies have been proposed that quantify the abundance of the identified proteins by means of *peptide spectra* or differential *isotope labeling* of test and control purifications in order to identify contaminant proteins based on their abundance and binding specificity.(22)

Similarly, false negative errors that may be incurred due to the washing of physiologically relevant, transiently attached proteins during purification or that may result from low abundance interactions can partly be mitigated by specialized protocols.(23)

(15) Puig et al. (2001), Rigaut et al. (1999)

(16) Babu et al. (2012), Ewing et al. (2007), Gavin et al. (2006), Krogan et al. (2006), Polanowska et al. (2004), Veraksa et al. (2005)

(17) Blagoev et al. (2003), Ranish et al. (2004, 2003)

(18) Chang (2006), Gavin et al. (2011)

(19) Havugimana et al. (2012), Kristensen et al. (2012), Wodak et al. (2009)

(20) Kapp et al. (2005), Perkins et al. (1999)

(21) Gavin et al. (2006, 2011), Krogan et al. (2006)

(22) Breitkreutz et al. (2010), Choi et al. (2012, 2011), Jäger et al. (2012), Lavallée-Adam et al. (2011), Ong et al. (2002), Sardiu et al. (2008)

(23) Stengel et al. (2012)

Interpreting protein interaction assays

APART FROM SUFFERING from specific false positive and false negative error profiles, purification data pose further challenges for interpretation:[24] even after removal of contaminants, the retrieved set of co-complexed proteins is stabilized by a mixture of direct and indirect physical interactions. It is therefore not easily possible to identify proteins that share a common binding interface, nor is it straightforward to identify proteins that might by part of multiple physiological protein complexes that are co-purified within one single purification. Although additional techniques such as binary protein interaction assays, perturbation protocols, and quantitative MS approaches involving peptide counts or isotope labeling may aid in answering these questions, no consensus on a method for the integration of these approaches has yet been established. Instead, *ad-hoc* methods for interpreting interactions within purifications are commonly used when deposited these interactions in public repositories.

These methods, termed *spoke* expansion (interactions between bait protein and all its preys) and *matrix* expansion (interactions between all proteins of a purification regardless of bait or prey status), are prone to false negative and false positive errors: the spoke expansion neglects possible interactions between prey proteins that are physiologically likely to stabilize the complex and furthermore assumes that all preys are in direct physical contact with the bait; the latter fact is unlikely for larger purifications due to mutually exclusive binding interfaces. For the same reason, the matrix model of expansion leads to high rates of false positive interactions since large complexes are unlikely to be fully connected. In addition, both expansion models do not delineate multiple physiological protein complexes that have been captured within the same purification.

While additional experimental approaches like reciprocal purifications, multiple orthogonal purification steps, or perturbation approaches may be conducted to obtain this missing information from protein purifications,[25] for example by cross-linking experiments,[26] these systems are cost-intensive and require specialized training. In addition, all known experimental methods, both high-throughput and low-throughput, influence cell physiology to different degrees and in different manners, resulting in experimental results that are significantly biased by the assay being used. This fact regularly results in low reproducibility between identical protein interaction assays run by different labs as well as low overlap between the results of different experimental approaches, thereby significantly complicating interpretation and comparison of protein interaction screens on genomic scales.[27]

[24] Gingras and Raught (2012)

[25] Gingras et al. (2007), Hyung and Ruotolo (2012), Moyer et al. (2006), Stengel et al. (2012)

[26] Leitner et al. (2010), Sinz (2010)

[27] Yu et al. (2011)

Computational interpretation and scoring methods. Due to these biases, even refined protein interaction protocols result in more measured protein interactions than would be expected given existing biological knowledge and comparisons with orthogonal protein interaction assays. Consequently, *in silico* methodologies have been developed in order to aid in the interpretation of the measured data and to support identification of false positive experimental results by means of statistical models, additional experimental variables, or external information such as known protein interactions, functional annotations, and additional assays measuring protein-protein or gene-gene associations as for example gene expression measurements or *synthetic lethality*[28] experiments.

Many computational methods that classify or quantify protein interactions with regard to their status as potential contaminants make use of a background set of known, high-confidence protein interactions deposited in public databases such as HPRD, DIP, IntAct, BioGRID, MINT, and BIND.[29] However, analyses have reported low levels of concordance between these databases, a fact that is likely due to differing methods of data curation.[30] Recent standards by the Proteomics Standard Initiative, data querying interfaces that are able to contact multiple databases, as well as meta-databases such as IRefIndex and IRefWeb have aimed to mitigate these discrepancies.[31]

While a broad (perhaps overly broad) range of protein interaction scoring schemes exists,[32] here we will concentrate on methods that perform primary data analysis of purifications originating from (T)AP/MS assays and do not utilize additional, external data sources such as public databases. These primary data analysis methods can be further subdivided into methods that only rely on the list of proteins identified within each purification experiment *(frequency-based approaches)* and methods that employ non-standard experimental data from the (T)AP/MS workflow and thus require special protocols *(spectral count approaches)*.[33] Two computational methods utilizing such additional (T)AP/MS data are ComPASS and SAINT, both of which use spectral counts and experimental replicates in combination with statistical models based on the Poisson distribution in order to detect false positive interactions within the purification data.[34]

On the other hand, frequency-based socio-affinity approaches apply statistical models directly to the purification data in order to detect unreliable or promiscuous interactions based on experimental replicates.[35] This class of approaches has lately been supplemented by more elaborate methods by Collins *et al.*, Guruharsha *el al.*, and Yu *et al.* that include clustering, supervised learning, and permutation-based approaches in order to increase sensitivity.[36] The last section of this chapter will discuss advantages and disadvantages of several of these methods in more detail.

[28] *Synthetic lethality*: a class of genetic interactions where where a combination of mutations in two or more genes is lethal for the cell while a mutation in a single gene is viable. Genes in a synthetic lethality reltionship may exhibit redundant functions within the cell. Synthetic lethality be measured in genome-wide screens by disabling pairs of gene by genetic process termed *double knockout*.

[29] Bader et al. (2003), Chatr-aryamontri et al. (2013), Kerrien et al. (2012), Keshava Prasad et al. (2009), Licata et al. (2012), Salwínski (2004)

[30] Turinsky et al. (2010)

[31] Aranda et al. (2011), Orchard et al. (2012), Razick et al. (2008), Turner et al. (2010)

[32] Armean et al. (2013)

[33] *Spectral counts:* the number of mass spectra assigned to a given protein. Given appropriate normalization, these counts can be used to estimate the abundance of proteins within purification. Spectral counts are regularly employed to identify contaminants and compare protein abundance across technical replicates.

[34] Choi et al. (2012, 2011), Sowa et al. (2009)

[35] Gavin et al. (2002), Hart et al. (2006)

[36] Collins et al. (2007), Guruharsha et al. (2011), Yu et al. (2009)

17 Detection of viral host factors

As discussed in previous sections, human pathogens in general and viruses in particular are relying on host-pathogen molecular interactions that are crucial for initiating and sustaining infection as well as for evading host immune responses.[1] *Successful establishment of infections requires networks of molecular interactions of a pathogen with its host that involve a variety of molecular agents such as proteins, nucleotides, small ligands, sugars, and fatty acids.*[2] *Better characterization of these networks may aid in identification of new targets for drugs or vaccines and further has the potential of facilitating development of therapeutic compounds that have higher genetic barriers to drug resistance or fewer side effects than presently available medicines.*[3] *Of the interacting agents mentioned, proteins have been identified as the factors of the highest medical and pharmacological interest and therefore their interactions are of primary importance for further investigation.*[4]

Experimental determination of host factors

Viruses are especially amenable to genome-wide protein interaction screens due to their relative small genome size; indeed, many viral pathogens displaying drug resistance consist of RNA viruses that encode only a limited number (on the order of 10) protein factors. Indeed, the smallest human infectious virus known, Hepatitis D, encodes only two mature proteins. While DNA viruses such as *herpesviridae* typically encode on the order of 200 proteins and viruses with up to 1,200 genes have been identified, these numbers are still dramatically lower than the sizes of proteomes of prokaryotes such as *E. coli* which contains more than 4,000 protein coding genes and a pan-genome of more than 16,000 such genes, similar in magnitude to the about 20,000 human protein-coding genes.[5]

Intra-viral interaction networks. Consequently, experimental measurements of host-pathogen PINs and subsequent analysis of their functional and structural principles have first been undertaken for viruses, followed by bacterial networks.[6] Following proof-of-principle investigations on bacteriophages, several protein interaction networks have been determined for HCV, herpesviruses such

[1] Durmuş Tekir and Ülgen (2013)

[2] Lengeling et al. (2001), Stebbins (2005)

[3] Barker (2006), Dyer et al. (2008), Münter et al. (2006)

[4] Thieu et al. (2012)

[5] de Chassey et al. (2012b), Dolan et al. (2006), International Human Genome Sequencing Consortium (2004), Zhaxybayeva and Doolittle (2011)

[6] Flajolet et al. (2000), McCraith et al. (2000), Rain et al. (2001)

as HSHV and EBV, and SARS coronavirus using predominantly Y2H assays.[7] Interactions thus identified have been deposited in general protein interaction database as well as in a rising number of more specialized repositories.[8]

Virus-host interaction networks. Due to the limited technology available at the time, only intra-virus protein interactions (i.e., excluding host factors) were investigated in the aforementioned studies. While this level of analysis allows for inference of basic network properties as well as identification of essential protein factors of a pathogen that may serve as therapeutic targets, only very limited knowledge can be gained about the role of these factors within the larger host protein interaction network. Therefore, and as a natural extension of these first intra-viral investigations, combined host-pathogen interaction studies were later proposed in order to interrogate the intricate interspecies interplay between the pathogen and components of the host immune, signaling, and nucleotide replication system.[9] While the first of these extensions have focused on computational methods due to missing experimental data,[10] more recent large-scale experimental setups target important human pathogens such as HCV, Influenza A, HIV, and EBV and include both human and viral protein factors.[11]

Structural approaches. In addition, structural studies have been undertaken that interrogate structural properties of host-pathogen interactions.[12] Using these structural data afforded a more detailed look at the specific modes of experimentally derived and predicted host-pathogen protein interactions by analyzing their physical binding interfaces. Since only relatively few pathogen protein structures are available at atomic resolution, these studies commonly employ *homology modeling*[13] to increase coverage of the host-pathogen PIN.[14]

Only few host-pathogen protein interactions with either atomic or homology models exist: a recent study identified only 53 such interactions.[15] While these numbers do not allow for high-coverage annotation or prediction of host-pathogen PINs, the available data are sufficient to derive general organizational principles of these networks. It is estimated that more than half of all proteins without known atomic structure share sufficient sequence similarity with a structurally resolved template protein, allowing the construction of at least partial homology models in most of these cases.[16] Structural coverage of endogenous protein interaction is significantly lower (on the order of 20% of all known human protein interactions are covered by at least partial models) and biased towards highly stable interactions.[17] However, it was noted that structural coverage may be significantly increased by using either more advances methods for homology modeling or by the additional use of templates with structural rather than sequence-based similarity to the query.[18]

[7] Bartel et al. (1996), Calderwood et al. (2007a), Dimitrova et al. (2003), Flajolet et al. (2000), Pan et al. (2008), von Brunn et al. (2007)

[8] Chatr-aryamontri et al. (2009), Navratil et al. (2009), Winnenburg et al. (2008)

[9] Stebbins (2005)

[10] Davis et al. (2007), Dyer et al. (2007), Evans et al. (2009), Lee et al. (2008)

[11] Calderwood et al. (2007a), de Chassey et al. (2008), Jäger et al. (2012), Khadka et al. (2011), Pichlmair et al. (2012), Rozenblatt-Rosen et al. (2012), Shapira et al. (2009), Uetz et al. (2006), von Schwedler et al. (2003)

[12] Franzosa and Xia (2011)

[13] *homology modeling*: for a given protein interaction, a query protein pair is searched against public databases containing 3D structural information in order to to identify pairs of template proteins with high similarities and known binding interfaces at atomic resolution. The most similar of these templates is then employed as a model of the binding modes of the query proteins.

[14] Martí-Renom et al. (2000), Russell et al. (2004)

[15] Franzosa and Xia (2011)

[16] Dalton and Jackson (2007), Franzosa et al. (2012), Madera et al. (2004), Martí-Renom et al. (2000)

[17] Franzosa et al. (2012), Kim et al. (2006)

[18] Kundrotas et al. (2012), Zhang et al. (2012)

Properties of host-pathogen interaction networks

ANALYSES OF THE TOPOLOGY of intra-virus interaction networks have demonstrated that viral proteins are often multi-functional and interact with a large number of distinct host factors, each of which in turn is often targeted by more than one viral factor.[19] Prior analyses of the about 144,000 known human endogenous protein interactions[20] have shown that the human PIN is scale-free and that its degree distribution (i.e., the distribution of the number of connections of all nodes within the network) follows a power law. This network contains only a limited number of highly connected nodes (*hubs*) and is dominated by a vast majority of nodes that have only few connections.[21] Scale-free topologies are characterized by functionally separated protein modules that are connected by a few hub proteins[22] As a consequence, the average number of connections between nodes is low, i.e., the network exhibits a high degree of signaling efficiency while still being very robust with respect to *random* removal of nodes; however, networks of this kind are highly sensitive to the removal of *hubs*.[23]

Properties of viral protein interaction networks. In contrast, topological analysis of intra-virus protein interaction networks indicated that viral proteins are highly coupled (i.e., are densely connected with many other viral proteins), resulting in relatively many hubs and only few peripheral nodes.[24] Detailed analyses of these viral network topologies showed that these networks are neither scale-free nor do they exhibit small-world properties, thus indicating their relatively high sensitivity to random perturbations or deliberate attacks on the connectivity of the network.[25] This apparent sensitivity of the viral protein interaction network is in stark contrast with the observed persistence of infection; indeed, even with an arsenal of only a few proteins, viruses regularly overcome immune responses and control cellular networks of vast size and complexity. It seems therefore likely that viral protein interaction networks obtain emergent properties within a native (i.e., host-embedded) context that provide them with robustness and versatility not inferrable by intra-viral experiments.[26]

Virally targeted host factors. Later investigations of host-pathogen interaction networks uncovered that viral proteins predominantly target highly connected (topological *hubs*, i.e., having a high degree within the interaction network) and bottleneck (topologically *central*, i.e., being included in many shortest paths between random proteins in the network) host proteins.[27] Host proteins thus targeted are themselves densely clustered into protein complexes and biological pathways that are essential to pathogen infection and propagation and that further are in close network proximity with each other.[28]

[19] Shapira et al. (2009)

[20] The human interactome is predicted to contain between 130,000 and 600,000 interactions. Of these, about 50,000 interactions are supported by high-confidence, small-scale experiments Bork et al. (2004), Stumpf et al. (2008), Venkatesan et al. (2009).

[21] Barabási and Oltvai (2004), Cohen and Havlin (2003), Stelzl and Wanker (2006)

[22] Barabási and Albert (1999), Maslov and Sneppen (2002)

[23] Albert et al. (2000), Cohen and Havlin (2003)

[24] Uetz et al. (2006)

[25] Meyniel-Schicklin et al. (2012)

[26] Meyniel-Schicklin et al. (2012)

[27] Calderwood et al. (2007a), de Chassey et al. (2008)

[28] Bushman et al. (2009), Dyer et al. (2008), Gulbahce et al. (2012), MacPherson et al. (2010), Navratil et al. (2011), Pichlmair et al. (2012)

Although experimental studies targeting host-pathogen interaction networks investigated viruses with very different life cycles, there is considerable overlap between the host factors identified $(5 - 20\%)$. In particular, viruses seem to commonly utilize host proteins and pathways concerning components of the innate immune defence such as the Toll-like receptor network that is responsible for eliciting interferon response or autophagy pathways which may trigger degradation of foreign factors by the lysosomal machinery.(29) While many of these host factors are under positive selection that may be evidence of pathogen-driven evolutionary selection pressures, they also tend to be highly conserved across closely related species, possibly highlighting components of the host cell that are especially vulnerable to viral perturbation.(30)

In order to further increase the power of host-pathogen interaction studies and allow for identification of driving principles of viral infection, more comprehensive approaches were undertaken that combined interaction data from several viral species. These analyses further confirmed the importance of host hub and bottleneck proteins such as transcription factors and proteins involved in cell cycle regulation, apoptosis, and cellular transport.(31) Interestingly, while bacteria and viruses share common strategies, such as attacking hub and bottleneck proteins and host factors involved in metabolic processes, viruses seem to favor host proteins with higher connectivity and centrality and tend to focus on cellular processes to a larger degree than bacteria, which in turn predominantly address the host immune system.(32)

Structural investigation of virus-host interactions. Structural investigations led to the conclusion that viral proteins mimic endogenous host protein interfaces, often without exhibiting any large-scale structural homology to the host protein thus mimicked. This tactic as well as observed overlaps of other existing endogenous host interfaces with exogenous viral interfaces may be a common strategy of viral proteins that facilitates accessing the host cell protein network, competing with host factors, and evading immune recognition. Host interfaces mimicked by viral factors are under positive selection, possibly indicating an evolutionary arms race between pathogen and host.(33) Reversely, viral interfaces are sometimes acquired by the host cell and re-purposed.(34)

Interestingly, viral protein binding interfaces seem to be significantly smaller than host interfaces, possibly indicating evolutionary pressure on viral genome size. Indeed, since viruses are obligate parasites that have to modulate or control the complex host environment in order to propagate, it is likely that most if not all physical interactions of the few viral proteins that can be efficiently encoded and replicated in the viral genome are also functionally relevant.(35) Additionally, smaller viral binding interfaces may be evidence for selection of less specific and more transient modes of interaction of viral proteins.(36) The latter assumption is further

(29) Abe et al. (2007), Grégoire et al. (2011), Meyniel-Schicklin et al. (2012), Navratil et al. (2010)

(30) Bozek and Lengauer (2010), Jäger et al. (2012), Pichlmair et al. (2012)

(31) Durmuş Tekir et al. (2012), Dyer et al. (2008)

(32) Durmuş Tekir et al. (2012)

(33) Franzosa and Xia (2011)

(34) Rappoport and Linial (2012)

(35) Franzosa et al. (2012)

(36) Franzosa and Xia (2011)

supported by the fact that host proteins targeted by viral interfaces are very often *date*-like, i.e., multiple binding interfaces of the same protein are not used simultaneously to form larger complexes but employed serially to bind to various partners across time. This fact further supports prior findings that host proteins targeted by viral factors are often hubs that mediate approximately twice as many protein interactions as an average human protein.[37] Such 'serial' hubs are often high-level regulatory host proteins that access and regulate the host cell interaction network, indicating that viruses specifically target these factors in order to efficiently control and perturb host network functions.

Approaches to inferring host-pathogen interactions

Computational approaches play a two-fold role in processing host-pathogen interaction data. First, since experimental data on host-pathogen interactions are comparatively rare, these approaches may predict high-confidence host-pathogen interactions based on existing data and thus prioritize these interactions for experimental investigation. Such predictive approaches can be roughly differentiated into homology-based, structural, protein region-based, and integrative methods.[38] Second, *in silico* methods can be utilized for assessing experimentally derived and predicted host-pathogen PINS based on reference protein interactions,[39] functional annotation,[40] and secondary experimental evidence such as RNAi.[41]

In the following, classes of experimental data are discussed that either may be utilized for directly validating host-pathogen protein interactions or that may serve as input for computational tools aiming to predict or validate such interactions. Subsequently, computational approaches that employ these data are discussed.

Approaches to validating physical interactions. Protein interactions in general and host-pathogen protein interactions in particular can be validated by several means. First and foremost is the assessment of query interactions by gold standard experimentally derived protein interactions. Such high-confidence interactions are only available for a few, well studied pathogens such as HIV[42] and are deposited in a range of specialized databases covering experimentally conformed interactions (PHI-base), interactions cited in literature (HIV-1, The Human Protein Interaction Database), VirusMINT, and interactions from mixed sources (VirHostNet).[43] All of these database have significant overlap with more general protein interaction database mentioned before and use the latter as sources of newly deposited interactions. Furthermore, meta-databases such as APID, iRefIndex, PHIDIAS, HPIDB, GPS-Prot, and PATRIC offer unified access to their member databases and also allow for limited data manipulation and visualization.[44]

[37] Calderwood et al. (2007a), Franzosa et al. (2012)

[38] Davis et al. (2007), Doolittle and Gomez (2010, 2011), Dyer et al. (2007, 2011), Evans et al. (2009), Lee et al. (2008), Qi et al. (2010), Rozenblatt-Rosen et al. (2012), Tastan et al. (2009)

[39] Davis et al. (2007), Doolittle and Gomez (2011), Dyer et al. (2011), Evans et al. (2009), Qi et al. (2010), Tastan et al. (2009)

[40] Davis et al. (2007), Doolittle and Gomez (2010, 2011), Wuchty (2011)

[41] Doolittle and Gomez (2010, 2011), Evans et al. (2009), Tastan et al. (2009)

[42] Fu et al. (2009)

[43] Chatr-aryamontri et al. (2009), Navratil et al. (2009), Peri et al. (2003), Winnenburg et al. (2008)

[44] Fahey et al. (2011), Gillespie et al. (2011), Kumar and Nanduri (2010), Prieto and De Las Rivas (2006), Razick et al. (2008), Xiang et al. (2007)

An interesting new alternative to traditional protein interaction assays that is able to measure transient or dynamic interactions is the use of molecular imaging techniques; in particular, live cell imaging microscopy approaches afford detection of individual molecules at high temporal resolution and have been employed for studying host-pathogen interactions.[45]

Genetic and genomic approaches. Viral host factors can further be identified by experimental RNA interference (RNAi), a protocol that employs short interfering and short hairpin RNAs to inhibit gene expression of specifically targeted factors. RNAi high-throughput screening resulted in several human proteins that are not essential to the host cell but whose inhibition proved lethal for viruses such as HIV, indicating that these proteins may be involved in virus-host protein interactions.[46] While the acceptance of RNAi is increasing, developing RNAi libraries with high coverage over host factors is still considered to be technically challenging and the relevant protocols may suffer from technical errors that limit reproducibility.[47]

While purely concentrating on analyzing variants within the host genome and correlating them with disease phenotypes using genome-wide association (GWAS) methods enabled highlighting of genomic loci that are predictive for the outcome of viral infections by HCV and HIV, these methods require large sample sizes especially for rare variants and can suffer from sample-selection biases and differences in population structures.[48]

Homology-based approaches to host-pathogen protein interaction prediction rely on the assumption that interaction patterns are conserved between protein homologs, in particular between orthologs.[49] This approach is often implemented by using template protein interactions from several databases containing protein and domain interactions and has predominantly been used to predict human-bacterial protein interactions due to the higher coverage of prokaryotic proteins and template protein interactions in public databases compared to their viral counterparts.[50] Further refinements of homology-based approaches that include additional filters to reduce false-positive interactions (for example, from protein homologs that are expressed in different cellular compartments in the target species and are thus unlikely to interact) have been proposed.[51]

Structural approaches. Analogous to the homology approach, host-pathogen protein pairs that bind via an interface similar to such already known to bind in other species are also suggested to represent true interactions. This approach is followed by several structural approaches using either protein sequences and corresponding SCOP super-families with known binding interfaces[52] or direct computation of structural similarity between atomic structures of pathogen and host proteins, assuming that pathogen proteins structurally similar to host proteins share binding partners.[53]

[45] Cristea et al. (2006), Lakadamyali et al. (2003)

[46] Brass et al. (2008), König et al. (2008), Yeung et al. (2009), Zhou et al. (2008a)

[47] Goff (2008), Shan (2010)

[48] Ge et al. (2009), Hsu and Spindler (2012), International HIV Controllers Study et al. (2010), Lambert and Black (2012)

[49] Matthews et al. (2001)

[50] Finn et al. (2004), Keshava Prasad et al. (2009), Licata et al. (2012), Matthews et al. (2009), Salwínski (2004)

[51] Wuchty (2011)

[52] Davis et al. (2007)

[53] Doolittle and Gomez (2010, 2011)

Proteins are composed of a finite set of protein structural domains that fold independently and are covalently and non-covalently bound to other domains within the same protein. Protein domains are modules of protein function and exhibit various patterns of well-defined activities, one of which being the stable binding of other protein domains in the same or other proteins. Protein domain families and their binding interfaces thus mediate and implement stable protein interactions in a regular and deterministic fashion.

Protein-region based approaches to analyzing host-pathogen protein interactions rely on obtaining patterns of domain interactions from either known protein interactions or public repositories.[54] Putative host-pathogen protein interactions are then believed to be real if they are supported by one of the known domain-domain interactions.[55] In principle, and due to the relatively low coverage of known or high-confidence predicted domain interactions, these methods currently have at most of 20% sensitivity.[56]

An extension of this approach consists of considering other protein regions, such as short linear motifs (SLiMs). SLiMs are short sequence motifs located in disordered protein regions that are believed to mediate transient protein interactions. Viruses that, due to their small proteomes, do not have physical space for as many binding domains as there are observed protein interactions, are believed to make extensive use of short linear motifs for interacting with host proteins.

This strategy is believed to be advantageous for viruses for several reasons: first, the short length and location in disordered protein regions allows quick evolution of these motifs without being limited by functional or structural constraints;

second, the broad specificity and transient interaction pattern of SLiMs makes them particularly suited for signaling purposes, for instance by modulating interaction patterns of cellular kinases;

third, the short length of linear motifs allows integration multiple such regions in a single viral protein chain, thus facilitating multi-functionality.[57] Indeed, experimental evidence suggest that viral proteins that target many host factors display a higher content of intrinsically disordered regions that are likely locations for linear motifs.[58] These locations seem to harbor clusters of short linear motifs that may be important determinants of virulence.[59]

Functional and integrative approaches Physical interactions, genetic attributes, and structural features of viral and host proteins often result in specific functional annotation of these factors in the Gene Ontology (GO). Analysis of GO terms in combination with scoring methods and approaches to functional enrichment analysis may thus serve as surrogates for experimental observables and can highlight high-confidence interactions for experimental validation.[60]

[54] Deng et al. (2002), Finn et al. (2004), Guimaraes et al. (2006), Ng et al. (2003), Schelhorn et al. (2008), Stein et al. (2011), Yellaboina et al. (2011)

[55] Dyer et al. (2007), Sprinzak and Margalit (2001)

[56] Prieto and De Las Rivas (2010)

[57] Evans et al. (2009)

[58] Meyniel-Schicklin et al. (2012)

[59] Sarmady et al. (2011), Yang (2012)

[60] Beissbarth and Speed (2004)

Similar to GO analysis, pathways annotations of genes or proteins are employed to highlight which host functions are preferentially targeted by pathogens; for instance, HIV alone is known to directly or indirectly interact with the majority of known human cellular pathways and these annotation data may thus indicate possible interactors.(61)

Finally, integrative approaches commonly use supervised or semi-supervised machine learning methods in order to classify and cluster protein interactions with regard to their similarity to known host-pathogen protein interactions. These models are trained with a wide array of proteomic, sequence-based, and functional features.(62)

(61) Singh et al. (2010), Zhao et al. (2011b)

(62) Dyer et al. (2011), Qi et al. (2010), Tastan et al. (2009)

Preliminary conclusion

In conclusion, this chapter has presented an overview of viral host factors that are essential for viral entry, replication, and immune evasion and that are cunningly subverted by all known viruses pathogenic to humans. Several classes of antivirals are currently employed to counteract viral disease, the use of many of which is hindered by side effects and emergence of viral drug resistance. Drugs targeting host factors promise to increase the repertoire of available drug targets and may even yield one of the most coveted results of antiviral research: a broad-spectrum antiviral. Identification of drug targets for anti-host factor drugs is depending on experimental assays that measure interactions between viral and host proteins. These assays commonly produce results that contain technical errors such as false positives that may suggest false drug targets. In order to curb technical error rates, computational post-processing schemes have been devised that increase the specificity of the measured interactions and furthermore increase the interpretability of one specific class of protein interaction assays. protein purifications.

18 Inferring physical protein contacts

As introduced earlier *in this chapter, protein purification data provide noisy representations of the dynamic protein complex landscape of the cell. In contrast to binary interaction assays such as Y2H and PCA, protein purification assays can capture complete protein complexes that retain most of their functional components, in principle. It can therefore be argued that the latter approch allows a more complete and naturalistic view on the cellular machinery. As a consequence of their suitability for high-throughput genome-wide applications, protein purification data can inform the systematic search for novel drug targets, for instance by determining the set of proteins that interact with viral peptides within human cell lines.*

This section presents a published manuscript of the author[1] that pertains to the analysis of genome-wide protein interaction assays. In particular, we introduce a novel approach for inferring binary protein interactions from protein purification data that is also well suited for detection of viral host factors. In particular, we offer two novel contributions to the analysis of protein purification data.

First, our approach focuses on inferring direct binary protein interactions from purification data rather than aiming to predict the full set of direct and indirect protein interactions existing within physiological protein complexes. As a consequence, our approach is better suited than comparable methods for inferring interacting protein regions (as, for instance, protein domains) and the result of our method are directly comparable to results of binary protein interaction assays such as Y2H or PCA.

Second, we employ a statistical method that aims to better incorporate repeated protein purifications, i.e., technical replicates, compared to related socio-affinity scores. This statistical approach allows for increased specificity at detecting physical contacts involving highly abundant or transiently interacting protein classes such as kinases and molecular chaperones that are of high relevance for antiviral research.

[1] Schelhorn et al. (2011)

Acknowledgements. The study presented in the next section has been conceived by Elena Zotenko (now at the Garvan Institute of Medical Research, Australia), Julián Mestre (University of Sydney, Australia), and Mario Albrecht (now at University Greifswald, Germany) in 2011. Part of this study was financially supported by the German National Genome Research Network (NGFN) and by the German Research Foundation (DFG), contract number KFO 129/1-2.

Computational implementations related to the study were performed by Sven-Eric Schelhorn. Writing of the manuscript, conceptual development of the main ideas, as well as statistical analyses were equally performed by Sven-Eric Schelhorn and Elena Zotenko. Mario Albrecht and Julián Mestre aided by reading the manuscript and providing conceptual advice.

In order to demonstrate the applicability of the presented methods for inferring host-pathogen protein interactions, the manuscript presented here has been supplemented by the author to include an analysis of physical contacts involving human and HIV proteins as well as an extended section focusing on the importance of molecular chaperones for antiviral research. In addition, the author appended an addendum at the end of this section discussing recent developments within the field of protein purification scoring methods. These extensions are not part of the original publication and consequently have not been approved by any of the other authors attributed.

Introduction

Abstract. Recent large-scale data sets of protein complex purifications have provided unprecedented insights into the organization of cellular protein complexes. Several computational methods have been developed to detect co-complexed proteins in these data sets. Their common aim is the identification of biologically relevant protein complexes. However, much less is known about the network of direct physical protein contacts within the detected protein complexes. Therefore, our work investigates whether direct physical contacts can be computationally derived by combining raw data from large-scale protein complex purifications. We assess four established scoring schemes and introduce a new scoring scheme that is specifically devised to infer direct physical protein contacts from protein complex purifications. The physical contacts identified by the five methods are comprehensively benchmarked against different reference sets that provide evidence for true physical contacts.

Our results show that raw purification data can indeed be exploited to determine high-confidence physical protein contacts within protein complexes. In particular, our new method outperforms competing approaches at discovering physical contacts involving proteins that have been screened multiple times in purification experiments. It also excels in the analysis of recent protein purification screens of molecular chaperones and protein kinases. In contrast to previous findings, we observe that physical contacts inferred from purification experiments of protein complexes can be qualitatively comparable to binary protein interactions measured by experimental high-throughput assays such as yeast two-hybrid. This suggests that computationally derived physical contacts might complement binary protein interaction assays and guide large-scale interactome mapping projects by prioritizing putative physical contacts for further experimental screens.

Introduction. Proteins often do not act in isolation, but cooperate in larger assemblies to fulfill their functions. The resulting protein complexes are essential in a variety of cellular processes.[2] Thus, the identification and annotation of protein complexes is currently the focus of both experimental and computational analyses.[3] Recent advances in experimental technologies for protein purification and identification,[4] such as tandem-affinity purification techniques, enabled high-throughput purification screens for protein complexes in several model organisms.[5] A typical high-throughput screen entails hundreds of purification experiments, where a single purification assays *prey* proteins that associate with a given *bait* protein through multi-protein complex formation.

Due to a variety of reasons, such as experimental noise, presence of non-specific interactors, or participation of the bait protein in multiple distinct protein complexes,[6] the experimentally obtained purifications are not directly interpretable as biologically

[2] Gavin and Superti-Furga (2003)

[3] Robinson et al. (2007)

[4] Dziembowski and Séraphin (2004)

[5] Collins and Choudhary (2008)

[6] Mackay et al. (2007)

relevant protein complexes. Therefore, computational methods are applied to infer these complexes from raw purification data by scoring protein interactions within the purifications. Publication of two independent large-scale screens of protein complexes in the yeast *Saccharomyces cerevisiae* triggered development of several such scoring schemes[7] and resulted in a revised catalog of manually curated yeast complexes.[8]

Proteins within a complex are connected by protein interactions. Here, protein interactions often refer to both direct physical contacts, in which two proteins share a common binding interface, and indirect, bridging interactions, in which the proteins do not contact each other directly. Established purification scoring schemes have been shown to perform well in determining the composition of protein complexes by identifying such protein interactions in the purification data. However, these scoring schemes do not discriminate between direct physical contacts and indirect protein interactions. Consequently, less is known about which proteins in large-scale protein purifications form direct physical contacts although this information is crucial for a deeper understanding of protein complex formation and organization.

Furthermore, the difficulty of identifying physical protein contacts within protein complex purifications has hampered the comparison with results of binary protein interaction experiments such as yeast two-hybrid assays. A recent comparison found substantially more true physical contacts from binary assays than purification experiments.[9] However, this analysis did not consider that protein complex purifications contain both direct physical contacts and indirect protein interactions in contrast to binary assays. Since this results in a lower enrichment with physical contacts, a comparison of the experimental assays that concentrates only on putative physical protein contacts would provide deeper insights into the relative merits of each experimental technology.

Even though several experimental and computational methods exist that produce structural models of protein complexes at various levels of resolution,[10] structural data required by these approaches is not readily available for the vast majority of complexes detected by large-scale protein purifications. Thus, the main objective of this work is to assess whether and how we can make use of the available purification screens to computationally infer the network of physical contacts within the assayed protein complexes.

Our guiding principle rests upon the observation that proteins forming physical contacts within a complex exhibit stronger associations and thus are more likely to survive purification procedures than proteins that do not form such contacts. A similar observation is central to a hybrid approach developed by the Robinson's lab in which individual protein complexes are perturbed by experimental techniques to discover physical contacts between proteins within these complexes.[11] We hypothesize that even though large-scale purification screens do not directly measure physical contacts

[7] Collins et al. (2007), Friedel et al. (2009), Gavin et al. (2006), Hart et al. (2007), Krogan et al. (2006), Pu et al. (2009)

[8] Pu et al. (2009)

[9] Yu et al. (2008)

[10] Alber et al. (2007), Aloy et al. (2004)

[11] Hernández et al. (2006)

within complexes, the resulting experimental data does contain sufficient information to reliably infer these interactions.The three main contributions of our work are as follows.

First, we propose an elegant computational method for scoring pairs of proteins based on their co-occurrence pattern within a combined set of purifications originating from multiple large-scale screens. In contrast to existing scoring schemes, which were originally developed and evaluated to detect co-complexed protein pairs regardless of their mode of interaction, our approach is tuned to detect protein pairs that form direct physical contacts and incorporates experimental replicates in a statistically sound fashion. As a consequence, our method can reliably detect true physical contacts even in the presence of many unspecific or highly transient protein interactions and especially outperforms existing scoring schemes if experimental replicates are available. These properties make our approach particularly suited for the joint analysis of physical contacts from multiple protein purification screens.

Second, we perform a comprehensive evaluation of our and four other published scoring methods on the task of detecting physical contacts. Each method scores purification data from two recent large-scale experiments in yeast.(12) The results of all scoring methods are benchmarked against several reference sets that represent complementary evidence for physical contacts. The reference sets are derived from experimentally determined physical interactions, three-dimensional structures of protein complexes, manually curated catalogs of protein complexes, and genetic interaction profiles. In particular, we assess the scoring methods by inferring specific physical contacts in two challenging and biologically relevant purification data sets containing repeated purifications of molecular chaperones or protein kinases.

Third, we compare top-ranking physical contacts inferred by our method to two recent high-throughput interaction data sets derived by the experimental techniques *yeast two-hybrid* (y2h)(13) and *protein fragment complementation assay* (pca)(14) and address intrinsic differences of high-throughput approaches to mapping physical interactomes.

(12) Gavin et al. (2006), Krogan et al. (2006)

(13) Yu et al. (2008)

(14) Tarassov et al. (2008)

Materials and methods

Large-scale yeast purification data. We utilized a combined set of purifications from two large-scale screens in the yeast *S. cerevisiae*.[15] Raw experimental data was obtained from the supporting web-site (`http://interactome-cmp.ucsf.edu`) of one of the existing scoring methods included in our evaluation, the PE method.[16] Purification data from the Gavin *et al.*[17] screen was taken as it is, while purification data from the Krogan *et al.*[18] screen was filtered as follows. The Krogan team used two different experimental protocols, LCMS and MALDI, for prey identification. For each purification, we retained preys having MALDI identification score of at least 1.25 and/or LCMS confidence score of at least 99%. We further filtered out preys that were identical to baits in their respective purifications from both screens. Last, we combined purifications from the two individual screens into one data set, which we denoted as LARGE-SCALE set of purifications. Table 18.1 summarizes purification data from individual screens as well as from the LARGE-SCALE set.

	GAVIN	KROGAN	LARGE-SCALE
purifications	1,912	3,999	5,911
baits	1,754	2,178	2,830
preys	1,813	3,505	3,759
avg. # preys	10.56	10.31	10.39
protein interactions (bait-prey pairs)	18,206	32,525	47,254
protein interactions (bait-prey and prey-prey pairs)	82,202	182,134	238,154

Table 18.1: *Summary of purification data.* Summary of purification data from two independent large-scale complex purification screens in yeast, denoted here as GAVIN and KROGAN, as well as for the combined LARGE-SCALE set. For each screen the number of purifications, the number of distinct bait proteins, the number of distinct prey proteins, the average number of preys per purification, and the number of distinct bait-prey and distinct bait-prey and prey-prey pairs are shown.

Large-scale host-pathogen purification data. In addition to the yeast purification data employed in the main validation of this study, we additional obtained large-scale host-pathogen protein purifications of HIV proteins and HIV poly-proteins in combination with human host factors.[19] These purifications were performed on two cell lines, HEK293 and Jurkat, and the resulting data sets are here denoted as HEK and JURKAT, respectively.

Protein kinase and phosphatase purification data. In addition to the LARGE-SCALE data set, we utilized a specialized purification data set focusing on *kinase* and *phosphatase* interactions in yeast from a recent experimental study by Breitkreutz *et al.*[20] The bait proteins in the Breitkreutz data were screened with three different tag systems (FLAG, HA, and TAP) and the purified prey proteins include information about peptide (spectral) counts that can be used as a semi-quantitative measure of absolute protein abundance. We obtained raw purification data for all three tag systems from the supporting website (`http://www.yeastkinome.org`). Subsequently, the purifications were filtered to (i) exclude tag-specific contami-

[15] Gavin et al. (2006), Krogan et al. (2006)

[16] Collins et al. (2007)

[17] Gavin et al. (2006)

[18] Krogan et al. (2006)

[19] Jäger et al. (2012)

[20] Breitkreutz et al. (2010)

nant proteins identified by control experiments in the original study and to (ii) remove unreliably identified prey proteins with Mascot scores $\leq$35. The resulting BREITKREUTZ (BK) purification data set is summarized in Table 18.2.

	BK-FLAG	BK-HA	BK-TAG
purifications	210	307	57
baits	129	207	48
avg. # preys	50.65	47.22	24.39

Table 18.2: *Summary of purification data from a recent purification screen.* Summary of purification data from the recent BREITKREUTZ (BK) purification screen by Breitkreutz et al. (2010) focusing on kinases and phosphatases in yeast. The screen was performed with three different tag systems, FLAG, HA, and TAG. For each tag system, the number of purifications, the number of distinct bait proteins, and the average number of preys per purification are shown.

High-confidence physical contacts in yeast. Binary gold standard (BGS) protein interactions used in a recent assessment of binary experimental methods[21] as well as the experimental yeast two-hybrid data (Y2H) generated in the same study were obtained from the CCSB interactome database. Note that, since no true gold standard for binary protein interactions is available, we decided to use the BGS naming convention from Yu et al. (2008) to allow for better comparability of our work. Further binary interactions measured by a recent protein-fragment complementation assay (PCA)[22] were extracted from the Saccharomyces Genome Database (SGD).[23]

Binary interactions originating from experimental assays that directly measure physical protein contacts were obtained from IntAct[24] and SGD. This set of interactions was filtered to exclude physical contacts that solely rely on evidence from the Y2H and PCA binary protein interaction data sets. The filtered data set was utilized to (i) define two reference sets: a CHAPERONE reference set containing only interactions involving yeast molecular chaperones and (ii) a KINASE reference set including solely interactions involving yeast kinases and phosphatases.

High-confidence host-pathogen physical contacts. Binary interactions representing physical contacts between human and HIV proteins as well as between pairs of HIV proteins were obtained from the publication of a recent large-scale purification screen involving these factors.[25] Since these validation data originally originated from the VirusMINT[26] database, we here denote the corresponding data set as VIRUSMINT.

Protein complexes and domain interactions. Protein complexes derived from Gene Ontology annotations as provided by SGD and manually curated protein complexes from the Munich Information Center for Protein Sequences (MIPS)[27] were imported from the websites of the respective organizations and yielded the SGD and MIPS reference sets, respectively. For the validation of inferred physical contacts on the level of protein domain interactions, a mapping table from SGD was obtained to assign UniProt accession numbers to all proteins in the LARGE-SCALE purification data. Subsequently, globular Pfam-A domain annotations for these proteins were obtained from the InterPro database.[28] We restricted the used annotations to globular domains since this type of protein regions is especially well characterized and known to be involved in stable protein interactions.[29] A list of Pfam-A domain interaction partners derived from structures of interacting proteins in the Protein Data Bank (PDB)[30] was obtained from the 3DID database.[31]

[21] Yu et al. (2008)

[22] Tarassov et al. (2008)

[23] Hong et al. (2007)

[24] Kerrien et al. (2007)

[25] Jäger et al. (2012)

[26] Chatr-aryamontri et al. (2009)

[27] Guldener et al. (2006)

[28] Mulder et al. (2007)

[29] Aloy and Russell (2006)

[30] Berman et al. (2000)

[31] Stein et al. (2005)

Genetic interaction profiles. Interaction confidence scores derived from *in-vivo* synthetic genetic interactions (GI) were obtained from the supplementary material of a recent large-scale functional study in yeast.[32] Of the several available data sets containing genetic interaction scores for pairs of proteins, we selected the *lenient cut-off interaction set* that offered the highest coverage of yeast proteins while, at the same time, including only statistically significant interactions. This data was used to reconstruct genetic interaction profiles for all proteins in the LARGE-SCALE purification data set. Consequently, genetic interaction profile similarities for pairs of proteins were generated by computing Pearson's correlation coefficients between the genetic interaction profiles of all protein pairs. Analogous to the original study, all protein pairs with a genetic interaction profile similarity $>= 0.2$ were used to form a functional map of the LARGE-SCALE purification data. Protein pairs in this map are denoted as the GI reference set.

[32] Costanzo et al. (2010)

Scoring methods. Let $\Phi = \{\phi_1, ..., \phi_N\}$ be a set of purifications where each purification ϕ_k is composed of a bait protein $bait_k$ and a set of prey proteins $preys_k$. We will use n_k to denote the number of preys in ϕ_k and designate $N = \sum_k n_k$ as the size of the *multi-set* of all $preys_k$. For a pair of proteins, i and j, let $S_{i\to j}$ be the number of times j is observed among preys in purifications performed with i as bait and $M_{i,j}$ be the number of times i and j are observed as preys in purifications performed with a third protein as a bait. In what follows, the experimental observations S and M are also denoted as *spoke observations* and *matrix observations*, respectively. Some scoring schemes combine S and M into one number $O_{i,j} = S_{i\to j} + S_{j\to i} + M_{i,j}$. We will denote by $S^{\text{null}}_{i\to j}$, $M^{\text{null}}_{i,j}$, and $O^{\text{null}}_{i,j}$ random variables representing these different types of observation counts under appropriate null models. Next, we briefly introduce the scoring methods that are assessed in this work.

Socio-affinity scores. The socio-affinity (SA) scoring scheme is one of the first approaches for scoring purification data and was developed by Gavin *et al.* to interpret the results of their large-scale screen.[33] The score is based on three counts $S_{i\to j}$, $S_{j\to i}$, and $M_{i,j}$ and is given by:

[33] Gavin et al. (2006)

$$\textsc{sa}(i,j) = \log \frac{S_{i\to j}}{E\left[S^{\text{null}}_{i\to j}\right]} + \log \frac{S_{j\to i}}{E\left[S^{\text{null}}_{j\to i}\right]} + \log \frac{M_{\{i,j\}}}{E\left[M^{\text{null}}_{\{i,j\}}\right]} \qquad (18.1)$$

The distributions of $S^{\text{null}}_{i\to j}$ and $M^{\text{null}}_{i,j}$ are modeled based on the assumption that purifications are drawn uniformly at random from the observed multi-set of preys. This means that ϕ^{null}_k is formed through n_k independent random selections of preys where the probability of selecting protein the prey protein j is equal to its relative frequency $f_j = |\{\phi_k \mid j \in \text{preys}_k\}| \cdot N^{-1}$. Under this null model, the expected values of $S^{\text{null}}_{i\to j}$ and $M^{\text{null}}_{i,j}$ are given by:

$$E\left[S_{i\to j}^{\text{null}}\right] = \sum_{k:i=\text{bait}_k} f_j n_k \qquad (18.2)$$

$$E\left[M_{i,j}^{\text{null}}\right] = \sum_k f_i f_j \binom{n_k}{2} \qquad (18.3)$$

Improved socio-affinity scores. In this work, we propose a modification of the SA scoring scheme termed *improved socio-affinity* (ISA) score. It makes full use of repetitive purifications and concentrates on spoke observations S to improve the detection of physical contacts. In particular, we adopt the null model used for SA scores and derive the ISA score as follows:

$$\text{ISA}(i,j) = -\log \Pr\left(S_{i\to j}^{\text{null}} \geq S_{i\to j}\right) - \log \Pr\left(S_{j\to i}^{\text{null}} \geq S_{j\to i}\right) \qquad (18.4)$$

To compute $\Pr(S_{i\to j}^{\text{null}} \geq S_{i\to j})$ and $\Pr(S_{j\to i}^{\text{null}} \geq S_{j\to i})$, we introduce an indicator random variable $X_{j,k}$ that corresponds to the selection of protein j into the set of preys of ϕ_k^{null}, thus $\Pr(X_{j,k}) = 1-(1-f_j)^{n_k}$. We then note that $S_{i\to j}^{\text{null}}$ is a sum of independent binary random variables: $S_{i\to j}^{\text{null}} = \sum_{k:i=\text{bait}_k} X_{j,k}$. Since $\Pr(X_{j,k})$ depends on the size of ϕ_k, it is, in general, not the same for different purifications performed with bait protein i. As a result, the distribution of $S_{i\to j}^{\text{null}}$ is not binomial. To alleviate this problem, we set $\Pr(X_{j,k}) = 1-(1-f_j)^{\hat{n}}$, where $\hat{n}$ is the average size of purifications performed with i, and use the binomial distribution to compute $\Pr(S_{i\to j}^{\text{null}} \geq S_{i\to j})$. To avoid a situation where a single observation with a rare prey protein receives very high scores, we adjust background prey frequencies f_j by adding a constant ϵ fraction of each prey to each purification. In this work we used $\epsilon = 0.0025$.

Purification enrichment scores. The purification enrichment (PE) scoring scheme was proposed by Collins *et al.*[34] as an alternative to the original SA scores to analyze the combined set of purifications from two recent large-scale screens in yeast. The authors adopted a more sophisticated statistical model to score evidence for each observation o separately:

[34] Collins et al. (2007)

$$\text{PE}(i,j) = \sum_o \log\left(\frac{\Pr(o \mid i \text{ and } j \text{ interact})}{\Pr(o \mid i \text{ and } j \text{ do not interact})}\right) \qquad (18.5)$$

The detailed description of the statistical model used to derive the probabilities above is beyond the scope of this paper and the interested reader is referred to the original publication.[35] We just note here that the estimation of parameters used by the model is not straightforward and requires a representative set of gold standard interactions.

[35] Collins et al. (2007)

Hart scores. Another scoring scheme was proposed by Hart *et al.*[36] with a particular emphasis on joint analysis of experimental data from several large-scale purification screens of protein complexes. In this approach, the scores are based on the combined number of observations, $O_{i,j}$, and are computed as:

$$\textsc{hart}(i,j) = -\log \Pr\left(O_{i,j}^{\text{null}} \geq O_{i,j}\right) \tag{18.6}$$

The distribution of $O_{i,j}^{\text{null}}$ is modeled based on the assumption that interactions of a protein are selected uniformly at random from the multi-set of all observed interactions. More precisely, for a pair of proteins i and j, $O_i = \sum_j O_{i,j}$ interactions are selected uniformly at random from the ground set of $O = \sum_{i,j} O_{i,j}$ interactions that contains $O_j = \sum_i O_{i,j}$ "relevant" (involving j) interactions and $O - O_j$ "irrelevant" (not involving j) interactions. The statistical significance of the observed $O_{i,j}$ is then assessed by determining the probability that at least $O_{i,j}$ "relevant" interactions are selected. Under this null model, $O_{i,j}^{\text{null}}$ has a hyper-geometric distribution and $\Pr\left(O_{i,j}^{\text{null}} \geq O_{i,j}\right)$ can be efficiently computed. We note that both the SA and HART null models are simple and parameter free, resulting in efficient computational procedures. However, the SA null model takes the structure of original purifications into account while the HART null model employs summary statistics of all pairwise interactions in the purification data and thus disregards the structure of the original purifications.

[36] Hart et al. (2007)

IDBOS scores. Recently, the IDBOS scoring scheme was proposed for scoring purification data with an emphasis on the prediction of direct physical protein interactions.[37] In this approach, the scores are based on the combined number of observations, $O_{i,j}$, and are computed as follows:

[37] Yu et al. (2009)

$$\textsc{idbos}(i,j) = \frac{O_{i,j} - E\left[O_{i,j}^{\text{null}}\right]}{S\left[O_{i,j}^{\text{null}}\right]} \tag{18.7}$$

The distribution of $O_{i,j}^{\text{null}}$ is modeled by assuming that the observed purifications are randomly permuted. The resulting null model is very similar to the one used by the SA and ISA approaches. The main difference is the accurate modeling of observed purifications – the IDBOS null model does not allow for random instances where a prey appears multiple times in a single purification. However, this small gain in accuracy comes at a high computational cost. Since the resulting distribution of $O_{i,j}^{\text{null}}$ is much more complex, extensive numerical simulations are required to estimate its properties. The authors perform 10^6 numerical randomization experiments to estimate the expected value of $O_{i,j}^{\text{null}}$, $E\left[O_{i,j}^{\text{null}}\right]$, and its standard deviation $S\left[O_{i,j}^{\text{null}}\right]$.

SAINT scores. The Significance Analysis of Interactome (SAINT) scoring scheme was recently introduced to detect non-specifically interacting proteins in the BREITKREUTZ purification data.[38] The method is depending on the use of peptide counts, an additional type of experimental data that was generated during the peptide identification phase of the BREITKREUTZ screen and can be utilized as a semi-quantitative measure of absolute protein abundance. SAINT employs a mixture of Poisson distributions to heuristically compute posterior probabilities of specific interactions between proteins based on the peptide counts. Due to the high complexity of the model, presentations of detailed theoretical underpinnings of SAINT are beyond the scope of this work. However, we note here that SAINT is a comparatively complex scoring method with many free parameters and that it requires the availability of experimental peptide count data. Both properties hinder its applicability to publicly available large-scale purification data.

Score implementations. SA, ISA, and HART scores were computed using in-house Python scripts on the LARGE-SCALE set of purifications. Due to the computational complexity of IDBOS and PE scores, these scores were obtained from the original publications. While PE scores were computed on the LARGE-SCALE set of purifications by the authors, IDBOS does not support the computation of scores based on multiple sources of purification data (personal communication with the first author of the Yu et al. (2009) publication). Therefore, we used the IDBOS scores computed on the GAVIN data since these showed the best performance among all scored data sets in the original publication.[39] Due to its reliance on peptide count data, SAINT is not applicable to the LARGE-SCALE set of purifications. SAINT scores for the BREITKREUTZ purification data were obtained from the supplementary materials of the original publication.[40]

Salama-Quade rank correlation. The Salama-Quade rank correlation coefficient measures similarity between two different rankings of a set of elements.[41] It was developed as an alternative to standard rank correlation measures, such as *Spearman's rho* and *Kendall's tau*, for situations where agreement in low ranks is more important than agreement in high ranks.

Let $\{1, ..., m\}$ be a set of elements, $R_1(i)$ be the rank of element i under the first method, and $R_2(i)$ be the rank of element i under the second method. The Salama-Quade (SQ) coefficient measures the agreement between rankings R_1 and R_2 and is given by $\text{SQ}(R_1, R_2) = \sum_{k=1}^{K} T_k / k$, where T_k is the number of elements having rank less or equal to k under both R_1 and R_2. For similarity values in Figure 18.1b we used $K = 10,000$ and normalization $\text{SQ}(R_1, R_2)/K$ in order to obtain rank similarities for the first 10,000 predicted interactions.

[38] Breitkreutz et al. (2010)

[39] Yu et al. (2009)

[40] Breitkreutz et al. (2010)

[41] Salama and Quade (1982)

Results and discussion

Scoring purification data. A purification represents the outcome of an experiment in which a single *bait protein* is tagged and biochemically co-purified with *prey proteins* that associate with the bait by forming of one or several protein complexes. For example, in the screen by Gavin *et al.*,[42] a specific purification using the α-subunit of clathrin adaptor complex AP-2 as bait contained 22 prey proteins. Three of the co-purified preys correspond to the other subunits of this hetero-tetrameric complex; the other preys might be either unknown subunits of the AP-2 complex, subunits of other complexes in which the α-subunit participates, or non-specific interactors.

Even though the interpretation of a single purification is limited, combined purifications from several large-scale screens contain repeated observations of associated proteins that may indicate true protein-protein interactions. Over the years, several computational approaches were developed to integrate experimental observations across purifications in order to infer pairs of interacting proteins. Four major approaches utilizing raw experimental data are socio-affinity scores (SA),[43] purification enrichment scores (PE),[44] scores developed by Hart *et al.*[45] (HART), and recently published ID-BOS scores.[46] However, none of these methods is ideally suited for identifying *direct physical contacts* between proteins within a complex through the joint analysis of purifications from several large-scale screens.

In this work we propose a novel scoring method specifically tailored to using repeated purifications. In the following, we briefly describe the main features of our new scoring method ISA (*improved socio-affinity* score).

A single purification provides two kinds of experimental evidence for protein interactions: *spoke observations* supporting interactions between the bait and each of the preys, and *matrix observations* supporting interactions between every pair of preys. In some cases, however, matrix observations are much less reliable than spoke observations. In particular, a large fraction of matrix observations from purifications containing several small complexes would support non-existing interactions between proteins in distinct complexes.

The distribution of protein complex sizes in manually curated catalogues of protein complexes in yeast suggests that the majority of complexes are small; about 64% of complexes in the MIPS catalogue, for example, have up to four subunits. In comparison, the average number of preys in the purifications in our data set is about 10 (see Table 18.1). It appears therefore that the majority of purifications are indeed composed of several complexes and thus provide many misleading matrix observations. While relying on potentially misleading matrix observations does not adversely affect scoring of co-complexed protein pairs, the task of identifying direct physical contacts is more sensitive towards misleading observations.

[42] Gavin et al. (2006)

[43] Gavin et al. (2006)

[44] Collins et al. (2007)

[45] Hart et al. (2007)

[46] Yu et al. (2009)

Consequently, ISA is cautious and derives interaction confidence scores solely from the more reliable spoke observations while completely discarding matrix observations. This is in stark contrast to the other four methods SA, PE, HART, and IDBOS that derive their scores from a mixture of both spoke and matrix observations.

Similar to related approaches, our method utilizes statistical techniques to derive interaction confidence scores from spoke observations contained in the experimental purification data. Specifically, for each pair of proteins, the number of spoke observations in the experimental data is compared to the number of such observations under an appropriate null model.

Our novel method ISA adopts the null model of Gavin *et al.* introduced in the context of the SA scoring method.[47] This null model considers size and content of the original purifications during computations, but selects prey proteins for each purification uniformly at random from the multi-set of preys.

Even though more sophisticated null models were proposed in the context of later scoring methods such as HART, we believe that the SA null model is ideally suited for scoring complex purification data. On the one hand, it is simple enough to allow for analytical derivation of statistical significances. On the other hand, it realistically models the observed data by preserving much of the structure of the original purifications such as the identity of bait proteins, purifications sizes, and frequency of prey proteins.

However, one of the main problems with the SA approach is that additional observations supporting a protein interaction result in a disproportionally small increase of the SA score. This poses a problem when purifications from several independent screens are jointly analyzed. Therefore, as a major improvement over the SA method, ISA scores are derived through statistical *p-value* computations which allows for attributing higher confidence to putative physical contacts with multiple supporting observations originating from experimental replicates.

Scoring two large-scale purification experiments in S. cerevisiae. We used the four established scoring schemes SA, HART, PE, and IDBOS as well as our own approach to score a combined set of purifications from two recent large-scale screens of protein complexes in *S. cerevisiae*.[48] In this section, we examine top-ranking inferred physical contacts between proteins by our method and relate them to results of the other four scoring methods.

Table 18.3 lists ten inferred physical contacts having the highest ISA scores. All but two interactions in the top-ten list are supported by small-scale experiments reported in the literature. Four top-ten physical contacts receive low scores under the SA method which highlights one of the main differences between the SA and ISA approaches.

[47] Gavin et al. (2006)

[48] Gavin et al. (2006), Krogan et al. (2006)

Consider, for example, the interaction between HAT2 and HHT2 proteins, which is ranked third by the ISA method and 56,934th by the SA method. The HAT2 protein appears as prey in 23 out of 27 purifications performed with HHT2. Still, the SA method assigns low weight to repeated observations of the kind 'HAT2 purifies HHT2', resulting in a high rank number of the corresponding physical contact.

In general, top-ranking physical contacts inferred by our method are expected to be enriched with interactions involving proteins that were purified multiple times. Indeed, if a given bait protein was purified repeatedly in multiple purifications, its interaction partners can and should be determined with increased confidence.

For instance, our experimental data contains 27 purifications performed with HHT2 as bait, which support a total of 124 protein interactions. Out of these 124 interactions, 15 have sufficiently high ISA scores to be included in the top-3,000 inferred physical contacts. Figure 18.1a depicts a network induced by the top-3,000 inferred physical contacts inferred by ISA. The network is sparse and modular, which agrees well with our intuition for the network of direct physical interactions within stable multi-protein complexes.

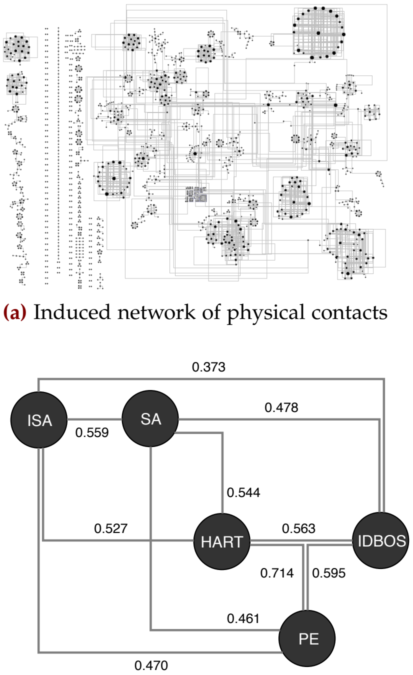
i	j	$S_{i\to j}$	$S_{j\to i}$	$M_{i,j}$	SA	PE	HART	IDBOS	Ref.
UBP2	RUP1	11	7	0	1568	403	114	67	2
RFA1	RFA2	16	3	13	1131	256	91	2371	3
HAT2	HHT2	0	23	0	56934	7624	435	NA	1
HIF1	HHT2	0	22	0	53035	6250	444	NA	0
HIF1	HAT1	7	14	32	1857	70	5	321	2
HHT2	HAT1	22	0	0	58294	8010	622	NA	1
SRB4	SRB5	5	15	14	1322	44	170	264	12
SPT16	POB3	16	2	55	2705	791	4	2326	6
HHT2	PSH1	18	0	0	47045	6862	702	NA	0
UBA2	AOS1	6	5	0	899	1113	412	NA	1

Table 18.3: *Top-10 physical contacts inferred by* ISA. A list of top-10 physical contacts inferred by the ISA score. For each physical contact, the number of supporting spoke observations ($S_{i\to j}$ and $S_{j\to i}$), number of supporting matrix observations ($M_{i,j}$), rank under the other scoring schemes, and number of distinct supporting SGD literature references (Ref.) are listed.

To assess the overall similarity among the five scoring methods we utilized the Salama-Quade rank correlation coefficient. The Salama-Quade coefficient belongs to a family of measures that assign greater weight to agreement in top-ranked elements than other correlations measures such as Spearman's rho and Kendall's tau and is thus more suitable for our purpose (see Materials and methods).

As we argue below, out of 238,154 possible physical contacts supported by raw purification data, only about 3,000 can be reliably scored. Therefore, the relative ordering of inferred physical contacts beyond this cutoff is less reliable and should affect the similarity score to a lesser extent. Figure 18.1b shows Salama-Quade rank correlation for every pair of methods. Ranking of inferred physical contacts induced by ISA scores is quite distinct from rankings induced by other scoring methods. Based on similarity scores, the methods can be grouped into two clusters, one including HART, PE, and IDBOS methods and another one containing SA and ISA methods.

We hypothesize that this grouping reflects a major difference among the scoring methods, namely, treatment of matrix observations. While the HART, PE, and IDBOS methods treat spoke and matrix observations equally, the ISA method completely ignores matrix observations. SA implicitly downplays the contribution of matrix observations by assigning low weight to repeated matrix observations and, as a result, is grouped together with the ISA method.



(a) Induced network of physical contacts

(b) Similarities of scoring methods

Figure 18.1: *Top-ranking physical contacts inferred by* ISA. Top-ranking physical contacts inferred by the ISA method and their relation to physical contacts inferred by other methods. Figure 18.1a: A network induced by the 3,000 protein interactions having the top ISA score ranks. Figure 18.1b: Similarity of the inferred physical contacts generated by the five scoring methods. Nodes represent different scoring schemes. Edges are labeled with the Salama-Quade correlation coefficient, which measures agreement in the ranking of inferred protein contacts induced by the scores of the corresponding methods.

Benchmarking against reference sets of physical contacts. In this section, we assess the ability of the four established scoring schemes and our new approach to detect physical contacts within protein purifications. Since, to the best of our knowledge, there is no comprehensive gold standard set of protein interactions that form physical contacts within protein complexes, we approach the evaluation task from four different directions. First, we compare top-ranked inferred physical contacts between proteins to protein interactions derived by experimental techniques that directly assay protein pairs able to physically interact. However, as we argue later on, interactions detected by these techniques represent only a small fraction of physical contacts present in protein complexes. Therefore, we resort to additional, albeit less direct, procedures to assess the performance of the scoring methods by (i) relying on three-dimensional structures of protein complexes, by (ii) utilizing manually curated catalogs of protein complexes, and by (iii) employing synthetic genetic interaction profiles.

Experimentally determined binary protein interactions. For the first evaluation, we compiled three reference sets (Y2H, PCA, and BGS) of experimentally validated binary protein interactions. The first two data sets originate from recent high-throughput interactome screens in yeast: one employing the yeast two-hybrid (Y2H) technique and another utilizing protein-fragment complementation assays (PCA).[49] The third data set, BGS, contains manually curated yeast interactions supported by literature and is taken from an extensive validation of the Y2H method.[50]

Figure 18.2 shows how well the scoring methods perform in identifying true physical contacts from the reference sets. Note that although all methods are able to infer a number of physical contacts beyond the depicted 10,000 ranks, physical contacts at these high cutoffs have only very low confidence and are thus omitted. Notably, while SA and ISA methods have comparable performance across assessments, both of them outperform other approaches on all three reference sets.

Moreover, the performance of all approaches starts to level off at about 3,000 to 4,000 ranks. We hypothesize that this number constitutes a reasonable limit on the number of physical contacts that can be reliably inferred given the available experimental data. This number of reliably inferable contacts also corresponds roughly to the number of direct binary interactions measured by high-throughput experimental techniques: the Y2H data set and the PCA data contain 2,930 and 2,616 interactions, respectively.

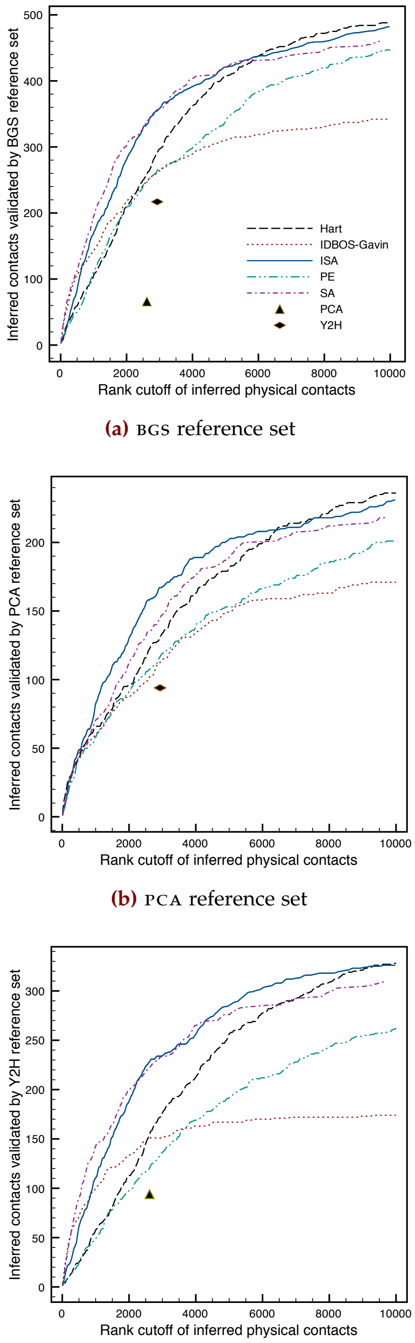
As can be derived from Figure 18.2a, Y2H and PCA data sets are less enriched in manually curated BGS interactions than an equivalent number of top-scoring interactions extracted from purification data. This suggests that physical contacts inferred by purification scoring schemes are at least qualitatively comparable and often superior to Y2H and PCA experimental data sets.

Three-dimensional structures of multi-protein complexes. In this assessment of inferred physical contacts, we rely on experimentally determined structures of protein complexes deposited in the Protein Data Bank (PDB).[51] Unfortunately, only crystal structures of about 250 interactions between proteins in yeast are available.[52] Therefore, the utilization of PDB structures for the assessment of putative physical contacts is not possible due to the low coverage of the validation set.

However, physical contacts in stable multi-protein complexes are typically formed by pairs of structural protein domains,[53] and members of an evolutionarily conserved domain family typically share a common set of domain binding partners. Accordingly, a set of protein interactions that correspond to physical contacts within yeast protein complexes should be enriched in domain pairs that are known to interact. Consequently, to achieve a higher coverage of true physical contacts, we perform this evaluation at the level of PDB-validated physical contacts between protein domains rather than at the level of interacting proteins.

Several resources exist that derive pairs of interacting domains from crystal structures of protein complexes in the PDB. In this work, we use the latest release of the 3DID database[54] to obtain interactions between domains that are annotated to at least one yeast gene. These interactions are denoted as 3DID reference set and compared to a set of domains induced by top-ranking inferred physical contacts. More specifically, for each method, all domain pairs were ranked according to the best-ranking inferred physical protein contact that could be formed by the domain pair.

The results of this evaluation are presented in Figure 18.3a. Again, the SA and ISA methods significantly outperform other approaches over the range of 3,000 to 4,000 inferred physical contacts that are reliably supported by experimental data. At the same time, both SA and ISA perform comparably to the Y2H binary experimental data set with about 240 true domain interactions at a rank cutoff of 4,000 physical contacts.



(a) BGS reference set

(b) PCA reference set

(c) Y2H reference set

Figure 18.2: *Assessment by binary reference sets.* Assessment of inferred physical protein contacts by five scoring methods against binary experimental reference sets that provide direct evidence for physical contacts.

[49] Tarassov et al. (2008), Yu et al. (2008)

[50] Yu et al. (2008)

[51] Berman et al. (2000)

[52] Mosca et al. (2009)

[53] Aloy and Russell (2006)

[54] Stein et al. (2009)

Functional similarity derived from genetic interaction profiles. To assess the functional similarity of proteins inferred to form physical contacts, we rely on *in-vivo* genetic interaction profiles measured by a recent, functionally unbiased large-scale screen.[55] While the results of this screen do not provide direct evidence for physical contacts, physically interacting proteins that carry out similar functions often exhibit strongly correlated genetic interaction profiles. We employed these profile data to define a GI reference set containing protein pairs with high genetic interaction profile similarities. This reference set was used to assess the functional similarity of inferred physical contacts from all five methods (see Fig. 18.3b). The assessment shows that the socio-affinity based methods SA and ISA significantly outperform other scoring methods as well as the binary experimental data sets PCA and Y2H in enriching for functionally similar protein pairs.

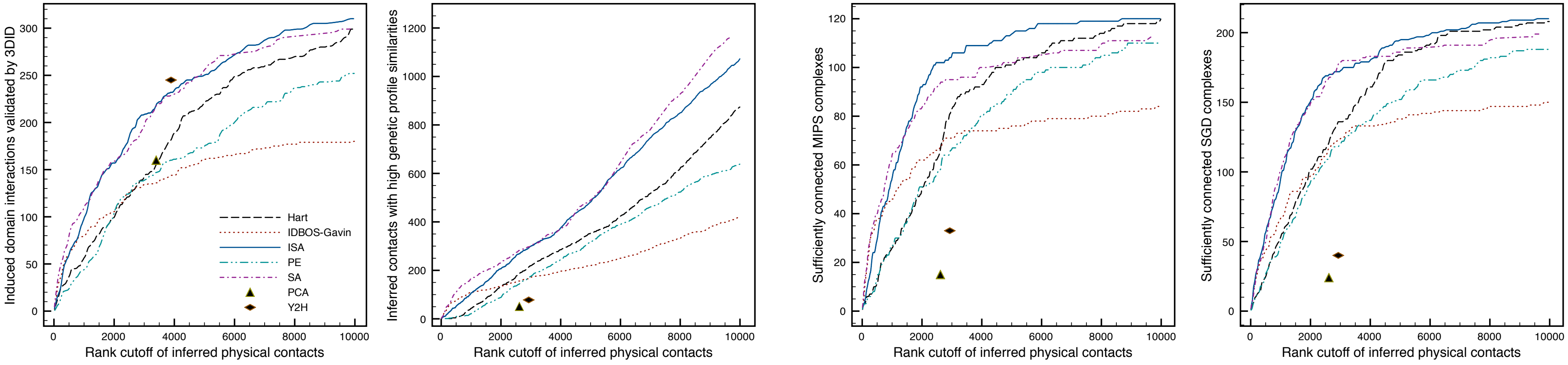
Manually curated catalogues of multi-protein complexes. Several catalogues of manually curated protein assemblies in yeast are publicly available, such as MIPS and SGD complexes. Unfortunately, these high-quality data sets only provide information on the protein composition of each assembly and do not include the network of physical contacts present within the complex. Therefore, we rely on the following assumption to assess the inferred physical contacts with the MIPS and SGD data sets: physical contacts within a complex connect all its member proteins. This means that, within a given complex, every protein is connected to every other protein through a network of physical contacts. Consequently, the quality of a set of inferred physical contacts can be estimated by assessing how well these physical contacts connect manually curated complexes. Figures 18.3c–18.3d depict how well top-ranking inferred physical contacts from different scoring methods connect complexes in the two manually curated catalogs. It is noticeable that the results generated by purification scoring schemes seem to be significantly better suited to connect these complexes than data sets originating from Y2H and PCA techniques, with the best performing scoring methods SA and ISA connecting more than three times as many complexes than the Y2H data at a rank cutoff of approximately 3,000 physical contacts.

Inferring physical protein contacts from repeated purifications. The use of experimental replicates in a statistically meaningful fashion to account for experimental errors is a current theme in interactomics research.[56] While the use of orthogonal assays is already well established in experimental protocols for binary protein interactions such as yeast two-hybrid,[57] it is less widespread in the analysis of protein purification experiments. Our novel method ISA aims at being a generally applicable scoring scheme for inferring physical protein contacts from repeated protein complex purification experiments.

[55] Costanzo et al. (2010)

[56] Braun et al. (2010), Chen et al. (2010)

[57] Lappe and Holm (2004), Schwartz et al. (2009)



(a) 3DID reference set

(b) GI reference set

(c) MIPS reference set

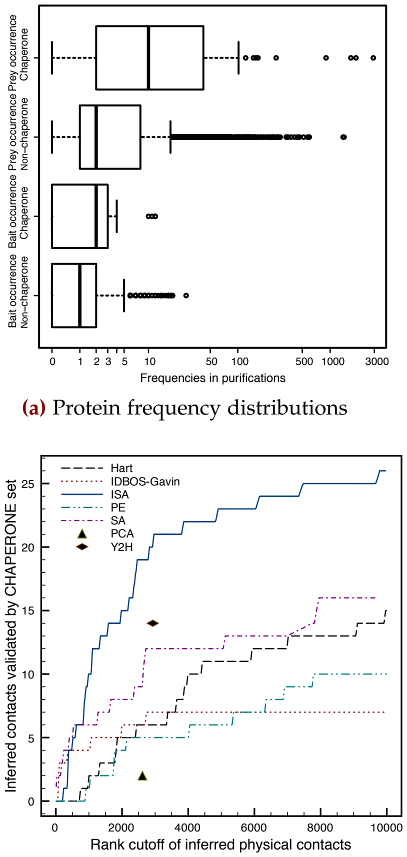
(d) SGD reference set

Figure 18.3: *Assessment by secondary reference sets.* Assessment of inferred physical protein contacts by five scoring methods against reference sets that provide indirect evidence for physical contacts. Inferred physical contacts are ranked by scores of the corresponding scoring method. Figure 18.3a: Physical contacts are evaluated by their enrichment in protein domains that are known to interact in crystal structures of protein complexes. Figure 18.3b: Functional similarity of proteins involved in inferred physical contacts is assessed by correlating genetic interaction profiles of these proteins. Figures 18.3c and 18.3d: Performance is measured by plotting the number of complexes that are *sufficiently connected* by top-ranking inferred physical contacts for different rank cutoffs. We consider a complex sufficiently connected by a set of inferred physical contacts if the physical contacts reduce the number of connected components within the complex to less than 50% compared to the unconnected complex.

While the ISA method performs consistently well in the assessment against experimentally determined physical contacts, its performance difference to the original SA method appears to be only minor. However, one of the main improvements of our novel method ISA is the enhanced null model that takes full advantage of additional evidence contained in repeated observations, depends on the presence of repeated purifications of the same bait protein in the experimental data. A closer analysis of the purifications in the LARGE-SCALE data reveals that proteins were used as baits at a median number of only one time (see bait frequency distribution of non-chaperone proteins in Figure 18.4a). Additionally, although both the GAVIN and the KROGAN experiments were performed on a genomic scale, only 1,102 of the overall 2,830 distinct bait proteins in the LARGE-SCALE data set were used as baits in both experiments. Therefore, the combination of the two experimental data sets resulted in relatively few repeated purifications.

In order to demonstrate the ability of ISA in utilizing repeated purifications to infer physical protein contacts with high confidence, we focused on two especially challenging purification data sets with high biological relevance that contain repeated experiments. The first analysis concentrates on inferring stable physical contacts involving *molecular chaperones* from the LARGE-SCALE data, while the second analysis aims at identifying specific interactions concerning *protein kinases* and *phosphatases* from a recently published purification data set.

Detecting stable interaction partners of molecular chaperones. As shown in the previous section, repeated purifications are not available for most bait proteins in the LARGE-SCALE data. Fortunately, however, a set of 63 known molecular chaperones in yeast were screened intentionally multiple times by the Krogan group to provide experimental data for a later study.[58] Chaperones are a broad class of heat shock proteins and protein-remodeling factors that mediate non-covalent protein folding, assembly of macro-molecular structures, protein transport, and degradation of misfolded proteins, thereby maintaining protein homeostasis.[59] It has recently become clear that membrane-associated chaperones are important factors of inter-cellular signaling and may trigger both innate and adaptive immune responses.[60] Misregulation of mutations within host chaperone pathways have been associated with cancer as well as with a number of human cardiovascular and neurodegenerative diseases, as well as with certain cancers.[61] In addition, chaperone expression is correlated with viral infections, either as host response to cellular stress or as a result of viral modulation.[62] Indeed, viruses regularly encode chaperones or subvert cellular chaperone pathways in order to to assist folding of viral proteins[63]: for example, heat shock proteins hsp70 and hsp40 have been identified as essential host factors for HCV replication.[64] Interestingly, drugs that target these chaperones have been shown to exhibit broad



(a) Protein frequency distributions

(b) CHAPERONE reference set

Figure 18.4: *Frequency distribution of chaperone-proteins and assessment of physical chaperone contacts.* 18.4a: Box plots showing the frequency distribution of chaperone-proteins and non-chaperone proteins in the LARGE-SCALE data set. Frequencies on the abscissa are scaled logarithmically. Boxes represent 50% of the data of a given distribution, while bold vertical lines denote the median. 18.4b: Assessment of inferred and experimentally obtained physical contacts involving molecular chaperones against the CHAPERONE reference set of experimentally confirmed binary chaperone interactions. The BGS reference set does not contain a sufficient number of chaperone interactions to allow validation.

[58] Gong et al. (2009)

[59] Sullivan and Pipas (2001), Wang and Li (2012)

[60] Beachy et al. (2007), Calderwood et al. (2007b), Pockley et al. (2008)

[61] Muchowski and Wacker (2005), Ni and Lee (2007)

efficacy against several RNA viruses and appeared to pose a sufficiently high genetic barrier as to not elicit selection of drug-resistant viral variants.[65] Consequently, chaperones may act as a prognostic biomarkers for viral infections such as HBV[66] and both cellular and viral chaperones are currently investigated as drug and vaccine targets, respectively, against infections by a broad range of viruses such as HCV, HSV-1, HBV, HIV, EBV, and Influenza.[67]

Due to their high prevalence and ability to assist in folding of diverse classes of proteins, molecular chaperones are expected to regularly appear in purifications with their substrate complexes. Indeed, the high abundance of this functional class of proteins results in their co-purification as preys at a median rate five times higher than non-chaperone proteins. As a consequence of selective screening in the LARGE-SCALE purifications, chaperones were used as baits twice as often as non-chaperone proteins, resulting in a strong bias of repeatedly purified proteins in the LARGE-SCALE experimental data towards molecular chaperones. (Compare the median number of bait and prey occurrences for chaperone and non-chaperone proteins in Figure 18.4a).

Due to their high abundance, for some chaperones, numerous, highly transient interactions with substrates obscure more permanent interactions with co-chaperones and other regulatory proteins. This makes detection of stable interactions a difficult task for any scoring method. In fact, in order to succeed in this task, a scoring method must take full advantage of repeated purifications with chaperone bait proteins contained in the experimental data. As such, the overrepresentation of molecular chaperones in the data provides an ideal opportunity to examine the ability of scoring methods to use repeated observations in the LARGE-SCALE purification data. To this end, we assess the performance of scoring methods in identifying stable physical contacts involving molecular chaperones from two different perspectives. First, we assess how the scoring methods perform in recovering direct physical contacts involving molecular chaperones that are confirmed by binary experimental assays. Second, we present a high-confidence network of inferred physical contacts involving molecular chaperones and investigate how well stable contacts between chaperones and their cofactors are recovered by the ISA method.

We investigated the performance of the five scoring schemes in recovering experimentally validated physical contacts involving chaperones by comparing the inferred contacts of the scoring methods to the CHAPERONE reference set (see Materials and methods section for a description of the reference data). As can be seen in Figure 18.4b, the ISA method excels in this validation and recovers 80% more physical contacts from the reference set than the SA approach at the previously determined high-confidence rank cutoff of 3,000 inferred physical contacts. ISA is the only approach that demonstrates a performance higher than the best-performing binary experimental assay Y2H at the same cutoff.

[62] Lewthwaite et al. (1998), Neckers and Tatu (2008)

[63] Kuciak et al. (2008), Spence and Pipas (1994), Thomas and Gorelick (2008), Xiao et al. (2010)

[64] Gonzalez et al. (2009)

[65] Geller et al. (2007), Ju et al. (2011)

[66] Lim et al. (2005), Zhu et al. (2004)

[67] Xiao et al. (2010)

Importantly, ISA does not simply promote interaction partners that have been screened repeatedly. Such an undifferentiated strategy would lead to many falsely inferred physical contacts since the highly abundant chaperones are involved in many interactions, of which only very few are likely to be reliable physical contacts. On the contrary, only 79 of the top 3,000 physical contacts inferred by the ISA method involve a molecular chaperone, much fewer than the upper limit of 7,315 spoke observations that pertain to chaperones and are present in the LARGE-SCALE experimental data. Of these 79 physical contacts, more than 20 are validated by the reference set as shown in Figure 18.4b. This indicates that the ability of the ISA method to make use of repeated observations in the data results in very selective promotion of true physical contacts that cannot be discovered by established scoring methods such as SA.

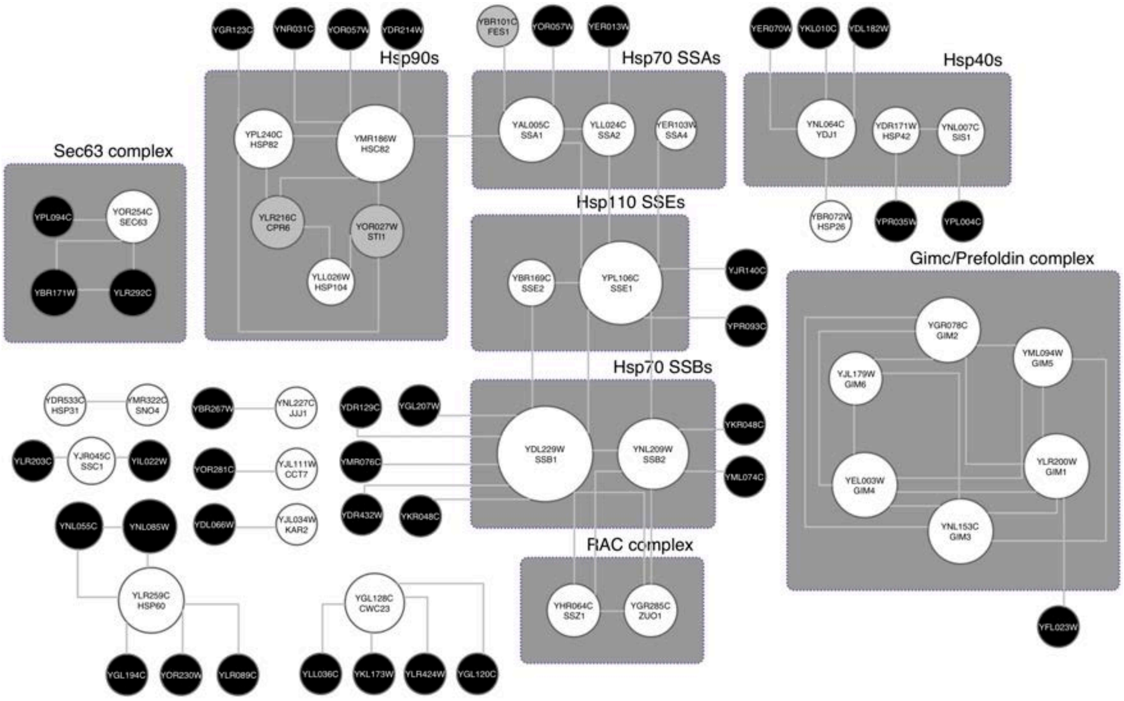


Figure 18.5: *Physical contacts involving molecular chaperones ofyeast.* All physical contacts involving molecular yeast chaperones extracted from the overall top-3,000 physical contacts as inferred by the ISA score. Nodes and edges denote proteins present in the LARGE-SCALE data set and their inferred physical contacts, respectively. The size of a node corresponds to its degree, that is, the number of physical contacts it is involved with. Chaperones are colored in white, while proteins with known chaperone-related function, such as co-chaperones, are colored in grey. Black nodes denote putative substrates of chaperones. Proteins belonging to known families or assemblies are grouped in grey rectangles.

Analysis of inferred chaperone interactions. To obtain a more detailed view on the relationships between molecular chaperones and their cofactors, we generated an interaction network induced by the top 3,000 physical contacts as inferred by the ISA method (see Figure 18.1a). From this network, we extracted all physical contacts that involve at least one molecular chaperone. The resulting interaction network is displayed in Figure 18.5. It contains 79 inferred physical contacts involving 31 of the 63 known yeast chaperones as well as their cofactors and putative substrates.

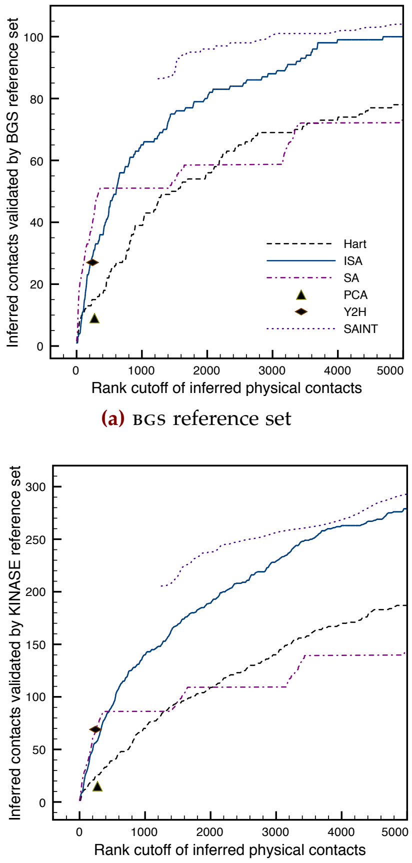
As can be seen in Figure 18.5, physical contacts inferred by the ISA method form a sparse network. Additionally, known protein assemblies, such as the RAC or Sec63 complexes, are connected by patterns of physical contacts and several fine-grained biological relationships between chaperone families are correctly recovered.

The Gimc complex, for instance, a hetero-oligomeric hexamer stabilizing non-native proteins, consists of a dimeric core consisting of α-subunits Gim2 and Gim5. The α-subunits form physical contacts with each other as well as with two of four possible β-subunits (Gim1,3,4 and 6) each.[68] While the ISA score correctly identifies the physical contact between the α-subunits by assigning it the top rank among all inferred physical contacts of that complex, it also infers contacts between the α-subunits and each of the four possible β-subunits at a slightly lower confidence.

Additionally, more intricate relationships, such as patterns of interactions between different families of chaperones, are correctly formed by the inferred physical contacts. Hsp110 homologs Sse1 and Sse2, for instance, form mutually exclusive, hetero-dimeric complexes with Hsp70 families SSA and SSB of the form SSA · SSE and SSB · SSE.[69] This relationship is correctly recovered by inferred physical contacts as depicted in Figure 18.5, where SSA · SSE and SSB · SSE interactions are partitioned and, correctly, no physical contacts between SSA and SSB chaperones are inferred.

Importantly, the presented network of chaperone interactions is unique among scoring methods. It is inherently difficult to infer physical contacts involving chaperones from purification data. This is due to the high abundance of this functional class of proteins that leads to a low signal-to-noise ratio for physical contacts that involve chaperones. As a result, established scoring schemes like the SA method tend to uniformly rank down chaperone interactions. Indeed, the SA method is unable to recover known biological relationships involving chaperones as shown in Figure 18.5 even among its top-10,000 inferred physical contacts.

Identifying specific interactions of protein kinases and phosphatases. The network of kinase and phosphatase interactions is an important component of cellular regulation and messaging. Therefore, these enzymes are highly relevant for understanding a wealth of cellular processes that are influenced by kinase signaling. Protein kinases are often targeted by infectious agents such as viruses due to the central role of these proteins within the cellular signaling network. Indeed, tumor viruses were instrumental for detection and characterization of human kinases and highlight the interconnection between cancer in viral infections in terms if subversion of cellular control pathways. More recently and on a genome-wide scale, genetic and phenotypic screens for viral factors targeting cellular kinases have been undertaken for HCV, Influenza A, and HIV.[70] While experimentally validated kinase and phosphatase interactions have been available only sparsely in public databases, a specialized large-scale purification screen of yeast kinases and phosphatases has recently been published by Breitkreutz *et al.*[71] Protein kinases are a challenging target for scoring schemes since kinases have a propensity towards binding to a large number of other proteins and it is therefore difficult to separate specific from



(a) BGS reference set

(b) KINASE reference set

Figure 18.6: *Assessment of kinase physical contacts.* Assessment of inferred and experimentally obtained physical contacts involving protein kinases and phosphatases against the BGS and KINASE reference set of experimentally confirmed binary kinase and phosphatase interactions. Note that the SAINT scoring scheme assigns identical score values for its top-$1,262$ inferred interactions. Therefore, SAINT performance curve seems to start at a later point than the curves of other methods.

[68] Leroux (1999)

[69] Shaner (2005), Yam (2005)

[70] Flores et al. (1999), Karlas et al. (2010), König et al. (2010), Suratanee et al. (2010), Zhou et al. (2008a)

[71] Breitkreutz et al. (2010)

non-specific interactions.[72] This is illustrated by the average purification size of the BREITKREUTZ purification data: it is more than twice as high as the corresponding sizes in the LARGE-SCALE data (compare Tables 18.1 and 18.2). One of the main contributions of the Breitkreutz *et al.*study thus is the introduction of SAINT, a computational method that identifies non-specific kinase interactors. SAINT and closely related methods such as COMPASS[73] primarily rely on peptide (spectral) counts, which constitute additional types of experimental data that can be interpreted as a semi-quantitative measure of protein abundance. To further increase the coverage and experimental confidence of their method, Breitkreutz *et al.* opted to perform their screen with three different tag systems, yielding multiple overlapping purifications with the same bait proteins. It is due to this intentional application of repeated purifications that we found the BREITKREUTZ data set especially suited for our ISA method.

We assessed the performance of the scoring methods SAINT, HART, ISA, and SA in inferring experimentally known kinase interactions from the BGS and KINASE reference sets (see Materials and methods section for a description of the reference data). Note that, due to the involved computations or unavailable implementations of the PE and IDBOS methods, these scores could not be evaluated here. However, it has been reported elsewhere that PE was not able to distinguish between true and false interactions in this setting.[74] As displayed in Figure 18.6, the ISA score outperforms other general-purpose purification scoring schemes on both reference sets by a large margin. Only the highly specialized SAINT scoring scheme can identify slightly more physical contacts in the data. Importantly, however, the peptide counts employed as an integral part of SAINT require additional processing during the experimental setup. Such counts are neither available for the LARGE-SCALE purifications nor for most other publicly available purification data. In contrast, the ISA scoring scheme is generally applicable to all raw purification data without the need for additional peptide count data.

Analysis of host-pathogen physical contacts. In order to demonstrate the general applicability of our scoring method to model organisms other than yeast, we utilize recently published large-scale purification data of a host-pathogen system.[75] Within the experimental setting that generated these data, genes corresponding to all HIV proteins were expressed in two different human cell lines and assayed using proteins purification methodology. Depending on the cell line thus assayed, two purification data sets here denoted as HEK and JURKAT were produced (see Materials and methods). Interactions within these purifications were scored by the authors of the original publication using MIST (mass spectrometry interaction statistics), a novel scoring scheme especially optimized for detecting host-pathogen interactions that employs both spectral counts and

[72] Breitkreutz et al. (2010)

[73] Sowa et al. (2009)

[74] Breitkreutz et al. (2010)

[75] Jäger et al. (2012)

computationally costly permutation statistics. We compared the MIST predictions on the two separate purification data sets HEK and JURKAT as well as SA, ISA, and HART scores of the combined HEK and JURKAT sets against validated physical contacts from VirusMINT[76] (see Figure 18.7). Similar to the previous validations, ISA performs significantly better than ISA and HART scores at recovering validated physical contacts. Indeed, ISA performs similar or better than MIST scores without relying on spectral count data or computationally costly permutation statistics.

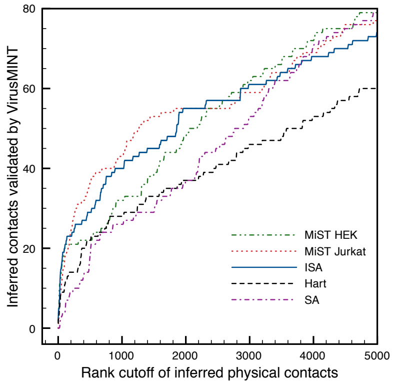


Figure 18.7: *Assessment of host-pathogen physical contacts.* Assessment of inferred physical contacts involving human and HIV proteins against the VIRUSMINT reference set of experimentally confirmed host-pathogen protein interactions.

Discussion. This work is first to investigate whether direct physical protein contacts can be extracted from raw purification data contained in a combined set of large-scale protein complex purifications. We analyzed four established scoring schemes and one new approach and assessed their ability to reliably detect physical contacts within assayed complexes. Top ranking inferred physical contacts from all five methods were benchmarked against reference sets based on binary experimental protein interactions, three-dimensional structures of interacting proteins, manually curated protein complexes, and genetic interaction profiles. Inclusion of these four complementary sources of validated physical contacts allowed us to investigate aspects of the scoring schemes that were not examined before.

The results of our evaluation showed that raw purification data, if scored correctly, can indeed be exploited to infer physical contacts within protein complexes. While established methods devised for inferring indirect protein interactions from purification data perform well in the task of identifying co-complexed protein pairs in reference protein complexes (data not shown), the performance of most of these methods in detecting direct physical protein contacts is considerably diminished. Only the two socio-affinity based methods SA and ISA consistently showed the best performance among all evaluated methods in inferring physical contacts.

We attribute this difference of performance between the methods to two facts. First, both SA and ISA share the concept of a simple, but elegant, null model to identify strongly associated proteins. Second, both methods have a high propensity towards using only direct, or spoke, associations between proteins as evidence for physical contacts, that is, they concentrate on interaction evidence between a bait protein and the prey proteins it purifies. This indicates that, while indirect, or matrix, observations are useful for detecting co-complexing protein pairs, direct observations are more informative for identifying physical contacts.

Besides intentionally concentrating on direct observations, the main innovation of the ISA method is an improved null model that allows for integration of repeated purifications using the same set of bait proteins in a statistically meaningful fashion. While improving the predictive power of our method in the presence of any repeated observations, this ability is especially relevant for inferring

[76] Chatr-aryamontri et al. (2009)

stable contacts involving highly abundant proteins, such as molecular chaperones or protein kinases, whose specific interactions are especially difficult to infer in the absence of repetitions. Intuitively, the approach taken by our ISA score is similar to strategies currently discussed for interactome mapping projects where repeated experiments can allow for an increased sensitivity and specificity of the resulting screens.[77]

Our analysis of the LARGE-SCALE purification data found that most proteins have been screened only once in experiments as baits. Few proteins, such as the functional class of molecular chaperones, have been used as baits multiple times. The assessment of chaperone interactions showed that the ISA method improves upon other scoring methods when repeated observations are available. Our method recovered a range of biologically significant relationships between chaperones and their cofactors that could not be detected by the second best performing SA method. This highlights the importance of correctly using repeated observations to gain statistical confidence in the inferred physical contacts.

The use of repeated observations for gaining statistical confidence in physical contacts is not limited to the general-purpose large-scale purification experiments, but can also be applied to specialized data sets aiming at specific biological targets as demonstrated by our analysis of a recent purification study focussing on protein kinases. Due to the intentional integration of repeated purifications in that study, ISA was available to infer significantly more physical contacts from the raw purifications than any other general-purpose scoring method.

Importantly, the mechanism of improved performance of ISA based on repeated observations is not depending on single purification data sets that feature repeated purifications. Instead, ISA is able to exploit repetitions across several distinct data sets. Therefore, we expect that physical contacts inferred by our method will further improve in quality compared to established scoring schemes once additional large-scale purification data sets become available. Considering the experimental replicates within the Breitkreutz *et al.*[78] data as well as current correspondences by experimentalists about integrating multiple orthogonal assays to increase confidence in the results,[79] there seem to be clear indications that repeated purifications within one data set will become more widespread in future.

An adequate comparison of interactions measured by high-throughput binary experimental approaches to physical contacts deduced from purification experiments has not been possible before. This is a result of the fact that binary experimental approaches are more straightforward to interpret: each physical interaction is measured directly, while in purification data only a small subset of all interactions are likely to be physical contacts. As a consequence, the whole set of interactions possible in purification data has previously been interpreted as putative physical contacts.

[77] Schwartz et al. (2009)

[78] Breitkreutz et al. (2010)

[79] Braun et al. (2010), Chen et al. (2010)

Since this resulted in a low enrichment of true physical contacts in the purification data, purification-based methodologies were determined of being less useful for measuring true physical contacts than high-throughput binary experimental techniques such as yeast two-hybrid.[80]

However, a closer analysis of the performance of all five scoring schemes and a comparative assessment of the inferred physical contacts with the results of binary experimental assays such as Y2H or PCA reveals several novel findings. First, the performance of all scoring schemes in recovering physical contacts from the reference sets levels off at about 3,000 top-ranking physical contacts, indicating that this number constitutes a limit on the number of interactions that can be reliably scored given current experimental data. Second, our results surprisingly suggest that, once correctly scored and ranked, physical protein contacts derived from complex purification experiments are qualitatively comparable to interactions measured by state-of-the-art Y2H and PCA techniques. Additionally, the purification scoring schemes perform significantly better than the Y2H and PCA data sets in connecting manually curated protein complexes. This suggests that physical contacts derived from purification data might be more relevant for interpreting protein complexes than interactions measured by these binary experimental techniques.

Besides offering new opportunities for the interpretation of purification data and the understanding of protein complexes, there are additional application scenarios of our scoring method. Since the ISA method is optimized to make best use of repeated observations in the data, experimental research groups can repeatedly perform small-scale purifications for proteins of interest, possibly involving perturbation experiments[81] or novel purification methodologies applicable to human cells.[82] These purifications can then be added to the LARGE-SCALE experimental data. Subsequently, the ISA method can be re-applied on this newly enlarged data set. The additional information contained in the repeated purifications will then allow for reliable inference of physical contacts involving the proteins of interest.

We further propose that physical contacts derived from purification data are applicable to large-scale interactome mapping projects. Several interactome screens are currently underway for model species such as *S. cerevisiae* or *D. melanogaster*. Meta-strategies for cost-effective mapping have been developed involving schemes for pooling, prioritization, and repetition of experiments to increase overall coverage and accuracy of the combined screens.[83] We suggest that physical contacts derived from purification data may be used complementary to binary interaction data sets by prioritizing high-confidence physical contacts for experimental validation by Y2H or PCA techniques. Such a prioritization strategy seems especially valuable as part of high-throughput large-scale interactome screening projects such as recently proposed by Schwartz *et al.*[84]

[80] Yu et al. (2008)

[81] Hernández et al. (2006)

[82] Malovannaya et al. (2010)

[83] Lappe and Holm (2004), Schwartz et al. (2009)

[84] Schwartz et al. (2009)

Addendum. Since publication of the preceding manuscript in 2011, several additional purification data sets and computational filtering schemes have been introduced.[85] Although the ISA score continues to offer specific and unique advantages compared to other affinity-based approaches such as SA and HART scores while also delivering on par performance with more involved spectral counting approaches such as MIST and SAINT, affinity-based approaches in general and ISA in particular have not seen increased adoption by the proteomics community.

This is likely due to three reasons: first, ISA has been developed for (and excels in) deriving physical protein contacts from protein purification data. While we could demonstrate that the quality of the interactions inferred by ISA are of similar or better quality than interactions measured by binary experimental methods such as Y2H and PCA, binary interactions are not a current focus of attention of biologists undertaking protein complex purification experiments. Instead, these experimentalists are more interested in recovering complete functional complexes, as demonstrated by the themes of several recent publications.[86]

Second, the application and validation of scoring schemes is not well standardized in the community. Rather, each publication of a novel purification data set is usually paralleled by a new scoring scheme and the choice of validation sets and the depth of validation of these novel schemes vary widely. While this development is partly justified by the high variability of data sets and biological targets under investigation that result in experimental setups of disparate sizes, coverages, target organisms, purification methods, and quantification schemes, it also results in low re-use of scoring schemes in general.

Third, and most importantly, all 'first line' affinity-based approaches (i.e., SA and HART) have been proposed by research groups that undertook genome-wide purification screens that included a large fraction of reciprocal experiments (i.e., the role of baits and preys was reversed in technical replicates). Results have indicated that affinity-based approaches excel under these conditions and produce results superior to spectral counting and clustering approaches if high (i.e., approaching genome-wide) coverage and reciprocal interactions are given; conversely, the performance of affinity-based methods suffers dramatically if these conditions are not met.[87] While our ISA score is able to incorporate incomplete purification sets and score repeated experiments in a statistically sound fashion, it is nevertheless bound by the same limitations as other socio-affinity approaches.

Generation of genome-wide purification of data is not a current focus of experimentalists who instead concentrate on small or mid-size subnetworks of particular interest for reasons of both better interpretability and increased cost-effectiveness. As a consequence, only one such genome-wide set is currently available for human[88] and the completion of the human protein complex interactome is

[85] Pardo and Choudhary (2012)

[86] Pardo and Choudhary (2012)

[87] Nesvizhskii (2012), Sowa et al. (2009)

[88] Ewing et al. (2007)

expected to rely on integration of these incomplete data sets rather than on concerted, genome-wide efforts.[89] Similarly, available host-pathogen interaction networks also tend to be of limited size and the availability of the high-coverage human-viral (Jäger et al., 2012) data set that was utilized in our study is the exception rather than the norm.

As a result of the lack of genome-wide data sets, socio-affinity based scoring schemes are currently less applicable on current data and alternative methods that are specialized towards incomplete purification screens are required. One way to control false positive rates and achieve sufficient specificity given these incomplete data is the use of additional data sources such as spectral counts and negative control experiments. Since publication of our manuscript in 2011, the necessary primary data analysis methods (i.e., the spectral counting) have been further refined and can now be generated in an automatized fashion. Consequently, spectral-count approaches such as ComPASS, SAINT and MIST that support this line of analysis and perform well with small or mid-size data sets currently dominate the landscape of protein complex purification analyses.

However, there is justified hope for a renewed focus on genome-wide approaches in future studies. As recently pointed out by Ideker and Krogan (2012), the high availability of genotypic data (e.g., originating from second generation sequencing approaches) has highlighted both the lack of correlated phenotypic data and the importance of differential approaches (i.e., comparison of samples given different biological conditions such as disease vs. normal or drug-treated vs. untreated). Almost all protein interaction networks examined to date, however, have been examined under static conditions that do not well reflect the dynamics of biological systems and thus limit the depth of scientific inquiry.

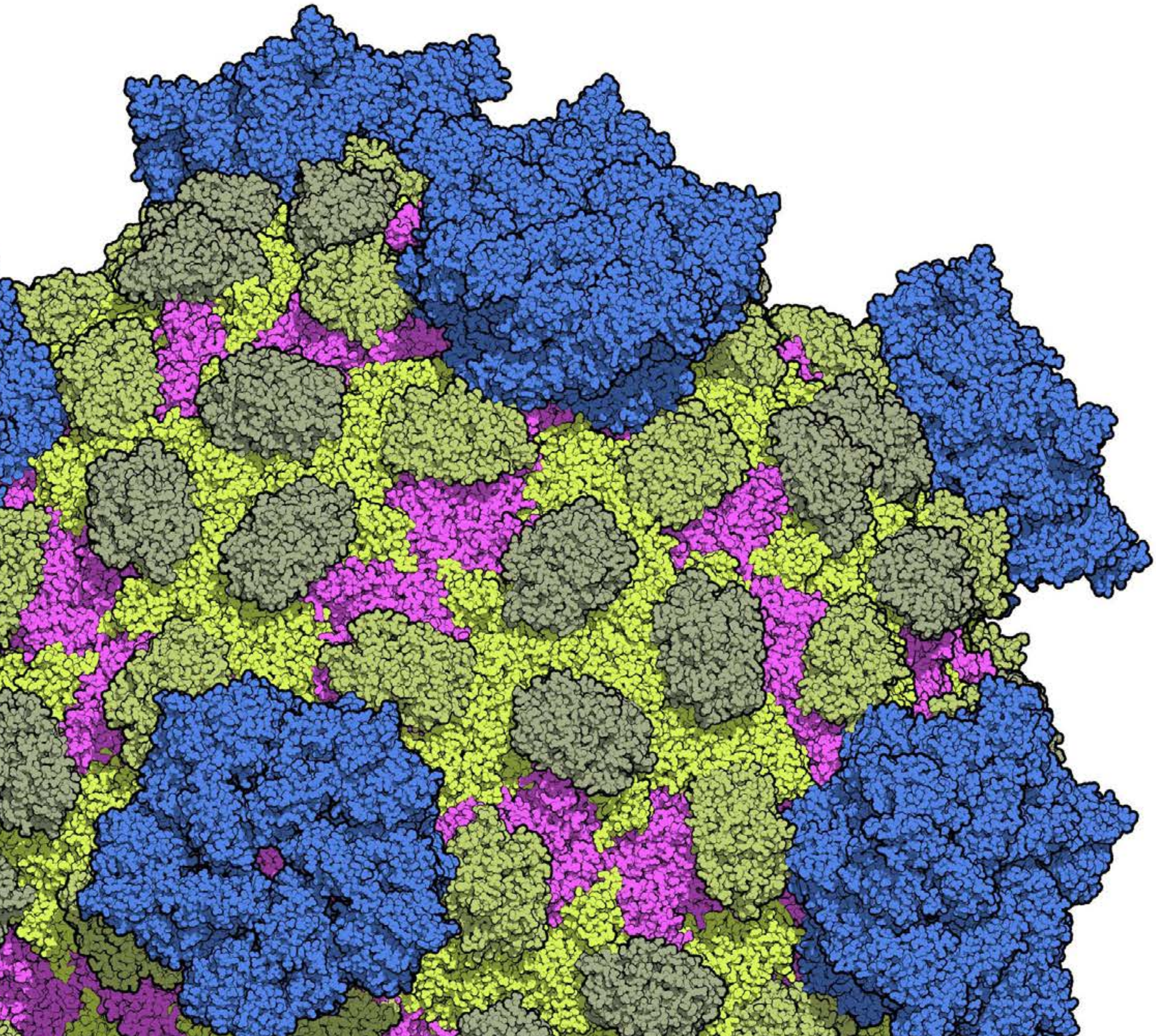
In order to gain deeper insights into biological properties of protein interaction networks, Ideker and Krogan argue that differential approaches for measuring physical protein interactions have to be undertaken. Besides aiming to investigate network changes given disease states such as cancerous transformation, host-pathogen interaction networks are of particular interest in this regard, as characterizing the viral subversion of host signalling pathways may be a particularly insightful application of differential methods.[90] Importantly, these approaches again necessity genome-wide screens in order to curb the rate of false-negatives that otherwise may confound differential approaches. It is due to these reasons that both socio-affinity approaches and the analysis of genome-wide host-pathogen interaction networks may well see a revival in the coming years.

[89] Nesvizhskii (2012)

[90] Ideker and Krogan (2012)

IV

Viral quasispecies



The last chapter of is concerned with real-world viral disease and closes with the description of a major antagonist of clinical treatment: the viral quasispecies. The existence of viral quasispecies populations, the recognition of their concerted evolution, and their role in antiviral therapy are the main points of interest in this chapter. Sections 19 to 22 introduce the concept of viral quasispecies, discuss the most important parameter governing its evolution, and exemplify these aspects in the HCV quasispecies. Subsequently, sections 23 to 23 review approaches to treating viral quasispecies by manipulating aspects of its evolution using antiviral drugs. The effects that these manipulations have on the genetic composition of the quasispecies are then quantified using sensitive deep sequencing technologies. Finally, all these aspects are combined in Section 24, where the first experimental study using in vivo deep sequencing data is introduced that investigates quasispecies behavior under monotherapy with ribavirin, the most important broad-spectrum antiviral and essential component of anti-HCV therapy.

19 Viral quasispecies

Among the epidemiologically *most relevant RNA viruses are highly divergent viral species such as HIV-1 and HCV, and Influenza-A, as well causative agents of the emerging viral diseases Ebola hemorrhagic fever, Dengue fever, and West Nile fever.*[1] *In particular, three viral species that cause chronic infections are associated with the highest disease burdens: HIV, HCV, and HBV. By conservative estimates, infections with these three viruses are afflicting 40 million, 400 million, and 200 million individuals worldwide, respectively, and may be collectively responsible for more than 3 million deaths per year.*[2]

Although these pathogens belong to different groups of the Baltimore classification and consequently have very different life cycles, they have at least three characteristics in common: these viruses employ RNA genomes for replication, display high rates of mutation of about 10^{-5} substitutions/site/copy, and chronically persist at very high abundances in patients (HIV: 10^{3}-10^{6} copies/mL, HBV: 10^{5}-10^{9} copies/mL, HCV: 10^{4}-10^{7} copies/mL, all in untreated hosts).[3] As a consequence of this high abundance and divergence, these viruses are believed to exist as a distribution of genetically distinct but closely related genotypes that are present at varying frequencies within the host.

This genotype distribution, also denoted as mutant distribution, mutant spectrum, or *viral quasispecies*, is induced by the lack of proofreading activity in viral RNA-dependent RNA polymerase (RdRp) and is results in high rates of mutations, or *variants*[4] within the viral progeny.

A short history of quasispecies theory

Phenotypic traits of viruses are a product of evolution, i.e., the consequence of genetic variation, reproduction, and subsequent selection of neutral or beneficial traits by the environment of the pathogen. In particular, traits that confer advantages with respect to the viral micro-environment, for instance with respect to the host immune response or tissue tropism, are rapidly selected for in viruses due to their short generation times and compact genomes. This micro-evolution is of particular clinical importance due to

[1] Geisbert and Jahrling (2004)

[2] Margeridon-Thermet and Shafer (2010)

[3] Margeridon-Thermet and Shafer (2010) An IU corresponds to about 1-5 RNA copies depending on the method of quantification, cf. Pawlotsky (2002).

[4] Within the following sections, we will denote as viral *variants* specific genome positions that differentiate viral genotypes from each other. These variants are commonly defined with respect to an external reference genome, the consensus sequence, or the *master sequence* of the quasispecies, i.e., the genome that is most abundant (has the highest frequency) in the genotype distribution. While the dominant genome is distinct from the consensus sequence, i.e., the average genome of the quasispecies, both of these sequences are often identical (are always identical if the frequency of the master sequence is above 50%). Since each genotype within a quasispecies is fully defined by the total of its variations from reference sequence, genotypes are often also simply denoted as *variants*. However, the latter choice of terminology predominantly applies to virology while in genomics and bioinformatics the term variant denotes the smallest unit of difference, in principle, such as single nucleotide variants (SNV, also termed SNP if the variant is common, e.g., has a frequency of $> 1\%$ within the population in case of the human genome) and indels, or structural variants such as copy number variations, duplications, recombinations, reassortments, and large insertions.

the selection of traits of therapeutic relevance, such as resistance to antiviral drugs, both within a single patient and across a whole multi-host population of pathogens.[5]

As a consequence of the close genetic relatedness of the quasispecies population, the whole quasispecies rather than the single virus is often considered to be the unit of natural selection [6] and traditional population genetics may not be well suited to model the behavior of RNA viruses.[7] Instead, the evolutionary dynamics of these pathogens is commonly described by what has been termed *quasispecies theory*.

Quasispecies theory has first been proposed by Manfred Eigen in 1971 as a purely mathematically concept describing an infinite number of replicons that regularly produce erroneous copies of a template molecule. Reportedly, the original research project that later culminated in this model was suggested to Eigen by Nobel laureate Francis Crick over breakfast.[8]

The model built on experimental work concerning the phage Qβ, an organism that displayed Darwinian behavior (i.e., mutation and selection) in serial transfer experiments.[9] By utilizing information theory as theoretical foundation, Eigen proposed the quasispecies theory as a model of the origin of life within a hypothetical, primitive RNA-world of self-organizing macromolecules (see Figure 19.1 for an introduction to the basic variables of this model).[10] Only later were these concepts extended to RNA viruses.[11]

$$\begin{aligned} \frac{dx_i}{dt} &= (A_i Q_i - D_i)x_i \qquad (19.1)\\ &+ \sum_{k=1,k\neq i}^{n} W_{ik}x_k - \Phi_i \\ v < v_{max} &= \frac{\ln \sigma_0}{1-\overline{q}} = \frac{\ln \sigma_0}{\overline{p}} \qquad (19.2)\\ p < p_{max} &= \frac{\ln \sigma_0}{v} \qquad (19.3) \end{aligned}$$

Figure 19.1: *Principle equations of quasispecies theory according to* Domingo et al. (2012), Eigen and Schuster (1977). The first equation models the concentration of genotypes i and k as functions of time $x_i(t)$ and $x_k(t)$. In particular, this formula describes the generation of a mutant i from a template population k due to erroneous replication. A_i and D_i are the rate parameters for the replication and degradation of i, respectively. Q_i is the fraction of faithful replicates that are produced from i during replication. W_{ik} is the rate for erroneous generation of i based on the template genotype k. Φ_i is a flux parameter that models the physical movement of molecules within the quasispecies environment. Equations 19.2 and 19.3 describe the error threshold relationship where v_{max} is the maximum genetic complexity maintainable through replication, σ_0 is the fitness advantage of the master sequence relative to minority genotypes, $\overline{q}$ is the average copying fidelity, and $1 - \overline{q} = \overline{p}$ represents the average error rate during replication. Based on quasispecies theory, replicative accuracy can only be maintained for sequences shorter than a maximal sequence length v_{max} for constant replication accuracy, or given a maximal error rate p_{max} provided a constant sequence length.

Application of quasispecies theory to viral populations. Viruses play a two-fold role in quasispecies theory: first, as direct subjects of enquiry that may divulge biologically and medically relevant mechanisms of viral replication and adaption; second, as explanatory devices for experimentally testing theories of population genetics.

Within the original theory of Eigen and Schuster, quasispecies describe replicon populations of infinite size within a *mutation-selection equilibrium*, also denoted as *mutation-selection balance*. At these equilibria, frequencies of replicon genotypes are balanced between production of new genotypes by mutation and pruning of unfit genotypes by selection; in particular, quasispecies theory predicts the existence of a well-adapted master sequence with high fitness and a distribution of mutated genotypes that display lower fitness, densely fill the genotypic space around the master sequence, and are constantly replenished by mutated progeny of the master sequence and of each other.

As a result of this *mutational coupling*, the quasispecies distribution as a whole rather than the individual genotype is believed to be the unit of selection.[12]

Theoretical considerations have demonstrated that quasispecies theory is generally compatible with the more traditional concept of selection-mutation (or Wright-Fisher) equilibrium known from classical population genetics, indicating the broad applicability of the quasispecies concept.[13] By utilizing models of population ge-

[5] Cannon et al. (2008), Mascolini et al. (2008), Wargo et al. (2007)

[6] Bull et al. (2005), Codoñer et al. (2006), Eigen and Schuster (1977)

[7] Holland et al. (1982)

[8] Ojosnegros et al. (2011)

[9] Mills et al. (1967)

[10] Eigen (1971), Eigen and Schuster (1977)

[11] Domingo (2006), Domingo and Wain-Hobson (2009), Holland (2006), Lauring and Andino (2010)

netics, quasispecies theory has been extended to better encompass features of experimental RNA virus populations that operate as finite populations and within dynamic fitness landscapes that prevent the population from achieving a permanent mutation-selection equilibrium.[14] These extensions granted the quasispecies model additional attributes that allowed it to explain medically relevant behaviors of viral populations, such as the emergence of resistance variants in drug-treated viral populations.[15]

Current state of viral quasispecies theory. Today, the term quasispecies commonly includes the aforementioned extensions and refers to

> *(...) distributions of non-identical but related genomes subjected to a continuous process of genetic variation, competition, and selection, and which act as a unit of selection.*[16]

More mechanistically, a viral quasispecies can be defined as populations of genetically diverse but related genotypes that are subjected to both internal (i.e., competition for resources between genotypes) and external (i.e., competition between genotypes and their environment) pressures that influence variation, replicative fitness, and selection within the genotype population.[17]

The experimental validity of these definitions is supported by several *in vitro* and *in vivo* experiments that relate antigenic heterogeneity to genetic variation (reviewed in Domingo et al. 2003), and identified mechanisms in viral quasispecies that are reminiscent of classical population genetics such as the competitive exclusion principle,[18] replicative fitness, and an evolutionary arms race[19] (reviewed in Novella (2003), see later sections for a more detailed description of these mechanisms).

Basic tenets of viral quasispecies.

Replication fidelity and mutation rates. Organisms existing in a static environment that is not dominated by changing selective forces are expected to evolve towards a global fitness optimum and low mutation rates in order to maximize the number of viable progeny[20] Conversely, genotypes within viral quasispecies exists in a constantly changing environment resulting comprising drug and immune pressure as well as competing genotypes. In order to uphold replicative fitness, it is assumed that quasispecies adapt to these conditions by modifying the phenotypic characteristics of their constituting genotypes.[21] This is enabled by inducing high rates of randomly mutated viral variants; although most of these variants are unlikely to be viable, a minority may, by chance, display higher replicative fitness than existing variants especially in adverse environmental conditions.

High rates of genetic variation, commonly quantified as the error rate (also termed *mutation rate*)[22] are be believed to significantly

[12] Bull et al. (2005), Codoñer et al. (2006), Eigen and Schuster (1977); further supported by supplemental theories on evolutionary dynamics stating that groups are more likely to be units of evolutionary selection if group fitness exceeds the average of group members (Lauring and Andino, 2010, Mayr, 1997).

[13] Musso (2012), Wilke (2005)

[14] Eigen (2000), Park et al. (2010), Saakian et al. (2009), Saakian and Hu (2006)

[15] Domingo (1989), Domingo and Holland (1992)

[16] Ojosnegros et al. (2011)

[17] Ojosnegros et al. (2011)

[18] *Competetive exclusion principle*: A principle of population genetics that states that, if no niche specialization is possible, less fit genotypes or genotype populations will be supplanted (or excluded from replication) by fitter genotypes that compete for the same resources. Note that, due to the continuous generation of distinct genotypes by mutation and the high abundance of the quasispecies in general, individual genotypes are unlikely to be replaced entirely. Rather, they are suppressed and maintained at very low frequencies.

[19] *Evolutionary arms race and Red Queen dynamics*: genotype populations within viral quasispecies are believed to be involved in a process of continuous maintenance or increase of fitness relative to co-evolving systems such as the host immune system, drug pressure, or viral genotypes competing for the same resources. As a net effect, viral quasispecies are believed to continuously increase their absolute fitness if measured in isolation; however, since the co-evolving systems also increase their respective fitness, the *relative fitness* of each system remains constant. These phenomena are reminiscent of the nuclear arms race of the superpowers during the cold war as well as of statements of the Red Queen in Lewis Carroll's "Through the Looking Glass" stating that "... *it takes all the running you can do to keep in the same place*".

[20] Jiang et al. (2010)

[21] Crill et al. (1999), Weaver et al. (1999)

increase the chances for generating beneficial mutations and may thus be maintained by RNA viruses in order to confer adaptability to rapidly changing environments.[23]

Indeed, RNA viruses display high mutation rates of about one mutation per genome and replication, values that are significantly increased compared to DNA viruses.[24] Due the compactness of viral genomes, however, most random variation is expected do be deleterious and thus decrease fitness.[25] RNA viruses utilize their high abundances and replication rates in order to increase the likelihood that progeny with higher replicative fitness will eventually be generated and evolutionary selected due to their replicative advantage.[26]

Optimal mutation rates. As a result of their importance for viral quasispecies, mutation rates themselves are believed to be a subject of selection, as for example demonstrated by the emergence of mutator and anti-mutator phenotypes in the HIV-1 quasispecies.[27] This evolutionary optimal mutation rate is believed to be specific to the selective and replicative context of the quasispecies and is determined by at least three factors:[28]

(1) Since most random mutations in RNA viruses are deleterious, selective pressure exists for reducing mutation rates in order to uphold replicative fitness; indeed, overly increased error rates may result in accumulation of deleterious mutations and irreversible loss of genetic information, a process also denoted as *genetic meltdown*.[29]

(2) Increased replication fidelity comes at an increased kinetic or energetic cost, thus limiting replication capacity; this fact is especially relevant for RNA viruses that rely on rapid infection cycles in order achieve high abundances before the host immune reaction sets in.[30]

(3) As noted before, elevated mutation rates confer increased adaptability, as has been demonstrated for poliovirus where increased replicative fidelity (and thus lower error rates) lead to a reduction in pathogenicity and tissue tropism of the virus, thus indicating a diminished ability to adapt to new cellular microenvironments.[31]

As a result of these conflicting factors influencing selective advantages of elevated mutation rates, a trade-off between mutation rates is assumed to exist that prevents that ensures sufficient adaptive variability to ensure survival of the viral quasispecies in the host microenvironment while also preventing the accumulation of deleterious mutations.[32]

The molecular basis of this trade-off may be implemented by selectively adapting the fidelity of viral polymerases to error rates just below a critical error threshold; this line of reasoning is supported by experimental evidence that relies on artificially increased mutation rates (reviewed in Anderson et al. 2004, Domingo et al. 2005) as well as by more recent theoretical models.[33] Indeed, ex-

[22] *Mutation rate*: Viral quasispecies are assumed to originate from a small founder population (conceivable only consisting of an individual viral genome) by error-prone replication that induces random mutations. The rate by which these mutations are introduced by the replicative molecule, generally a polymerase, is termed *mutation rate*, or *error rate*. It is often determined by sequencing the viral quasispecies and is quantified as the estimated rate at which an individual genomic site is expected to mutate either at a single replicative event or across a timespan of a year of continuous replication. *Mutation frequency* then is the average rate at which a single genomic site differs from the consensus sequence of the viral quasispecies after selective events of the environment have removed genomes with lethal mutations. Importantly, therefore, mutation rate does not generally equal mutation frequency: the latter takes environmental factors into account while the former does not; on the other hand, mutation frequency is considerably simpler to measure experimentally because it can be determined by the present state of the quasispecies.

[23] de Visser (2002), Domingo and Holland (1997), Sanjuán et al. (2010)

[24] Drake (1999)

[25] Fay et al. (2001), Sanjuán et al. (2004)

[26] Coffey et al. (2011), Coffin (1995), Perales et al. (2011b), Sanjuán et al. (2004)

[27] Burch and Chao (2000), Cases-Gonzalez et al. (2000), Eigen and Schuster (1977), Johnson (1999), Nowak (1992), Schuster and Swetina (1987), Taddei et al. (1997)

[28] Sniegowski et al. (2000)

[29] Anderson et al. (2004), Sanjuán et al. (2004)

[30] Coffin (1995), Dawson (1998)

[31] Pfeiffer and Kirkegaard (2005a), Vignuzzi et al. (2006)

[32] Johnson and Barton (2002), Orr (2000)

[33] Bull et al. (2007), Manrubia et al. (2010), Ochoa (2005), Takeuchi and Hogeweg (2007)

perimental data indicate that mutation rates of RNA viruses can only be increased slightly: 2 to 3-fold increases of mutation rates for non-retroviral RNA viruses and about 13-fold increased rates for retroviruses are sufficient for viral fitness to degrade significantly; in contrast, bacteria are able to cope with more than 1,000-fold higher increases or error rates.[34]

Characteristic mutation rates. This balancing tradeoff between increasing adaptability on the one hand and reducing the number of inviable progeny on the other hand is theorized to lead to specific error rates characteristic for each RNA virus species.[35] Such optimal rates were determined for several special cases using modeling approaches; for instance (Kamp et al., 2003) have undertaken to derive optimal viral error rates required for escaping the host immune answer, arguably one of the major determinants of viral fitness. Similar to viral quasispecies, B-cell receptor sequences are associated with closely related receptors that result from somatic hypermutation of B-cells.[36] Consequently, the virus is believed to evade immune response by mutating antigenic epitopes and avoid eliciting proliferation processes of the corresponding immune receptor.

Within the Kamp et al. model, competition between viral quasispecies and the adaptive immune response features an asymmetric coupling where the immune response is attracted by (i.e., proliferates in the presence of) matching viral epitopes while viral quasispecies populations are selected that are different from the high-profile master sequence and thus provide increased means of immune escape. This process resembles a predator-prey dynamic as demonstrated for HIV.[37] Modeling can be utilized to derive optimal mutation rates of immune receptors and the viral quasispecies, respectively.[38]

Error threshold and error catastrophe. One of the most interesting and controversial implications of this theory is the existence of a phase transition that be triggered at relatively trivial increases in mutation rates and that causes dramatic changes in the genotype distribution of the quasispecies. The critical mutation rate is also denoted as *error threshold* below which

> *"populations equilibrate in a traditional mutation–selection balance and above which the population experiences an* error catastrophe*, that is, the loss of the favored genotype through frequent deleterious mutations."*[39]

The concept of an error catastrophe is one of the main features of quasispecies theory and is of considerable importance for antiviral treatments.[40] In this context, mutagenic drugs may be employed to destabilize the viral quasispecies by elevating mutation rates, a process that may result in the quasispecies undergoing *genetic meltdown*, i.e., the irrecoverable loss of genetic information, and extinction by *lethal mutagenesis*.

[34] Camps et al. (2003), Holland et al. (1990), Pathak and Temin (1990)

[35] Drake et al. (1998)

[36] Harris et al. (1999)

[37] Kamp and Bornholdt (2002)

[38] While a complete description of the Kamp et al. model is omitted here for sake of brevity, we will briefly discuss one of the major results: given that the host immune system adapts to a new viral subpopulation with in a time-span t_i and the virus replicates within a time $1/\sigma_v$, the optimal genomic mutation rate $\mu_v \approx 1/(\sigma_v t_i)$ can be derived by approximation from the model. The ratio between these time scales represents the duration of one generation of virus in units of the response time of the immune system; consequently, the the optimal (i.e., minimal) viral error rate sufficient for immune escape is one mutation per genome within the time the immune system requires for adaption. Intuitively, this result that every viral genome produced within that time frame is different from the viral subpopulation that the immune system adapted. Assuming adaption times of the immune system of 7-14 days, the predicted error rates are well within the range of experimentally measured generation times and mutation rates for several RNA viruses, including HIV.

[39] Bull et al. (2005)

[40] Anderson et al. (2004), Jonsson et al. (2005)

While it is commonly stated that lethal mutagenesis is identical to (or an an instance of) an error catastrophe,[41] newer theoretical results argue differently.[42] However, a detailed discussion of these arguments require prior knowledge of additional fundamental principles such as genotypic spaces and viral fitness and is therefore postponed until later in this chapter.

Challenges of quasispecies theory

While quasispecies theory is conceptually appealing and seems to be well in accordance with experimentally observable attributes of RNA viruses, there are controversies concerning its applicability to several viral systems that are currently proposed to exhibit quasispecies properties. Specifically, critics raise concerns about overusing the quasispecies concept and point out that the original theory was developed by Eigen *at al.* to define primordial RNA replicators and not to describe features of viral pathogens. Quasispecies theory may therefore not generally provide more explanatory power to describe clinical phenomena than existing models of population genetics.[43] In particular, well-understood concepts of mutation, genetic drift, and natural selection as represented in classical population genetics may serve to describe many phenomena of viral populations without making explicit use of quasispecies theory.[44]

Population geneticists further point out that at its core, quasispecies theory diverges from classical population genetics in only two ways: first, in contrast to the species concept in population genetics, quasispecies genotypes are not independent but tightly coupled in genotype-phenotype space; thus, the "*entire population forms a cooperative structure that evolves as a single unit*" upon which selection acts.[45] Second, due to large effective viral population sizes, small genomes, and high mutation rates, the genotypic space around viral genotypes with high fitness is assumed to be completely explored by viral genotypes, hence preventing genetic drift (see later sections of this chapter for a detailed discussion of genotypic spaces and mutational coupling).[46]

Both of these defining features are imparted by quasispecies theorists to real-world viral populations while being contested by population geneticists; in particular, population geneticists raise the point that viral populations are subjected to immune selection that results in many genotypes being inviable; as a consequence, tight genetic coupling between all genotypes (as in a continuous genetic distribution of viral genotypes) is impossible, thus questioning the concept of combined selection as the defining attribute of viral quasispecies[47] If, however, combined selection is indeed not a property of real-world ensembles of viruses, then observed features of RNA-viral populations are purported to be equally well explainable by classical population genetics and random drift, thus possibly making quasispecies theory unnecessary to explain the experimental data.[48]

[41] Crotty et al. (2001), Graci and Cameron (2002), Jonsson et al. (2005)

[42] Bull et al. (2005, 2007), Wilke (2005)

[43] Holmes and Moya (2002)

[44] Jenkins et al. (2001), Morse (1992), Moya et al. (2000)

[45] Holmes and Moya (2002)

[46] Eigen (1987)

[47] Holmes and Moya (2002)

[48] Holmes and Moya (2002)

20 Quasispecies dynamics

Genotypes of viral quasispecies *are constantly undergoing negative and positive selection based on the characteristics of their environment. Positive selection rewards phenotypic traits that confer replicative advantages and typically results in increased abundances of the selected genotype. However, positive selection does not automatically result in the selected genotypes becoming the dominant subpopulation (or master sequence) of the viral quasispecies since multiple genotypes with distinct genetic makeup may share the trait that is selected for. Instead, these genotypes compete based on their relative fitness and the resulting frequency distribution of the viral quasispecies can be interpreted as a steady-state of these competing fluxes and environmental constraints. This section introduces the concept of viral fitness and discusses particular aspects of fitness that relate to viral quasispecies such as bottleneck events and fitness landscapes.*

Viral fitness and selection

Commonly, viral variants are associated with an ability to survive and replicate in a given environment. These phenotypic characteristics are often summarized in terms of *fitness* (literally derived from "fitting into the environment") that denote the ability of viral quasispecies genotypes or the whole quasispecies to survive and procreate.

While absolute fitness may be defined by the number of replicates a viral genotype produces in a given unit of time in isolation, the concept of fitness is highly relative. For instance, modest positive selection of a genotype (or, more precisely, of a genetic variant encoded by a genotype that results in an advantageous phenotypic trait) in absolute terms may be regarded as negative selection in relative terms if another genotype population is undergoing extremely strong positive selection.

By the same token, selective pressures in quasispecies almost never result in absolute dominance or absolute extinction of quasispecies genotypes, even if fitness is low; instead, selection is considered to work in a ’soft’ rather than in a ’hard’ manner and low-fitness genotypes may be maintained indefinitely as low-frequency minority genotypes (see later sections for a discussion of viral minorities).[1]

[1] Ruiz-Jarabo et al. (2000)

As a result of the relativity of the concept, viral fitness can be regarded as a complex parameter that is often quantified in absolute by surrogate variables such as the number of the copy number of viral genomes within a given volume of tissue. This *viral load*,[2] is often correlated with the severity of disease and is thus commonly employed to quantify viral fitness in clinical settings.[3]

Alternatively, relative measures of viral fitness may be conducted based on the replication capacity of a virus relative to a reference virus using competitive growth assays.[4]

Finally, sequence analysis of viral quasispecies is often employed to determine fitness by measuring the frequencies of genotype populations within the viral quasispecies. This method is motivated by the fact that higher frequencies of viral genotypes typically are associated with increased replication rates, thus implicitly entailing fitness gains.[5]

It is important to note the differences between these three approaches: while competitive growth experiments allow for quantification of relative fitness but include only a limited number of reference strains and are undertaken *in vitro* in a controlled and observable environment, viral load is directly based on patient samples but only reports the absolute fitness of a whole viral quasispecies; however, viral load measurements produce clinically actionable results and are well suited for comparisons between different patients. Sequence analysis of viral genomes may report the full genotype distribution of a quasispecies based on either patient samples or *in vitro* samples and allows for deduction of relative fitness based on the relative frequencies of genotypes; however, it does not yield an absolute quantification of viral fitness.

Bottleneck events. Fitness is often investigated with respect to *bottleneck events*, i.e., rapid changes in the quasispecies environment that reduce the effective population size of the quasispecies due to strong selective forces that spare only few genotypes of the viral quasispecies that are adapted to the new environment, for example, by virtue of enabling immune escape or conferring drug resistance. Bottlenecks regularly occur during initial transmission of viruses between hosts and are correlated with significant changes in the viral quasispecies[6] that can have dramatic consequences for clinical outcomes.[7]

Since only few viral variants are expected to encode phenotypic features compatible with surviving the bottleneck event, the surviving part of the viral quasispecies typically features lower abundance and reduced variability compared to the pre-bottleneck viral population.[8] This is due to the fact that quasispecies genotypes with high fitness are likely to be removed by the bottleneck due to sampling effects; since fitness of a viral population is largely determined by its fittest genotype, the replicative capacity of the post-bottleneck population is decreased, as has been demonstrated experimentally[9] as well as by stochastic simulations.[10]

[2] quantified either in terms of viral genomes copies per mL or as standardized quantitative units (IU) per mL; IUs are an artificial measure based on standardized samples that are used to calibrate quantitative assays for viral genome detection. Calibration is necessary since results from assays of different vendors exhibit distinct dynamic ranges and thus report varying quantifications. For HCV, a viral load of 2×10^6 genome copies/mL (the accepted decision threshold to differentiate low viral load from high viral load in therapy naive patients) corresponds to about 8×10^5 IU/mL in commonly used assays. However, the relation between copies/mL and IU/mL is nonlinear and depends on the calibration curve of the assay (Pawlotsky et al., 2000).

[3] Mellors et al. (1996), Mueller et al. (2008), Srikiatkhachorn and Green (2010)

[4] Domingo and Holland (1997), Martinez-Picado and Martínez (2008), Nájera et al. (1995), Orr (2009), Quiñones-Mateu and Arts (2006)

[5] Escarmís et al. (1999), Lorenzo-Redondo et al. (2011), Novella et al. (1995)

[6] Ali et al. (2006), Betancourt et al. (2008), Bull et al. (2011), Escarmís et al. (2006), Haaland et al. (2009), Li and Roossinck (2004), Quer et al. (2005), Vignuzzi et al. (2006)

[7] Escarmís et al. (2006), Manrubia et al. (2005)

[8] Gerrish and García-Lerma (2003)

[9] Ali and Roossinck (2010), Clarke et al. (1993), Novella et al. (2008), Vignuzzi et al. (2008)

[10] Gerrish and García-Lerma (2003)

Sequence spaces. The viral quasispecies is often conceptualized dynamically as a swarm of genotype populations that moves through an abstract, high-dimensional *sequence space*, also termed *genotypic space*. This representation serves as a means for conceptualizing the the interactive and interdependent processes of viral variability, fitness, and selective pressures. Coordinates within sequence spaces represent specific viral genotypes as determined by genomic nucleotide sequences. Thus, geometric distances within sequence space represent sequence dissimilarity between viral genotypes.

Since phenotypic characteristics of a virus are highly dependent on its genotypic identity, each coordinate in sequence space is commonly associated with a specific fitness value. These fitness values represent the ability of a specific viral phenotype to procreate in the current environment. Since higher fitness of a viral genotypes typically also entails higher frequencies of this genotype within the viral quasispecies, the density of genotypes at a given coordinate is assumed to be directly correlated with relative fitness of the corresponding viral phenotype.

Fitness values are commonly depicted as elevations of the sequence space; if represented in a simplified 3-space that allows for two genotypic dimensions and one fitness dimension, the resulting manifold bears similarity to natural landscape with riffs and valleys, plateaus and peaks. As a consequence, sequence spaces are termed to induce *fitness landscapes*. Within these landscapes, genotype populations (as identified by their sequence and fitness coordinates) form a spatially connected swarm that clusters near peaks of the fitness landscape.(11)

Fitness landscapes are believed to be rough (i.e., non-continuous) due to the deleterious effects of most mutations and may display considerable dynamics, both in terms of changing fitness values associated with specific genotypes, and in terms of which parts of the fitness landscape are currently occupied by viral quasispecies genotypes(12)

Exploration of fitness landscapes. The adaptive capacity of viruses is mediated by several parameters of the viral quasispecies. Among these are the mutation rate and associated genetic divergence of the viral quasispecies that facilitate rapid exploration of sequence space and are believed to be important determinants of viral adaptability. Replication-competent viruses will constantly generate offspring that differ from the parent by a number of nucleotide positions according to the mutation rate. As a consequence, viral progeny will slightly deviate from the ancestral genotypic coordinates in sequence space.

The continuous generation of viral progeny can be represented by discrete movements of viral genotypes in sequence space. As a result, viral genotypes are constantly "on the move": while advantageous movements in sequence space may lead to the occupation

(11) Lauring and Andino (2010, 2011), Smith-Tsurkan et al. (2010)

(12) Domingo and Holland (1997)

of new "peaks" on the fitness landscape and thus in increased numbers of progeny, disadvantageous movements may lead into fitness valleys and thus to loss of replicative ability.

Sequence spaces of viral quasispecies are tremendously large; for a viruses such as HCV with a genome length of approximately 10 kbp, $4^{10,000}$ ($\approx 10^{6,020}$) sequence states are possible, significantly more than the number of atoms in the observable universe (estimated to be on the order of 10^{82} atoms, based on combined stellar mass of the observable universe.[13]

In order to enable highly parallel exploration of the sequence space, large viral population sizes are required; these high abundances are supported by the short genome lengths of RNA viruses that allow for efficient replication and decrease the probability of accumulating several deleterious mutations within an individual replicate. In addition, high rates of replication also entail advantages with respect to the host: since the immune system cannot clear viral genomes at a sufficiently high rate to overcome new infections of host cells, large viral population sizes also promote chronicity of infection.

Mutational pathways and genetic barriers. Acquiring new phenotypes that are advantageous for replicating within new host micro-environments requires following a number of discrete steps (i.e., mutations) in sequence space. These *mutational pathways* are believed to be guided by fitness gradients that facilitate evolution towards advantageous phenotypes. Mutational pathways may require a large number of steps or the traversal of low-fitness "valleys" in the rough fitness landscape. The set of limitations a viral genome has to overcome in order to overcome a selective constraint such as an antiviral drug by a mutational pathway is often denoted as the *genetic barrier* of the selective constraint.

The "height" of the genetic barrier is generally quantified as the length of the mutational pathways that has to be traversed.[14] In principle, each pair of coordinates in sequence space is connected by a mutational pathway that is shorter than the viral genome length. In practice, however, much shorter distances consisting of only a few mutations may already have high genetic barriers, likely due to the roughness of the fitness landscape that results in intermediate steps exhibiting only low fitness. For instance, while single nucleotide transitions are associated with low barriers in viral quasispecies since they are readily reached by only one random mutation, nucleotide transversions that are less likely to occur randomly as well as pathways requiring two or three mutations in order to elicit changes on the protein sequence level are already considered to have comparatively high genetic barriers.

[13] Kragh (1999)

[14] Beerenwinkel et al. (2003, 2005) have introduced the more fitting notion of probability for the virus to escape. Here the mutational steps are weighed according to the advance that they bring in overcoming the barrier.

Error threshold and catastrophe

Since, by definition, viral quasispecies arise from a background of high mutation rates, regions in sequence space that confer high fitness are densely occupied by viral genotypes while genotypes that are distant from high-fitness regions are pruned by negative selection. However, as a consequent of the geometric properties of the sequence space and the inherent dynamics of fitness landscape, genotypes with lower fitness can be maintained if they are near in sequence space to high-fitness peaks. Since the progeny of high-fitness genotypes will commonly be distributed across a domain of sequence space that includes regions with lower fitness, the high-fitness genotype continuously replenishes surrounding low-fitness genotypes.

Similarly, low-fitness genotypes also replenish the high-fitness genotype as well as each other with their mutated progeny, although at a decreased rate due to their lower replicative fitness. This effect is termed *genotypic coupling* and can be generalized as to include all genotypes within connected and viable regions of the fitness landscape. The domain that encompasses all genotypes benefiting from such a shared region of high fitness is denoted as a *neutral network* or *contingent neutrality*.[15]

[15] Wilke (2005)

Advantages of genotypic coupling. It is in this context that broad exploration of a sequence space by viral quasispecies has important advantages: due to the high dimensionality of the space and the associated close connectivity, each genotype inhabiting a particular coordinate is connected to many other genotypes by just a single mutation. As a consequence of this tight coupling, a large hypersphere around a central master sequence with the currently highest fitness can be densely occupied by viral genotypes. This allows for continuous dissemination of new variants by the master sequence while also providing protection from rapid environmental changes: if a fitness peak shifts within the hypersphere due to environmental changes, the quasispecies already has established genotypes that encode many adapted phenotypes, one of which may be selected to become the new master sequence.

Occupying such a large sequence space with sufficient coupling, however, requires high population sizes; indeed, given the HCV example, 3×10^4 viral genotypes are within distances of a single mutation with respect to the master sequence. This number is well below the population size of RNA viruses inhabiting a single infected host, thus ensuring dense coupling and consequently high adaptability of the quasispecies. In contrast, larger mammalian genomes may generate many more variants (on the order of 10^{10}), in principle, thus seemingly offering more potential for variability and adaptability. However, this is offset by the considerably smaller population sizes of mammals which usually is well below the 1-step hypersphere around the master sequence.

Error catastrophe, revisited. Based on the previously introduced concepts pertaining to viral sequence spaces and genotypic coupling, we are now in a position to better appreciate recent finding on the distinctiveness of error catastrophe and lethal mutagenesis. Bull et al. was one of the first to argue, based on both theoretical models and experimental data, that lethal mutagenesis and error catastrophe are not equivalent, i.e., that error catastrophe is not a particular form of lethal mutagenesis and that the prior may even impede the latter.(16)

Bull et al. employed a minimal theoretical model of viral quasispecies theory as proposed by Eigen consisting of a two-genotype quasispecies population. The model assumes a population A and a mutant population B, the latter of which is constantly replenished by mutated progeny of A at a mutation rate μ_a. B also creates mutated progeny at a mutation rate μ_b, but all of its descendants are assumed to be inviable. The directionality of mutation is assumed to be asymmetric, i.e., descendants of A can mutate to B but not vice versa. Both populations A and B have their own relative fitness values w_a and w_b, respectively, and $w_a > w_b$.

It follows from quasispecies theory that both populations will achieve a mutation-selection equilibrium where mutations continuously create B from A and natural selection purges B by removing inviable descendants, thus decreasing the relative abundance of B in favor of A.

An explanatory model for the error catastrophe. This equilibrium between A and B is subject to change if mutation rates or relative fitness values are altered; in particular, lowering the fitness of A (w_a) or increasing the mutation rate of A progeny (μ_a) will lead to a constant decrease of A abundance in the quasispecies. Let $w_x(1-\mu_x)$ be the number of unaltered X progeny originating from the X population. We will denote this value as the *replacement rate* of X and assume that it has a value greater one. Since, by definition, the fitness of A is higher than the fitness of B, the replacement rate of A is greater than that of B at $\mu_a = \mu_b$. This relation maintains both A and B at an equilibrium: A because of the higher replacement rate, and B due to mutation of A progeny to B.

If however, μ_a unilaterally increases (17) until $w_b(1-\mu_b) > w_a(1-\mu_a)$, A has a lower replacement rate than B and thus is gradually replaced by B. The μ_a at which $w_b(1-\mu_b) = w_a(1-\mu_a)$ is the *error threshold* associated with that particular error catastrophe. Since A is not supported by mutated progeny of B, the frequency of A is reduced to 0 in this model. This elimination of A is termed an *error catastrophe* and is the endpoint of a gradual decrease of the frequency of A compared to B.(18)

(16) Bull et al. (2005, 2007)

(17) for example, by introducing a parameter μ_0 that represents the base mutation rate of the quasispecies and setting $\mu_a = k\mu_0$ and $\mu_b = \mu_0$, with $k > 1$

(18) Notably, error thresholds are further modulated by finite population sizes and the frequency of back mutations, the latter of which we disregarded in the model; while finite population sizes are likely to increase the conversion of A to B due to the accumulation of deleterious mutations (Nowak and Schuster, 1989), back mutations can retain A at infinitesimal population sizes; however, given analytical results, back mutations are only relevant if occurring at large frequencies (Bull et al., 2005). Also, in real-world settings A may not necessarily become extinct but may be maintained as a low-frequency minority genotype (Ruiz-Jarabo et al., 2000).

Quasispecies interactions and mutational robustness.

As noted upon in previous sections, viral mutation rates themselves are subject to selection.[19] Quasispecies theory adds to these propositions by further suggesting that mode selection optimizes the mutation rate of quasispecies in order to confer robustness with respect to mutation and thus allow for the quasispecies to exist at relatively high mutation rates.[20]

In combination with high rates of genetic coupling that is characteristic for viral quasispecies, it follows that (1) selection acts upon population of genotypes within the quasispecies that are close in sequence space and (2) that the selective process maximizes the average fitness of these groups rather than the fitness of single genotypes. As a result, less fit genotypes that occupy theoretical, elevated regions in the fitness landscape that provide moderate fitness can, as a connected population, generate more progeny than other viral populations that occupy narrow, high-fitness peaks[21] and whose progeny is more likely to be deleterious given even slight changes in the fitness landscape.[22]

This hypothetical phenomenon is denoted as "survival of the flattest" and can well be illustrated with the minimal quasispecies model introduced earlier. As can be derived from the previous definitions, an error catastrophe of A can only take place if A suffers a stronger mutational loss of its progeny than B. While the differing sensitivity to mutations may seem contrived in this model setting, the phenomenon is well in accordance with supplementary evidence indicating that error thresholds are believed to exist due to deleterious mutations may vary between populations (c.f. Wilke 2005). Importantly, after crossing of the error threshold of A and its extinction due to error catastrophe, further deleterious impact on the remaining population B is impeded: since B is more robust against mutations in the model (due to lower μ_b), the replacement rate of B is less affected than the replacement rate of A was prior to its error catastrophe.

Interestingly, this varying robustness is not necessarily a result of differing fidelities of the replication processes of genotype populations; instead, robustness may also depend on the particularities of the fitness landscape that is populated by a given genotypic population. In particular, high genotypic coupling of quasispecies populations can convey additional robustness to genotypic populations that can support each other by mutated progeny (as B was produced by A in the model presented). Such relations between subpopulations are denoted as *neutral networks*. Large neutral networks are assumed to pose additional robustness towards elevated mutation rates since neighboring positions in sequence space are more likely to be viable and populations within these network are constantly supplemented by mutated progeny of related genotypes.[23]

[19] Johnson (1999), Taddei et al. (1997)

[20] Eigen and Schuster (1977), Nowak (1992), Schuster and Swetina (1987)

[21] Sardanyés et al. (2008), Wilke et al. (2001)

[22] Graci et al. (2012), Sardanyés et al. (2008, 2009)

[23] Burch et al. (1999), Fontana and Schuster (1987), Huynen (1996), Huynen et al. (1996)

Two model of mutational robustness. These robustness theories, while notoriously difficult to prove experimentally due to the complexity of finding suitable model systems, have been supported by tentative experimental results demonstrating the selective advantages of interacting subpopulations.[24] In addition, simulation studies have beed devised in which simulated, asexual organisms compete for resources within a computational model environment.[25]

The simulation extends the previously introduced, minimal quasispecies model (A/B) and consists of two populations of organisms that originate from the same common ancestor but were evolved at differing mutation rates. While population L adapted to low rates of mutations, achieved high absolute fitness (in terms of simulated replication rates), and tended to populate small, high-fitness regions of the fitness landscape, population H was evolved at (and thus adapted to) higher mutation rates, displayed lower absolute fitness, and occupied elevated plateaus ("flat peaks") of the fitness landscape. L and H were then mixed in equal proportions and competed for resources across a ranges of mutation rates.

As could be expected, L replaced H at low mutation rates while H excluded L at higher rates of mutation within the simulation. Interestingly, this ability of a low-fitness population to replace a high-fitness population is neither due to absolute fitness increases of the H population in high-mutation environments (for instance, by increasing replication fidelity), nor due to absolute fitness decrease of the L population in the same, high-mutation environment; rather, the absolute fitness of H as measured in replicates per simulated organism per time unit is constant and remains below the fitness of L also in the high mutation environment. Similarly, L is well able to replicate indefinitely also in presence of high mutation rates if H is not present.

The shape of fitness peaks occupied by the respective populations was quantitatively confirmed. The relative fitness of a population and its relation to mutation rate and peak shape can be approximated by a realized growth rate (i.e., the relative fitness under different loads of mutations) $w(\mu) = w_0 \exp(-a\mu - b\mu^2)$ that depends on the genomic mutation rate μ, and the absolute replication rate w_0 of the populations in their original, non-competitive environment. a and b correspond to mutational robustness parameters that are fit to each population based on characteristics of the model environment, in particular the peak regions occupied by a population.

Based on these approximations, the reason for the success of H could be determined to be based on the shape of the fitness peaks that were occupied: while L occupied narrow peak regions of high absolute fitness that could no support a neutral network, H evolved towards occupying flatter peak and thus gained support from a neutral network, resulting in increased protection from high mutation rates.

[24] Burch and Chao (2000)

[25] Wilke et al. (2001)

Interpretation of minimal quasispecies models. The two models described (mutational robustness and error catastrophe: A/B, shape of the fitness landscape: H/L), while simple, are well in accordance with quasispecies theory and convey several messages of practical importance.

First, the A/B model can be extended to arbitrarily many genotypes that are connected by mutations. Multiple error thresholds are therefore possible if subpopulations have increasing error thresholds due to increasing robustness against mutations.(26) Such error thresholds could be arranged sequentially in a way that the highest-fitness, lowest-robustness error threshold is lowest and is succeeded by increasingly higher thresholds of more robust genotypic populations; as a consequence, increased error rates would be elicit continuous replacement of less robust genotypes with more robust genotypes.(27) In addition to being well covered by quasispecies theory, this sequential structure results in increasing resistance (or adaptation) to rising mutation rates, a phenomenon that may explain the therapeutic failure of monotherapies with mutagens against RNA virus populations (see below).

Second, from existence of multiple error thresholds becomes clear that an error catastrophe do not necessarily imply lethal mutagenesis and extinction of the complete population: only if the mutation rate exceeds the error thresholds specific to all genotypic populations, mutagenesis is truly lethal and the total quasispecies become extinct. Therefore, the concepts of error catastrophe and lethal mutagenesis are distinct; the error threshold only refers to "the loss of the highest fitness genotype in favor of other genotypes with lower fitness but greater mutational robustness"(28) and takes place when one population's replacement rate exceeds that of another population. In contrast, lethal mutagenesis and extinction occurs if no population can maintain a replacement rate of at least one ($w_i(1-\mu_i) < 1$ for all populations i). The smallest μ_i for which this is true can then be considered the last error threshold or *extinction threshold*.

Finally, the emergence of neutral networks at elevated mutation rates ("survival of the flattest") is likely to play an important role for robustness of the viral quasispecies. The phenomenon is especially relevant in the context of the use of mutagenic drugs against infections with RNA viruses. Intuitively, a modest increase of the mutation rate of the quasispecies that does not immediately lead to lethal mutagenesis as may be considered to be detrimental to the host since it will increase the viral adaptability.

However, assuming the existence of successive error thresholds and of neutral networks, an increased mutation rate may result in error catastrophes that is paralleled by a shift to more robust and less fit viral subpopulations. While the viral infection may not be cleared in this manner, treatment may still result in decreased fitness and constrained adaption, thus representing a viable therapy option, in principle.(29).

(26) Bull et al. (2005)

(27) Tannenbaum and Shakhnovich (2004)

(28) Bull et al. (2005)

(29) Holmes (2003)

21 *Quasispecies and antiviral therapy*

The genetic variability *and the population mechanics of viral quasispecies enable it to undergo rapid cycles of mutation and selection in order to adapt to novel environmental conditions. As a result, the quasispecies nature of many human pathogens constitutes a major hindrance to antiviral therapy.*[1] *The association of viral sequence variation with therapy success as an intensively studied field of current research*[2] *and detection of viral minorities, here exemplified by tracking the subclonal evolution of viral quasispecies with regard to drug resistance*[3] *and immune escape*[4] *has important consequences for developing and implementing rational drug treatment regimes.*[5] *This section discusses the role of genetic variants within the viral quasispecies and their relation to drug resistance and immune escape. In addition, effects of increasing mutation rates on the quasispecies are reviewed in a context of mutagenic drugs.*

[1] Domingo (1989), Figlerowicz et al. (2003)

[2] Luciani and Alizon (2009), Westby et al. (2006)

[3] Archer et al. (2010), Tsibris et al. (2009)

[4] Bull et al. (2011), Henn et al. (2012)

[5] Beerenwinkel (2003), Hinkley et al. (2011), Thielen and Lengauer (2012)

Resistance and compensation

As noted in the previous section, it follows from quasispecies theory that individual viral genotypes within a quasispecies are not the units of selection; rather, populations of these genotypes are subjected to group selection under consideration of inter-population relations such as neutral networks. It has therefore become clear that antiviral therapies should not target individual genotypes of the quasispecies but specific populations in the quasispecies.

Resistance-associated variants. Strong selective constraints such as immune pressure or antiviral drugs may have significant effects on genotype distributions within the quasispecies, for instance by drastically reducing the fitness of the current master sequence. In such cases, minority genotypes that previously had low replicative fitness or were suppressed by the highly competitive master sequence may display high relative fitness under the new selective constraints. These minority genotypes are commonly termed *escape mutants*[6] or, in the context of antiviral therapy, *resistance-associated variants* (RAVs).

[6] Melnick et al. (1961)

RAVs are mutations within the viral genome that confer resistance to specific or groups of closely related antiviral drugs, usually by altering viral protein binding interfaces that these drugs target to inhibit pathways of the viral replicative cycle. As a result of their protective effect against antivirals, RAVs confer additional phenotypic advantages to viral populations that evolve under pressure of these drugs. Viral genomes containing RAVs therefore often evolve according to a selection of mutational pathways that are characterized by serial and interdependent accumulation of RAVs.[7] The time a viral population requires to traverse such pathways to drug resistance is often also conceptualized as genetic barrier with respect to the development of resistance.

Importantly, viral genotypes are not limited to acquire resistance to only one antiviral at a time; instead, *cross-resistance*, i.e. resistance against several drugs, may be achieved my serially acquiring multiple RAVs and is especially pronounced with regard to drugs that employ identical modes of action.[8]

Compensatory mutations. Often, mutations such as RAVs with respect to the viral wild type entail additional fitness costs that limit either reproducibility or infectability of the affected viral species. This phenomenon is due to the fact that large structural changes in the viral proteins that confer high levels of resistance also inhibit normal functions of these proteins, for example by altering catalytic centers or binding interfaces, resulting in deleterious effects on viral replicative ability. Still, RAVs observed in clinical settings are generally advantageous for the viral quasispecies in spite of reduced fitness simply because they prevent extinction of the quasispecies by the antiviral.

Often, genotypes encoding RAVs that directly confer drug resistance evolve additional RAVs that restore viral fitness. These supporting RAVs are denoted *compensatory mutations* and do not themselves confer drug resistance. Rather, they induce compensatory structural changes in viral proteins, thus rescuing viral replicative ability in the presence of high-level drug-resistance RAVs. Commonly, these compensatory mutations are located in viral proteins targeted by directly acting antivirals, such as viral proteases or polymerases.[9] are are evolved subsequently to drug resistance mutations.

The concept of rapidly acquiring protective phenotypes is not limited to drug resistance: similar mechanisms are also active in viral evasion of antibody recognition. Therapies with neutralizing antibodies that target conserved viral epitopes are an active field of pharmacological research; however, similar to emerging drug resistance, viral genotypes encoding variants in epitope regions that do not encode antigenic epitopes are quickly selected for. These escape mechanisms are complemented by convergence towards epitopes that emulate sequence motifs of host factors and are therefore not readily targetable by antibodies due to immune tolerance.[10]

[7] Domingo et al. (2012)

[8] Harrigan and Larder (2002), Wyles (2013)

[9] Nijhuis et al. (1999)

[10] Novella et al. (1993)

Mutational pathways to drug resistance. As a consequence of the complex networks of escape variants and compensatory mutations as well as the various structural mechanism that may confer drug resistance on the protein level, often multiple mutational pathways exist that confer resistance to a single drug.[11] Each of these pathways may be associated with different levels of drug resistance, fitness costs, and compensatory mutations.[12]

RAVs induced within the HCV genome, for instance, are associated with various levels of resistance as quantified by IC50 assays[13] and corresponding losses of fitness as determined by growth competition and colony formation assays.[14]

The role of minority variants. Genotypes within viral quasispecies that occur at frequencies of lower than 20% (for Sanger sequencing) or lower than single-digit percentages (for NGS sequencing) are denoted as *viral minorities* and sequence variants that distinguish them from genotypes with higher frequencies are termed *minority variants*. Based on quasispecies theory of population equilibria, these genotypes are likely to display lower fitness if compared to the high-frequency genotypes such as the dominant species within the viral quasispecies (i.e., the master sequence). Minorities are of high clinical importance due to their crucial role in the development of drug resistance and immune escape variants by (1) providing high sequence diversity to the viral quasispecies that encode beneficial phenotypes for yet-to-be encountered selective constraints, and (2) by allowing for storage of formerly high-fitness genotypes for later use, for instance by encoding resistance-associated variants from prior drug treatment episodes.[15] This phenomena will be briefly discussed here.

The first feature has gained considerable clinical importance: due to the high variability of viral quasispecies, all possible single-mutations genotypes (compared to the master sequence) are likely to exist within a viral quasispecies at any given time. Some of these genotypes may encode for drug resistant phenotypes that exist at low frequencies of 10^{-3} to 10^{-5} even prior to therapy.[16] Upon antiviral therapy, these resistant genotypes are selected for and rapidly increase in frequency within the viral quasispecies. This phenomenon may result in rapid treatment failure during early phases of therapy and the characterization of the baseline genotype population prior to therapy has therefore gained considerably in clinical importance.

Even after treatment with an antiviral inhibitor has ended, viral minorities encoding resistance-associated variants may persist within the quasispecies. These "stored" genotypes are hypothesized to implement a *molecular memory* of formerly highly fit variants that may reemerge once selective pressures change again in their favor.[17] These memory mechanisms are associated with specific clinical phenomena such as resistant viral phenotypes that are present either before therapy (and possibly originate from infection

[11] Spyrakis et al. (2011), Tzeng and Kalodimos (2011)

[12] Agudo et al. (2010), Verbinnen et al. (2010)

[13] Viral drug resistance is commonly measured using *in-vitro* replication assays using the IC50 (inhibitory concentration) measure, e.g. the quantity of an inhibitor (a drug) that is required to inhibit a given biological process (for example, viral replication) by half. The IC50 can be utilized to compare the replicative competence of a wild type and a resistant virus by measuring the inhibitory concentration of both assays. Resistance is commonly quantified as the resistance factor (RF), defined as $\log_{10}$-fold-change in susceptibility to the drug relative to a wild type reference virus. Susceptibility can thus be directly based on the respective IC50 values. The IC50 is related to the EC50 and the CC50 with quantify the time-dependent potency of small-molecule drugs and the cytotoxicity of a compound, respectively.

[14] Koev and Kati (2008)

[15] Metzner et al. (2009), Mitsuya et al. (2008), Nájera et al. (1994), Ruiz-Jarabo et al. (2000), Varghese et al. (2009)

[16] Domingo and Holland (1992)

[17] Interestingly, the concept of molecular memory does not originate from quasispecies dynamics but from another diverse population: T-memory cells of the adaptive immune system that confer enduring immunity against certain antigen patterns and are maintained by the host long after initial infection Ahmed and Gray (1996), Fazilleau et al. (2007)

by another treatment-experienced host) or re-emergence of highly fit viral genomes after successful therapy.[18] This process may be further aided by viral compartmentalization, i.e., low-level replication or maintenance of these memory quasispecies in cell types not readily available to immune clearance. Memory genomes are under constant competitive pressure by other genotypes of the viral quasispecies and have been shown to be eliminated by bottleneck events,[19] indicating that antiviral treatment targeting viral replication may also be successful in counteracting memory mechanisms.

Mutagenesis and error catastrophe

As touched upon earlier, a high rate of genomic variation is a crucial factor for the evolution of viral quasispecies. It serves as a bedrock for high adaptability and fitness in of the viral population, especially in environments that are dominated by strong selective forces as for example induced by the host immune system or by antiviral therapy. The high rate of variability in RNA viruses is believed to be a consequence of the lack of error-correcting proofreading activity in viral RNA-dependent RNA polymerases, as well as due to missing mechanism of post-transcriptional repair of viral genomes.

Mutagenic drugs, especially nucleoside analogs such as *ribavirin*,[20] have been employed as antivirals since the 1970's and their effect on viral quasispecies is an active field of current research (see Table 21.1). Mutagenesis may be achieved by mutagenic drugs by several means, for example directly by genetic damage to viral genomes, or indirectly by either biasing intracellular nucleotide concentrations or by incorporation of non-complementary nucleotides, terminating nucleotides, or nucleotide analogs with incorrect base-pairing during viral genome synthesis.[21]

1990	Error catastrophe firstly investigated in experimental settings; Poliovirus has been shown to be affected by mutagenesis Holland et al. (1990).
1999	Mutagens are shown to impair HIV-1 replication *in vitro*; first use of the term "lethal mutagenesis" Loeb et al. (1999)
2000	First evidence that RNA viruses can indeed be extinguished by mutagenic drugs if viral load and viral fitness is low Sierra et al. (2000).
2000	The mutagen Ribavirin is shown to be effective against poliovirus in a clinically relevant setting Crotty et al. (2000).
2005	Experimental evidence for the lethal defection activity of mutagenic drugs, i.e., clinically relevant impairment of viral replication even at low concentrations of mutagenic drugs by generation of "interfering" genomes that can replicate but actively inhibit formation of infectious particles Grande-Pérez et al. (2005).
2009	First studies on therapy optimization and drug interactions involving mutagenic drugs; sequential and combination therapies are shown to a have relative advantages depending on the distribution of the viral quasispecies Perales et al. (2009a).

Table 21.1: *Timeline of viral mytagenesis.* Derived from Moreno et al. (2012).

Nucleoside analogs. Of the known mutagenic compounds, nucleoside analogs are the most well studied mutagens of RNA viruses. These compounds are often intracellularly metabolized into nucleoside-triphosphates, which are incorporated into viral genomes during polymerase-mediated nucleotide elongation. In contrast to antagonistic nucleoside/nucleotide inhibitors as used for example in anti-HIV therapy, nucleoside analogs do not result in chain termination as their main mode of action.

Instead, these compounds are incorporated into the nascent nucleotide chain and allow, due to their ambiguous molecular structure, allow for pairings with complementary purines and pyrimidines upon later use of the modified nucleotide chain as either substrate or template of the viral replication machinery. This unspecific pairing results in nucleotide substitution patterns that increase the viral mutation rate.[22]

[18] Briones et al. (2006)

[19] Arias et al. (2004), Ruiz-Jarabo et al. (2000)

[20] Clay et al. (2011), Harki et al. (2006)

[21] Anderson et al. (2004)

[22] Anderson et al. (2004), Domingo et al. (2005), Graci and Cameron (2008), Perales et al. (2011a)

Mutagens and error catastrophe. As a consequence of the increased mutation rate, random mutations are believed to be accumulated within viral genomes until the error threshold of the viral quasispecies is exceeded. This loss of genetic information may lead the quasispecies into one or more error catastrophes, a process that can theoretically lead to the extinction of the viral population by lethal mutagenesis (see previous sections).[23]

Importantly, the mutation frequency is not required to increase linearly during successful mutagenic treatments and elimination of the virus by mutagenic drugs is not required to be associated with large increases of mutational complexity within the viral quasispecies; instead, several error thresholds may be exceeded during treatment and the viral quasispecies may undergo multiple rounds of adaption towards more robust populations and decreased fitness without showing signs of large-scale deterioration.[24]

Indicators for mutagenesis. Interestingly, the main mode of action of widely used mutagens such as ribavirin is not well understood. As later sections of this chapter will discuss in detail, mutagenesis is a complex process that may involve indirect means of actions such as effects of mutagenic compounds on the intracellular nucleoside metabolism and on immune pathways. In general, however, a few direct indicators for the effectiveness of mutagenic drugs can be formulated:[25]

(1) mutagenic drugs should increase the mutation frequency in the viral progeny, (2) the mutagen or its metabolite should be incorporated into viral genomes by the viral polymerase, (3) mutation rates *in vitro* should depend on the concentration of the mutagen, (4) mutagenesis can be experimentally facilitated by increasing incorporation of the mutagen or by decreasing proof reading/genomic repair, and last (5), reduction of viral load should increase effectiveness of the mutagen due to proposed effects of Muller's ratchet.[26]

Since several modes of action may be causal for the observed efficacy of mutagens, the ability of mutagenesis to independently cause viral extinction is currently contested.[27] Partly, this circumstance is due to the fact that lethal mutagenesis is hard to prove experimentally, especially since current experiments cannot differentiate ongoing lethal mutagenesis from a sequence of nonlethal error catastrophes.[28] Because both mechanisms follow the same evolutionary pathways, only the total extinction of a viral quasispecies can provide conclusive evidence for lethal mutagenesis.[29] However, since viral clearance may also be a result of an immune response becoming more effective due to lower viral fitness, the distinction is very hard to make *in vivo*.

[23] Interestingly, the proposed induction of error catastrophes by mutagens parallels the action of components of the innate immune system: specific host factors of the APOBEC (*apolipoprotein B mRNA editing complex*) family are known to also have antiviral functions by deaminating viral RNA, a process denoted as hypermutagenesis. In particular, the human cytidine deaminase APOBEC3G was shown to target HIV and HBV ribonucleotides (Holmes et al., 2007, Mangeat et al., 2003). The fact that components of the immune system employ such strategies suggest that mutagenic therapies may indeed be based on sensible foundations

[24] Grande-Pérez et al. (2002), Ojosnegros et al. (2008)

[25] Anderson et al. (2004)

[26] Muller's ratchet: hypothetical genetic process that describes the accumulation of harmful mutations and the loss of genetic diversity in asexual populations due to the higher likelihood of harmful rather than beneficial mutations (Haigh, 1978). Especially relevant in populations with small effective population size such as post-treatment viral quasispecies. (Domingo et al., 1996, Duarte et al., 1992, 1993, Loeb et al., 1999, Yuste et al., 1999)

[27] Bull et al. (2007)

[28] Bull et al. (2007)

[29] Crotty et al. (2001), Grande-Pérez et al. (2005, 2002), Loeb et al. (1999)

Therapy optimization using mutagens

Monotherapy with mutagens often has only limited effects on viral quasispecies and generally does not result in viral extinction at non-toxic concentrations of the mutagen. One avenue for explaining this phenomenon is the great abundance of the viral quasispecies that allows for adaptation to increased mutation rates as discussed previously in the context of error catastrophes and mutational robustness. As a consequence, mutagens are likely not useful in monotherapy but should be administered in combination with other antivirals such as directly acting antivirals that drastically reduce viral population size and thus induce a bottleneck event.(30)

While not effective in monotherapy, mutagenic drugs are interesting candidates for combination therapy with directly acting antivirals. Directly acting antivirals induce strong population bottlenecks and abundance and fitness of the post-bottleneck viral population is predictive for therapy success.(31) As we we will argue in the following, the lower abundance and replicative fitness of the persisting, drug-resistant viral population offers an opportunity for treatment with mutagenic drug in order to further destabilize the viral quasispecies and increase treatment efficacy.

The efficacy of combination therapy with mutagens and directly acting antivirals is assumed to depend on the specific sequence of the treatment.(32) Parallel combination therapy involving a specifically acting drug and a mutagenic antiviral may not always be optimal for supporting viral extinction: while administering both drugs simultaneously is indicated in settings of high viral fitness in order to maximally constrain viral replication,(33) sequential dosage of first the specifically acting inhibitor followed by the mutagen may be advantageous in order to increase the effects of each individual drug.

In particular, administration of the mutagen prior to treatment with directly acting antivirals may increase the frequency of resistance-conferring viral variants, thus increasing surviving population size. On the other hand, the persistent population may also carry additional harmful mutations in addition to resistance mutations, thus resulting in lower numbers of viable genotypes after treatment. Stochastic simulations that model the effects of bottleneck events and acquired drug resistance on viral population dynamics indicate that the latter effect prevails at high mutation rates and especially with smaller persistent population sizes.(34)

Similarly, mutagens may be discontinued after directly antiviral treatment, thus slowing the development of compensatory mutations,(35) or be continued in order to further destabilize the viral population by enhancing the effect of Muller's ratchet. Tentative experimental results and simulations indicate, that continuing the mutagen may make sense if the persistent population is sufficiently small.(36)

(30) Tapia et al. (2005)

(31) Ioannidis et al. (2000), Sierra et al. (2000)

(32) Gerrish and García-Lerma (2003)

(33) Pariente et al. (2001), Tapia et al. (2005)

(34) Gerrish and García-Lerma (2003)

(35) Handel et al. (2006)

(36) Gerrish and García-Lerma (2003), Grande-Pérez et al. (2002), Nowak and Schuster (1989)

The applicability of treatment serialization strategies involving mutagens is currently explored experimentally in combination therapies against foot-and-mouth disease virus.[37] Results of these studies indicate that the use of a mutagen and a directly acting inhibitor increased treatment efficacy compared to the mutagenic agent alone and that sequential treatment protocols of first the directly acting inhibitor followed by the mutagen were more successful than combination treatments. These findings were further corroborated by models of viral dynamics at different doses of inhibitor and mutagen that stressed the importance of the inhibitor-induced bottleneck event for success of the mutagen.[38]

[37] Pariente et al. (2001, 2003), Perales et al. (2009b)

[38] Iranzo et al. (2011b)

22 The HCV quasispecies

Of the approximately 170 million infected individuals that are chronically infected with HCV, twenty percent will develop liver cirrhosis of which up to 2.5% will progress to hepatocellular carcinoma (HCC).[1] *Apart from displaying a wide range of side effects, the efficacy of the current standard of care (SOC) therapy is highly dependent on viral genotype and may result in viral clearance in only 50-60% of all patients in large subgroups of the patient population.*[2] *Several indicators of highly adaptive quasispecies as discussed previously are present in HCV mutants, such as the rapid emergence of drug resistance and compensatory mutations,*[3] *the shaping of viral populations by epidemiological bottleneck events,*[4] *as well as increased variability of viral genes that are associated with immune escape.*[5] *This section discusses several aspects of viral quasispecies at the example of HCV. In particular, the high diversity and abundance of this virus as well as modes of disease progression and immune evasion are reviewed in detail. Finally, antiviral treatment options are reviewed, both in relation to the current standard of care therapy as well as in a context of recently approved directly acting inhibitors.*

Diversity and abundance

The existence of viral quasispecies in HCV was proposed almost immediately after the discovery of the pathogen and is now an integral part of anti-HCV treatment.[6] Based on experimental data originating from both Sanger and deep sequencing techniques, the normal mutation rate of HCV can be estimated at about 10^{-2} substitutions per site and year,[7] a relatively high value compared to other RNA viruses. HCV features a very narrow host cell tropism as well as a limited host range and almost exclusively infects liver cells of humans and chimpanzees.[8] Due to its reliance on RNA genomes with short half-lifes and the lack of latency mechanisms such as proviral integration, HCV is depending on continuous replication for persistence.

[1] Bowen and Walker (2005)

[2] Pearlman (2004)

[3] Halfon and Locarnini (2011), Kato et al. (2006), Pokrovskii et al. (2011)

[4] Laskus et al. (2004), Quer et al. (2005)

[5] Farci (2011), Simmonds (2004)

[6] Farci (2011), Martell et al. (1992)

[7] Lutchman et al. (2007)

[8] Transgenic mouse models have recently been developed but require expression of viral entry factors in mouse hepatocytes, cf. Dorner et al. (2011)

HCV variability. Similar to other RNA viruses, HCV features high replication kinetics of up to 10^{12} virions per day in untreated patients.[9] Since both humoral and cellular arms of the immune system exert selection pressure on the quasispecies,[10] deleterious variants are constantly pruned from the population. Still, the remaining diversity of the quasispecies quantified as average genetic distances between genomes regularly exceeds 10%.[11]

RNA copies produced by the HCV RNA-dependent RNA-polymerase NS5B have an estimated error rate of approximately 10^{-5} per genomic site and copy, a rate typical for non-retroviral RNA viruses.[12] Given a genomic length of approximately 9.6 kbp, experimental data suggest that about 0.096 mutations are introduced per replicated viral genome, on average.[13],[14] While these rates do not incorporate additional effects that influence genomic variation such as recombination by polymerase template switching, the latter is believed to be a low-frequency process in HCV and is therefore neglected here.[15]

Given these estimates and assuming a binomial model of mutations[16] within copied viral genomes, each new virion has a probability of 0.91 to be free of mutations and probabilities of 0.087, 0.0042, and 0.00013 to harbor one, two, or three substitutions, respectively. Based on the large number of virions that are produced each day, on the order of 10^{10} and 10^{9} single- and double mutants, respectively, are generated daily, effectively covering the space of theoretically possible single and double mutants more than 10-fold, even given the low average life span of the virus of about 4 hours.[17]

As a consequence of this dense coverage of genotypic space, it is statistically likely that all viable single and double mutations that confer drug resistance already exist within the HCV quasispecies prior to antiviral treatment.[18] While only a small fraction (on the order of 10^{-5}) of the space of all possible triple mutants is explored by the HCV quasispecies each day, resistance-conferring triple mutations can be acquired by the virus through sequential accumulation of single mutants. Even given the most potent anti-HCV therapeutics that decrease HCV RNA to 10^{-5} and thus reduce the number of virions produced per day to 10^{7}, a sufficiently high number of single mutants remain to achieve resistance phenotypes within days.

Resistance-associated variants. Indeed, resistance-associated variants can be detected rapidly after HCV monotherapy with *directly acting antivirals* (DAA) as part of *specifically acting antiviral therapy for hepatitis C* (STAT-C).[19] As a consequence, the clinical outcome of disease treatment can be significantly impaired if no alternative therapy options are available. The early detection of RAVs is therefore a critical component of rational therapy optimization.[20]

[9] Neumann et al. (1998), Ramratnam et al. (1999)

[10] Erickson et al. (2001), Mondelli et al. (2003), Sheridan et al. (2004), Söderholm et al. (2006)

[11] Cristina et al. (2007), Fan et al. (2009), Manzin et al. (1998), Martell et al. (1992)

[12] Powdrill et al. (2010)

[13] Zeuzem et al. (1998)

[14] Due to the strandedness of the RNA molecule, two rounds of low fidelity replication are required to generate the plus strand for a new virion; however, a single round of replication is conservatively assumed in the inference of mutation rates Rong et al. (2010)

[15] Simmonds (2004)

[16] Given a genome of length g and an error rate e, the probability of having exactly s substitutions within a replication is $P_s = \binom{g}{s} e^s (1-e)^{g-s}$ and there are a total of $\binom{g}{s} 3^s$ possible sequences with that many substitutions. The binomials can be approximated by the Poisson distribution.

[17] Neumann et al. (1998)

[18] Rong et al. (2010)

[19] Rong et al. (2010)

[20] Koev and Kati (2008)

Importantly, however, RAVs are not only a product of evolution of the viral quasispecies within the host; instead, due to the high variability of the viral quasispecies as well as a result of possible infections with viral strains originating from pretreated hosts (see previous sections on *molecular memory*), viral genomes infecting treatment-naive patients may already contain minority RAVs.[21] These baseline escape mutations have been experimentally measured at minority frequencies of about 1-2% in *in vitro* fluorescence experiments involving replicon assays as well as by direct sequencing.[22]

Disease progression and immune evasion

HCV infection is diagnosed by the detection of HCV RNA as well as anti-HCV antibody sero-conversion, both of which are detectable 4–10 weeks after exposure to the virus, in principle. HCV acutely infects humans within the first six months after transmission and a minority of 20% of patients is able to clear the virus spontaneously. Since infection is mostly asymptomatic at the acute state, patients frequently remain undiagnosed and clinical progression at this stage is severely understudied.[23] The remaining 80% of infected individuals develops chronic liver diseases such as cirrhosis as indicated by increased levels of serum alanine aminotransferase (ALT), an enzyme used as a surrogate for liver damage.

HCV is the dominant factor for both the development of virally induced liver carcinoma, which occurs in about 1-2% of chronic patients, and the main cause of liver transplantations. While there is evidence for extra-hepaticular compartmentalization of HCV by low replication in immuno-privileged brain cells, the relevance of these data is currently contested.[24] Possibly as a result of this compartmentalization, liver grafts are recurrently infected by HCV, often leading to highly progressive liver disease in organ acceptors.[25]

Complexity of viral quasispecies has been correlated with therapy response and clinical progression of HCV.[26] The HVR-1 region in particular is a common surrogate marker for HCV quasispecies heterogeneity. As it is located within the viral envelope protein, HVR-1 may be relevant for both cellular entry and evasion from antibody recognition.[27]

Fittingly, while newly infected HCV patients tend to have low HVR-1 variability, possibly as a result of bottleneck events associated with transmission,[28] chronic HCV patients usually feature high variability and differences in mutant frequencies in this region, possibly indicating complex interactions of the quasispecies genotypes with the immune system. By way of supporting this assumption, quasispecies complexity in the HVR-1 has been shown to be low in individuals with suppressed immune system, possibly indicating low selective pressure.[29]

[21] Robinson et al. (2011)

[22] Robinson et al. (2011), Sarrazin and Zeuzem (2010), Verbinnen et al. (2010)

[23] Edlin (2011)

[24] Farci (2011)

[25] Berenguer et al. (2001), Pessoa et al. (1999)

[26] Domingo and Gómez (2007), Farci et al. (2000), Le Guen et al. (1997), López-Labrador et al. (1999), Más et al. (2004), Morishima et al. (2006), Rothman et al. (2005)

[27] Bartosch et al. (2005)

[28] Laskus et al. (2004), Quer et al. (2005)

[29] Bernini et al. (2011), Odeberg et al. (1997)

Humoral immunity and HCV infection. Similar to other viral infections, both humoral and cellular components of the adaptive immune system play important roles in modulating the pathogenesis of HCV infections and are inhibited by different viral mechanisms.[30] Successful clearance of initial infection is associated with an increased production of CD8$^+$ and CD4$^+$ T cells as well as with the initiation of immunologic memory that confers modest protection upon reinfection.[31]

HCV is very apt at evading adaptive immune response both in the initial acute phase of infection as well as during chronic persistence. Interestingly, HCV immune evasion is highly specific and does not seem to affect host immune responses to other pathogens. While HCV-specific antibodies are detectable soon after infection,[32] it is still unclear if these immunoglobulins are indeed neutralizing as the occurrence of HVR-1 escape mutations in the HCV E2 proteins of chronically infected individuals may suggest.[33] If the latter is the case, then these mutations are likely evidence for viral quasispecies adaption to humoral and cellular (CD8$^+$) immune responses by antigenic variation and epitope loss.

While epitope loss seems to be a primary mechanism of escape from neutralizing antibodies,[34] a more active role of the viral hypervariable regions has been proposed that may serve as decoy for neutralizing antibodies or physical hindrance to antibody binding, thus protecting conserved viral epitopes.[35] Similarly, lipoproteins and glycans of the HCV virion may have additional roles concerning antibody masking, for instance by inducing conformational changes in neutralizing antibodies or by preferably binding to non-neutralizing antibodies that may mask epitopes.[36]

Loss of viral epitopes by hypermutation during the acute phase of infection in dependence on the host HLA type seems to determine immune failure and disease chronicity.[37] These mutations predominantly occur at the HLA binding anchors of the epitope peptide as well as in the centr of the epitope and may inhibit HLA binding and T-cell recognition, respectively. Their frequency and highly clustered nature on the genome seems to represent a trade-off between maintaining replicative fitness on the one hand and disabling broad cross-recognition of the epitope by T-cells on the other hand.[38]

Cellular immunity and HCV infection. In contrast to humoral responses, cellular immune response to HCV is currently better understood and the initial delay of T-cell response as well as the failure of chronic patients to generate adequate immune answers is an active field of research.[39] Experimental results have indicated that both CD8$^+$ and CD4$^+$ T cells of successful therapy responders target several MHC-I and MHC-II epitopes within HCV proteins[40] while patients with progressive disease express lower numbers of T-cells that target significantly fewer epitopes.[41]

[30] Rehermann (2009), Thimme et al. (2012), Walker (2010)

[31] Mehta et al. (2002)

[32] Gretch (1997)

[33] Farci et al. (2000)

[34] Farci et al. (1996)

[35] Bankwitz et al. (2010), von Hahn et al. (2007)

[36] Falkowska et al. (2007), Helle et al. (2007)

[37] Erickson et al. (2001), Timm et al. (2007)

[38] Söderholm et al. (2006), Thimme et al. (2012)

[39] Cooper et al. (1999)

[40] Diepolder et al. (1995), Grakoui et al. (2003), Lechner et al. (2000), Missale et al. (1996)

[41] Koziel et al. (1993), Lauer et al. (2004), Takaki et al. (2000)

Proposed explanations for this inability of the adaptive immune system to generate adequate responses include impairment of antigen presentation dendritic cells (DC) by viral modulation of the innate immune response or inhibition of DC maturation, as well as antigenic variation and functional impairment of CD8$^+$ T cells.[42] In particular, HCV may modulate T-cell response by up-regulating inhibitory receptors such as *PD-1* and *Tim-3* on HCV-specific CD8$^+$ T-cells, thereby inhibiting proliferation and predisposing these cells to apoptosis.[43] While these pathways may indicate promising new drug targets for host-factor based anti-HCV medicines, the precise molecular mechanisms of active HCV immune evasion are currently unknown.[44]

Anti-HCV therapy

The current standard of care (SOC) therapy against HCV is based on a combination treatment of the immunoregulatory cytokine interferon-α (IFN-α), an important modulator of the host immune system, and the antiviral prodrug ribavirin (RBV). The latter compound is known to display several modes of action against other RNA viruses, including interference with RNA metabolism and mutagenic activity. However, its effective mechanism against HCV is presently unknown.[45]

Both the effects of host genotype and of viral genotype have been associated with the mode of action of interferon-α. Viral determinants of interferon response are probably multi-factorial and may include the multifunctional HCV NS5A protein as well as structural E2 (envelope) and core (capsid) HCV factors.[46] Of these proteins, especially the aptly named interferon sensitivity-determining region (ISDR) of NS5A and the hypervariable region (HVR-1) of the E2-coding region may be involved in inhibition of interferon-inducible protein kinase R (PKR) and immune escape by antigenic variability,[47] respectively.

In addition to viral determinants of interferon response, the detection of host genes that significantly effect anti-HCV treatment response has permanently shaped HCV therapy towards a personalized (or precision) paradigm of medicine. A number of polymorphisms within the IL28B gene on human chromosome 19 that are associated with interferon secretion and were shown to be predictive for SVR.[48],[49] Similarly, IP-10, a host factor involved in chemo-attraction that may by at variation in non-responding patients compared to responding patients, has been demonstrated to be a promising new factor in HCV pharmacogenomics.[50] Finally, the human inosine triphosphatase (ITPA) gene, a relative of IMDPH, as well as the ENT1 nucleoside transporter are correlated with therapy response.[51]

[42] Gruener et al. (2001), Jinushi et al. (2004), Khakoo et al. (2004), Wedemeyer et al. (2002)

[43] Golden-Mason et al. (2009), Radziewicz et al. (2008)

[44] Thimme et al. (2012)

[45] Fried et al. (2002), Manns et al. (2001)

[46] Chayama and Hayes (2011), Enomoto et al. (1996), Pawlotsky et al. (1998)

[47] Farci (2011), Gale et al. (1997)

[48] *SVR*: Sustained virological response; defined as non-detectable ($<$ 50 IU/mL) serum HCV RNA by the end of the treatment period or 24 weeks after treatment.

[49] Ge et al. (2009), Honda et al. (2010)

[50] Casrouge et al. (2011)

[51] Fellay et al. (2010), Morello et al. (2010)

Treatment efficacy. While the SOC interferon-α plus ribavirin therapy achieves a SVR against most the viral genotypes, it is not effective against all patients infected with HCV genotype 1, the most frequent genotype in western countries. In particular, treatment of HCV genotype 1 patients results in SVR of only about 50% of patients (in contrast to more than 80% of patients for genotypes 2 and 3).[52] As a consequence, novel, directly acting anti-HCV compounds have recently been approved or are currently in late stages of development.

While the combination therapy of pegylated interferon-α and ribavirin is still considered to be the standard of care therapy against HCV, these recommendations are in the process of being changed as DAAs are introduced in clinical therapy. Two protease inhibitors, *telaprevir* and *boceprevir*, are currently FDA approved and at least 20 additional compounds including nucleoside (chain-terminating) and non-nucleoside (allosterically functioning) inhibitors of the polymerase, an inhibitor of the membrane-associated HCV phosphoprotein NS5A, and a cyclophilin inhibitor are in current testing.[53] Peptidomimetic protease inhibitors constituted the first line of DAAs that underwent development,[54] and have been shown to dramatically increase SVR rates especially for genotype 1 patients.[55]

However, due to their direct mode of action, these new antivirals are prone to the development of viral resistance and several RAVs are currently known for all inhibitors currently used in clinical settings.[56] Since one to two mutations are required to acquire the resistance phenotype for the currently used DAAs,[57] these estimates suggest that treatment failure for HCV monotherapy is inevitable. While some classes of drugs, such as HCV polymerase inhibitors may feature a higher genetic barrier against evolving drug resistance, combination therapy may still be required in order to raise the genetic barrier of therapy to at least three and better four mutations.[58]

HCV therapy optimization. Monotherapy with currently approved DAAs leads to development of resistance-associated variants within the inhibited viral proteins, resulting in viral breakthrough and drastically reduced efficacy of therapy.[59] Subsequent analysis of the viral genotypes confirmed the existence of RAVs at frequencies of 4-20% in the viral quasispecies within just days of therapy, significantly more rapid than observed in monotherapy against HIV and HBV.[60] This rapid emergence of resistance can be explained by the high variation and large numbers of progeny of HCV that allow for efficient exploration of genotypic space in search for beneficial mutations.[61] HCV replication requires multiple rounds of low-fidelity RNA copying, a process that increases error rates in comparison to other positive-sense single-stranded RNA viruses.

[52] Manns et al. (2006)

[53] Gao et al. (2010), Naggie et al. (2010), Sarrazin and Zeuzem (2010), Watkins et al. (2010)

[54] de Francesco and Migliaccio (2005), Manns et al. (2007)

[55] Lamarre et al. (2003)

[56] Halfon and Locarnini (2011), Halfon and Sarrazin (2012), Hiraga et al. (2011), Kieffer et al. (2010), Sarrazin and Zeuzem (2010), Verbinnen et al. (2010)

[57] Zhou et al. (2008b, 2007)

[58] Ali et al. (2008), He et al. (2008), Manns et al. (2007), McCown et al. (2008), Rong et al. (2010)

[59] Reesink et al. (2006), Sarrazin et al. (2007)

[60] Kieffer et al. (2007), Soriano et al. (2008)

[61] Cox et al. (2005), Mao et al. (2001), Ray et al. (2000)

In addition, the HCV genome does not face an important functional constraints to evolvability: while genomes of other RNA viruses such as HIV and HBV feature overlapping reading frames and mutations are thus likely to have functional impact on multiple viral genes, the HCV has only a single reading frame (possible exceptions exist in specific viral genotypes)[62] and should therefore be able to accumulate mutations with less deleterious effects, in principle.[63]

In order to avoid or postpone eventual treatment failure, antiviral approaches that counteract the selection of both RAVs and compensatory mutations while also hindering the viral population at overcoming the genetic barrier of resistance are of crucial clinical importance. Several classes of strategies for modulating the accumulation of RAVs are currently in clinical use. Among these are drug combination therapies that increase the genetic barrier by requiring a viral population to evolve RAVs and compensatory mutations against several, orthogonally acting drugs in parallel, a process that is statistically less likely the more drugs are involved.[64]

Similarly, the use of drugs that target host factors instead of viral factors is also believed to increase the genetic barrier against HCV drug resistance. In order to gain resistance against host-factor targeting drugs, the virus has to acquire mutations that makes it less dependent on a host factor that is essential for viral replication. This feat may necessitate interactions with alternative host factors featuring distinct binding interfaces and thus require considerable more changes to viral proteins than evolving resistance mutations against DAA binding viral proteins would require.[65]

Clinical studies have demonstrated that anti-HCV treatment should start early and with maximal effectiveness in order to inhibit the recovery of intra-host fitness from bottleneck events induced by transmission and initial adaption to the the immune system.[66] Ideally, such protocols should be highly personalized and the choice of drugs as well as the timing of the intervention should take into account viral genotype, the composition of the viral quasispecies, as well as genetic host genotypic factors that may influence immune response or pharmacokinetics. Future anti-HCV therapies will thus likely exclude interferon due to its intravenous application and high side effects and will instead consists of an all-oral therapy of combinations of three or four direct acting anti-HCV agents.[67] These therapies may optionally be followed up with with mutagenic drugs that further reduce the fitness of the viral quasispecies and thus increase the likelihood of immune clearance (cf. next section). However, while promising, none of these approaches is currently considered to result in unsurmountable genetic barriers against antiviral resistance.[68]

[62] Morice et al. (2009)

[63] Neumann et al. (1998)

[64] Domingo et al. (2012)

[65] Coelmont et al. (2010)

[66] Amador-Cañizares and Dueñas-Carrera (2010), Ho (1995)

[67] Burney and Dusheiko (2011), Paeshuyse et al. (2011)

[68] Domingo et al. (2012)

23 *Anti-HCV efficacy of ribavirin*

Ribavirin *(1-β-D-ribofuranosyl-1H-1,2,4-triazole-3-carboxamide or $C_8H_{12}N_4O_5$) is a guanosine analogue with antiviral activity against many RNA viruses.[1] In addition to its mutagenic effects, the compound is directly modulating different aspects of nucleotide synthesis and metabolism while also indirectly supporting proliferation of T-helper cells towards immune response profiles beneficial for viral clearance. At its conception in 1972,[2] ribavirin was the first synthetic nucleoside that exhibited broad-spectrum antiviral activity. It is currently used in monotherapy against highly aggressive zoonotic Arenaviruses such as Lassa virus[3] and is also being indicated against mild infections of viruses that lack other antiviral such as caused by the respiratory syncytial virus.[4] While mutagenic activity of ribavirin is experimentally supported in several viral species, evidence in that regard concerning HCV is conflicted. This section aims to review the proposed modes of action of ribavirin both in HCV and in other viral species with a particular focus on mutagenesis. Finally, this section discusses ways of experimentally quantifying mutagenesis in viral quasispecies.*

Ribavirin monotherapy. Ribavirin monotherapy of HCV is associated with mildly beneficial effects on treatment progress such as transiently lowered aminotransferase levels and modest improvement in hepatic histology.[5] Ribavirin induces a modest direct antiviral effect mainly in the first weeks of treatment,[6] and a transient decrease in viral load associated with ribavirin monotherapy in patients by the end of 8-48 weeks of treatment has been confirmed by a Cochrane meta analysis of 11 randomized trials.[7] However, the drug causes no lasting decrease of HCV viral load, the most important indicator for sustained treatment success.[8] The drug is therefore not used in HCV monotherapies but in combination with pegylated interferon-α, where by reasons that are still not completely clear it significantly improves sustained virological response, primarily by reduced relapse rates.[9]

Significant long-term beneficial effects of ribavirin in combination therapy with interferon have also been confirmed by meta analyses.[10] Interestingly, the synergistic effect of ribavirin with other antivirals is not limited to interferon; instead, ribavirin combination therapy with several novel specifically acting antivirals have resulted in dramatically improved sustained virological response

[1] Wu et al. (2003)

[2] Sidwell et al. (1972)

[3] Andrei and de Clercq (1993), Huggins (1989)

[4] Krilov (2001)

[5] Bodenheimer et al. (1997), Dusheiko et al. (1996), Pawlotsky et al. (2004)

[6] Dixit et al. (2004), Herrmann et al. (2003)

[7] Brok et al. (2006)

[8] Bodenheimer et al. (1997), Cattral et al. (1999), Di Bisceglie et al. (1995), Hoofnagle et al. (1996), Pawlotsky et al. (2004)

[9] National Institutes of Health (2002), Pawlotsky (2005)

[10] Brok et al. (2005)

rates (SVRs) compared to using the direct inhibitor in monotherapy[11] possibly due to ribavirin-mediated inhibition of resistance mutations.[12]

While in principle RAVs against mutagenic drugs seem to pose a viable escape strategy for various viral species confronted with mutagens, for example by increasing the fidelity or selectivity of the viral polymerase in order to reject or compensate for mutagenic drugs,[13] it is presently unknown if such mutagen-resistant viruses are indeed evolving in clinical settings.[14] Although indicators for ribavirin-induced RAVs have been observed in cell culture and monotherapy,[15] these mutations are associated with multiple viral genes and their precise effects on therapy success are currently debated.

Ribavirin side effects. However, treatment of Ribavirin induces significant dose-limiting side effects as for example haemolytic anaemia in patients due to accumulation of ribavirin metabolites in erythrocytes, thus limiting dosage and broad applicability of the drug.[16] Indeed, 10-22% of treated patients develop significant anaemia of which up to 52% experience a decrease in quality of life.[17] In addition, ribavrin has been shown to act as teratogen and carcinogen animal models, highlighting its toxicity in certain circumstances.[18] As a consequences of these side effects, other ribavirin-like molecules are in development that promise to display better metabolic properties or increased specificity of targeting liver compartments, thus facilitating tolerance.[19] In particular, these compounds are aimed at combination therapies involving a variety of novel anti-HCV drugs including protease, helicase, and polymerase inhibitors, antagonists of the internal ribosome entry site and of Ithenosine-5′-monophosphate dehydrogenase (IMPDH), as well as small interfering RNAs, ribozymes, and new interferons that are currently under development.[20]

Overview of proposed modes of action

The mode of action of ribavirin within anti-HCV therapy is currently now well understood.[21] However, several possibilities have been proposed, most of which are supported by evidence from the use of ribavirin in other RNA-viral diseases such as polio or foot-and-mouth disease. Among these proposed modes of action are the ability of ribavirin to influence metabolic pathways within the cell, immune-related modes of action, antagonistic inhibition of the viral polymerase, directly mutagenic activity, modulation of guanylyl- and methyltransferase activity, interference with viral transcript capping, inhibition of the viral polymerase by binding close to the active site and interfering with elongation reaction.[22] Several of these modes of action alone may serve to explain the antiviral activity of ribavirin in HCV. Interestingly, however, these modes of action may also have mutually supportive roles.[23]

[11] Gane et al. (2013), Lange and Zeuzem (2013), Poordad et al. (2013)

[12] Feld (2012), Rotman et al. (2013)

[13] Agudo et al. (2010), Arias et al. (2008), Pfeiffer and Kirkegaard (2005a), Vignuzzi et al. (2006)

[14] Mullins et al. (2011)

[15] Agudo et al. (2010), Coffey et al. (2011), Levi et al. (2010), Pfeiffer and Kirkegaard (2003)

[16] De Franceschi et al. (2000), Lin et al. (2004), Lindahl et al. (2005), Sulkowski (2003)

[17] Birgegård et al. (2005), De Franceschi et al. (2000)

[18] Kochhar et al. (1980)

[19] Benhamou et al. (2009), Markland et al. (2000)

[20] McHutchison and Dev (2004), McHutchison and Patel (2002)

[21] National Institutes of Health (2002), Pawlotsky (2005)

[22] Cecere et al. (2004), Crotty et al. (2001), Eriksson et al. (1977), Fang et al. (2000), Goswami et al. (1979), Leyssen et al. (2005, 2006), Maag et al. (2001), Vo et al. (2003)

[23] Moreno et al. (2011)

Metabolic mode of action. Due to the rapid replication kinetics of HCV, *de novo* nucleotide synthesis rather than salvage of cellular nucleosides is the predominant source for viral RNAs in HCV infected cells. During *de novo* synthesis, Inosine-5′-monophosphate dehydrogenase (IMPDH) catalyses of ionosine monophosphate into xanthine monophosphate has been shown to be the rate-delimiting step in the production of guanine nucleotides, especially for rapid viral replication.[24]

Ribavirin in its native form is a prodrug that is metabolized into ribavirin 5′-mono-, di-, and triphosphates (RMP, RDP, and RTP, respectively) by cellular kinases within erythrocytes.[25] RMP is a competitive inhibitor of both isoforms of IMPDH and thus decreases cellular levels of guanosine triphosphate (GTP) by about 50%.[26] Nucleoside triphosphates are important precursors for RNA synthesis. In theory, the perturbation of their intracellular concentrations may negatively affect the fidelity of the RNA polymerase by increasing the probability of misinsertions, primarily by suppression of proofreading mechanisms (c.f. Kumar et al. (2011a) for an investigation of the proposed mechanisms in the DNA polymerase). IMPDH inhibition results in negatively effecting rapid viral replication as shown for yellow fever virus and human parainfluenza virus,[27] while being of lesser but still sufficient effectiveness in HCV.[28]

Immune-related mode of action. Successful immune control of HCV infection is associated with strong intrahepatic T-helper cell response, in particular with Th_1 proliferation and the corresponding cytokine profile.[29] Chronicity, on the other hand, is often paralleled with insufficient Th_1 activation, possibly due to HCV-core dependent subversion of the interleukin system.[30] Ribavirin is assumed to shape the adaptive immune response against HCV infected cells by enhancing the effect $CD4^+$ T-helper cells Th_1 cytokine profiles, resulting in increased production of tumor necrosis factor-α, interleukin-2, γ-interferon and IFN-α.[31] In addition to T-helper mediated effects, ribavirin may also potentiate inflammatory defenses through specific activation of IFN stimulated genes (ISGs), thus modulating the expression of host Toll-like receptor 7 (TLR7) and ISG15, the latter of which is implied in HCV immune evasion (c.f. Paeshuyse et al. (2011) for a review).

Mutagenic mode of action. While modulation of intracellular nucleotide pools by competitive interaction with IMPDH remains the experimentally most well supported means of action of ribavirin, it is notable that not all IMPDH inhibitors have antiviral activity[32] and it is thus likely that ribavirin and related compounds also have other antiviral activities. The monophosphate form of ribavirin (RMP) can be phosphorylated to form ribavirin triphosphate (RTP), which inhibits viral polymerase[33] and may be non-specifically incorporated into viral RNA during replication opposite to either

[24] Markland et al. (2000)

[25] Lin et al. (2004), Wu et al. (2003)

[26] Müller et al. (1977)

[27] Leyssen et al. (2005, 2006)

[28] McHutchison et al. (2005), Mori et al. (2011)

[29] Abonyi and Lakatos (2005)

[30] Cecere et al. (2004), Vollmer et al. (2004)

[31] Cecere et al. (2004), Fang et al. (2000)

[32] Crotty et al. (2000), Lanford et al. (2001)

[33] Vo et al. (2003)

cytidine (in place of guanine, by *Watson-Crick pairing*) or uridine (in place of adenosine, by *wobble pairing*) at roughly one copy per HCV genome. This non-specific incorporation is due to the two different possible configurations of the carboxamide moiety of the ribavirin pseudobase[34] and results in increased rates of C ⇀ U and G ⇀ A and, to a lesser extent, U ⇀ C and A ⇀ G transitions.[35] This erroneous incorporation causes viral genome mutations at subsequent cycles of replication of the erroneous template, resulting in increased mutation rate of the viral quasispecies.[36]

Experimental evidence for mutagenesis. Although first experiments employing radio-labeled ribavirin have indicated only low rates of incorporation,[37] later studies that compared mutation frequencies in test and control samples as a surrogate variable for ribavirin incorporation could determine significant substitution rates of about 0.015 substitutions per base in *in vitro* experiments. These substitutions resulted in about three ribavirin molecules per poliovirus genome per replication cycle, causing more than three times as many mutations per genome under high doses of ribavirin of 400 μM than without medication.[38] Due to the promiscuous base pairing of ribavirin (both cytidine and uridin are templated with about equal efficiency), incorporation was shown to particularly increase the frequency of transition mutations in viral replicates.[39] The observed increases in mutation rates were associated with reductions of viral infectivity and fitness to about 30% of the untreated population when the mutation frequency exceeded two mutations per genome, and to about 4% at more than 6 mutations per genome.[40]

These results parallel similar investigations pertaining to HIV and vesicular stomatitis virus that also reported about threefold increases in mutation rates drastically reducing viral viability.[41] Interestingly, however, mutagen monotherapy did not result in viral extinction and could only be achieved by additional measures that reduce the fitness of the viral quasispecies, for instance by serial passaging or addition of specifically acting antivirals.[42]

Anti-ribavirin drug resistance. Another type of evidence for the efficacy of ribavirin in HCV is the emergence of drug resistance agains ribavirin. Of particular interest in this regard is the identity of viral genes that are associated with the emergence of anti-ribavirin RAVs. Since HCV genes are specialized towards one or more, non-overlapping functions, the predominant mode of action of ribavirin in HCV should be associated with the effected genes, in principle.

[34] Graci and Cameron (2002)

[35] Vo et al. (2003)

[36] Crotty et al. (2001)

[37] Graci and Cameron (2002)

[38] Crotty et al. (2001, 2000)

[39] Maag et al. (2001)

[40] Crotty et al. (2001)

[41] Holland et al. (1990), Loeb and Mullins (2000)

[42] Loeb and Mullins (2000), Pariente et al. (2001)

The HCV NS5B RNA-dependent RNA polymerase is similar to other viral polymerase in that is contains finger, palm, and thumb domains with 7 conserved domains, including a unique 12-amino acid β-hairpin in the thumb domain.[43] A resistance mutation (F415Y) within the β-hairpin against ribavirin in the HCV polymerase that could narrow the putative RNA-binding pocket (and thus disable ribavirin incorporation) was reported based on *in vivo* data of non-responders. This finding is generally interpreted as supporting a direct inhibitory effect of ribavirin on the polymerase.[44] These results were confirmed in *in vitro* experiments[45] and additional mutations were proposed that possibly are induced by ribavirin resistance (V85I, K124E, and V85I/K124E).[46] In addition, low level resistance to ribavirin has also been associated with NS5A mutations G404S and E442G.[47]

However, while the aforementioned results as well as research concerning other viruses indeed indicate the emergence of anti-ribavirin drug resistance,[48] it remains unclear to which precise mode of action of ribavirin this resistance relates. In particular, it is still open if the resistance refers to the mutagenic activity rather than the inhibitory activity of the compound.[49] The matter is further complicated by the fact that viruses may implement several mechanisms of ribavirin resistance: while picornaviruses employ RAVs in order to increase the fidelity of the polymerase,[50] experimental evidence on other viral families suggests that ribavirin resistance is implemented by restricting ribavirin incorporation.[51]

As discussed previously, the question whether a mutagen has predominantly inhibitory or mutagenic properties is highly relevant for therapy optimization: while a dominantly inhibitory component would suggest combination therapy with other specifically acting antivirals, a mutagenic components is an indicator for sequential administration of specific antiviral and mutagen. The question of therapy serialization in the context of highly adaptive viral quasispecies is an active field of research; while this research currently focuses on non-retroviruses due to their simpler replication dynamics, extensions to other viral species are also investigated.[52]

Evidence for ribavirin-induced mutagenesis in HCV

In HCV, the mutagenic mode of action of ribavirin has been extensively investigated using both *in vitro* and *in vivo* experiments.[53] However, these results were not conclusive especially for *in vivo* experiments,[54] only some of which reported increased mutation rates in HCV NS5A/B genes. Especially analyses of ribavirin-monotherapy in comparison to an untreated placebo group, arguably the most meaningful way of isolating ribavirin activity, have resulted in conflicting results ranging from low or transient increases in mutation rates to more than two-fold increases.[55]

[43] Lesburg et al. (1999), Lutchman et al. (2007)

[44] Cao et al. (2011), Young et al. (2003)

[45] Hmwe et al. (2010)

[46] Nakamura et al. (2008)

[47] Pfeiffer and Kirkegaard (2005b)

[48] Feigelstock et al. (2011), Pfeiffer and Kirkegaard (2005b), Young et al. (2003)

[49] Domingo et al. (2012)

[50] Pfeiffer and Kirkegaard (2005a), Vignuzzi et al. (2006)

[51] Agudo et al. (2010), Arias et al. (2008), Sierra et al. (2007)

[52] Iranzo et al. (2011a)

[53] Contreras et al. (2002), Hofmann et al. (2007), Maag et al. (2001), Vallejo et al. (2004), Zhou et al. (2003)

[54] Asahina et al. (2005), Chevaliez et al. (2007), Schinkel et al. (2003)

[55] Cuevas et al. (2009), Hofmann et al. (2007), Lutchman et al. (2007), Young et al. (2003)

In vitro studies. Experiments employing *in vitro* technology are commonly based on the ribavirin-treated HCV replicon assays or, more recently, cell assays or cultures, followed by Sanger sequencing. While some studies support increases in mutation frequencies in these systems after ribavirin treatment,[56] other studies resulted in ambiguous or negative results.[57]

In all cases where evidence for ribavirin-associated increases in mutation rates could be identified *in vitro*, the corresponding fold increases were roughly comparable (2 to 3.5-fold) to related experiments on poliovirus. Critics of *in vitro* approaches to measuring the mutagenic effect of ribavirin in HCV point out that the transferability of *in vitro* results to *in vivo* conditions is severely limited;[58] in addition, the *in vitro* mutagenic effect of ribavirin was only observed at high concentrations between 400 and 1000 μM, far above the tolerable plasma concentration in patients of about 9 μM.[59]

In vivo studies. However, *in-vivo* studies for determining mutagenic effects of ribavirin on HCV are similarly conflicting:[60] while some investigations reported an increased mutation rate in individual viral genes,[61] other studies found only insignificant or no measured increases of mutations rates.[62] All studies presented here are inherently limited due to their restriction to only parts of the viral genome due to artifacts induced by the replicon system or the cost-aware use of sequencing technology that lead researchers to prioritize only few HCV genes such as NS3, NS5A, and NS5B. Although even modest increases of mutation rates of about 5% as identified by Hofmann et al. (2007), Young et al. (2003) may be sufficient to trigger viral error catastrophes,[63] depth of sequencing in experimental setups is often restricted to 20-80 clones per sample, thus severely limiting the ability to detect minority variants.

Shortcomings of present studies. All relevant *in vivo* studies on the mutagenic effect of ribavirin on the HCV quasispecies known to the author have significant flaws that limit the informative value of the results. A highly cited study[64] that aims to disprove ribavirin-induced mutagenesis in HCV may serve as illustration. This study investigated only four patients under ribavirin monotherapy and did not employ an untreated control group for comparison. The sequencing approach utilized 20 clones of the HCV NS3 and NS5A genes per sample. As a result of the low sequencing depth, the study is severely limited in resolution, especially with respect to the detection of minority variants. As a consequence of the few patients that were participating in the study, the results feature large within-sample variances of mutation frequencies, a fact that may be indicative of insufficient statistical power. Many or all of these flaws, most prominently the limitation to only about 10 sequence clones per sample, are also recurring in other studies that did find no or only weak evidence for ribavirin-mediated mutagenesis.[65]

[56] Brochot et al. (2007), Contreras et al. (2002), Hofmann et al. (2007), Lanford et al. (2003), Tanabe et al. (2004)

[57] Kato et al. (2005), Zhou et al. (2003).

[58] Cao et al. (2011)

[59] Dixit and Perelson (2006)

[60] Pawlotsky (2005)

[61] Asahina et al. (2005), Hofmann et al. (2007)

[62] Chevaliez et al. (2007), Lutchman et al. (2007), Mori et al. (2011), Schinkel et al. (2003), Young et al. (2003)

[63] Krieger et al. (2001), Lohmann et al. (2001)

[64] Chevaliez et al. (2007)

[65] Lutchman et al. (2007), Mori et al. (2011), Young et al. (2003)

Quantification of mutagenic effects

These varying outcomes of experimental studies aiming to investigate the mutagenic effect of ribavirin reflect the large difficulties that are associated with the detection of minute changes in mutation frequencies in sequencing data.

Since the true mutation rate of the viral polymerase cannot be conveniently observed, surrogate measures are employed that aim to quantify increased mutation frequencies of the viral quasispecies in drug-treated versus control subjects based on sequencing data. Several of these quantization methods are in current use.

Means of quantization. Mutation frequency is estimated in different ways from the sequencing data. In general, both *within-sample* methods, i.e., the the characterization of sequence diversity within a single sequence sample, and *between-sample* method, i.e., the computation of changes of sequence diversity between time points are employed to derive estimates for viral mutation frequencies at each viral genome position such as overall mutation frequencies or overall error rates (see side note for a more detailed description of these methods).[66]

Subsequently to averaging across all genome positions and normalization by the time difference between sampling points, the fold-changes of these estimates can be compared between mutagen-treated and control groups by means of inferential statistics such as χ^2 in order to determine the effects of ribavirin treatment.

Increasing the resolution of quantization. While the within-sample and between-sample methods are simple to compute and interpret, they either are not well suited for comparing mutation frequencies between time points (within-sample method) or are disregarding the nucleotide frequency distribution at the baseline time point (between-sample method). Both of these drawbacks have consequences on the qualities of the estimates that can be derived.

Omitting the comparison of nucleotide frequencies between time points (as the within-sample method does) can be considered to be a flawed approach for determining the effects of mutagens insofar as the mutation rates of viral quasispecies can differ between subjects and thus confound further analyses. As a consequence, nucleotide frequencies should be normalized by differential analysis of pre-treatment and post-treatment samples of the same subject.

Similarly, relying only on the consensus nucleotide of the baseline time point and disregarding the full nucleotide frequency distribution at this time point (the strategy of the between-sample method) is perilous since a shift in consensus sequence is not a good indicator for changes in sequence diversity, in general (see previous sections). In addition, this approach does not into consideration presence of minority genotypes at the baseline time point, thus hindering an unbiased estimation of mutation frequencies.

[66] Within both methods, absolute counts of sequence reads are converted to frequencies of nucleotides at each viral genome position i. This results in a discrete distribution F of frequencies f for the four nucleotides $n \in \{A, C, G, T\}$ (for instance, $F^i = \{A : 0.1, C : 0.4, G : 0.3, T : 0.2\}$).

Using the within-sample method, this nucleotide distribution is employed to estimate the mutation frequency m_n^i for each nucleotide based on a single sequence sample by first determining the *frequency of the consensus nucleotide* v^i, i.e., the mode of F^i, and then determining the *partial mutation frequencies* for each nucleotide $m_n^i = v^i - f_n^i | n \in \{A, C, G, T\}$ as well the *overall mutation frequency* $M^i = \sum_{n \in \{A,C,G,T\}} m_n^i$. Intuitively, the overall mutation frequency is zero at a position if the nucleotide distribution is clonal and only contains the consensus nucleotide; conversely, the maximal overall mutation frequency is 0.75.

This approach can be directly extended to compute the *error rate* e_i by the *between-sample* method; computations are analogous to the within-sample method but includes sequence data of two sampling time points t_1 and t_2, where t_1 is prior to treatment with the mutagen (*baseline*) while t_2 is after (or during) treatment (*treatment*). The error rate is then computed based on the nucleotide distribution $F_{t_2}^i$ (i.e., at the time of therapy) while the frequency of the consensus nucleotide is derived from the frequency distribution of the baseline data, i.e., $v_{t_1}^i$. From this follow the *partial error rates* $e_i = v_{t_1}^i - f_n^i | n \in \{A, C, G, T\}$ as well as the *overall error rate* E_i, again as a sum of the partials over all nucleotides (see above).

In order to confirm specific nucleotide substitution patterns for mutagens (e.g., C to T and G to A for ribavirin (Crotty et al., 2000, Vo et al., 2003)), s_n^i and e_n^i are only evaluated at positions where the consensus nucleotide corresponds to C or G and n corresponds to T or A, respectively. While in fact it is not clear if any shift in frequencies is indeed a result of nucleotide substitutions from the the consensus nucleotide to one of the target nucleotides T and A, this simplifying assumption is regularly made in related work (Chevaliez et al., 2007, Lutchman et al., 2007, Mori et al., 2011, Young et al., 2003)

The author and colleagues therefore recently proposed an alternative approach that employs the full nucleotide frequency distributions at both time points in order to quantify minute frequency changes with less dependence on the consensus nucleotide (see Section 24) We denote a substitution of a *source* nucleotide k with a *target* nucleotide l with $(k,l)|k \in \{A,C,G,T\}, l \in \{A,C,G,T\}, k \neq l$. In case of ribavirin, the substitutions of interest are (C,T) and (G,A).[67] By denoting two subsequent sampling time points as t_1 and t_2 where t_1 corresponds to pre-treatment sequence data and t_2 corresponds to post-treatment sequence data, the observed frequencies of a nucleotide n at a time point t and genome position i can be denoted by $f^t_{n,t}$. We define a quantifier δ for the change in relative frequencies of these substitutions as follows:

$$\delta^i_{(k,l)} = (f^i_{l,t_2} - f^i_{k,t_2}) - (f^i_{l,t_1} - f^i_{k,t_1}) \qquad (23.1)$$

Intuitively, δ ranges from 0 to 1 its value increases if the frequency of the target nucleotide in relation to the frequency of the source nucleotide increases at the second time point compared to the first time point. In order to consider multiple substitution patterns, we average $\delta^i_{(k,l)}$ across all such substitution patterns ((C,T) and (G,A) in the case of ribavirin, see Equation 24.2 in the next section for an example). This averaged value can then interpreted as an estimate of the mutagen-induced mutation frequency of the viral quasispecies. This estimate can further be aggregated across genome positions and subsequently compared between mutagen-treated and control groups using inferential statistics in order to statistically infer mutagenic effects of a purported mutagen.

Confounding effects of selection. Viral genotypes that are heavily mutated due to mutagenic action have a higher likelihood of being deleterious and are thus preferably removed from the population by selective forces. The observable rates of mutation frequencies are therefore bound to be significantly lower than the original mutation rate and thus do not accurately reflect the effect of the mutagen.

At first sight, a viable strategy for dealing with these purifying effects of natural selection may be based on determining frequencies of functionally neutral mutations, e.g., mutations that are synonymous at the amino-acid level.[68] Since the amino-acid sequences viral proteins remain constant even if synonymous positions mutate, these positions may exhibit fewer effects on fitness (and, therefore, less selection bias) than using non-synonymous positions.

However, since genomes of RNA viruses are densely packed with functional elements and often possess multiple reading frames or noncoding regions that determine properties of RNA secondary structure, even synonymously coding sites in these viruses are not truly neutral. In addition, the accumulation of neutral mutations is also depending on replication cycles, a number that is expected to be dependent on treatment response.[69]

[67] Crotty et al. (2000), Vo et al. (2003)

[68] Cuevas et al. (2009)

[69] Cuevas et al. (2009)

Importance of minority variants. Quantification of mutagenic effects based on sequencing if further complicated by the fact that population sequencing, i.e., determination of the *average* genotype of a viral quasispecies as still commonly done on viral diagnostics, is not informative for quantification of viral mutation frequencies.

For instance, high rates of nucleotide substitutions within the consensus sequence do not necessarily imply increased mutation rates of the underlying quasispecies; instead, they may merely indicate an increased fixation of selectively advantageous variants in the consensus sequence.[70] This phenomenon is a consequence of the fact that changes observable on the level of the consensus sequence are a product of not only mutation but more importantly of competition and selection within the quasispecies population: since deleterious variants are regularly removed from the population, the true viral mutation rate is likely to be significantly higher than the mutation frequency or the consensus sequence may indicate.

Conversely, low rates of observed change of the consensus sequence at rates of 10^{-1} substitutions per site and year (common values for well adapted viral quasispecies) are not a sign of reduced mutation rates. This has been demonstrated in the context of of mutagenic drugs where invariant consensus sequences where measured in spite of significant changes in viral mutation rates.[71]

These phenomena generally confound approaches to quantifying mutation rates that employ low-sensitivity Sanger technology. Instead, a deeper look at the viral population structure is required in order to measure minute mutagenic effects. This is currently only possible if low-frequency minority variants of the viral quasispecies, as, for example, determined by deep sequencing, are considered during the analysis.

Difficulty of estimating minority variants. Increases in low-frequency mutation frequencies are notoriously hard to measure precisely with limited-dilution Sanger sequencing, a method which exhibits limited sensitivity at the numbers of clones that are usually analyzed.[72] Next-generation or deep sequencing offers significantly higher sensitivity than Sanger approaches but displays high technical error rates that may mask ribavirin activity (cf. Chapter II).

The background mutation rate of the HCV quasispecies is about 10^{-5} substitutions per site and copy.[73] Experimental evidence from other organisms such as polio report clinically relevant ribavirin-induced increases of mutation frequencies of about 3-fold. A similar analysis of the HCV quasispecies would require detection sensitivities about equal to the mutation rate of 10^{-5}. However, even advanced statistical approaches that aim to detect increases of minority frequencies based on comparative approaches and model sequencing error require sequencing depths of more than the inverse of the expected frequency for successful detection.[74] Consequently, sequencing depths of about 10,000 fold are required to reliably detect ribavirin-induced mutations in HCV.

[70] Cuevas et al. (2009)

[71] González-López et al. (2005), Grande-Pérez et al. (2005), Iranzo and Manrubia (2008)

[72] Palmer et al. (2004), Salazar-Gonzalez et al. (2008), Zhang et al. (1991), Zhu et al. (1993)

[73] Zeuzem et al. (1998)

[74] Gerstung et al. (2012)

Differential analysis of minority variants. In order to quantify changes in mutation frequencies as a function of ribavirin administration, the viral quasispecies of multiple patient groups (treated versus placebo) and time points (prior to administration versus at end of therapy) are commonly sequenced and mutation frequencies are quantified.

Importantly, sequencing data may include technical errors that may mask true signatures of ribavirin incorporation (see following section for a more detailled discussion of error modes). Consequently, downstream analysis should utilize statistical models that consider the error rates of the sequencing process in order to (1) detect minority variants within the quasispecies and delineate these variants from technical errors, and to (2) statistically compare subsequent time points within each patient in order to decide if a observed increase in frequency is due to technical noise or true biological variation.

Finally, these estimates of increased frequencies and their statistical confidences have to be employed to statistically compare ribavirin-treated groups from placebo-treated groups in order to associate the mutagenic effect with the treatment.

Determination of minority variants. Viral minority variants are usually determined by deep sequencing,[75] a technology enables the parallel measurement of billions of nucleotide bases at low cost.[76] In contrast to *vanilla* Sanger technologies, deep sequencing does not result in an average genotype of a given heterogeneous population of genomes. Instead, it affords the determination of the identity of individual DNA fragments, thus providing a digital view on a large portion of that population. However, due to the elevated error rates of current, second-generation sequencing technology on the order of 1%, statistical and algorithmically approaches that differentiate technical sequencing error from true minority variants are required.[77]

In particular, the study of subclonal human cancers has highlighted the importance of minority variants for clinically relevant phenotypes associated with growth, resistance, and proliferation of cancer.[78] Novel mutations that occurred during tumorgenesis and proliferation, so called *somatic* mutations, are of interest for tumor subtyping and the estimation of its probably evolutionary trajectory – information highly important for precise diagnostics and therapy optimization.[79] As a consequence of this interest in heterogeneous cancers, a variety of computational approaches have been developed for identifying minority variants in mixed samples. Many of these approaches are also applicable to viral quasispecies and are thus discussed briefly here.

[75] Domingo (2006), Domingo and Wain-Hobson (2009), García-Arriaza et al. (2007), Más et al. (2010), Webster et al. (2009), Wright et al. (2011), Zagordi et al. (2010)

[76] Shendure and Ji (2008)

[77] Beerenwinkel et al. (2012), Eriksson et al. (2008)

[78] Ding et al. (2012), Gerlinger et al. (2012), Harismendy et al. (2011), Inda et al. (2010)

[79] Boyd (2013)

Sequencing errors. Deep sequencing approaches afford high throughput and theoretically also high sensitivity; the latter, however, is confounded by the relatively high error rate resulting from library preparation and sequencing artifacts.[80] In combination, these error sources result in per-base error rates of the whole sequencing process of about 1%.[81]

Interestingly, local characteristics of genome sequences induce errors also in multiple experimental replicates that have been processed independently.[82] These sequence-specific and strand-specific errors have been identified, for instance around G-rich motifs. While greater sequencing depth may facilitate detection of many errors due to the associated increased power of statistical approaches, in principle, sequence-specific errors tend to occur consistently at any depth and are therefore difficult to remove by any approach.[83] In addition to considering platform-specific errors, modeling these local sequence characteristics has especially important consequences for successful identification of genetic variation.[84] In particular, variants predominantly occurring on one strand have been shown to more often correspond to sequencing errors than equally distributed SNVs.[85]

Use of specialized sequencing libraries. Microbiological and chemical techniques[86] may be employed to reduce the error rate of the library generation and sequencing process. Artifacts such as PCR duplicates can be addressed by random tagging approaches that discriminate between multiple copies of the same molecule, thus allowing for the correction of errors and better identification of minority variants, in principle.[87] As a result, considerably reduced error rates of about 0.001% are achievable in special cases.[88] Still, this error rate is at least 1,000-fold higher than the mutation rates of human cells,[89] the latter being only recently approximated by independently tagging and sequencing each of the two strands of a DNA duplex.[90] However, these approaches require skilled personnel and may exhibit reduced scalability compared to established high-throughput library preparation techniques. As a consequence, identification of tumor minority variants is commonly undertaken on standard deep sequencing sequencing libraries.

Minority variant calling In order to reliably estimate minority variants from noisy second-generation deep sequencing data, approaches towards modeling intratumor heterogeneity as resulting from sample impurity and genomic subclonality have been proposed.[91] These *comparative*, *subclonal*, or *somatic* variant callers employ more sensitive statistical models than standard variation callers that are limited to differentiating homozygous variants from heterozygous variants in diploid genomes.

[80] Goren et al. (2010), Kanagawa (2002), Meyerhans et al. (1990), Shendure and Ji (2008)

[81] Gundry and Vijg (2012)

[82] Harismendy et al. (2009)

[83] Meacham et al. (2011)

[84] Harismendy et al. (2009), Pop and Salzberg (2008)

[85] Chapman et al. (2011), Gerstung et al. (2012), Varela et al. (2011)

[86] Flaherty et al. (2012), Kozarewa et al. (2009)

[87] Hamady et al. (2008), Hiatt et al. (2010, 2013), Kinde et al. (2011), McCloskey et al. (2007), Miner et al. (2004)

[88] Kinde et al. (2011)

[89] Cervantes et al. (2002), Roach et al. (2010)

[90] Schmitt et al. (2012)

[91] Lee et al. (2010), Ley et al. (2008)

Interestingly, even in relatively simple scenarios that exclusively involve sequence data of diploid organisms and that lacksubclonal variation, concordance between diploid variant callers is low (about 60%). This indicates the inherent difficulty of distinguishing sequencing errors from true variation in general.[92]

Deducing minority variants occurring at frequencies comparable to technical error rates of the library generation and sequencing process seems to be daunting and indeed has been denoted as impractical in the general sense.[93] These error rates differ not only between sequencing platforms but also between sequencing runs. Furthermore, physical placement of the sequencing product on the machine (lanes or flowcells), multiplexing technology, properties of the sequenced genomes, genomic location of a read, and error type.[94] Due to this variety of factors involved, error rates are highly non-uniform with respect to their location on sequencing reads and genomic positions.

Approaches to variant calling. By exploiting this non-uniformity, approaches that employ high sequencing coverage and adapt error models to the sequencing data may resolve many minority variants at frequencies lower that the average error rate.[95] In addition to error modeling and employing more sensitive statistical models compared to diploid variant callers, comparative variant callers employ a third trick to increase sensitivity: the combined sequencing and analysis of a *test* sample and a *control* sample of the same patient. This analysis affords both the determination of local error rates of the sequencing process and the identification of variants that are unique to the test sample.[96]

While in cancer sequences test and control samples correspond to tumor tissue and healthy tissue, respectively, virological applications often employ viral samples from two different time points or with different degrees of subclonality. Estimated error rates are then compared with the frequencies of putative minority variants by binomial[97] and Bayesian likelihood models.

Comparative variant calling tools. In the last three years, several comparative variant callers have been proposed for estimating minority variants at low frequencies of about 1% from sequencing data of subclonal cancers[98] that employ related analysis principles and provide increased sensitivity with regard to variant callers designed for diploid genomes.[99] These comparative variation callers claim accurate calls of minority variants of frequencies less than 0.05%,[100] significantly less than the average error rate of sequencing.

A particularly interesting comparative variant caller that displays best-in-class performance at recovering minority variants based on deep sequencing data is DeepSNV.[101] As it is one of the few tools specifically designed for inferring SNVs in subclonal populations with an unknown number of clones, the software is also applicable to viral quasispecies, in principle.

[92] 1000 Genomes Project Consortium et al. (2012), Mills et al. (2006)

[93] Wright et al. (2011)

[94] Minoche et al. (2011), Nakamura et al. (2011), Suzuki et al. (2011)

[95] Wilm et al. (2012)

[96] Cibulskis et al. (2013), Gerstung et al. (2012), Koboldt et al. (2012)

[97] Or often also by beta-binomial models that directly work with count data, better account for the high variance of biological data (overdispersion). The hyperparameters of the betabinomial model can be estimated from the data using likelihood techniques

[98] Christoforides et al. (2013), Cibulskis et al. (2013), Gerstung et al. (2012), Goya et al. (2010), Kim et al. (2013), Koboldt et al. (2012), Larson et al. (2012), Saunders et al. (2012), Shiraishi et al. (2013), Wei et al. (2011), Wilm et al. (2012), Yost et al. (2013)

[99] Li (2011), McKenna et al. (2010), Xu et al. (2012a)

[100] Wilm et al. (2012)

[101] Gerstung et al. (2012)

Similar to related methods, DeepSNV relies on a control, for instance originating from pre-treatment samples or a clonal aliquot of the same virus, in order to infer change rates of nucleotide frequencies at each genomic position. Utilization of a clonal control sample is desirable in cases where the existence of minority variants in the test sample is the target of inquiry. However, the statistical approach in general also allows for selection of non-clonal samples if changes in clonality at specific sequence positions are investigated with respect to a combined background of technical errors and variation in the control sample. Indeed, similar approaches are suggested by the authors for the use on whole exome sequencing data.

DeepSNV models nucleotide counts based on deep sequencing read data of test and control samples at each genomic position using a betabinomial model. This betabinomial model is parametrized by an overdispersion factor that is derived from the control data using a maximum likelihood approach. SNV frequencies in the test data that differ significantly from corresponding frequencies in the control set are determined based on binomial models of each sequence strand by a likelihood ratio test statistic.

Since the test statistic is χ^2 distributed, P-values can be computed for nucleotide frequencies of both nucleotide strands and for data of both samples. For each nucleotide but the consensus nucleotide, P-values of each strand and sample are then combined into a single P value using Fisher's method followed by correction for multiple testing. The combined P-value reflects the uncertainty of an increased variant frequency at a specific genomic position.

Preliminary conclusion

In conclusion, this chapter has presented a view on the viral quasispecies as highly adaptable and robust system that is evolutionary optimized to withstand antiviral treatment. However, by combining drugs with different modes of action that influence the structure of the quasispecies population, novel therapy options seem possible that may increase viral response rates. Optimization of these combination treatments requires a detailed characterization of the molecular modes of action of the drugs employed. One class of drugs whose mode of action presently is not well understood are antiviral mutagens such as *ribavirin*. Evaluation of the proposed mutagenic effect of ribavirin on the HCV quasispecies requires ultradeep sequencing, a technology that exhibits error rates of the same order of magnitude as the purported mutagenic effects of the drug. These error rates can be significantly lowered by bioinformatics analyses that employ sensitive statistical models and suitable control samples.

24 HCV and mode of action of ribavirin

This last section *synthesizes much of the content of this chapter in order to best introduce an investigation for detecting minute changes of population structures of HCV viral genomes upon treatment with the therapeutic mutagen ribavirin. Ribavirin displays several modes of action in other highly divergent viral species, including mutagenic activity. As discussed previously, the mutagenic effect of antivirals is important for informing therapy decisions and coordinating the administration of drug combinations involving directly acting antivirals in order to limit emergence of drug resistance and maximize therapy efficacy. However, a possible mutagenic effect of ribavirin on the HCV quasispecies is contested and rational therapy optimization involving ribavirin is therefore impeded.*

We undertook the first deep-sequencing analyses of HCV samples in order to detect mutagenic activity based on time-series data of patients under ribavirin monotherapy. In contrast to related studies, we firstly sequenced the HCV quasispecies of patients under ribavirin monotherapy at sufficient depth in order to detect minority variants at significantly lower frequencies than was previously possible using clonal Sanger sequencing. In addition, we employed amplicon sequencing of the full HCV genome at multiple time points and using two different sequencing platforms in order to obtain a broad overview on mutation rates across genomic sites. Finally, we performed a statistical filtering procedure that compares quasispecies distributions across time points in a statistically meaningful fashion and takes sequencing errors into account.

Acknowledgments. The study presents joint work of S.-E. Schelhorn with J. Dietz, D. Fitting, U. Mihm, S. Susser, M.-W. Welker, C. Füller, Martin Däumer, G. Teuber, H. Wedemeyer, T. Berg, T. Lengauer, S. Zeuzem, E. Herrmann and C. Sarrazin and has been published in the Journal of Virology. J. Dietz and S.-E. Schelhorn are joint first authors of this publication.[1] Apart from formatting changes and few stylistic corrections such as shifting citation marks to the end of sentences for better readability, the manuscript presented here is identical to the published version. The study was supported by a DFG grant (Klinische Forschergruppe, KFO 129). S.-E. Schelhorn was responsible for designing and performing all bioinformatical investigations as well as informing, but not conducting, the subsequent statistical analyses. In addition, S.-E. Schelhorn co-wrote the manuscript together with J. Dietz.

[1] Dietz et al. (2013)

Introduction

Abstract. THE PREEMINENT MODE OF ACTION of the broad-spectrum antiviral nucleoside ribavirin in the therapy of chronic hepatitis C is currently unresolved. Particularly under contest are possible mutagenic effects of ribavirin that may lead to viral extinction by lethal mutagenesis of the hepatitis C virus (HCV) genome. We applied ultradeep sequencing to determine ribavirin-induced sequence changes in the HCV coding region (nucleotides [nt] 330 to 9351) of patients treated with 6-week ribavirin monotherapy ($n = 6$) in comparison to placebo (n = 6). Baseline HCV RNA levels maximally declined on average by -0.8 or -0.1 $\log_{10}$ IU/ml in ribavirin- versus placebo-treated patients. No general increase in rates of nucleotide substitutions in ribavirin-treated patients was observed. However, more HCV genome positions with high G-to-A and C-to-U transition rates were detected between baseline and treatment week 6 in ribavirin-treated patients in comparison to placebo-treated patients (rate of 0.0041 transitions per base pair versus rate of 0.0022 transitions per base pair; p = 0.049). Similarly, the sensitive detection of low-frequency minority variants by statistical filtering indicated significantly more positions with G-to-A and C-to-U transitions in ribavirin-treated patients than in placebo-treated patients (rate of 0.0331 transitions versus rate of 0.0186 transitions per G/C-containing position at baseline; p = 0.018). In contrast, non-ribavirin-associated A-to-G and U-to-C transitions were not enriched in the ribavirin group (p = 0.152). We conclude that ribavirin exerts a mutagenic effect on the virus in patients with chronic hepatitis C by facilitating G-to-A and C-to-U nucleotide transitions.

Introduction. The guanosine analogue ribavirin (1-β-D-ribofuranosyl-1,2,4-triazole-3-carboxamide) displays broad antiviral activity against RNA and DNA viruses in vitro and is used for the treatment of hepatitis C virus (HCV), respiratory syncytial virus (RSV), and Lassa fever virus infections.[2] Monotherapy of patients with chronic HCV infection with ribavirin causes declining serum aminotransferase levels but leads to only a moderate and transient reduction of HCV RNA levels.[3] In spite of the inefficacy of ribavirin monotherapy, the drug acts synergistically with (pegylated) alpha interferon (IFN-α) during anti-HCV therapy, resulting in roughly 3-fold-enhanced sustained virologic response (SVR) rates compared to those resulting from interferon monotherapy.[4] Similar effects of ribavirin were observed for triple therapies of HCV genotype 1-infected patients that included direct antiviral agents. Here the addition of ribavirin to a NS3/4A protease inhibitor and pegylated IFN-α (PEG-IFN-α) resulted in increased antiviral efficacy and reduced viral breakthroughs associated with resistant viral variants.[5] Furthermore, even in the context of a combination of two direct antiviral agents in patients with chronic hepatitis C which led to frequent treatment failure, the addition of ribavirin

[2] Crotty et al. (2002), Sidwell et al. (1972), Streeter et al. (1973)

[3] Di Bisceglie et al. (1992), Pawlotsky et al. (2004)

[4] McHutchison et al. (1998), Poynard et al. (1998)

[5] Hézode et al. (2009), Kwo et al. (2010), Shimakami et al. (2009)

without PEG-IFN-α enhanced the initial virus decline and reduced the number of viral breakthrough events.[6]

The preeminent mode of action of ribavirin in the therapy of chronic hepatitis C is unresolved. It was recently shown that ribavirin in conjunction with IFN-α/PEG-IFN-α induces the expression of specific interferon-stimulated genes (ISGs) in vitro and in vivo, thereby potentiating the anti-HCV effect of interferon.[7] Other possible mechanisms of ribavirin action include the following models: strengthening of the adaptive antiviral immune response, impairment of the cellular enzyme IMP dehydrogenase (IMPDH), direct inhibition of the HCV nonstructural 5B (NS5B) RNA-dependent RNA polymerase, and exertion of a mutagenic effect on RNA viruses and the resulting error catastrophe.[8] The lack of proofreading activity of the HCV polymerase results in a population of divergent but closely related viruses, termed viral quasispecies, that optimizes viral evolutionary fitness by maximizing genetic variation. Quasispecies are assumed to exist at the edge of a genomic error threshold.[9] Exceeding this error threshold may lead the quasispecies into a sequence of error catastrophes, termed lethal mutagenesis, that results in viral extinction.[10] The application of a model poliovirus polymerase showed that the incorporation of ribavirin templates the incorporation of cytidine and uridine, leading to a mutagenic effect which coincided with reduced poliovirus infectivity as well as with the observation of G-to-A and C-to-U transitions in mutagenized genomes.[11]

Ribavirin-induced mutagenesis of the HCV genome has been demonstrated in vitro based on sequenced isolates from ribavirin-treated HCV cell culture systems.[12] Furthermore, HCV cell culture experiments revealed the selection of several HCV mutations[13] as well as changes in the cell line,[14] both conferring ribavirin resistance. In vivo analyses of ribavirin-induced mutations in the HCV genome, on the other hand, remain inconclusive. Several studies reported selective mutations in NS5A/B as well as increased nucleotide substitution rates consistent with G-to-A and C-to-U nucleotide transitions in NS3/NS5B in patients under ribavirin monotherapy.[15] Also, in patients undergoing therapy with IFN-α plus ribavirin, an increased mutation rate and a mutational spectrum with increased G-to-A and C-to-U transitions were detected during treatment based on E1/E2 and NS5A sequencing studies.[16]

[6] Zeuzem et al. (2012)

[7] Feld et al. (2007), Rotman et al. (2013), Thomas et al. (2011)

[8] Feld and Hoofnagle (2005), Hofmann et al. (2008)

[9] Eigen (1993a)

[10] Anderson et al. (2004), Hofmann et al. (2008)

[11] Crotty et al. (2000)

[12] Brochot et al. (2007), Contreras et al. (2002), Kanda et al. (2004)

[13] Feigelstock et al. (2011), Pfeiffer and Kirkegaard (2005b)

[14] Pfeiffer and Kirkegaard (2005b)

[15] Asahina et al. (2005), Hofmann et al. (2007), Young et al. (2003)

[16] Cuevas et al. (2009)

A greater nucleotide sequence variation with an increase in C-to-U transitions within NS3 and NS5B was also observed for patients treated with a NS3 inhibitor and a NS5B inhibitor in combination with ribavirin-containing regimens.(17) In contrast, other results showed only transient(18) or no(19) increases in substitution rates in NS5B and NS3/4A, respectively. Selection for a ribavirin-associated resistance mutation in NS5B that is suggestive of a mutagenic effect of ribavirin(20) was not verified in nonresponder patients who were treated with ribavirin in combination with IFN-α or PEG-IFN-α.(21) In this study, we employed ultradeep sequencing of the HCV coding region (nucleotides [nt] 330 to 9351) in order to investigate the ribavirin-induced mutagenesis of the viral quasispecies in detail over time in 12 patients under ribavirin monotherapy versus placebo.

Materials and methods

Patients. In a prospective, randomized, placebo-controlled study, 68 patients with chronic HCV genotype 1 infection were randomized and treated at the University Hospitals in Frankfurt, Berlin, Hannover, and Homburg/Saar, Germany, as well as at an Independent Medical Center in Frankfurt between 2007 and 2010. Initially, for 6 weeks, patients received either placebo or ribavirin at 1,000 to 1,200 mg per day according to body weight or 180 μg PEG-IFN-α2a per week. Subsequently, all patients received antiviral treatment according to the standard of care (180 μg PEG-IFN-α-2a once weekly plus $1,000$ to $1,200$ mg ribavirin daily, body weight adapted).(22) For deep-sequencing analysis of HCV quasispecies, serum samples during monotherapy were selected from 12 patients (HCV subtype 1b $[n = 6]$, ribavirin $[n = 6]$, and placebo) at baseline (before treatment) and at treatment day 42. Enrollment in the clinical study as well as the usage of patient serum samples for HCV sequencing studies were approved by the local ethics committee, and written informed consent was obtained from all patients. Quantitative HCV RNA measurement was performed by using a commercially available assay (COBAS AmpliPrep/COBAS TaqMan HCV test; Roche Diagnostics).

HCV RNA extraction, reverse transcription, and PCR. For HCV RNA extraction, 140 μ L of serum was used (QIAamp viral RNA minikit; Qiagen, Hilden, Germany). cDNA synthesis was performed in triplicates by using SuperScript III reverse transcriptase (Invitrogen) with random or specific primers and 8μ L of viral RNA corresponding to 0.5 μ g of viral RNA on average. Amplification of the HCV genome from the 5′ N-terminal region (NTR) to NS5B (nt 145 to 9351) occurred in 5′ overlapping PCR amplicons with gene-specific primers for outer and inner nested PCRs. Patient-specific primers were designed after sequencing and alignment of the primer binding regions of all patients. Nested PCRs were con-

[17] Hebner et al. (2011)
[18] Lutchman et al. (2007)
[19] Chevaliez et al. (2007)
[20] Young et al. (2003)
[21] Ward et al. (2008)
[22] Mihm et al. (2013)

ducted with 1/20 of cDNA or outer PCR product, using the Expand High FidelityPlus PCR system (Roche Applied Science) containing a DNA polymerase and a proofreading protein. The resulting amplicons were analyzed for correct size and purity on 0.8% agarose gels stained with ethidium bromide.

Deep sequencing. For a first analysis, four patients (ribavirin-treated patients 1 and 2 and placebo-treated patients 7 and 8) were selected for 454 deep sequencing. Based on the results for the comparison of mutational frequencies among baseline, day 7, day 21, and day 42 of treatment with ribavirin versus placebo, eight additional patients (ribavirin-treated patients 3 to 6 and placebo-treated patients 9 to 12) were subsequently enrolled for deep sequencing analysis (Illumina technology) at baseline and day 42. Per patient and time point, five HCV amplicons were generated for deep sequencing analysis. All amplicons were purified by using Agencourt CleanSeq beads on a BioMek NX workstation (Beckman Coulter), quantified fluorometrically on a FluoStar Optima instrument (BMG Labtech) by using Quant-iT Picogreen double-stranded DNA (dsDNA) reagent (Invitrogen), and sample-specific amplicons were pooled equimolarly for library preparation.

For 454 deep sequencing, amplicons were fragmented by nebulization. Next, a sizing solution was applied to remove fragments shorter than 400 bp on a BioMek NX workstation, and the size distribution of the DNA was confirmed by a 2100 Bioanalyzer (Agilent Technologies). DNA adaptors containing primer binding sites for deep sequencing as well as multiplex identifiers (MIDs) for sample bar coding were ligated into the purified DNA fragments by using a GS FLX Titanium Rapid Library Preparation kit (Roche Applied Science). The library was subjected to emulsion PCR for clonal amplification of DNA fragments on water-in-oil emulsion microreactors followed by enrichment and counting of DNA containing beads (GS FLX Titanium LV emPCR kit [Lib-L] and GS FLX Titanium emPCR breaking kit LV/MV 12pc; Roche Applied Science). Subsequently, microbeads were collected and loaded onto the PicoTiter plate of the FLX Genome Sequencer (Roche Applied Science). 454 FLX second-generation sequencing technology with an average read length of 400 bp was performed according to the manufacturer's protocols by using the GS FLX Titanium sequencing kit (Roche Applied Science).

For the preparation of libraries for Illumina deep sequencing, equimolarly pooled amplicons were "tagmentated" (fragmented and tagged) by using a Nextera DNA sample preparation and index kit (Illumina) according to the manufacturer's manual. DNA fragments shorter than 400 bp were removed as described above. Resulting libraries were quantified on a 2100 Bioanalyzer (Agilent Technologies) and diluted to 10 pM for cluster generation and subsequent sequencing on an Illumina MiSeq platform using the paired-end sequencing protocol for 2 × 250 bp runs.

The theoretical mean coverages were calculated to be approximately 4,900 reads by using the 454 platform and 8,200 reads by using Illumina deep sequencing technology, assuming optimal quantity and length of generated reads (see Tables 24.1 and 24.2 for the actual coverages). To compare 454 and Illumina deep sequencing, one patient sample was sequenced at baseline and at day 42 with both platforms, and the generated reads were compared. Although variants determined by the 454 and Illumina platforms were highly correlated, we computed substitution and nucleotide transition rates exclusively on pairs of samples that underwent identical library preparation and that were sequenced on the same platform and the same sequencing run. After nucleotide transition and substitution rates were thus generated in a platform-dependent manner, these quantities were combined and subjected to further statistical analysis.

Patient	Time	Core	E1	E2	P7	NS2	NS3	NS4A	NS4B	NS5A	NS5B	all regions
Rbv Pat1	bl	7.883	4.792	4.189	4.416	4.582	6.285	2.743	7.849	4.643	6.722	5.410
Rbv Pat1	d42	5.321	3.533	4.542	6.245	5.762	7.562	6.154	11.095	6.336	9.365	6.592
Rbv Pat2	bl	3.972	2.944	3.554	4.636	4.310	4.720	1.866	5.567	3.988	4.883	4.044
Rbv Pat2	d42	4.571	3.560	5.171	7.536	6.679	6.425	2.510	8.563	6.595	9.506	6.112
Rbv	total	5.437	3.707	4.364	5.708	5.334	6.248	3.318	8.268	5.390	7.619	5.539
Plac Pat7	bl	4.411	2.905	2.662	2.841	3.350	3.041	1.131	4.917	3.504	3.690	3.245
Plac Pat7	d42	6.249	3.892	3.579	4.279	4.552	5.342	1.932	7.846	5.708	6.155	4.953
Plac Pat8	bl	4.618	3.246	3.500	4.756	4.726	5.370	2.157	6.654	4.593	5.389	4.501
Plac Pat8	d42	4.412	3.519	3.577	4.330	4.110	4.306	1.657	5.271	4.080	5.390	4.065
Plac	total	4.923	3.390	3.329	4.051	4.184	4.515	1.719	6.172	4.471	5.156	4.191
total	total	5.180	3.549	3.847	4.880	4.759	5.381	2.519	7.220	4.931	6.388	4.865

Table 24.1: *Average nucleotide depths after 454 deep sequencing according to HCV regions.*

Mapping of deep sequencing data. A standard flowgram format (SFF) file containing the nucleotide sequence reads was generated for each sequenced sample by the 454 sequencing software (GS Run Processor; Roche Applied Sciences) provided with the instrument. Reads produced by Illumina software were provided in FASTQ format. Subsequently, Phred quality scores were extracted from the reads, and primer sequences were removed from the start of the reads. For 454 reads, bases succeeding the first base call with a Phred quality score below 10 were trimmed, as reported in previous studies,[23] and reads produced by the Illumina platform were trimmed to a Phred quality score of 20. We note that Phred scores have a slightly different interpretation on both platforms, denoting the probability of a miscalled base and incorrect read length for Illumina and 454 technologies, respectively. Reads with no "N" base calls and a length of more than 25 bases were retained, and quality control metrics were computed for the trimmed reads to ensure consistent high quality across all samples. The resulting high-quality reads were mapped to the HCV-J reference genome[24] with the gapped read mapper SMALT (Wellcome Trust Sanger Center), the successor of the popular long-read mapper SSAHA2[25] that

[23] Zagordi et al. (2010)

[24] Kato et al. (1990)

[25] Ning et al. (2001)

displays increased specificity and sensitivity according to both the original authors and external validations.[26] Base insertions and deletions, both common errors in deep sequencing, were detected and marked in the alignment and were not included in subsequent evaluations. Subsequently, nucleotide variant distributions were called for each variant position in the alignment (nt 330 to 9351) by the SAMtools suite.[27] The uniformity of the base distribution at each alignment position was quantified with an entropy measure. HCV sequences from patient samples taken at baseline and during ribavirin monotherapy and placebo treatment, respectively, were compared.

Patient	Time	Core	E1	E2	P7	NS2	NS3	NS4A	NS4B	NS5A	NS5B	all regions
Rbv Pat3	bl	3.002	8.074	12.337	19.464	12.725	6.413	5.423	9.409	13.625	8.631	9.910
Rbv Pat3	d42	880	1.911	7.055	20.546	14.905	9.164	9.558	5.323	4.088	385	7.382
Rbv Pat4	bl	3.297	3.266	8.126	19.160	14.929	7.437	5.275	3.326	1.241	6.049	7.211
Rbv Pat4	d42	4.271	4.814	11.583	27.182	18.961	10.082	7.321	9.476	13.450	2.679	10.982
Rbv Pat5	bl	4.694	5.648	11.416	35.134	24.446	6.857	3.722	5.319	6.643	6.082	10.996
Rbv Pat5	d42	6.744	7.835	9.763	21.851	14.837	3.761	1.449	2.162	3.120	19.824	9.135
Rbv Pat6	bl	3.418	4.434	5.725	6.721	5.354	4.522	4.402	3.252	3.573	11.733	5.313
Rbv Pat6	d42	798	929	2.070	4.860	3.924	1.483	1.172	1.191	1.872	4.678	2.298
Rbv	total	3.388	4.614	8.509	19.365	13.760	6.215	4.790	4.932	5.951	7.508	7.903
Plac Pat9	bl	12.861	21.154	22.951	16.042	12.538	5.619	4.530	6.145	8.555	8.967	11.936
Plac Pat9	d42	6.295	9.426	11.272	11.607	8.896	7.683	8.200	8.190	8.157	7.131	8.686
Plac Pat10	bl	2.177	2.816	3.297	5.927	4.545	1.594	1.312	2.210	3.093	6.871	3.384
Plac Pat10	d42	1.663	2.422	3.507	6.820	5.499	3.286	3.532	2.827	2.247	2.712	3.452
Plac Pat11	bl	3.975	5.496	5.208	5.586	4.747	3.375	3.678	4.153	4.677	1.746	4.264
Plac Pat11	d42	10.822	15.457	12.204	4.172	3.259	3.740	3.615	8.037	10.592	5.842	7.774
Plac Pat12	bl	5.380	6.708	5.942	3.426	3.654	3.374	3.170	4.218	4.726	11.623	5.222
Plac Pat12	d42	1.475	1.807	4.499	10.077	8.968	4.735	3.687	4.351	5.219	23.974	6.879
Plac	total	5.581	8.161	8.610	7.957	6.513	4.176	3.966	5.017	5.908	8.608	6.450
total	total	4.484	6.387	8.560	13.661	10.137	5.195	4.378	4.974	5.930	8.058	7.176

Table 24.2: *Average nucleotide depths after Illumina deep sequencing according to HCV regions.*

Variant calling. For each base position of an aligned read set (HCV coding region at nt 330 to 9351) at two time points, t_1 and t_2 (baseline and day 42, respectively), we consider the frequencies of the individual bases A, C, G, and T. The predominant substitution rate, S, at a specific position describes the dominant base mutation comparing time points t_1 and t_2:

$$S(t_1, t_2) = max\Big(|\text{Freq}_{t1}(B) - \text{Freq}_{t2}(B)|\Big) \qquad (24.1)$$

(where B is A, C, G, or U).

Based on previous approaches,[28] at each position, an indicator, T, of ribavirin-supported transition rates comparing time points t_1 and t_2 is defined as follows:

[26] Oliver (2012)

[27] Li et al. (2009a)

[28] Hofmann et al. (2007), Young et al. (2003)

$$\begin{aligned} T(t_1,t_2) = \frac{1}{2} \Big(& \left[\mathrm{Freq}_{t_2}(A) - \mathrm{Freq}_{t_2}(G)\right] \\ - & \left[\mathrm{Freq}_{t_1}(A) - \mathrm{Freq}_{t_1}(G)\right] \\ + & \left[\mathrm{Freq}_{t_2}(U) - \mathrm{Freq}_{t_2}(C)\right] \\ - & \left[\mathrm{Freq}_{t_1}(U) - \mathrm{Freq}_{t_1}(C)\right]\Big) \end{aligned} \tag{24.2}$$

Transitions from G to A and C to U are considered to be facilitated by ribavirin.[29] A cutoff value of 0.4 for the evaluation of substitution rates and transition rates was chosen from a first evaluation of a first patient treated with ribavirin compared with a patient treated with placebo by comparing several possible values. Unfortunately, the limited number of patients in this study does not allow for a refined optimization of the cutoff levels.

Investigation of deep sequencing error rates. In order to determine the technical error rates of library preparation and deep sequencing using 454 and Illumina platforms, amplicon 3 (2,228 bp fragment ranging from NS3 to NS4B) of one patient and time point was cloned into a pSC-A-amp/kan vector and transformed into Escherichia coli competent cells (StrataClone PCR cloning kit; Agilent Technologies). The plasmid DNA was purified (QIAprep Spin miniprep kit; Qiagen) and Sanger sequenced with M13 forward and reverse as well as internal template-specific primers according to the manufacturer's protocol (BigDye Terminator v1.1 cycle sequencing kit; Applied Biosystems), on an ABI Prism 3130xl genetic analyzer (Applied Biosystems). The insert of one sequenced plasmid clone was excised from the vector with the restriction enzyme EcoRI and was deep sequenced in parallel with the other PCR amplicon samples. In accordance with other studies,[30] the technical error rate was calculated by counting all nucleotide variants of the plasmid reads in the alignment that did not correspond to the sequence of the clone determined by Sanger sequencing. While insertion errors were subject to automatic removal during the mapping of the sequencing reads to the HCV-J reference genome,[31] deletions with respect to the reference sequence were detected during the mapping and quantified as errors but excluded from all further analyses.

Minority variant calling based on a statistical filtering procedure. Low-frequency nucleotide variants were distinguished from technical errors by utilizing the statistical filtering procedure deepSNV (where SNV is single-nucleotide variants).[32] The filtering procedure estimates sequence-specific and strand-specific error rates to derive a position-wise test statistic for reliably identifying low-frequency minority variants that display increased frequencies in a test sample compared to a given control. Using this method, we computed FDR (false discovery rate)-corrected p-values for all nucleotide variants. To calculate ribavirin-supported G-to-A and C-

[29] Asahina et al. (2005), Contreras et al. (2002), Crotty et al. (2000), Cuevas et al. (2009), Pfeiffer and Kirkegaard (2005b), Young et al. (2003)

[30] Gilles et al. (2011), Hedskog et al. (2010), Kagan et al. (2012), Niu et al. (2010), Wang et al. (2007), Zagordi et al. (2010)

[31] Kato et al. (1990)

[32] Gerstung et al. (2012)

to-U nucleotide transitions between two time points, only positions within the HCV coding region (nt 330 to 9351) containing a G or C in the consensus baseline sequence were considered. At these positions, variants were considered for further analysis if they showed a significant increase of A or U frequencies, respectively, between two time points, as characterized by a position-wise p-value of the statistical filtering procedure below a cutoff of 0.001. The computation of non-ribavirin-associated A-to-G and U-to-C transitions occurred analogously to the calculation described above. Thereby, variants with a significant increase of G or C between both time points were considered at positions with an A or U in the consensus baseline sequence.

Statistical analysis. The predominant substitution rate, S, and the nucleotide transition rate, T, of ribavirin- and placebo-treated patients were compared by random-effect models for these rates and testing for group effects. Statistical analysis was conducted based on the metaprop procedure of the meta package by Guido Schwarzer for R software (R Foundation for Statistical Computing, Vienna, Austria). p-values of less than 0.05 were considered significant.

Deep sequencing data access. The deep sequencing data from this study have been deposited in the Sequence Read Archive (SRA) under accession number ERP001566.

Results and discussion

HCV RNA kinetics and analysis of deep sequencing. For the investigation of ribavirin-induced mutations, we used clinical serum samples from HCV genotype 1-infected patients (see Materials and Methods). We selected an overall number of 12 patients ($n = 6$ for ribavirin; $n = 6$ for placebo) with the same HCV genotype (GT1b) and similar HCV RNA concentrations (between around 1×10^6 and 1.5×10^7 IU/ml) before initiation of therapy (baseline) for deep sequencing analysis of the HCV quasispecies. As it is not known whether a mutagenic effect of ribavirin would be reflected by a decline of the HCV RNA concentration, we chose patients under ribavirin monotherapy or placebo treatment who displayed various HCV kinetics (Fig. 24.1A).

On average, baseline HCV RNA levels maximally decreased by -0.8 or $-0.1\,\log_{10}$ IU/ml in ribavirin- versus placebo-treated patients. 454 deep sequencing of the HCV quasispecies from four patients (see Materials and Methods) during the 6-week monotherapy resulted in an average of $160,250$ sequenced reads per sample, with a mean read length of 316.6 bp after removing low-quality bases. A total of 99.1% of high-quality reads could be unambiguously aligned to the HCV-J reference genome, resulting in a mean minimum coverage (averaged over all patients and positions) of $4,394.3$

(range, 597 to 17, 542) reads per position for the 454 platform (Fig. 24.1B).

Samples from further patients included in the study (ribavirin-treated patients 3 to 6 and placebo-treated patients 9 to 12) were sequenced by using the Illumina platform, which produced an average of 527,968 high-quality reads per sample, with a mean read length of 133.4 bp after clipping of bases with low quality. The alignment of 94.9% of high-quality reads to the HCV-J reference genome[33] gave rise to a mean minimum coverage of 4,649.2 (range, 0 to 27,033) reads per position (Fig. 24.1C). We observed a decreased coverage of the Illumina sequencing reads between positions 1500 and 1550 (E2) and at position 8627 (NS5B). The reduced coverage within E2 is attributed mainly to hypervariable region 1 (HVR-1), which is known to produce difficulties during the alignment process due to its high variability. Furthermore, the Nextera method used for library preparation for Illumina deep sequencing was shown to produce biased data sets.[34] As position 8627 is located before a large G/C stretch, this may influence the recovery of these sequences during library preparation. Nonetheless, during statistical filtering, positions with low coverages are rejected from further analyses.

To compare the sequencing results produced by 454 and Illumina deep sequencing, two samples (baseline and day 42) from one patient were each processed with both platforms. The nucleotide ratios of these samples were compared between the 454 and the Illumina data, and an R^2 correlation resulted in a correlation of >0.995 for the ratios of all four nucleotides (A, C, G, and T). Alignment of the deep sequencing reads produced nucleotide counts of the four possible bases (A, C, G, and U) at every position of the HCV-J reference genome.[35] These nucleotide counts were analyzed for each sample and compared between baseline and day 42 in order to investigate a potential ribavirin-induced increase in the number of nucleotide substitutions.

Ribavirin does not generally induce mutations in the HCV genome. To determine whether ribavirin treatment is associated with the fixation of mutations in the HCV genome, we estimated S, the change rate of the predominant substitution between two consecutive sampling time points (see Materials and Methods). S values exceeding a 40% threshold, indicating large changes of the HCV quasispecies, were observed at slightly more HCV genome positions in patients undergoing ribavirin monotherapy than in patients receiving placebo between baseline and day 42 (rate of 0.0085 substitutions per base pair versus rate of 0.0052 substitutions per base pair) (Table 24.3 and Fig. 24.2). However, no significant differences between the ribavirin group and the placebo group were detected ($p = 0.230$), and no local clustering of S values within the HCV genome of patients treated with ribavirin was detected (Fig. 24.2). This indicates that ribavirin does not lead to a generally increased

[33] Kato et al. (1990)

[34] Quail et al. (2012)

[35] Kato et al. (1990)

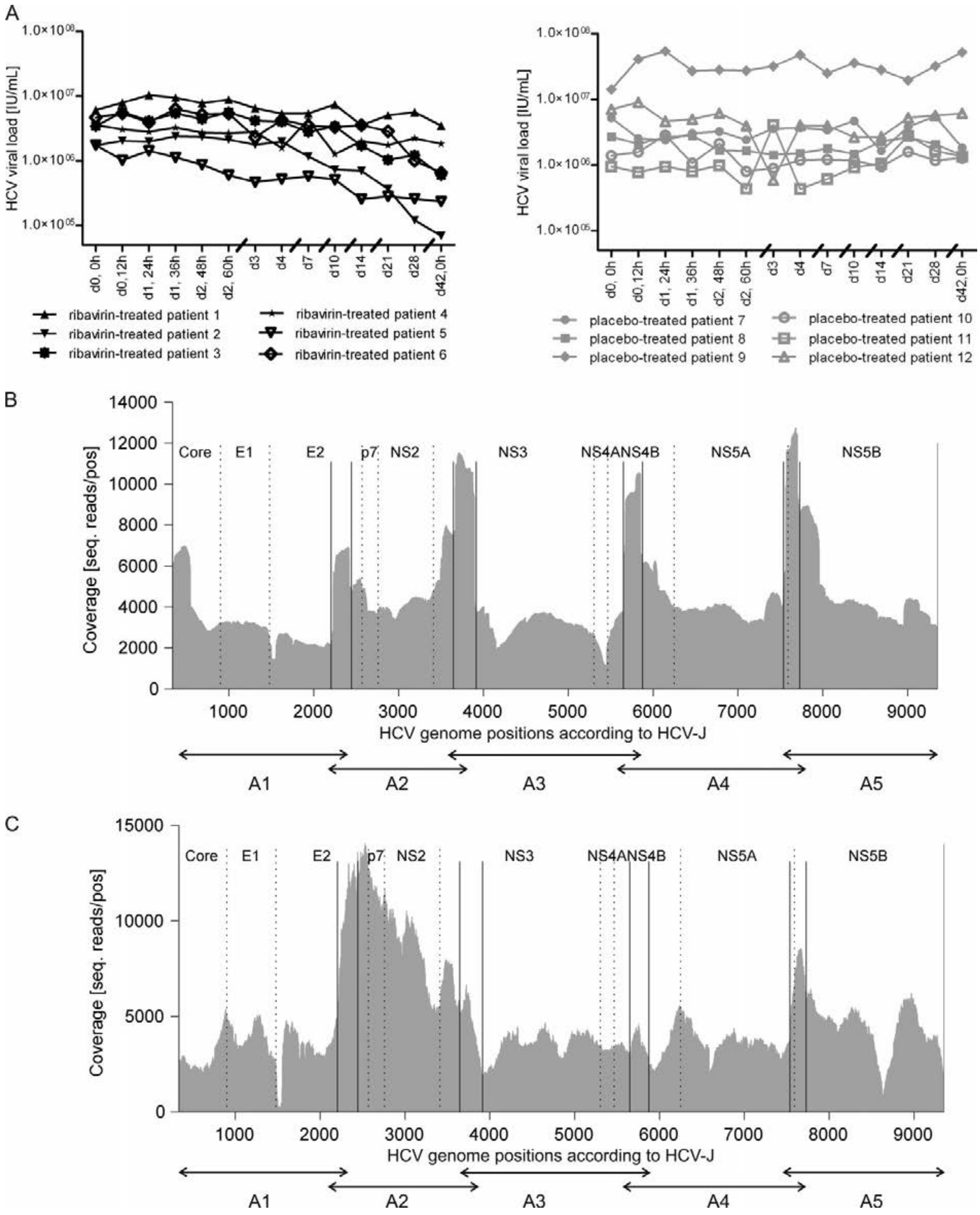


Figure 24.1: *Deep sequencing of HCV quasispecies.* Deep sequencing of HCV quasispecies of 12 patients during the 6-week monotherapy phase (n = 6 for ribavirin; n = 6 for placebo). For each patient, samples were sequenced at two time points (baseline and day 42) during monotherapy. (A) HCV RNA kinetics for selected patients during monotherapy. (B) Minimal coverage of the HCV genome (nt 330 to 9351) after 454 deep sequencing of PCR amplicons (A1 to A5) averaged over two patients who received ribavirin monotherapy and two patients who received placebo. (C) Minimal coverage of the HCV genome (nt 330 to 9351) after Illumina deep sequencing of PCR amplicons (A1 to A5) of ribavirin-treated (n = 4) and placebo-treated (n = 4) patients.

mutation rate.

Patient[a]	S (per 9,022 nt)[b]	T (per 9,022 nt)[b]
Rbv Pat1	98	47
Rbv Pat2	143	72
Rbv Pat3	94	42
Rbv Pat4	18	4
Rbv Pat5	134	62
Rbv Pat6	48	24
Plac Pat7	57	24
Plac Pat8	39	14
Plac Pat9	21	10
Plac Pat10	156	44
Plac Pat11	41	19
Plac Pat12	35	18

Statistical measure	Group[a]	Value
Avg. rate of S (per bp)	Rbv	0.0085
Avg. rate of T (per bp)	Rbv	0.0041
95% confidence interval for S	Rbv	0.0056-0.0127
95% confidence interval for T	Rbv	0.0027-0.0063
Avg. rate of S (per bp)	Plac	0.0052
Avg. rate of T (per bp)	Plac	0.0022
95% confidence interval for S	Plac	0.0027-0.0102
95% confidence interval for T	Plac	0.0014-0.0034
p-value for S	Rbv vs. Plac	0.230
p-value for T	Rbv vs. Plac	0.049

Table 24.3: *Comparison of predominant substitution rates and transition rates of HCV quasispecies.* Comparison of predominant substitution rates and transition rates of HCV quasispecies between patients treated with ribavirin and those treated with placebo between baseline and day 42 of the study. S, substitution rate; T, transition rate. [a] Rbv Pat, patient given ribavirin; Plac Pat, patient given placebo. [b] Indicated is the number of HCV genome positions with changes in the respective parameter exceeding a 0.4 cutoff threshold referring to the entire analyzed HCV coding region (nt 330 to 9351).

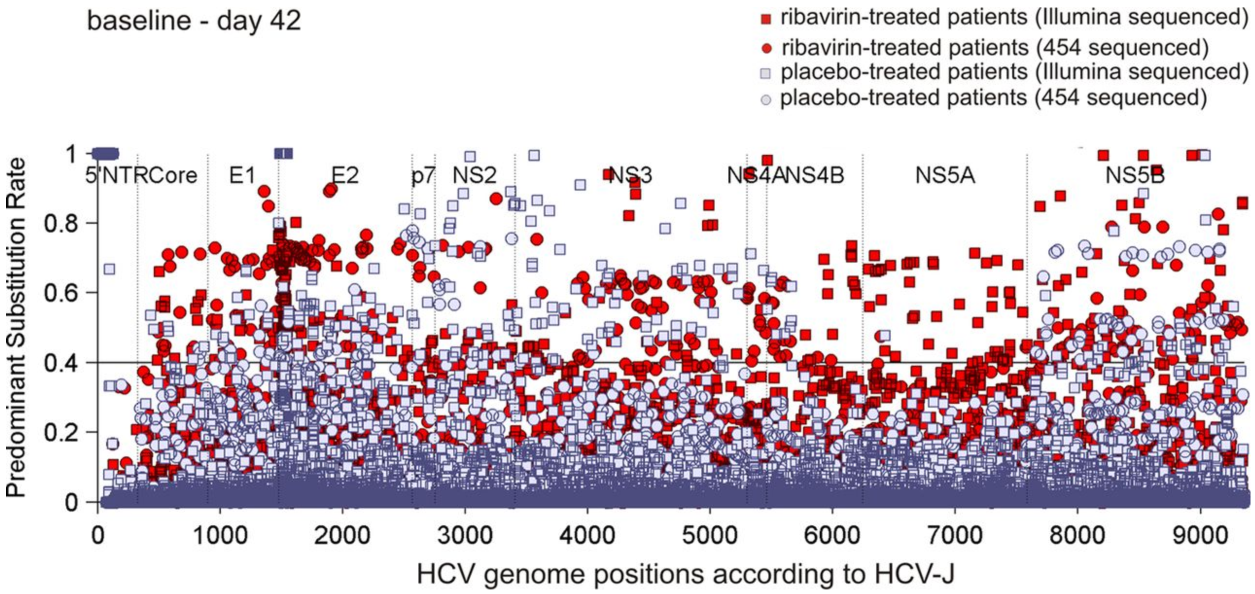


Figure 24.2: *HCV genome positions for ribavirin- versus placebo-treated patients.* HCV genome positions for ribavirin- versus placebo-treated patients with changes in predominant substitution rates. The horizontal line at the 0.4 level represents the cutoff used for the evaluation of the predominant substitution rate, S, between baseline and day 42.

Accumulation of ribavirin-induced nucleotide transitions. Misincorporation of the guanosine analogue ribavirin in the viral RNA results in mutation of the viral genome by acting as a nonspecific nucleotide template for the incorporation of both cytidine and uridine. This mutagenic effect may lead the viral quasispecies into error catastrophe and lethal mutagenesis.[36] Increased nucleotide

[36] Crotty et al. (2000), Hofmann et al. (2007)

substitution rates and accumulation of G-to-A and C-to-U transitions within NS3 and NS5B of patients with chronic HCV infection receiving ribavirin monotherapy support this hypothesis.[37] We computed the change in substitution rates of ribavirin-associated transitions, T (G to A and C to U), between baseline and day 42 by expanding on approaches introduced in previous work[38] (see Materials and Methods). Patients under ribavirin monotherapy exhibited significantly more HCV genome positions with large ($T > 40\%$) increases in rates of ribavirin-associated transitions than the placebo group (rate of 0.0041 transitions per base pair versus rate of 0.0022 transitions per base pair; $p = 0.049$) (Table 24.3 and Fig. 24.3) between baseline and day 42. An enhancement of ribavirin-associated transitions in ribavirin-treated patients was still observable after small alterations of the 0.4 cutoff level. To characterize the mutation spectrum at baseline, quasispecies entropy was calculated for all samples at baseline. Entropy was determined to be highly similar for the ribavirin and placebo groups, indicating similar mutation spectra at baseline. Furthermore, the frequency of positions with a C/G in the baseline consensus sequences was comparable between ribavirin- and placebo-treated patients (data not shown). Moreover, we did not observe any favored localization of ribavirin-associated transitions within the HCV genome of ribavirin-treated patients (Fig. 24.3).

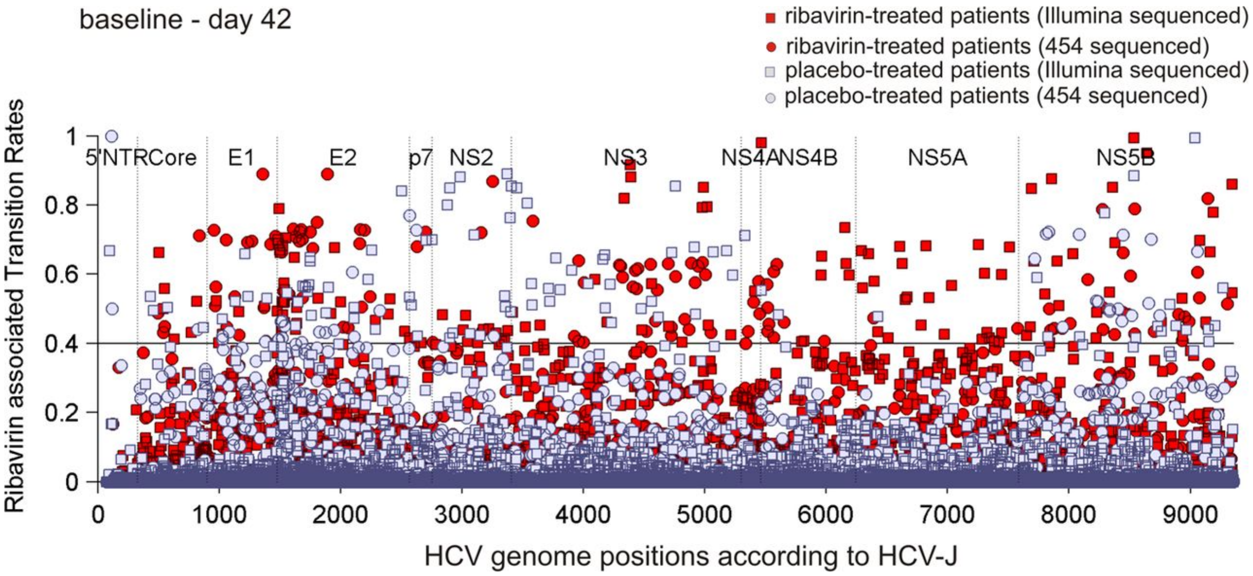


Figure 24.3: *Comparison of HCV genome positions with changes in nucleotide transition rates.* Comparison of HCV genome positions with changes in nucleotide transition rates, T, comprising G-to-A and C-to-U transitions between patients receiving ribavirin and those receiving placebo, respectively. The cutoff used for the assessment of the nucleotide transition rate, T, between baseline and day 42 is depicted by the horizontal line at the 0.4 level.

Analysis after statistical filtering procedure. In order to differentiate low-frequency nucleotide variants from technical errors, we applied an involved statistical filtering procedure that allows for the identification of such minority variants with high sensitivity and specificity.[39] Using this method, p-values were computed for each HCV genome position, quantifying the significance of occurring nu-

[37] Hofmann et al. (2007)

[38] Hofmann et al. (2007), Young et al. (2003)

[39] Gerstung et al. (2012)

cleotide substitutions between two time points. Ribavirin-supported G-to-A and C-to-U nucleotide transitions were calculated by considering only positions which exhibited a G or a C in the consensus sequence at baseline and displayed a significant ($\alpha = 0.001$) increase in A and U frequencies, respectively, between baseline and day 42. Based on the application of this strategy for filtering of deep sequencing data, we confirmed that patients under ribavirin monotherapy display significantly more HCV genome positions with significant G-to-A and C-to-U transitions than patients treated with placebo (rate of 0.0331 transitions versus rate of 0.0186 transitions per G/C-containing position at baseline; $p = 0.018$) (Table 24.4). Furthermore, ribavirin-induced transitions did not display a preferred nucleotide signature in ribavirin-treated patients in comparison to placebo-treated patients, indicating that these transitions do not occur in a certain context of surrounding nucleotides (data not shown).

Patient[a]	Rbv transitions[b]
Rbv Pat1	148/5,307
Rbv Pat2	105/5,319
Rbv Pat3	339/5,307
Rbv Pat4	111/5,334
Rbv Pat5	209/5,360
Rbv Pat6	234/5,292
Plac Pat7	60/5,271
Plac Pat8	91/5,267
Plac Pat9	106/5,289
Plac Pat10	87/5,273
Plac Pat11	174/5,255
Plac Pat12	96/5,276

Statistical measure	Group[a]	Value
Avg. rate of Rbv transitions (per bp)[c]	Rbv	0.0331
Avg. rate of Rbv transitions (per bp)[c]	Plac	0.0186
95% confidence interval for rate	Rbv	0.0228-0.0479
95% confidence interval for rate	Plac	0.0138-0.0250
p-value	Rbv vs. Plac	0.018

Table 24.4: *Number of HCV genome positions from patients treated with ribavirin and placebo after the application of statistical filtering.* Statistical filtering considers positions with a significant increase ($\alpha = 0.001$) in A and U frequencies between baseline and day 42 related to the number of positions showing a G or C, respectively, in the baseline consensus sequence. [a] Rbv Pat, patient given ribavirin; Plac Pat, patient given placebo. [b] Number of ribavirin- associated transitions per G/C-containing position. [c] Rate of ribavirin-associated transitions per G/C-containing bp.

We intended to determine whether differences from ribavirin-supported transitions originate from many small or from large transitions. Therefore, we analyzed the magnitude and the corresponding frequencies of G-to-A and C-to-U transitions and compared the results before and after application of the data-filtering strategy (Fig. 24.4A and 24.4B). Both before and after filtering of deep sequencing data, ribavirin-treated patients displayed an overall increase in the number of G-to-A and C-to-U transitions in comparison to placebo-treated patients between baseline and day 42 that encompassed small, medium, and large changes in ribavirin-associated transitions, T. Application of the statistical filtering procedure especially eliminated variants with nonsignificant small changes, affirming the importance of this method for differentiating genuine variants from technical errors. However, when small nucleotide transition changes (including changes of <40%) were included in the analysis without statistical filtering, we observed no significant differences between the ribavirin and the placebo groups, indicating that ribavirin-associated transitions are generated predominantly via medium and large nucleotide changes. Transitions in the reverse direction (A-to-G and U-to-C) were not significantly enhanced in ribavirin-treated patients in comparison to placebo-treated patients (rate of 0.0184 transitions versus rate of 0.0128 transitions per A/U-containing position at baseline; $p = 0.152$) (Table 24.5) between baseline and day 42. Therefore, the mutagenic effect of ribavirin is attributed mainly to the generation of G-to-A and C-to-U transitions.

Patient[a]	Rbv transitions[b]
Rbv Pat1	85/3,712
Rbv Pat2	48/3,698
Rbv Pat3	98/3,711
Rbv Pat4	25/3,684
Rbv Pat5	108/3,656
Rbv Pat6	76/3,725
Plac Pat7	19/3,748
Plac Pat8	56/3,753
Plac Pat9	48/3,732
Plac Pat10	82/3,686
Plac Pat11	71/3,766
Plac Pat12	34/3,744

Statistical measure	Group[a]	Value
Avg. rate of Rbv transitions (per bp)[c]	Rbv	0.0184
Avg. rate of Rbv transitions (per bp)[c]	Plac	0.0128
95% confidence interval for rate	Rbv	0.0131-0.0256
95% confidence interval for rate	Plac	0.0090-0.0183
p-value	Rbv vs. Plac	0.152

Table 24.5: *Number of HCV genome positions from patients treated with ribavirin and placebo after application of statistical filtering.* Statistical filtering considers positions with a significant increase ($\alpha = 0.001$) in G and C frequencies between baseline and day 42 related to the number of positions showing an A or U, respectively, in the baseline consensus sequence. [a] Rbv Pat, patient given ribavirin; Plac Pat, patient given placebo. [b] Number of ribavirin- associated transitions per A/U-containing position. [c] Rate of ribavirin-associated transitions per A/U-containing bp.

Determination of deep sequencing error rates. In order to estimate the technical error rate of our sequencing approach, a plasmid-amplified clonal sample of a fragment ranging from NS3 to NS4B was sequenced by using 454 and Illumina technologies with the same parameters as those used for samples from the main study. Sequencing reads and nucleotide distributions from this positive control were compared with the Sanger sequence corresponding to the clone, which also matched the consensus sequences generated from the 454 and Illumina reads. Sanger sequencing is commonly used as a comparator for the determination of deep sequencing errors[(40)] and is expected to have an error rate of 0.01% to 1%, depending on the software applied.[(41)] All sequenced nucleotides differing from the Sanger sequence were considered technical errors attributable to library preparation and the sequencing process. Following this protocol, we estimated a technical error rate of 0.507% erroneous substitutions per sequenced nucleotide for 454 deep sequencing and an error rate of 0.036% erroneous substitutions per sequenced nucleotide for Illumina deep sequencing. These rates include mismatches (454, 0.158%; Illumina, 0.035%) and deletions

[(40)] Hedskog et al. (2010), Kagan et al. (2012), Wang et al. (2007), Zagordi et al. (2010)

[(41)] Hoff (2009), Keith et al. (1993), Noguchi et al. (2006)

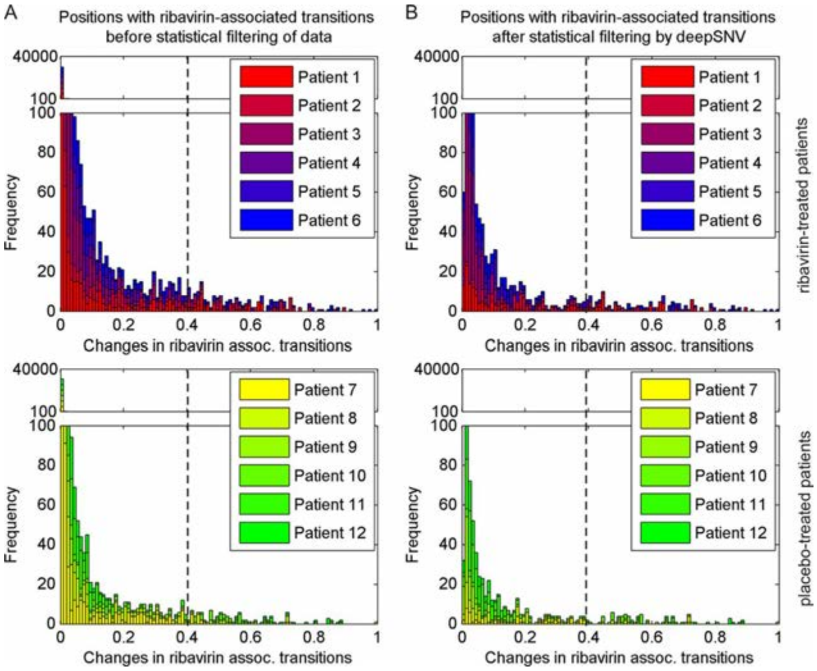


Figure 24.4: *Occurrence of ribavirin-associated G-to-A and C-to-U transitions.* Occurrence of ribavirin-associated G-to-A and C-to-U transitions between baseline and day 42 in the HCV genome of ribavirin- and placebo-treated patients. The vertical dashed line at the 0.4 level displays the cutoff, which was applied to differentiate small versus medium/large changes in ribavirin-associated transitions. (A) Frequency of HCV genome positions with changes in ribavirin-associated transitions before statistical filtering of deep sequencing data. All HCV genome positions with transition changes are shown. (B) Frequency of HCV genome positions with changes in ribavirin-associated transitions after application of statistical filtering. HCV genome positions with significant transition changes are represented.

(454, 0.349%; Illumina, 0.001%) while excluding insertion errors due to the particularities of the alignment process (see Materials and Methods).

Discussion. Existing studies on the preeminent mode of action of ribavirin in HCV therapy and particularly the significance of ribavirin-induced mutagenesis remained inconclusive. Several in vitro analyses of poliovirus, GB virus B, hantaan virus, and foot-and-mouth disease virus[42] showed that ribavirin exhibits mutagenic properties. In contrast, studies investigating ribavirin-induced HCV mutagenesis in patients receiving ribavirin monotherapy concentrated mainly on the analysis of small regions of the HCV genome (NS3, NS5A, and NS5B) via standard clonal sequencing or direct sequencing of PCR products at a depth of about 30 clones per time point, which yielded contradictory results.[43] We performed deep sequencing of the complete HCV coding region of patients with chronic hepatitis C undergoing ribavirin monotherapy in order to analyze systematically and with high sensitivity whether ribavirin induces nucleotide substitutions. An increased mutation rate may indicate an error catastrophe, which is followed by viral extinction. Alternatively, it is also conceivable that viruses which are generated in the presence of ribavirin display decreased infectivity and reduced replication capacities,[44] which could lead, in combination with interferon, to the antiviral effect against HCV.[45] Our results do not reveal significantly higher frequencies of predominant substitutions in ribavirin-treated patients, indicating that the number of mutations is not generally increased. As during ribavirin monotherapy, only a slight reduction of the HCV load is

[42] Chung et al. (2007), Crotty et al. (2001, 2000), Lanford et al. (2001), Pfeiffer and Kirkegaard (2005b), Sierra et al. (2007)

[43] Asahina et al. (2005), Chevaliez et al. (2007), Hofmann et al. (2007), Lutchman et al. (2007), Young et al. (2003)

[44] Dixit and Perelson (2006)

[45] Dixit et al. (2004)

detectable, a continuous accumulation of mutations with a consecutive increasing probability for the generation of defective viruses is presumably not taking place. The mutagenic effect of ribavirin seems to lead rather to the continuous production of new viruses with increased numbers of ribavirin-induced mutations incorporated, which exhibit slightly reduced replication capacity and/or infectivity, thereby explaining the weak HCV load reduction. This may result in a continuous process with a constant exchange of viral variants. In the present study, we demonstrate that during the 6-week ribavirin monotherapy, the rate of G-to-A and C-to-U transitions, which are specifically induced by the guanosine analogue ribavirin, is enhanced within the HCV genome of ribavirin-treated patients in comparison to placebo-treated patients. In a subgroup of four patients, additional time points between baseline and day 42 (day 7 and day 21) were investigated. In line with our hypothesis, here we observed significant mutational differences between each time point without an overall constant increase in the rate of mutations from baseline to day 42 (data not shown).

The observed transitions were not limited to NS3 and NS5B, as reported previously,[46] but occurred throughout the full HCV genome, with no preferential location of transitions in certain HCV genome regions.

In principle, ultradeep sequencing allows for the sensitive detection of HCV variants. Nevertheless, it is critical to differentiate minor variants from technical errors associated with library preparation and the sequencing process. By sequencing a clonal fragment, we determined for the 454 platform and the Illumina platform error rates in good accordance with data from previous studies.[47] Although specialized software for additional error correction exists,[48] we refrained from employing it in this setting due to the risk of discarding genuine minority variants. Instead, we applied an alternative approach that explicitly incorporates error modeling into the calculation of changes in variant frequencies between two time points (see below). Furthermore, a quality control of deep sequencing reads ensured that only reads with high-quality scores and sufficient read length were used for mapping and downstream analyses (see Materials and Methods). In order to discriminate low-frequency nucleotide variants from technical errors, we employed deepSNV, a statistical filtering procedure that calculates the significance of changes in variant frequencies between two time points for each HCV genome position. Although this method was validated on virus data generated by the Illumina platform, the underlying mathematical model of deepSNV does not make assumptions about the sequencing platform or platform-specific error patterns and is thus also generally applicable to 454 data (N. Beerenwinkel, personal communication). To estimate the genuine number of ribavirin-supported G-to-A and C-to-U transitions, only positions with significant increases in these transitions were considered. This additional analysis confirms our results, further

[46] Hofmann et al. (2007)

[47] Gilles et al. (2011), Kagan et al. (2012), Loman et al. (2012), Niu et al. (2010), Quail et al. (2012)

[48] Hedskog et al. (2010), Zagordi et al. (2010)

supporting the conclusion that patients under ribavirin monotherapy exhibit significantly more HCV genome positions with G-to-A and C-to-U transitions than do placebo-treated patients. Moreover, non-ribavirin-associated A-to-G and U-to-C transitions were not significantly enriched in ribavirin-treated patients, indicating that ribavirin specifically induces G-to-A and C-to-U transitions.

Both analyses (with and without statistical filtering) show that ribavirin-supported transitions occur as small, medium, and large changes in nucleotide substitutions between two time points. Differences in medium- and high-frequency transitions (>40% change) between both patient groups could readily be detected without statistical filtering of the deep sequencing data. In ribavirin-treated patients, small transition changes (including changes of <40%) were not significantly enriched, suggesting that ribavirin-associated transitions originate mainly from medium and large nucleotide changes. However, statistical filtering was essential for distinguishing low-level transitions from technical errors in both patient groups, underlining the importance of the statistical filtering method, particularly for the sensitive detection of minority variants.

Nevertheless, our study has several limitations. Due to difficulties of full-length HCV genome amplification and the high costs of deep sequencing of the entire HCV genome, we were able to sequence the HCV genomes of only a limited number of patients. Therefore, our results do not allow for a direct conclusion on the extent of the effect of ribavirin-induced mutagenesis on virologic response. The clinical results of the study in which the patients analyzed here were included are reported elsewhere.[49] Furthermore, we have not yet performed functional analyses to clarify the potential mechanism of action of ribavirin-induced mutagenesis.

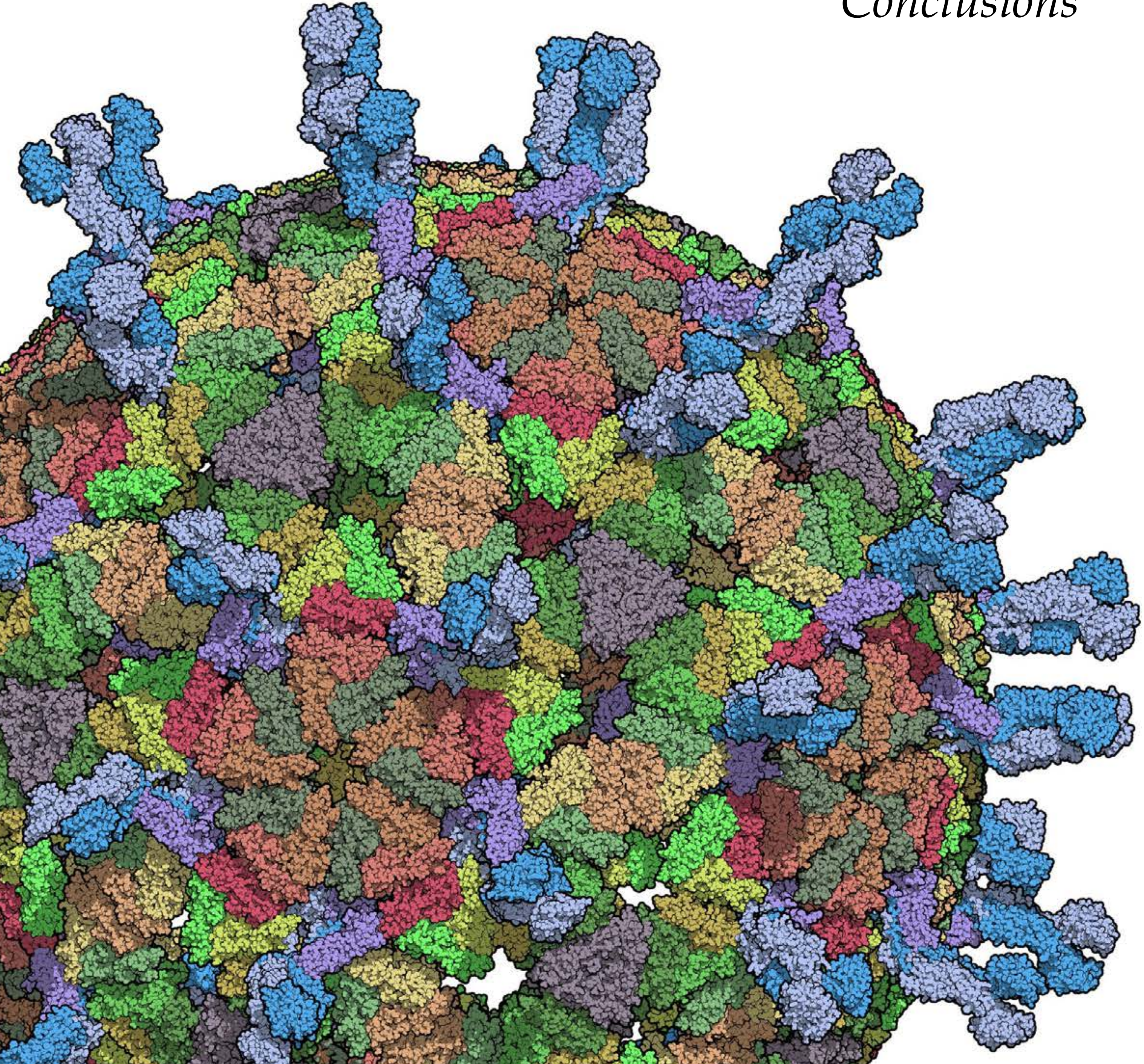
Finally, our findings do not readily explain the enhanced virological response rates for treatment with IFN-α in combination with ribavirin. It is conceivable that the clinical effect of ribavirin may be at least partially mediated by resetting the IFN responsiveness in the liver.[50] In conclusion, this explorative study analyzed ribavirin-induced mutations with high sensitivity and demonstrated that ribavirin induces nucleotide transitions during monotherapy. This effect seems to be a relevant factor for the antiviral activity of ribavirin, which is independent of the additional application of PEG-IFN-α and may explain the efficiency of ribavirin in combination with other direct-acting antivirals. The observed mutagenesis of the HCV genome in patients undergoing ribavirin monotherapy is based on the generation of G-to-A and C-to-U transitions. Ribavirin-induced mutagenesis does not explain the normalization of serum aminotransferase levels observed in patients undergoing ribavirin monotherapy, which we also detected in the patients analyzed here. Therefore, other mechanisms of action are likely also involved in the antiviral activity of ribavirin.

[49] Mihm et al. (2013)

[50] Rotman et al. (2013)

V

Conclusions



25 *Summary*

This thesis *presented an integrated view on the analysis of virological data that spans (1) early detection of emergent viral pathogens, (2) antiviral drug target discovery, and (3) the clinical treatment of genetically complex viral disease. Three organizing principles were proposed by the author to structure this work. First among these themes was the aforementioned integrated view on antiviral data analysis. Second was the introduction of particularly sensitive methods for the detection of faint biological signatures such as viral nucleotide sequences, human-viral protein interactions, or low-frequency mutations in viral genomes amidst experimental noise. The last and most speculative theme concerned subtle but reoccurring similarities of viral infections to another heterogeneous and adaptable disorder: cancer. After summarizing the main contributions of this thesis in the context of the first two themes, we will discuss the third theme in more detail before providing an outlook of the field as the author perceives it.*

Introduction to viruses and viral disease

In the beginning of this thesis in Chapter I, characteristics of viruses and their life cycles were introduced and the changing definition of these entities was discussed. We learned that viruses are taxonomically ill-defined and that their status of living entities is sometimes affirmed but more often denied. Several schemes to classify viruses were put forward, the importance of the Baltimore scheme was highlighted, and an overview of the known numbers of viral families and viral species was provided.

Next, the evolutionary origins of viruses were detailed and several interesting hypotheses regarding the association between the respective origins of viral and cellular organisms were proposed. One of these hypotheses argued that viruses were essential for the origin of DNA-based cellular life. The importance of viruses for the human species was further highlighted by results on ancient proviral integration events, some of which may have had beneficial effects for the human species. Subsequently, the detrimental effects of viral infection, in particular in the context of public health, were estimated and current trends of infectious disease burdens were analyzed.

Finally, an outlook on viral diversity and zoonotic emergence was provided and estimates for the number of novel human pathogenic viruses, most of which originate from animal hosts, were presented. At the end of the first chapter, a picture of viruses emerged as complex, highly diverse, and abundant entities that are intricately linked to cellular and human life.

Detecting tumor viruses in human cancers

Chapter II introduced a burgeoning field of current research: microbial *metagenomics*. By way of continuing the theme of viral variety and zoonotic emergence of the introductory chapter, an initial introduction to the basic tenets of metagenomics and an exposition of the astonishingly high abundance of microbial organisms in general and viruses in particular was provided. Next, deep sequencing approaches and associated error sources were introduced, both of which serve as major methodological backdrops for the remainder of the thesis. Within the context of these methods, recent results of human and viral metagenomics as well as computational aspects of metagenomic analyses were discussed, highlighting the difficulty of taxonomic annotation based on short read data.

Subsequently, the focus of the chapter shifted towards biology and tumor viruses and their molecular mechanisms of cellular transformation were introduced. Tumor viruses inadvertently act as cofactors of cancers either due to their manipulations of cellular environments and signaling networks or by integration into the host genome. Epidemiological indicators of likely cancer-virus associations as well as the difficulty of establishing causal relationships between pathogens and delayed oncogenesis were discussed and systematic metagenomics approaches for the detection of novel cancer causing viruses were proposed.

Finally, the chapter introduced a novel computational method for inferring known and novel viruses in deep sequencing data of human cancers that offered increased sensitivity and interpretability compared to related approaches. These advantages were realized by adapting a strategy of mapping reads to viral and human references in a parallel fashion and using sensitive read mapping and taxonomic annotation procedures. As a result, viruses with high sequence divergence from known references or close homology to human factors could be reliably detected, in principle. In addition, this study is the first to suggest that oncogenic viruses are not causative for metastatic neuroblastoma, a common tumor of infants.

Inferring viral host factors from protein purifications

Chapter III focused on approaches for inferring the phenotypic interaction patterns of viral and host proteins in order to inform the search for antiviral drug targets. After introducing the concept of viral host factors and their importance as targets of viral manipulation, a variety of human-viral interaction patterns involving viral entry factors and components of the human immune system were discussed.

After elucidating how measurement of host-pathogen protein interaction patterns may aid in the discovery of novel antivirals, the history and current state of development of several classes of antivirals was presented and experimental approaches for the measurement of protein interactions were introduced.

Finally, a method for the statistical inference of physical protein interactions from protein purification assays was presented that makes better use of technical replicates than related approaches. This feature enables the method to detect reliable interaction patterns involving transient interactions or highly abundant protein classes such as molecular chaperones and protein kinases, both of which are frequent viral host factors.

Treatment of divergent viral quasispecies

The last core chapter of this thesis, Chapter IV, presented a clinically relevant topic with additional merits for basic research: the characterization and treatment of highly diverse and adaptable viral populations. After a short introduction to the history of quasispecies theory and its basic tenets abundance, mutation, fitness, and selection, a particular focus was directed at factors that influence the dynamics and robustness of viral genotype populations.

Subsequently, approaches for reducing the fitness of viral populations by mutagenesis were discussed and related to techniques of therapy optimization such as drug combination therapy. Quasispecies attributes and treatment options of a specific, highly divergent virus, HCV, were introduced and the proposed mode of action of a specific mutagenic drug, *ribavirin*, against HCV was introduced. In order to quantify the faint mutagenic action of ribavirin based on *in vivo* sequencing data, we introduced methods for genotyping, quantification, and statistical comparison of minority variants that can serve as indicators for ongoing mutagenesis.

Finally, these methods were applied in the first deep sequencing study on HCV patients under ribavirin monotherapy in order to demonstrate the mutagenic effect of ribavirin on the HCV quasispecies.

26 *Viral infection and malignant disease*

Viral infectious disease and cancer *share several subtle similarities that are of theoretical and practical interest for virology and oncology. Both maladies are essentially genetic diseases, in which "selfish" genetic elements manipulate cellular pathways for their unique evolutionary purposes. This relatedness spans a range of concepts such as common molecular mechanisms, evolutionary behavior, treatment targets, and therepeutic strategies.*

As discussed at some length in Chapter II, viruses are masterful manipulators of cellular growth control as well as of innate and adaptive immunity. Many molecular mechanisms leading to cellular transformation by tumor viruses, such as perturbation of the genomic environment of oncogenes, misregulation of cell cycle control, and subversion of host immune and apoptosis pathways, are also implicated in spontaneous (i.e., non-viral) oncogenesis. The investigation of transforming RNA and DNA viruses in particular served as a Rosetta stone unlocking the secrets of the mitogenic pathway and thus provided the conceptual basis for understanding the molecular mechanisms of cancer.[1] As a consequence, the study of tumor viruses may suggest drug targets such as kinases that are commonly targeted by anti-cancer drugs and also play a major role as viral host factors (cf. Chapter III). Also, viruses may act directly as therapeutics against cancer, as exemplified by oncolytic viruses that specifically infect cancer cells or induce anti-tumor immune responses.[2]

Cancer, as we now understand it, represents a diverse family of genetic diseases that is characterized by a loss of the usual cooperativity between cells in exchange for the gain of proliferative attributes that confer selective growth advantages to the individual cancer cell.[3] Similar to viral infection, cancer may arise from a low number of founder cells (tumor stem cells) whose progeny follows an evolutionary trajectory that maximizes replicative fitness by clonal expansion, progressive accumulation of mutations, and selection of genotypes with advantageous phenotypes.[4] These mechanisms lead the cancer cell population to diversification into heterogeneous genotypes that follow separate mutational pathways within the host.[5]

[1] Butel (2000)

[2] Chiocca (2002), Zeyaullah et al. (2012)

[3] Merlo et al. (2006), Nicolson (1987), Nowell (1976)

[4] Greenman et al. (2007), Hanahan and Weinberg (2011)

[5] Campbell et al. (2010), Navin et al. (2011), Snuderl et al. (2011), Yachida et al. (2010)

While high genomic mutation rates of viral quasispecies are probably unparalleled among other biological systems and the abundance and mutational coupling of cancer cells is dramatically lower than that of a viral population, eukaryotic cancer cells are also known to generate considerable genetic heterogeneity and may exist as a population distribution within the patient.[6] Efforts concerning the characterization of cancer variability and evolution are currently undertaken and seem to qualitatively confirm previous assumptions of cancers as evolving Darwinian systems that are similar to viral quasispecies in many regards.[7] Analogous to the master sequence of viral quasispecies, tumor stem cells have been proposed to act as stabilizing elements in the cancer genotype distribution from which highly adaptive mutant genotypes arise.[8]

Similar to the diversity of RNA virus populations, genetic heterogeneity of cancer has been linked to escape mechanisms against host immune answer and targeted anti-cancers drugs.[9] Also, both cancer cells and virus populations employ latency mechanisms in which the pathogen either lies dormant or withdraws to reservoirs that lack efficient immune surveillance in order to persist in the host. These insights have led to the development of predictive models for cancer progression[10] and the design of anti-cancer treatments based on highly active antiretroviral therapy (HAART), both of the latter propose combination treatment in order to overcome drug resistance.[11] In the same context, the concept of error catastrophe has also been formulated for cancer and the use of mutagenic drugs for the treatment of cancer is an ongoing field of research.[12]

[6] Snijder and Pelkmans (2011)

[7] Brumer et al. (2006), Deisboeck and Couzin (2009), Ding et al. (2013), Merlo et al. (2006), Nowell (1976)

[8] Solé et al. (2008)

[9] González-García et al. (2002)

[10] Maley et al. (2006)

[11] Al-Lazikani et al. (2012), Bock and Lengauer (2012), Glickman and Sawyers (2012), Komarova and Wodarz (2005), Misale et al. (2012), Szakács et al. (2006), Yap et al. (2013)

[12] Fox and Loeb (2010)

27 Outlook

All of the research topics *discussed in this thesis are currently being inovated to different degrees by technical advancements. This section aims, based on the limited perspective of the author, to predict development of the next five years in the respective fields.*

The use of metagenomics techniques (Chapter II) is likely to expand in the near future and may soon represent a general and cost-effective approach for clinical diagnostics of infectious diseases as well as for disease surveillance in tropic regions in order to identify zoonotic or pre-zoonotic pathogens. In line with results of this treatise, metagenomics will furthermore be useful for identifying viral cofactors of certain cancers as well as of metabolic and gastroenterological diseases with unknown etiology.

In addition, metagenomics analyses of highly abundant microbial habitats like sea water and soil may significantly extend our knowledge of functional genetic elements, a fact that may have interesting consequences for industrial biotechnology applications. While presently the assembly of metagenomes and their taxonomic annotation present significant hurdles for interpreting metagenomic data, long-read deep sequencing technologies and more efficient algorithmic approaches will soon allow near complete characterization of highly abundant microbial habitats.

Advances in studying host-pathogen protein interactions (Chapter III) are presently still limited by available technology to measure protein interactions on a large scale; even given automatized approaches, the high error rates of presently available methodologies make determining the set of all human protein interactions extremely challenging.

If the inherent dynamics of protein interaction networks given different cell types and disease conditions are added to the picture, it becomes clear that the complete characterization of the protein interaction phenotype of a cell is considerably more challenging than determining its genotype. However, these phenotypic characterizations are urgently needed in order to understand and treat genetically (and thus phenotypically) divergent diseases such as cancer and viral infection.

While protein arrays and molecular imaging techniques are interesting technologies that can measure physical interactions of cellular components in a high-throughput and high-accuracy manner, in principle, both technologies are not sufficiently matured to provide viable solutions to the problem. Also, it seems unlikely at present that a concerted effort like the Human Genome Project is conducted for determination of the cellular interaction phenotype. Instead, only incremental advances in both experimental and computational methodologies can be expected.

Still, the availability of genome-wide protein interaction screens will likely increase as technologies mature and differential network analyses gain in interest. These screens as well as host-pathogen protein interaction studies of smaller scale will continue to identify new drug targets for antivirals and antibiotics and it is likely that drugs targeting host factors as well as biologicals such as recombinant vaccines and monoclonal antibodies will be of particular interest in this regard.

Finally, developments in treating highly divergent viruses (Chapter IV) are likely to be significantly influenced by advances in sequencing technology. While it is currently difficult or impossible to genetically characterize a viral quasispecies in its entirety due to the comparably short read lengths and high error rates of second generation sequencing approaches, upcoming third generation sequencing platforms such as the Pacific Biosciences RS offer read lengths spanning complete viral genomes at very high accuracy. Provided that these technologies further increase sequencing throughput, the identification and evolution of minority variants as well as characterization of genetic interactions within viral genomes will simplify dramatically.

In addition, further advances to clinical virology will probably originate from metagenomics and oncology. These fields currently attract considerable attention and are likely to yield novel experimental and computational methods that will also be applicable to measuring viral diversity and optimizing antiviral therapies.

Publications

Schatz, M. C., Taylor, J., & **Schelhorn, S.-E.** (2013). The DNA60IFX contest. *Genome Biol*, 14(6), 124.

Schelhorn, S.-E., Fischer, M., Tolosi, L., Altmüller, J., Nürnberg, P., Pfister, H., Lengauer, T., Berthold, F. (2013). Sensitive Detection of Viral Transcripts in Human Tumor Transcriptomes. *PLoS Comput Biol*, 9(10), e1003228.

Dietz, J., **Schelhorn, S.-E.**,[1] Fitting, D., Mihm, U., Susser, S., Welker, M.-W., Füller, C., Däumer, M., Teuber, G., Wedemeyer, H., Berg, T., Lengauer, T., Zeuzem, S., Herrmann, E., Sarrazin, C. (2013). Deep sequencing reveals mutagenic effects of ribavirin during monotherapy of hepatitis C virus genotype 1-infected patients. *J Virol*, 87(11), 6172–6181.

Schelhorn, S. E., Mestre, J., Albrecht, M., & Zotenko, E. (2011). Inferring Physical Protein Contacts from Large-Scale Purification Data of Protein Complexes. *Mol Cell Prot*, 10(6), M110.004929-M110.004929.

Blankenburg, H., Finn, R. D., Prlić, A., Jenkinson, A. M., Ramirez, F., Emig, D., **Schelhorn, S.-E.**, Büch, J., Lengauer, T., Albrecht, M. (2009). DASMI: exchanging, annotating and assessing molecular interaction data. *Bioinformatics*, 25(10), 1321–1328.

Schelhorn, S.-E., Lengauer, T., Albrecht, M. (2008). An integrative approach for predicting interactions of protein regions. *Bioinformatics*, 24(16), i35–41.

Assenov, Y., Ramirez, F., **Schelhorn, S.-E.**, Lengauer, T., Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics*, 24(2), 282–284.

Schelhorn, S.-E., Griego, J., & Schmid, U. (2007). Transformational and derivational strategies in analogical problem solving. *Cognitive processing*, 8(1), 45–55.

[1] Shared first authorship.

Bibliography

1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

Abe, T., Kaname, Y., Hamamoto, I., Tsuda, Y., Wen, X., Taguwa, S., Moriishi, K., Takeuchi, O., Kawai, T., Kanto, T., Hayashi, N., Akira, S., and Matsuura, Y. (2007). Hepatitis C virus nonstructural protein 5A modulates the toll-like receptor-MyD88-dependent signaling pathway in macrophage cell lines. *J Virol*, 81(17):8953–8966.

Abonyi, M. E. and Lakatos, P. L. (2005). Ribavirin in the treatment of hepatitis C. *Anticancer Res*, 25(2B):1315–1320.

Abroi, A. and Gough, J. (2011). Are viruses a source of new protein folds for organisms? - Virosphere structure space and evolution. *Bioessays*, 33(8):626–635.

Agudo, R., Ferrer-Orta, C., Arias, A., de la Higuera, I., Perales, C., Pérez-Luque, R., Verdaguer, N., and Domingo, E. (2010). A multi-step process of viral adaptation to a mutagenic nucleoside analogue by modulation of transition types leads to extinction-escape. *PLoS Pathog*, 6(8):e1001072.

Ahmed, R. and Gray, D. (1996). Immunological memory and protective immunity: understanding their relation. *Science*, 272(5258):54–60.

Akira, S., Uematsu, S., and Takeuchi, O. (2006). Pathogen recognition and innate immunity. *Cell*, 124(4):783–801.

Al-Lazikani, B., Banerji, U., and Workman, P. (2012). Combinatorial drug therapy for cancer in the post-genomic era. *Nat Biotechnol*, 30(7):679–692.

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Rout, M. P., and Sali, A. (2007). Determining the architectures of macromolecular assemblies. *Nature*, 450(7170):683–694.

Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382.

Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, 92(3):291–294.

Alcami, A. (2003). Viral mimicry of cytokines, chemokines and their receptors. *Nat Rev Immunol*, 3(1):36–50.

Alff, P. J., Sen, N., Gorbunova, E., Gavrilovskaya, I. N., and Mackow, E. R. (2008). The NY-1 hantavirus Gn cytoplasmic tail coprecipitates TRAF3 and inhibits cellular interferon responses by disrupting TBK1-TRAF3 complex formation. *J Virol*, 82(18):9115–9122.

Ali, A., Li, H., Schneider, W. L., Sherman, D. J., Gray, S., Smith, D., and Roossinck, M. J. (2006). Analysis of genetic bottlenecks during horizontal transmission of Cucumber mosaic virus. *J Virol*, 80(17):8345–8350.

Ali, A. and Roossinck, M. J. (2010). Genetic bottlenecks during systemic movement of Cucumber mosaic virus vary in different host plants. *Virology*, 404(2):279–283.

Ali, S., Leveque, V., Le Pogam, S., Ma, H., Philipp, F., Inocencio, N., Smith, M., Alker, A., Kang, H., Najera, I., Klumpp, K., Symons, J., Cammack, N., and Jiang, W.-R. (2008). Selected replicon variants with low-level in vitro resistance to the hepatitis C virus NS5B polymerase inhibitor PSI-6130 lack cross-resistance with R1479. *Antimicrob Agents Chemother*, 52(12):4356–4369.

Allander, T., Andreasson, K., Gupta, S., Bjerkner, A., Bogdanovic, G., Persson, M. A. A., Dalianis, T., Ramqvist, T., and Andersson, B. (2007). Identification of a third human polyomavirus. *J Virol*, 81(8):4130–4136.

Allander, T., Tammi, M. T., Eriksson, M., Bjerkner, A., Tiveljung-Lindell, A., and Andersson, B. (2005). Cloning of a human parvovirus by molecular screening of respiratory tract samples. *Proc Natl Acad Sci USA*, 102(36):12891–12896.

Aloy, P., Böttcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A.-C., Bork, P., Superti-Furga, G., Serrano, L., and Russell, R. B. (2004). Structure-based assembly of protein complexes in yeast. *Science*, 303(5666):2026–2029.

Aloy, P. and Russell, R. B. (2002). Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci USA*, 99(9):5896–5901.

Aloy, P. and Russell, R. B. (2006). Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol*, 7(3):188–197.

Alper, T., Cramp, W. A., Haig, D. A., and Clarke, M. C. (1967). Does the agent of scrapie replicate without nucleic acid? *Nature*, 214(5090):764–766.

Altmann, A., Däumer, M., Beerenwinkel, N., Peres, Y., Schülter, E., Büch, J., Rhee, S.-Y., Sönnerborg, A., Fessel, W. J., Shafer, R. W., Zazzi, M., Kaiser, R., and Lengauer, T. (2009). Predicting the response to combination antiretroviral therapy: retrospective validation of geno2pheno-THEO on a large clinical database. *J Infect Dis*, 199(7):999–1006.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402.

Amador-Cañizares, Y. and Dueñas-Carrera, S. (2010). Early interferon-based treatment after detection of persistent hepatitis C virus infection: a critical decision. *J. Interferon Cytokine Res.*, 30(11):817–824.

Amann, R. I., Ludwig, W., and Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.*, 59(1):143–169.

Amara, R. R., Villinger, F., Altman, J. D., Lydy, S. L., O'Neil, S. P., Staprans, S. I., Montefiori, D. C., Xu, Y., Herndon, J. G., Wyatt, L. S., Candido, M. A., Kozyr, N. L., Earl, P. L., Smith, J. M., Ma, H. L., Grimm, B. D., Hulsey, M. L., Miller, J., McClure, H. M., McNicholl, J. M., Moss, B., and Robinson, H. L. (2001). Control of a mucosal challenge and prevention of AIDS by a multiprotein DNA/MVA vaccine. *Science*, 292(5514):69–74.

Ambrose, H. E. and Clewley, J. P. (2006). Virus discovery by sequence-independent genome amplification. *Rev. Med. Virol.*, 16(6):365–383.

Anderson, D. R. D., Carthy, C. M. C., Wilson, J. E. J., Yang, D. D., Devine, D. V. D., and McManus, B. M. B. (1997). Complement component 3 interactions with coxsackievirus B3 capsid proteins: innate immunity and the rapid formation of splenic antiviral germinal centers. *J Virol*, 71(11):8841–8845.

Anderson, J. P., Daifuku, R., and Loeb, L. A. (2004). Viral error catastrophe by mutagenic nucleosides. *Annu Rev Microbiol*, 58:183–205.

Anderson, N. G., Gerin, J. L., and Anderson, N. L. (2003). Global screening for human viral pathogens. *Emerging Infect. Dis.*, 9(7):768–774.

Andersson, M., Pääbo, S., Nilsson, T., and Peterson, P. A. (1985). Impaired intracellular transport of class I MHC antigens as a possible means for adenoviruses to evade immune surveillance. *Cell*, 43(1):215–222.

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. P., Chothia, C., and Murzin, A. G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*, 32(Database issue):D226–9.

Andrei, G. and de Clercq, E. (1993). Molecular approaches for the treatment of hemorrhagic fever virus infections. *Antiviral Res*, 22(1):45–75.

Angly, F., Rodriguez-Brito, B., Bangor, D., McNairnie, P., Breitbart, M., Salamon, P., Felts, B., Nulton, J., Mahaffy, J., and Rohwer, F. (2005). PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics*, 6:41.

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., Chan, A. M., Haynes, M., Kelley, S., Liu, H., Mahaffy, J. M., Mueller, J. E., Nulton, J., Olson, R., Parsons, R., Rayhawk, S., Suttle, C. A., and Rohwer, F. (2006). The marine viromes of four oceanic regions. *PLoS Biol*, 4(11):e368.

Angly, F. E., Willner, D., Prieto-Davó, A., Edwards, R. A., Schmieder, R., Vega-Thurber, R., Antonopoulos, D. A., Barott, K., Cottrell, M. T., Desnues, C., Dinsdale, E. A., Furlan, M., Haynes, M., Henn, M. R., Hu, Y., Kirchman, D. L., McDole, T., McPherson, J. D., Meyer, F., Miller, R. M., Mundt, E., Naviaux, R. K., Rodriguez-Mueller, B., Stevens, R., Wegley, L., Zhang, L., Zhu, B., and Rohwer, F. (2009). The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol*, 5(12):e1000593.

Anthony, S. J., Epstein, J. H., Murray, K. A., Navarrete-Macias, I., Zambrana-Torrelio, C. M., Solovyov, A., Ojeda-Flores, R., Arrigo, N. C., Islam, A., Ali Khan, S., Hosseini, P., Bogich, T. L., Olival, K. J., Sanchez-Leon, M. D., Karesh, W. B., Goldstein, T., Luby, S. P., Morse, S. S., Mazet, J. A. K., Daszak, P., and Lipkin, W. I. (2013). A strategy to estimate unknown viral diversity in mammals. *MBio*, 4(5).

Antia, R., Regoes, R. R., Koella, J. C., and Bergstrom, C. T. (2003). The role of evolution in the emergence of infectious diseases. *Nature*, 426(6967):658–661.

Antonelli, G. and Turriziani, O. (2012). Antiviral therapy: old and current issues. *Int. J. Antimicrob. Agents*, 40(2):95–102.

Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F. S. L., Ceol, A., Chautard, E., Dana, J. M., De Las Rivas, J., Dumousseau, M., Galeota, E., Gaulton, A., Goll, J., Hancock, R. E. W., Isserlin, R., Jimenez, R. C., Kerssemakers, J., Khadake, J., Lynn, D. J., Michaut, M., O'Kelly, G., Ono, K., Orchard, S., Prieto, C., Razick, S., Rigina, O., Salwinski, L., Simonovic, M., Velankar, S., Winter, A., Wu, G., Bader, G. D., Cesareni, G., Donaldson, I. M., Eisenberg, D., Kleywegt, G. J., Overington, J., Ricard-Blum, S., Tyers, M., Albrecht, Mario, and Hermjakob, H. (2011). PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat Methods*, 8(7):528–529.

Archer, J., Baillie, G., Watson, S. J., Kellam, P., Rambaut, A., and Robertson, D. L. (2012). Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator II. *BMC Bioinformatics*, 13:47.

Archer, J., Rambaut, A., Taillon, B. E., Harrigan, P. R., Lewis, M., and Robertson, D. L. (2010). The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time–an ultra-deep approach. *PLoS Comput Biol*, 6(12):e1001022.

Arias, A., Arnold, J. J., Sierra, M., Smidansky, E. D., Domingo, E., and Cameron, C. E. (2008). Determinants of RNA-dependent RNA polymerase (in)fidelity revealed by kinetic analysis of the polymerase encoded by a foot-and-mouth disease virus mutant with reduced sensitivity to ribavirin. *J Virol*, 82(24):12346–12355.

Arias, A., Ruiz-Jarabo, C. M., Escarmís, C., and Domingo, E. (2004). Fitness increase of memory genomes in a viral quasispecies. *J Mol Biol*, 339(2):405–412.

Ariumi, Y., Kuroki, M., Abe, K.-i., Dansako, H., Ikeda, M., Wakita, T., and Kato, N. (2007). DDX3 DEAD-box RNA helicase is required for hepatitis C virus RNA replication. *J Virol*, 81(24):13922–13926.

Armean, I. M., Lilley, K. S., and Trotter, M. W. B. (2013). Popular computational methods to assess multiprotein complexes derived from label-free affinity purification and mass spectrometry (AP-MS) experiments. *Mol Cell Proteomics*, 12(1):1–13.

Arnaud, F., Murcia, P. R., and Palmarini, M. (2007). Mechanisms of late restriction induced by an endogenous retrovirus. *J Virol*, 81(20):11441–11451.

Arnesen, T. and Kapiriri, L. (2004). Can the value choices in DALYs influence global priority-setting? *Health Policy*, 70(2):137–149.

Arron, S. T., Ruby, J. G., Dybbro, E., Ganem, D., and DeRisi, J. L. (2011). Transcriptome sequencing demonstrates that human papillomavirus is not active in cutaneous squamous cell carcinoma. *J. Invest. Dermatol.*, 131(8):1745–1753.

Asahina, Y., Izumi, N., Enomoto, N., Uchihara, M., Kurosaki, M., Onuki, Y., Nishimura, Y., Ueda, K., Tsuchiya, K., Nakanishi, H., Kitamura, T., and Miyake, S. (2005). Mutagenic effects of ribavirin and response to interferon/ribavirin combination therapy in chronic hepatitis C. *J Hepatol*, 43(4):623–629.

Aschoff, M., Hotz-Wagenblatt, A., Glatting, K.-H., Fischer, M., Eils, R., and König, R. (2013). SplicingCompass: differential splicing detection using RNA-Seq data. *Bioinformatics*, 29(9):1141–1148.

Ashkenazi, A. and Dixit, V. M. (1998). Death receptors: signaling and modulation. *Science*, 281(5381):1305–1308.

Assenov, Y., Ramirez, F., Schelhorn, S.-E., Lengauer, T., and Albrecht, Mario (2008). Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284.

Atkinson, A., Siegel, V., Pakhomov, E. A., Jessopp, M. J., and Loeb, V. (2009). A re-appraisal of the total biomass and annual production of Antarctic krill. *Deep Sea Research Part I: Oceanographic Research Papers*, 56(5):727–740.

Awadalla, P. (2003). The evolutionary genomics of pathogen recombination. *Nat Rev Genet*, 4(1):50–60.

Babu, M., Vlasblom, J., Pu, S., Guo, X., Graham, C., Bean, B. D. M., Burston, H. E., Vizeacoumar, F. J., Snider, J., Phanse, S., Fong, V., Tam, Y. Y. C., Davey, M., Hnatshak, O., Bajaj, N., Chandran, S., Punna, T., Christopolous, C., Wong, V., Yu, A., Zhong, G., Li, J., Stagljar, I., Conibear, E., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2012). Interaction landscape of membrane-protein complexes in Saccharomyces cerevisiae. *Nature*, 489(7417):585–589.

Bacon, B. R., Shiffman, M. L., Mendes, F., Ghalib, R., Hassanein, T., Morelli, G., Joshi, S., Rothstein, K., Kwo, P., and Gitlin, N. (2009). Retreating chronic hepatitis C with daily interferon alfacon-1/ribavirin after nonresponse to pegylated interferon/ribavirin: DIRECT results. *Hepatology (Baltimore, Md)*, 49(6):1838–1846.

Bader, G. D., Betel, D., and Hogue, C. W. V. (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, 31(1):248–250.

Bae, S.-E. and Son, H. S. (2010). Classification of viral zoonosis through receptor pattern analysis. *BMC Bioinformatics*, 12:96–96.

Baker, M. (2012). De novo genome assembly: what every biologist should know. *Nat Methods*, 9(4):333–337.

Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature*, 226(5252):1209–1211.

Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriol Rev*, 35(3):235–241.

Balzer, S., Malde, K., and Jonassen, I. (2011). Systematic exploration of error sources in pyrosequencing flowgram data. *Bioinformatics*, 27(13):i304–i309.

Bamford, D. H., Grimes, J. M., and Stuart, D. I. (2005). What does structure tell us about virus evolution? *Curr Opin Struct Biol*, 15(6):655–663.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*, 19(5):455–477.

Bankwitz, D., Steinmann, E., Bitzegeio, J., Ciesek, S., Friesland, M., Herrmann, E., Zeisel, M.-B., Baumert, T.-F., Keck, Z.-Y., Foung, S. K. H., Pécheur, E.-I., and Pietschmann, T. (2010). Hepatitis C virus hypervariable region 1 modulates receptor interactions, conceals the CD81 binding site, and protects conserved neutralizing epitopes. *J Virol*, 84(11):5751–5763.

Bannert, N. and Kurth, R. (2004). Retroelements and the human genome: new perspectives on an old relation. *Proc Natl Acad Sci USA*, 101 Suppl 2:14572–14579.

Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X., and Song, Y.-Q. (2011). Evaluation of next-generation sequencing software in mapping and assembly. *J Hum Genet*, 56(6):406–414.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113.

Baranowski, E. (2001). Evolution of Cell Recognition by Viruses. *Science*, 292(5519):1102–1105.

Barker, J. J. (2006). Antibacterial drug discovery and structure-based design. *Drug Discov. Today*, 11(9-10):391–404.

Barouch, D. H., Stephenson, K. E., Borducchi, E. N., Smith, K., Stanley, K., McNally, A. G., Liu, J., Abbink, P., Maxfield, L. F., Seaman, M. S., Dugast, A.-S., Alter, G., Ferguson, M., Li, W., Earl, P. L., Moss, B., Giorgi, E. E., Szinger, J. J., Eller, L. A., Billings, E. A., Rao, M., Tovanabutra, S., Sanders-Buell, E., Weijtens, M., Pau, M. G., Schuitemaker, H., Robb, M. L., Kim, J. H., Korber, B. T., and Michael, N. L. (2013). Protective Efficacy of a Global HIV-1 Mosaic Vaccine against Heterologous SHIV Challenges in Rhesus Monkeys. In *Cell*, pages 531–539. Cell Press.

Barral, P. M., Morrison, J. M., Drahos, J., Gupta, P., Sarkar, D., Fisher, P. B., and Racaniello, V. R. (2007). MDA-5 is cleaved in poliovirus-infected cells. *J Virol*, 81(8):3677–3684.

Barrios-Rodiles, M., Brown, K. R., Ozdamar, B., Bose, R., Liu, Z., Donovan, R. S., Shinjo, F., Liu, Y., Dembowy, J., Taylor, I. W., Luga, V., Przulj, N., Robinson, M., Suzuki, H., Hayashizaki, Y., Jurisica, I., and Wrana, J. L. (2005). High-throughput mapping of a dynamic signaling network in mammalian cells. *Science*, 307(5715):1621–1625.

Bartel, P. L., Roecklein, J. A., SenGupta, D., and Fields, S. (1996). A protein linkage map of Escherichia coli bacteriophage T7. *Nat Genet*, 12(1):72–77.

Bartosch, B., Verney, G., Dreux, M., Donot, P., Morice, Y., Penin, F., Pawlotsky, J.-M., Lavillette, D., and Cosset, F.-L. (2005). An interplay between hypervariable region 1 of the hepatitis C virus E2 glycoprotein, the scavenger receptor BI, and high-density lipoprotein promotes both enhancement of infection and protection against neutralizing antibodies. *J Virol*, 79(13):8217–8229.

Bateman, A., Coin, L., Durbin, R., Finn, R., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E., Studholme, D., Yeats, C., and Eddy, S. R. (2003). The Pfam protein families database. *Nucleic Acids Res*, 32(Database issue):D138–41.

Baumgartner, C. K. and Malherbe, L. P. (2010). Regulation of CD4 T-cell receptor diversity by vaccine adjuvants. *Immunology*, 130(1):16–22.

Bazinet, A. L. and Cummings, M. P. (2012). A comparative evaluation of sequence classification programs. *BMC Bioinformatics*, 13(1):92.

Beachy, S. H., Kisailus, A. J., Repasky, E. A., Subjeck, J. R., Wang, X. Y., and Kazim, A. L. (2007). Engineering secretable forms of chaperones for immune modulation and vaccine development. *Methods*, 43(3):184–193.

Beck, J. C., Hansen, T. H., Cullen, S. E., and Lee, D. R. (1986). Slower processing, weaker beta 2-M association, and lower surface expression of H-2Ld are influenced by its amino terminus. *J. Immunol.*, 137(3):916–923.

Beck, S. S. and Barrell, B. G. B. (1988). Human cytomegalovirus encodes a glycoprotein homologous to MHC class-I antigens. *Nature*, 331(6153):269–272.

Beerenwinkel, N. (2003). Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res*, 31(13):3850–3855.

Beerenwinkel, N., Däumer, M., Oette, M., Korn, K., Hoffmann, D., Kaiser, R., Lengauer, T., Selbig, J., and Walter, H. (2003). Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res*, 31(13):3850–3855.

Beerenwinkel, N., Günthard, H. F., Roth, V., and Metzner, K. J. (2012). Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front Microbiol*, 3:329.

Beerenwinkel, N., Rahnenführer, J., Kaiser, R., Hoffmann, D., Selbig, J., and Lengauer, T. (2005). Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics*, 21(9):2106–2107.

Beissbarth, T. and Speed, T. P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465.

Belyi, V. A., Levine, A. J., and Skalka, A. M. (2010). Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. *PLoS Pathog*, 6(7):e1001030.

Benhamou, Y., Afdhal, N. H., Nelson, D. R., Shiffman, M. L., Halliman, D. G., Heise, J., Chun, E., and Pockros, P. J. (2009). A phase III study of the safety and efficacy of viramidine versus ribavirin in treatment-naïve patients with chronic hepatitis C: ViSER1 results. *Hepatology (Baltimore, Md)*, 50(3):717–726.

Bennett, E. M. E., Bennink, J. R. J., Yewdell, J. W. J., and Brodsky, F. M. F. (1999). Cutting edge: adenovirus E19 has two mechanisms for affecting class I MHC expression. *J. Immunol.*, 162(9):5049–5052.

Bennett, N. J., May, J. S., and Stevenson, P. G. (2005). Gammaherpesvirus latency requires T cell evasion during episome maintenance. *PLoS Biol*, 3(4):e120–e120.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2013). GenBank. *Nucleic Acids Res*, 41(Database issue):D36–42.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2006). GenBank. *Nucleic Acids Res*, 35(Database issue):D21–D25.

Beral, V., Peterman, T. A., Berkelman, R. L., and Jaffe, H. W. (1990). Kaposi's sarcoma among persons with AIDS: a sexually transmitted infection? *Lancet*, 335(8682):123–128.

Berenguer, M., López-Labrador, F. X., and Wright, T. L. (2001). Hepatitis C and liver transplantation. *J Hepatol*, 35(5):666–678.

Bergh, O., Børsheim, K. Y., Bratbak, G., and Heldal, M. (1989). High abundance of viruses found in aquatic environments. *Nature*, 340(6233):467–468.

Berk, A. J. (2005). Recent lessons in gene expression, cell cycle control, and cell biology from adenovirus. *Oncogene*, 24(52):7673–7685.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–242.

Bernardin, F., Operskalski, E., Busch, M., and Delwart, E. (2010). Transfusion transmission of highly prevalent commensal human viruses. *Transfusion*, 50(11):2474–2483.

Bernini, F., Ebranati, E., De Maddalena, C., Shkjezi, R., Milazzo, L., Lo Presti, A., Ciccozzi, M., Galli, M., and Zehender, G. (2011). Within-host dynamics of the hepatitis C virus quasispecies population in HIV-1/HCV coinfected patients. *PLoS ONE*, 6(1):e16551.

Betancourt, M., Fereres, A., Fraile, A., and García-Arenal, F. (2008). Estimation of the effective number of founders that initiate an infection after aphid transmission of a multipartite plant virus. *J Virol*, 82(24):12416–12421.

Bexfield, N. and Kellam, P. (2011). Metagenomics and the molecular identification of novel viruses. *Vet. J.*, 190(2):191–198.

Bhaduri, A., Qu, K., Lee, C. S., Ungewickell, A., and Khavari, P. A. (2012). Rapid identification of non-human sequences in high-throughput sequencing datasets. *Bioinformatics*, 28(8):1174–1175.

Bill, C. A. and Summers, J. (2004). Genomic DNA double-strand breaks are targets for hepadnaviral DNA integration. *Proc Natl Acad Sci USA*, 101(30):11135–11140.

Bimber, B. N., Dudley, D. M., Lauck, M., Becker, E. A., Chin, E. N., Lank, S. M., Grunenwald, H. L., Caruccio, N. C., Maffitt, M., Wilson, N. A., Reed, J. S., Sosman, J. M., Tarosso, L. F., Sanabani, S., Kallas, E. G., Hughes, A. L., and O'Connor, D. H. (2010). Whole-genome characterization of human and simian immunodeficiency virus intrahost diversity by ultradeep pyrosequencing. *J Virol*, 84(22):12087–12092.

Birgegård, G., Aapro, M. S., Bokemeyer, C., Dicato, M., Drings, P., Hornedo, J., Krzakowski, M., Ludwig, H., Pecorelli, S., Schmoll, H., Schneider, M., Schrijvers, D., Shasha, D., and Van Belle, S. (2005). Cancer-related anemia: pathogenesis, prevalence and treatment. *Oncology*, 68 Suppl 1:3–11.

Bishop, J. M. (1991). Molecular themes in oncogenesis. *Cell*, 64(2):235–248.

Blagoev, B., Kratchmarova, I., Ong, S.-E., Nielsen, M., Foster, L. J., and Mann, M. (2003). A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nat Biotechnol*, 21(3):315–318.

Blaise, S. S., de Parseval, N. N., Bénit, L. L., and Heidmann, T. T. (2003). Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. *Proc Natl Acad Sci USA*, 100(22):13013–13018.

Blankenburg, H., Finn, R. D., Prlić, A., Jenkinson, A. M., Ramirez, F., Emig, D., Schelhorn, S.-E., Büch, J., Lengauer, T., and Albrecht, Mario (2009). DASMI: exchanging, annotating and assessing molecular interaction data. *Bioinformatics*, 25(10):1321–1328.

Boccardo, E. and Villa, L. L. (2007). Viral origins of human cancer. *Curr. Med. Chem.*, 14(24):2526–2539.

Bock, C. and Lengauer, T. (2012). Managing drug resistance in cancer: lessons from HIV therapy. *Nat Rev Cancer*, 12(7):494–501.

Bodenheimer, H. C., Lindsay, K. L., Davis, G. L., Lewis, J. H., Thung, S. N., and Seeff, L. B. (1997). Tolerance and efficacy of oral ribavirin treatment of chronic hepatitis C: a multicenter trial. *Hepatology (Baltimore, Md)*, 26(2):473–477.

Boisvert, S., Laviolette, F., and Corbeil, J. (2010). Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J Comput Biol*, 17(11):1519–1533.

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., and Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol*, 13(12):R122.

Boname, J. M. J. and Stevenson, P. G. P. (2001). MHC Class I Ubiquitination by a Viral PHD/LAP Finger Protein. *Immunity*, 15(4):10–10.

Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I., and Marcotte, E. M. (2004). Protein interaction networks from yeast to human. *Curr Opin Struct Biol*, 14(3):292–299.

Borozan, I., Wilson, S., Blanchette, P., Laflamme, P., Watt, S. N., Krzyzanowski, P. M., Sircoulomb, F., Rottapel, R., Branton, P. E., and Ferretti, V. (2012). CaPSID: A bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes. *BMC Bioinformatics*, 13:206.

Bowen, D. G. and Walker, C. M. (2005). The origin of quasispecies: cause or consequence of chronic hepatitis C viral infection? *J Hepatol*, 42(3):408–417.

Bowie, A. G. and Unterholzner, L. (2008). Viral evasion and subversion of pattern-recognition receptor signalling. *Nat Rev Immunol*, 8(12):911–922.

Boyd, M. T. M., Simpson, G. R. G., Cann, A. J. A., Johnson, M. A. M., and Weiss, R. A. R. (1993). A single amino acid substitution in the V1 loop of human immunodeficiency virus type 1 gp120 alters cellular tropism. *J Virol*, 67(6):3649–3652.

Boyd, S. D. (2013). Diagnostic applications of high-throughput DNA sequencing. *Annu Rev Pathol*, 8:381–410.

Boyer, M., Madoui, M.-A., Gimenez, G., La Scola, B., and Raoult, D. (2010). Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4 domain of life including giant viruses. *PLoS ONE*, 5(12):e15530.

Boyer, M., Yutin, N., Pagnier, I., Barrassi, L., Fournous, G., Espinosa, L., Robert, C., Azza, S., Sun, S., Rossmann, M. G., Suzan-Monti, M., La Scola, B., Koonin, E. V., and Raoult, D. (2009). Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc Natl Acad Sci USA*, 106(51):21848–21853.

Bozek, K. and Lengauer, T. (2010). Positive selection of HIV host factors and the evolution of lentivirus genes. *BMC Evol Biol*, 10:186.

Braciale, T. J., Morrison, L. A., Sweetser, M. T., Sambrook, J., Gething, M.-J., and Braciale, V. L. (1987). Antigen Presentation Pathways to Class I and Class II MHC-Restricted T Lymphocytes. *Immunol Rev*, 98(1):95–114.

Brady, A. and Salzberg, S. L. (2009). Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods*, 6(9):673–676.

Braithwaite, A. W. and Prives, C. L. (2006). p53: more research and more questions. *Cell Death Differ.*, 13(6):877–880.

Brand, S. R., Kobayashi, R., and Mathews, M. B. (1997). The Tat protein of human immunodeficiency virus type 1 is a substrate and inhibitor of the interferon-induced, virally activated protein kinase, PKR. *J Biol Chem*, 272(13):8388–8395.

Brass, A. L., Dykxhoorn, D. M., Benita, Y., Yan, N., Engelman, A., Xavier, R. J., Lieberman, J., and Elledge, S. J. (2008). Identification of host proteins required for HIV infection through a functional genomic screen. *Science*, 319(5865):921–926.

Bratbak, G., Egge, J. K., and Heldal, M. (1993). Viral mortality of the marine alga Emiliania huxleyi (Haptophyceae) and termination of algal blooms. *Marine Ecology Progress Series*, (93):39–48.

Braun, P. (2012). Interactome mapping for analysis of complex phenotypes: Insights from benchmarking binary interaction assays. *Proteomics*, 12(10):1499–1518.

Braun, P., Tasan, M., Cusick, M., Hill, D. E., and Vidal, M. (2010). Reply to "Exhaustive benchmarking of the yeast two-hybrid system". *Nat Methods*, 7(9):668.

Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J. M., Murray, R. R., Roncari, L., de Smet, A.-S., Venkatesan, K., Rual, J.-F., Vandenhaute, J., Cusick, M. E., Pawson, T., Hill, D. E., Tavernier, J., Wrana, J. L., Roth, F. P., and Vidal, M. (2009). An experimentally derived confidence score for binary protein-protein interactions. *Nat Methods*, 6(1):91–97.

Breitbart, M., Felts, B., Kelley, S., Mahaffy, J. M., Nulton, J., Salamon, P., and Rohwer, F. (2004). Diversity and population structure of a near-shore marine-sediment viral community. *Proceedings of the Royal Society B: Biological Sciences*, 271(1539):565–574.

Breitbart, M., Haynes, M., Kelley, S., Angly, F., Edwards, R. A., Felts, B., Mahaffy, J. M., Mueller, J., Nulton, J., Rayhawk, S., Rodriguez-Brito, B., Salamon, P., and Rohwer, F. (2008). Viral diversity and dynamics in an infant gut. *Res. Microbiol.*, 159(5):367–373.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P., and Rohwer, F. (2003). Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.*, 185(20):6220–6223.

Breitbart, M. and Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends in microbiology*, 13(6):278–284.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., Azam, F., and Rohwer, F. (2002). Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA*, 99(22):14250–14255.

Breitkreutz, A., Choi, H., Sharom, J. R., Boucher, L., Neduva, V., Larsen, B., Lin, Z.-Y., Breitkreutz, B.-J., Stark, C., Liu, G., Ahn, J., Dewar-Darch, D., Reguly, T., Tang, X., Almeida, R., Qin, Z. S., Pawson, T., Gingras, A.-C., Nesvizhskii, A. I., and Tyers, M. (2010). A global protein kinase and phosphatase interaction network in yeast. *Science*, 328(5981):1043–1046.

Brennan, P. A. P. and Kendrick, K. M. K. (2006). Mammalian social odours: attraction and individual recognition. *Philos Trans R Soc Lond, B, Biol Sci*, 361(1476):2061–2078.

Briese, T., Paweska, J. T., McMullan, L. K., Hutchison, S. K., Street, C., Palacios, G., Khristova, M. L., Weyer, J., Swanepoel, R., Egholm, M., Nichol, S. T., and Lipkin, W. I. (2009). Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. *PLoS Pathog*, 5(5):e1000455.

Briones, C., de Vicente, A., Molina-París, C., and Domingo, E. (2006). Minority memory genomes can influence the evolution of HIV-1 quasispecies in vivo. *Gene*, 384:129–138.

Brochot, E., Duverlie, G., Castelain, S., Morel, V., Wychowski, C., Dubuisson, J., and François, C. (2007). Effect of ribavirin on the hepatitis C virus (JFH-1) and its correlation with interferon sensitivity. *Antivir Ther (Lond)*, 12(5):805–813.

Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W. L., Russ, C., Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008). Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res*, 18(5):763–770.

Brodeur, G. M. (2003). Neuroblastoma: biological insights into a clinical enigma. *Nat Rev Cancer*, 3(3):203–216.

Brodsky, F. M., Lem, L., Solache, A., and Bennett, E. M. (1999). Human pathogen subversion of antigen presentation. *Immunol Rev*, 168:199–215.

Brok, J., Gluud, L. L., and Gluud, C. (2005). Effects of adding ribavirin to interferon to treat chronic hepatitis C infection: a systematic review and meta-analysis of randomized trials. *Arch. Intern. Med.*, 165(19):2206–2212.

Brok, J., Gluud, L. L., and Gluud, C. (2006). Ribavirin monotherapy for chronic hepatitis C infection: a Cochrane Hepato-Biliary Group systematic review and meta-analysis of randomized trials. *Am. J. Gastroenterol.*, 101(4):842–847.

Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., and Brom, T. H. (2012). A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. *arXiv*.

Bruinsma, R., Gelbart, W., Reguera, D., Rudnick, J., and Zandi, R. (2003). Viral Self-Assembly as a Thermodynamic Process. *Phys Rev Lett*, 90(24):248101.

Brumer, Y., Michor, F., and Shakhnovich, E. I. (2006). Genetic instability and the quasispecies model. *J. Theor. Biol.*, 241(2):7–7.

Bruno, A. E., Miecznikowski, J. C., Qin, M., Wang, J., and Liu, S. (2013). FUSIM: a software tool for simulating fusion transcripts. *BMC Bioinformatics*, 14(1):13.

Brussow, H. (2009). The not so universal tree of life or the place of viruses in the living world. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1527):2263–2274.

Buechen-Osmond, C. and Dallwitz, M. (1996). Towards a universal virus database - progress in the ICTVdB. *Arch Virol*, 141(2):392–399.

Bull, J. J., Meyers, L. A., and Lachmann, M. (2005). Quasispecies made simple. *PLoS Comput Biol*, 1(6):e61.

Bull, J. J., Sanjuán, R., and Wilke, C. O. (2007). Theory of lethal mutagenesis for viruses. *J Virol*, 81(6):2930–2939.

Bull, R. A., Luciani, F., McElroy, K., Gaudieri, S., Pham, S. T., Chopra, A., Cameron, B., Maher, L., Dore, G. J., White, P. A., and Lloyd, A. R. (2011). Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. *PLoS Pathog*, 7(9):e1002243.

Burch, A. D. A., Ta, J. J., and Fane, B. A. B. (1999). Cross-functional analysis of the Microviridae internal scaffolding protein. *J Mol Biol*, 286(1):10–10.

Burch, C. L. C. and Chao, L. L. (2000). Evolvability of an RNA virus is determined by its mutational neighbourhood. *Nature*, 406(6796):625–628.

Burgert, H. G. H. and Kvist, S. S. (1985). An adenovirus type 2 glycoprotein blocks cell surface expression of human histocompatibility class I antigens. *Cell*, 41(3):987–997.

Burney, T. and Dusheiko, G. (2011). Overview of the PROVE studies evaluating the use of telaprevir in chronic hepatitis C genotype 1 patients. *Expert review of anti-infective therapy*, 9(2):151–160.

Burton, D. R. (2002). Antibodies, viruses and vaccines. *Nat Rev Immunol*, 2(9):706–713.

Burton, D. R., Desrosiers, R. C., Doms, R. W., Koff, W. C., Kwong, P. D., Moore, J. P., Nabel, G. J., Sodroski, J., Wilson, I. A., and Wyatt, R. T. (2004). HIV vaccine design and the neutralizing antibody problem. *Nat. Immunol.*, 5(3):233–236.

Bushman, F. D., Malani, N., Fernandes, J., D'Orso, I., Cagney, G., Diamond, T. L., Zhou, H., Hazuda, D. J., Espeseth, A. S., König, R., Bandyopadhyay, S., Ideker, T., Goff, S. P., Krogan, N. J., Frankel, A. D., Young, J. A. T., and Chanda, S. K. (2009). Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS Pathog*, 5(5):e1000437.

Butel, J. S. (2000). Viral carcinogenesis: revelation of molecular mechanisms and etiology of human disease. *Carcinogenesis*, 21(3):405–426.

Butel, J. S. and Lednicky, J. A. (2000). Response to more about: cell and molecular biology of simian virus 40: implications for human infections and disease. *J. Natl. Cancer Inst.*, 92(6):496–497.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008). ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Res*, 18(5):810–820.

Calderwood, M. A., Venkatesan, K., Xing, L., Chase, M. R., Vazquez, A., Holthaus, A. M., Ewence, A. E., Li, N., Hirozane-Kishikawa, T., Hill, D. E., Vidal, M., Kieff, E., and Johannsen, E. (2007a). Epstein-Barr virus and virus human protein interaction maps. *Proc Natl Acad Sci USA*, 104(18):7606–7611.

Calderwood, S. K., Mambula, S. S., Gray, P. J., and Theriault, J. R. (2007b). Extracellular heat shock proteins in cell signaling. *FEBS Lett*, 581(19):3689–3694.

Callahan, R. (1996). MMTV-induced mutations in mouse mammary tumors: their potential relevance to human breast cancer. *Breast Cancer Res. Treat.*, 39(1):33–44.

Campbell, P. J., Yachida, S., Mudie, L. J., Stephens, P. J., Pleasance, E. D., Stebbings, L. A., Morsberger, L. A., Latimer, C., McLaren, S., Lin, M.-L., McBride, D. J., Varela, I., Nik-Zainal, S. A., Leroy, C., Jia, M., Menzies, A., Butler, A. P., Teague, J. W., Griffin, C. A., Burton, J., Swerdlow, H., Quail, M. A., Stratton, M. R., Iacobuzio-Donahue, C., and Futreal, P. A. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, 467(7319):1109–1113.

Camps, M., Naukkarinen, J., Johnson, B. P., and Loeb, L. A. (2003). Targeted gene evolution in Escherichia coli using a highly error-prone DNA polymerase I. *Proc Natl Acad Sci USA*, 100(17):9727–9732.

Cannon, J. S., Hamzeh, F., Moore, S., Nicholas, J., and Ambinder, R. F. (1999). Human herpesvirus 8-encoded thymidine kinase and phosphotransferase homologues confer sensitivity to ganciclovir. *J Virol*, 73(6):4786–4793.

Cannon, N. A., Donlin, M. J., Fan, X., Aurora, R., Tavis, J. E., and Group, V.-C. S. (2008). Hepatitis C virus diversity and evolution in the full open-reading frame during antiviral therapy. *PLoS ONE*, 3(5):e2123.

Cao, F., Donlin, M. J., Turner, K., Cheng, X., and Tavis, J. E. (2011). Genetic and biochemical diversity in the HCV NS5B RNA polymerase in the context of interferon plus ribavirin therapy. *J Viral Hepat*, 18(5):349–357.

Carbonell, P., Nussinov, R., and del Sol, A. (2009). Energetic determinants of protein binding specificity: insights into protein interaction networks. *Proteomics*, 9(7):1744–1753.

Carrillo-Tripp, M., Shepherd, C. M., Borelli, I. A., Venkataraman, S., Lander, G., Natarajan, P., Johnson, J. E., Brooks, C. L., and Reddy, V. S. (2009). VIPERdb2: an enhanced and web API enabled relational database for structural virology. *Nucleic Acids Res*, 37(Database issue):D436–42.

Cases-Gonzalez, C. E. C., Gutierrez-Rivas, M. M., and Ménendez-Arias, L. L. (2000). Coupling ribose selection to fidelity of DNA synthesis. The role of Tyr-115 of human immunodeficiency virus type 1 reverse transcriptase. *J Biol Chem*, 275(26):19759–19767.

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Mol. Microbiol.*, 49(2):277–300.

Casrouge, A., Decalf, J., Ahloulay, M., Lababidi, C., Mansour, H., Vallet-Pichard, A., Mallet, V., Mottez, E., Mapes, J., Fontanet, A., Pol, S., and Albert, M. L. (2011). Evidence for an antagonist form of the chemokine CXCL10 in patients chronically infected with HCV. *J Clin Invest*, 121(1):308–317.

Cattral, M. S., Hemming, A. W., Wanless, I. R., Al Ashgar, H., Krajden, M., Lilly, L., Greig, P. D., and Levy, G. A. (1999). Outcome of long-term ribavirin therapy for recurrent hepatitis C after liver transplantation. *Transplantation*, 67(9):1277–1280.

Cecere, A., Marotta, F., Vangieri, B., Tancredi, L., and Gattoni, A. (2004). Progressive liver injury in chronic hepatitis C infection is related to altered cellular immune response and to different citokine profile. *Panminerva Med*, 46(3):171–187.

Centurion-Lara, A., LaFond, R. E., Hevner, K., Godornes, C., Molini, B. J., Van Voorhis, W. C., and Lukehart, S. A. (2004). Gene conversion: a mechanism for generation of heterogeneity in the tprK gene of Treponema pallidum during infection. *Mol. Microbiol.*, 52(6):1579–1596.

Cervantes, R. B., Stringer, J. R., Shao, C., Tischfield, J. A., and Stambrook, P. J. (2002). Embryonic stem cells and somatic cells differ in mutation frequency and type. *Proc Natl Acad Sci USA*, 99(6):3586–3590.

Cesar Ignacio-Espinoza, J., Solonenko, S. A., and Sullivan, M. B. (2013). The global virome: not as big as we thought? *Curr Opin Vir*.

Chang, I.-F. (2006). Mass spectrometry-based proteomic analysis of the epitope-tag affinity purified protein complexes in eukaryotes. *Proteomics*, 6(23):6158–6166.

Chang, Y., Cesarman, E., Pessin, M. S., Lee, F., Culpepper, J., Knowles, D. M., and Moore, P. S. (1994). Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science*, 266(5192):1865–1869.

Chapman, M. A., Lawrence, M. S., Keats, J. J., Cibulskis, K., Sougnez, C., Schinzel, A. C., Harview, C. L., Brunet, J.-P., Ahmann, G. J., Adli, M., Anderson, K. C., Ardlie, K. G., Auclair, D., Baker, A., Bergsagel, P. L., Bernstein, B. E., Drier, Y., Fonseca, R., Gabriel, S. B., Hofmeister, C. C., Jagannath, S., Jakubowiak, A. J., Krishnan, A., Levy, J., Liefeld, T., Lonial, S., Mahan, S., Mfuko, B., Monti, S., Perkins, L. M., Onofrio, R., Pugh, T. J., Rajkumar, S. V., Ramos, A. H., Siegel, D. S., Sivachenko, A., Stewart, A. K., Trudel, S., Vij, R., Voet, D., Winckler, W., Zimmerman, T., Carpten, J., Trent, J., Hahn, W. C., Garraway, L. A., Meyerson, M., Lander, E. S., Getz, G., and Golub, T. R. (2011). Initial genome sequencing and analysis of multiple myeloma. *Nature*, 471(7339):467–472.

Chatr-aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L., Reguly, T., Breitkreutz, A., Sellam, A., Chen, D., Chang, C., Rust, J., Livstone, M., Oughtred, R., Dolinski, K., and Tyers, M. (2013). The BioGRID interaction database: 2013 update. *Nucleic Acids Res*, 41(Database issue):D816–23.

Chatr-aryamontri, A., Ceol, A., Peluso, D., Nardozza, A., Panni, S., Sacco, F., Tinti, M., Smolyar, A., Castagnoli, L., Vidal, M., Cusick, M. E., and Cesareni, G. (2009). VirusMINT: a viral protein interaction database. *Nucleic Acids Res*, 37(Database issue):D669–73.

Chatterji, S., Yamazaki, I., Bai, Z., and Eisen, J. A. (2008). CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4955 LNBI:17–28.

Chatterji, U., Lim, P., Bobardt, M. D., Wieland, S., Cordek, D. G., Vuagniaux, G., Chisari, F., Cameron, C. E., Targett-Adams, P., Parkinson, T., and Gallay, P. A. (2010). HCV resistance to cyclosporin A does not correlate with a resistance of the NS5A-cyclophilin A interaction to cyclophilin inhibitors. *J Hepatol*, 53(1):50–56.

Chayama, K. and Hayes, C. N. (2011). Hepatitis C virus: How genetic variability affects pathobiology of disease. *J. Gastroenterol. Hepatol.*, 26 Suppl 1:83–95.

Chen, L. P., Thomas, E. K., Hu, S. L., Hellström, I., and Hellström, K. E. (1991). Human papillomavirus type 16 nucleoprotein E7 is a tumor rejection antigen. *Proc Natl Acad Sci USA*, 88(1):110–114.

Chen, M. M. and Bouvier, M. M. (2007). Analysis of interactions in a tapasin/class I complex provides a mechanism for peptide selection. *EMBO Journal*, 26(6):1681–1690.

Chen, W. and Dimitrov, D. S. (2012). Monoclonal antibody-based candidate therapeutics against HIV type 1. *AIDS Res Hum Retroviruses*, 28(5):425–434.

Chen, Y.-C., Rajagopala, S. V., Stellberger, T., and Uetz, P. (2010). Exhaustive benchmarking of the yeast two-hybrid system. *Nat Methods*, 7(9):667–668.

Chevaliez, S., Brillet, R., Lázaro, E., Hézode, C., and Pawlotsky, J.-M. (2007). Analysis of ribavirin mutagenicity in human hepatitis C virus infection. *J Virol*, 81(14):7732–7741.

Chien, C. T., Bartel, P. L., Sternglanz, R., and Fields, S. (1991). The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc Natl Acad Sci USA*, 88(21):9578–9582.

Chikhi, R. and Rizk, G. (2012). Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *WABI, to appear*.

Chiocca, E. A. E. (2002). Oncolytic viruses. *Nat Rev Cancer*, 2(12):938–950.

Choi, H., Glatter, T., Gstaiger, M., and Nesvizhskii, A. I. (2012). SAINT-MS1: protein-protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments. *J Proteome Res*, 11(4):2619–2624.

Choi, H., Larsen, B., Lin, Z.-Y., Breitkreutz, A., Mellacheruvu, D., Fermin, D., Qin, Z. S., Tyers, M., Gingras, A.-C., and Nesvizhskii, A. I. (2011). SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nat Methods*, 8(1):70–73.

Christoforides, A., Carpten, J. D., Weiss, G. J., Demeure, M. J., Von Hoff, D. D., and Craig, D. W. (2013). Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. *BMC Genomics*, 14(1):302.

Chua, K. B., Goh, K. J., Wong, K. T., Kamarulzaman, A., Tan, P. S. K., Ksiazek, T. G., Zaki, S. R., Paul, G., Lam, S. K., and Tan, C. T. (1999). Fatal encephalitis due to Nipah virus among pig-farmers in Malaysia. *The Lancet*, 354(9186):1257–1259.

Chung, D.-H., Sun, Y., Parker, W. B., Arterburn, J. B., Bartolucci, A., and Jonsson, C. B. (2007). Ribavirin reveals a lethal threshold of allowable mutation frequency for Hantaan virus. *J Virol*, 81(21):11722–11729.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*, 31(3):213–219.

Ciehanover, A., Hod, Y., and Hershko, A. (1978). A heat-stable polypeptide component of an ATP-dependent proteolytic system from reticulocytes. *Biochem. Biophys. Res. Commun.*, 81(4):1100–1105.

Claesson, M. J., Jeffery, I. B., Conde, S., Power, S. E., O'Connor, E. M., Cusack, S., Harris, H. M. B., Coakley, M., Lakshminarayanan, B., O'Sullivan, O., Fitzgerald, G. F., Deane, J., O'Connor, M., Harnedy, N., O'Connor, K., O'Mahony, D., van Sinderen, D., Wallace, M., Brennan, L., Stanton, C., Marchesi, J. R., Fitzgerald, A. P., Shanahan, F., Hill, C., Ross, R. P., and O'Toole, P. W. (2012). Gut microbiota composition correlates with diet and health in the elderly. *Nature*, 488(7410):178–184.

Clarke, D. K., Duarte, E. A., Moya, A., Elena, S. F., Domingo, E., and Holland, J. (1993). Genetic bottlenecks and population passages cause profound fitness differences in RNA viruses. *J Virol*, 67(1):222–228.

Claverie, J.-M. (2006). Viruses take center stage in cellular evolution. *Genome Biol*, 7(6):110.

Claverie, J.-M. and Abergel, C. (2010). Mimivirus: the emerging paradox of quasi-autonomous viruses. *Trends Genet*, 26(10):431–437.

Clay, P. G., McRae, M., and Laurent, J.-P. (2011). Safety, Tolerability, and Pharmacokinetics of KP-1461 in Phase I Clinical Studies: A Single Oral Dose Study in Non-HIV-Infected Adults, and a 14-Day Dose-Escalating Study in Highly Antiretroviral-Experienced HIV-Infected Adults. *J Int Assoc Physicians AIDS Care (Chic)*, 10(4):232–238.

Cleaveland, S., Laurenson, M. K., and Taylor, L. H. (2001). Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. *Philos Trans R Soc Lond, B, Biol Sci*, 356(1411):991–999.

Cochrane, G., Alako, B., Amid, C., Bower, L., Cerdeno-Tárraga, A., Cleland, I., Gibson, R., Goodgame, N., Jang, M., Kay, S., Leinonen, R., Lin, X., Lopez, R., McWilliam, H., Oisel, A., Pakseresht, N., Pallreddy, S., Park, Y., Plaister, S., Radhakrishnan, R., Rivière, S., Rossello, M., Senf, A., Silvester, N., Smirnov, D., Ten Hoopen, P., Toribio, A., Vaughan, D., and Zalunin, V. (2013). Facing growth in the European Nucleotide Archive. *Nucleic Acids Res*, 41(Database issue):D30–5.

Codoñer, F. M., Darós, J.-A., Solé, R. V., and Elena, S. F. (2006). The fittest versus the flattest: experimental confirmation of the quasispecies effect with subviral pathogens. *PLoS Pathog*, 2(12):e136.

Coelmont, L., Hanoulle, X., Chatterji, U., Berger, C., Snoeck, J., Bobardt, M., Lim, P., Vliegen, I., Paeshuyse, J., Vuagniaux, G., Vandamme, A.-M., Bartenschlager, R., Gallay, P., Lippens, G., and Neyts, J. (2010). DEB025 (Alisporivir) inhibits hepatitis C virus replication by preventing a cyclophilin A induced cis-trans isomerisation in domain II of NS5A. *PLoS ONE*, 5(10):e13687.

Coffey, L. L., Beeharry, Y., Bordería, A. V., Blanc, H., and Vignuzzi, M. (2011). Arbovirus high fidelity variant loses fitness in mosquitoes and mice. *Proc Natl Acad Sci USA*, 108(38):16038–16043.

Coffin, J. M. (1992). Superantigens and endogenous retroviruses: a confluence of puzzles. *Science*, 255(5043):411–413.

Coffin, J. M. (1995). HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science*, 267(5197):483–489.

Coffin, J. M. (1997). *Retroviruses*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

Coffin, J. M., Hughes, S. H., Varmus, H. E., Boeke, J. D., and Stoye, J. P. (1997a). *Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelements*. Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).

Coffin, J. M., Hughes, S. H., Varmus, H. E., Rosenberg, N., and Jolicoeur, P. (1997b). *Retroviral Pathogenesis*. Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).

Cohen, R. and Havlin, S. (2003). Scale-free networks are ultrasmall. *Phys Rev Lett*, 90(5):058701.

Coiras, M., López-Huertas, M. R., Pérez-Olmeda, M., and Alcamí, J. (2009). Understanding HIV-1 latency provides clues for the eradication of long-term reservoirs. *Nat Rev Microbiol*, 7(11):798–812.

Collins, K. L. K., Chen, B. K. B., Kalams, S. A. S., Walker, B. D. B., and Baltimore, D. D. (1998). HIV-1 Nef protein protects infected primary cells against killing by cytotoxic T lymphocytes. *Nature*, 391(6665):397–401.

Collins, M. O. and Choudhary, J. S. (2008). Mapping multiprotein complexes by affinity purification and mass spectrometry. *Curr. Opin. Biotechnol.*, 19(4):324–330.

Collins, S. R., Kemmeren, P., Zhao, X.-C., Greenblatt, J. F., Spencer, F., Holstege, F. C. P., Weissman, J. S., and Krogan, N. J. (2007). Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics*, 6(3):439–450.

Colman, P. M., Laver, W. G., Varghese, J. N., Baker, A. T., Tulloch, P. A., Air, G. M., and Webster, R. G. (1987). Three-dimensional structure of a complex of antibody with influenza virus neuraminidase. *Nature*, 326(6111):358–363.

Colman, P. M., Varghese, J. N., and Laver, W. G. (1983). Structure of the catalytic and antigenic sites in influenza virus neuraminidase. *Nature*, 303(5912):41–44.

Colot, V. and Rossignol, J. L. (1999). Eukaryotic DNA methylation as an evolutionary device. *Bioessays*, 21(5):402–411.

Conley, A. B., Piriyapongsa, J., and Jordan, I. K. (2008). Retroviral promoters in the human genome. *Bioinformatics*, 24(14):1563–1567.

Contreras, A. M., Hiasa, Y., He, W., Terella, A., Schmidt, E. V., and Chung, R. T. (2002). Viral RNA mutations are region specific and increased by ribavirin in a full-length hepatitis C virus replication system. *J Virol*, 76(17):8505–8517.

Contreras-Galindo, R., Kaplan, M. H., Leissner, P., Verjat, T., Ferlenghi, I., Bagnoli, F., Giusti, F., Dosik, M. H., Hayes, D. F., Gitlin, S. D., and Markovitz, D. M. (2008). Human endogenous retrovirus K (HML-2) elements in the plasma of people with lymphoma and breast cancer. *J Virol*, 82(19):9329–9336.

Conway, T., Wazny, J., Bromage, A., Zobel, J., and Beresford-Smith, B. (2012). Gossamer–a resource-efficient de novo assembler. *Bioinformatics*, 28(14):1937–1938.

Cooper, S., Erickson, A. L., Adams, E. J., Kansopon, J., Weiner, A. J., Chien, D. Y., Houghton, M., Parham, P., and Walker, C. M. (1999). Analysis of a successful immune response against hepatitis C virus. *Immunity*, 10(4):439–449.

Cornelissen, M., van der Kuyl, A. C., van den Burg, R., Zorgdrager, F., van Noesel, C. J. M., and Goudsmit, J. (2003). Gene expression profile of AIDS-related Kaposi's sarcoma. *BMC Cancer*, 3:7.

Coscoy, L. and Ganem, D. (2000). Kaposi's sarcoma-associated herpesvirus encodes two proteins that block cell surface display of MHC class I chains by enhancing their endocytosis. *Proc Natl Acad Sci USA*, 97(14):8051–8056.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L. Y., Toufighi, K., Mostafavi, S., Prinz, J., St Onge, R. P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F. J., Alizadeh, S., Bahr, S., Brost, R. L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z.-Y., Liang, W., Marback, M., Paw, J., San Luis, B.-J., Shuteriqi, E., Tong, A. H. Y., van Dyk, N., Wallace, I. M., Whitney, J. A., Weirauch, M. T., Zhong, G., Zhu, H., Houry, W. A., Brudno, M., Ragibizadeh, S., Papp, B., Pál, C., Roth, F. P., Giaever, G., Nislow, C., Troyanskaya, O. G., Bussey, H., Bader, G. D., Gingras, A.-C., Morris, Q. D., Kim, P. M., Kaiser, C. A., Myers, C. L., Andrews, B. J., and Boone, C. (2010). The genetic landscape of a cell. *Science*, 327(5964):425–431.

Cox, A. L., Mosbruger, T., Mao, Q., Liu, Z., Wang, X.-H., Yang, H.-C., Sidney, J., Sette, A., Pardoll, D., Thomas, D. L., and Ray, S. C. (2005). Cellular immune selection with hepatitis C virus persistence in humans. *J Exp Med*, 201(11):1741–1752.

Cox-Foster, D. L., Conlan, S., Holmes, E. C., Palacios, G., Evans, J. D., Moran, N. A., Quan, P. L., Briese, T., Hornig, M., Geiser, D. M., Martinson, V., vanEngelsdorp, D., Kalkstein, A. L., Drysdale, A., Hui, J., Zhai, J., Cui, L., Hutchison, S. K., Simons, J. F., Egholm, M., Pettis, J. S., and Lipkin, W. I. (2007). A Metagenomic Survey of Microbes in Honey Bee Colony Collapse Disorder. *Science*, 318(5848):283–287.

Craig, A. G. and Scherf, A. (2003). Antigenic Variation. Academic Press.

Crick, F. H. and Watson, J. D. (1956). Structure of small viruses. *Nature*, 177(4506):473–475.

Crill, W. D. W., Wichman, H. A. H., and Bull, J. J. J. (1999). Evolutionary reversals during viral adaptation to alternating hosts. *Genetics*, 154(1):27–37.

Cristea, I. M., Carroll, J.-W. N., Rout, M. P., Rice, C. M., Chait, B. T., and MacDonald, M. R. (2006). Tracking and elucidating alphavirus-host protein interactions. *J Biol Chem*, 281(40):30269–30278.

Cristina, J., del Pilar Moreno, M., and Moratorio, G. (2007). Hepatitis C virus genetic variability in patients undergoing antiviral therapy. *Virus Res*, 127(2):185–194.

Crotty, S., Cameron, C., and Andino, R. (2002). Ribavirin's antiviral mechanism of action: lethal mutagenesis? *J Mol Med*, 80(2):86–95.

Crotty, S., Cameron, C. E., and Andino, R. (2001). RNA virus error catastrophe: direct molecular test by using ribavirin. *Proc Natl Acad Sci USA*, 98(12):6895–6900.

Crotty, S., Maag, D., Arnold, J. J., Zhong, W., Lau, J. Y., Hong, Z., Andino, R., and Cameron, C. E. (2000). The broad-spectrum antiviral ribonucleoside ribavirin is an RNA virus mutagen. *Nat Med*, 6(12):1375–1379.

Cuevas, J. M., González-Candelas, F., Moya, A., and Sanjuán, R. (2009). Effect of ribavirin on the mutation rate and spectrum of hepatitis C virus in vivo. *J Virol*, 83(11):5760–5764.

Cui, J. and Holmes, E. C. (2012). Endogenous RNA viruses of plants in insect genomes. *Virology*, 427(2):77–79.

Culley, A. I., Lang, A. S., and Suttla, C. A. (2006). Metagenomic analysis of coastal RNA virus communities. *Science*, 312(5781):1795–1798.

Dalton, J. A. R. and Jackson, R. M. (2007). An evaluation of automated homology modelling methods at low target template sequence similarity. *Bioinformatics*, 23(15):1901–1908.

Dalton-Griffin, L. and Kellam, P. (2009). Infectious causes of cancer and their detection. *J Biol*, 8(7):67.

Damon, I. I., Murphy, P. M. P., and Moss, B. B. (1998). Broad spectrum chemokine antagonistic activity of a human poxvirus chemokine homolog. *Proc Natl Acad Sci USA*, 95(11):6403–6407.

D'Angio, G. J., Evans, A. E., and Koop, C. E. (1971). Special pattern of widespread neuroblastoma with a favourable prognosis. *Lancet*, 1(7708):1046–1049.

Dasgupta, A. A., Hammarlund, E. E., Slifka, M. K. M., and Früh, K. K. (2007). Cowpox virus evades CTL recognition and inhibits the intracellular transport of MHC class I molecules. *J. Immunol.*, 178(3):1654–1661.

Daskalogianni, C., Apcher, S., Candeias, M. M., Naski, N., Calvo, F., and Fåhraeus, R. (2008). Gly-Ala repeats induce position- and substrate-specific regulation of 26 S proteasome-dependent partial processing. *J Biol Chem*, 283(44):30090–30100.

Datta, A., Sinha-Datta, U., Dhillon, N. K., Buch, S., and Nicot, C. (2006). The HTLV-I p30 interferes with TLR4 signaling and modulates the release of pro- and anti-inflammatory cytokines from human macrophages. *J Biol Chem*, 281(33):23414–23424.

Daubin, V. and Ochman, H. (2004). Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. *Genome Res*, 14(6):1036–1042.

Davies, J. and Davies, D. (2010). Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.*, 74(3):417–433.

Davis, F. P., Barkan, D. T., Eswar, N., McKerrow, J. H., and Sali, A. (2007). Host pathogen protein interactions predicted by comparative modeling. *Protein Sci*, 16(12):2585–2596.

Davison, A. J., Eberle, R., Ehlers, B., Hayward, G. S., McGeoch, D. J., Minson, A. C., Pellett, P. E., Roizman, B., Studdert, M. J., and Thiry, E. (2009). The order Herpesvirales. *Arch Virol*, 154(1):171–177.

Dawkins, R. (2006). *The Selfish Gene: 30th Anniversary Edition*. OUP Oxford.

Dawson, K. J. K. (1998). Evolutionarily Stable Mutation Rates. *J. Theor. Biol.*, 194(1):15–15.

de Chassey, B., Meyniel-Schicklin, L., Aublin-Gex, A., André, P., and Lotteau, V. (2012a). Genetic screens for the control of influenza virus replication: from meta-analysis to drug discovery. *Mol Biosyst*, 8(4):1297–1303.

de Chassey, B., Meyniel-Schicklin, L., Aublin-Gex, A., André, P., and Lotteau, V. (2012b). New horizons for antiviral drug discovery from virus-host protein interaction networks. *Curr Opin Vir*, 2(5):606–613.

de Chassey, B., Navratil, V., Tafforeau, L., Hiet, M. S., Aublin-Gex, A., Agaugué, S., Meiffren, G., Pradezynski, F., Faria, B. F., Chantier, T., Le Breton, M., Pellet, J., Davoust, N., Mangeot, P. E., Chaboud, A., Penin, F., Jacob, Y., Vidalain, P. O., Vidal, M., André, P., Rabourdin-Combe, C., and Lotteau, V. (2008). Hepatitis C virus infection protein network. *Mol Syst Biol*, 4:230.

de Clercq, E. (2007). Three decades of antiviral drugs. *Nat Rev Drug Disc*, 6(12):941–941.

de Clercq, E. (2011). *Antiviral Drug Strategies*. John Wiley & Sons.

De Franceschi, L., Fattovich, G., Turrini, F., Ayi, K., Brugnara, C., Manzato, F., Noventa, F., Stanzial, A. M., Solero, P., and Corrocher, R. (2000). Hemolytic anemia induced by ribavirin therapy in patients with chronic hepatitis C virus infection: role of membrane oxidative damage. *Hepatology (Baltimore, Md)*, 31(4):997–1004.

de Francesco, R. and Migliaccio, G. (2005). Challenges and successes in developing new therapies for hepatitis C. *Nature*, 436(7053):953–960.

De Las Rivas, J. and de Luis, A. (2004). Interactome data and databases: different types of protein interaction. *Comp Funct Genomics*, 5(2):173–178.

de Villiers, E.-M., Fauquet, C., Broker, T. R., Bernard, H.-U., and zur Hausen, H. (2004). Classification of papillomaviruses. *Virology*, 324(1):17–27.

de Visser, J. A. G. M. J. (2002). The fate of microbial mutators. *Annu Rev Microbiol*, 148(Pt 5):1247–1252.

Deamer, D. W. and Fleischaker, G. R. (1994). *Origins of life*. the central concepts. Jones & Bartlett Pub, Boston.

Deane, C. M., Salwiński, , Xenarios, I., and Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, 1(5):349–356.

DeDuve, C. (1991). *Blueprint for a Cell: The Nature and Origin of Life*. Carolina Biological Supply Co, 1 edition.

Deichmann, A., Brugman, M. H., Bartholomae, C. C., Schwarzwaelder, K., Verstegen, M. M. A., Howe, S. J., Arens, A., Ott, M. G., Hoelzer, D., Seger, R., Grez, M., Hacein-Bey-Abina, S., Cavazzana-Calvo, M., Fischer, A., Paruzynski, A., Gabriel, R., Glimm, H., Abel, U., Cattoglio, C., Mavilio, F., Cassani, B., Aiuti, A., Dunbar, C. E., Baum, C., Gaspar, H. B., Thrasher, A. J., von Kalle, C., Schmidt, M., and Wagemaker, G. (2011). Insertion sites in engrafted cells cluster within a limited repertoire of genomic areas after gammaretroviral vector gene therapy. *Mol. Ther.*, 19(11):2031–2039.

Deisboeck, T. S. and Couzin, I. D. (2009). Collective behavior in cancer cell populations. *Bioessays*, 31(2):190–197.

Delang, L., Vliegen, I., Froeyen, M., and Neyts, J. (2011). Comparative study of the genetic barriers and pathways towards resistance of selective inhibitors of hepatitis C virus replication. *Antimicrob Agents Chemother*, 55(9):4103–4113.

Delobel, P., Sandres-Sauné, K., Cazabat, M., Pasquier, C., Marchou, B., Massip, P., and Izopet, J. (2005). R5 to X4 switch of the predominant HIV-1 population in cellular reservoirs during effective highly active antiretroviral therapy. *J. Acquir. Immune Defic. Syndr.*, 38(4):382–392.

Delwart, E. (2013). A roadmap to the human virome. *PLoS Pathog*, 9(2):e1003146.

Delwart, E. L. (2007). Viral metagenomics. *Rev. Med. Virol.*, 17(2):115–131.

Dempsey, P. W., Allison, M. E., Akkaraju, S., Goodnow, C. C., and Fearon, D. T. (1996). C3d of complement as a molecular adjuvant: bridging innate and acquired immunity. *Science*, 271(5247):348–350.

Deng, M., Mehta, S., Sun, F., and Chen, T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Res*, 12(10):1540–1548.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43(5):491–498.

Der, C. J. (1987). Cellular oncogenes and human carcinogenesis. *Clin. Chem.*, 33(5):641–646.

Der, C. J., Krontiris, T. G., and Cooper, G. M. (1982). Transforming genes of human bladder and lung carcinoma cell lines are homologous to the ras genes of Harvey and Kirsten sarcoma viruses. *Proc Natl Acad Sci USA*, 79(11):3637–3640.

Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G., and Fertit, B. (1999). Genomic signature: Characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol*, 16(10):1391–1399.

Desnues, C., Rodriguez-Brito, B., Rayhawk, S., Kelley, S., Tran, T., Haynes, M., Liu, H., Furlan, M., Wegley, L., Chau, B., Ruan, Y., Hall, D., Angly, F. E., Edwards, R. A., Li, L., Thurber, R. V., Reid, R. P., Siefert, J., Souza, V., Valentine, D. L., Swan, B. K., Breitbart, M., and Rohwer, F. (2008). Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature*, 452(7185):340–343.

Dethlefsen, L., McFall-Ngai, M., and Relman, D. A. (2007). An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature*, 449(7164):811–818.

Devaraj, S. G., Wang, N., Chen, Z., Chen, Z., Tseng, M., Barretto, N., Lin, R., Peters, C. J., Tseng, C.-T. K., Baker, S. C., and Li, K. (2007). Regulation of IRF-3-dependent innate immunity by the papain-like protease domain of the severe acute respiratory syndrome coronavirus. *J Biol Chem*, 282(44):32208–32221.

Devare, S. G., Reddy, E. P., Law, J. D., Robbins, K. C., and Aaronson, S. A. (1983). Nucleotide sequence of the simian sarcoma virus genome: demonstration that its acquired cellular sequences encode the transforming gene product p28sis. *Proc Natl Acad Sci USA*, 80(3):731–735.

Dewannieux, M., Harper, F., Richaud, A., Letzelter, C., Ribet, D., Pierron, G., and Heidmann, T. (2006). Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Res*, 16(12):1548–1556.

Di Bartolo, D. L., Cannon, M., Liu, Y.-F., Renne, R., Chadburn, A., Boshoff, C., and Cesarman, E. (2008). KSHV LANA inhibits TGF-beta signaling through epigenetic silencing of the TGF-beta type II receptor. *Blood*, 111(9):4731–4740.

Di Bisceglie, A. M., Conjeevaram, H. S., Fried, M. W., Sallie, R., Park, Y., Yurdaydin, C., Swain, M., Kleiner, D. E., Mahaney, K., and Hoofnagle, J. H. (1995). Ribavirin as therapy for chronic hepatitis C. A randomized, double-blind, placebo-controlled trial. *Annals of Internal Medicine*, 123(12):897–903.

Di Bisceglie, A. M., Shindo, M., Fong, T. L., Fried, M. W., Swain, M. G., Bergasa, N. V., Axiotis, C. A., Waggoner, J. G., Park, Y., and Hoofnagle, J. H. (1992). A pilot study of ribavirin therapy for chronic hepatitis C. *Hepatology (Baltimore, Md)*, 16(3):649–654.

Diepolder, H. M., Zachoval, R., Hoffmann, R. M., Wierenga, E. A., Santantonio, T., Jung, M. C., Eichenlaub, D., and Pape, G. R. (1995). Possible mechanism involving T-lymphocyte response to non-structural protein 3 in viral clearance in acute hepatitis C virus infection. *Lancet*, 346(8981):1006–1007.

Dietz, J., Schelhorn, S.-E., Fitting, D., Mihm, U., Susser, S., Welker, M.-W., Füller, C., Däumer, M., Teuber, G., Wedemeyer, H., Berg, T., Lengauer, T., Zeuzem, S., Herrmann, E., and Sarrazin, C. (2013). Deep sequencing reveals mutagenic effects of ribavirin during monotherapy of hepatitis C virus genotype 1-infected patients. *J Virol*, 87(11):6172–6181.

Dimitrova, M., Imbert, I., Kieny, M. P., and Schuster, C. (2003). Protein-protein interactions between hepatitis C virus nonstructural proteins. *J Virol*, 77(9):5401–5414.

Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., Ritchey, J. K., Young, M. A., Lamprecht, T., McLellan, M. D., McMichael, J. F., Wallis, J. W., Lu, C., Shen, D., Harris, C. C., Dooling, D. J., Fulton, R. S., Fulton, L. L., Chen, K., Schmidt, H., Kalicki-Veizer, J., Magrini, V. J., Cook, L., McGrath, S. D., Vickery, T. L., Wendl, M. C., Heath, S., Watson, M. A., Link, D. C., Tomasson, M. H., Shannon, W. D., Payton, J. E., Kulkarni, S., Westervelt, P., Walter, M. J., Graubert, T. A., Mardis, E. R., Wilson, R. K., and Dipersio, J. F. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510.

Ding, L., Raphael, B. J., Chen, F., and Wendl, M. C. (2013). Advances for Studying Clonal Evolution in Cancer. *Cancer Lett.*, pages –.

Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., Furlan, M., Desnues, C., Haynes, M., Li, L., McDaniel, L., Moran, M. A., Nelson, K. E., Nilsson, C., Olson, R., Paul, J., Brito, B. R., Ruan, Y., Swan, B. K., Stevens, R., Valentine, D. L., Thurber, R. V., Wegley, L., White, B. A., and Rohwer, F. (2008). Functional metagenomic profiling of nine biomes. *Nature*, 452(7187):629–632.

Dixit, N. M., Layden-Almer, J. E., Layden, T. J., and Perelson, A. S. (2004). Modelling how ribavirin improves interferon response rates in hepatitis C virus infection. *Nature*, 432(7019):922–924.

Dixit, N. M. and Perelson, A. S. (2006). The metabolism, pharmacokinetics and mechanisms of antiviral activity of ribavirin against hepatitis C virus. *Cell Mol Life Sci*, 63(7-8):832–842.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

Doerfler, W. (1991). Patterns of DNA methylation–evolutionary vestiges of foreign DNA inactivation as a host defense mechanism. A proposal. *Biol. Chem. Hoppe-Seyler*, 372(8):557–564.

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*, 36(16):e105–e105.

Dolan, A., Addison, C., Gatherer, D., Davison, A. J., and McGeoch, D. J. (2006). The genome of Epstein-Barr virus type 2 strain AG876. *Virology*, 350(1):164–170.

Doll, R. and Hill, A. B. (1956). Lung cancer and other causes of death in relation to smoking; a second report on the mortality of British doctors. *Br Med J*, 2(5001):1071–1081.

Domingo, E. (1989). RNA virus evolution and the control of viral disease. *Prog Drug Res*, 33:93–133.

Domingo, E., editor (2006). *Quasispecies: Concept and Implications for Virology*, volume 299 of *Current Topics in Microbiology and Immunology*. Springer-Verlag, Berlin/Heidelberg.

Domingo, E., Escarmís, C., Baranowski, E., Ruiz-Jarabo, C. M., Carrillo, E., Núñez, J. I., and Sobrino, F. (2003). Evolution of foot-and-mouth disease virus. *Virus Res*, 91(1):47–63.

Domingo, E., Escarmís, C., Lázaro, E., and Manrubia, S. C. (2005). Quasispecies dynamics and RNA virus extinction. *Virus Res*, 107(2):129–139.

Domingo, E., Escarmís, C., Sevilla, N., Moya, A., Elena, S. F., Quer, J., Novella, I. S., and Holland, J. J. (1996). Basic concepts in RNA virus evolution. *FASEB J*, 10(8):859–864.

Domingo, E. and Gómez, J. (2007). Quasispecies and its impact on viral hepatitis. *Virus Res*, 127(2):131–150.

Domingo, E. and Holland, J. J. (1992). Complications of RNA heterogeneity for the engineering of virus vaccines and antiviral agents. *Genet. Eng. (N.Y.)*, 14:13–31.

Domingo, E. and Holland, J. J. (1997). RNA virus mutations and fitness for survival. *Annu Rev Microbiol*, 51:151–178.

Domingo, E., Sheldon, J., and Perales, C. (2012). Viral quasispecies evolution. *Microbiol. Mol. Biol. Rev.*, 76(2):159–216.

Domingo, E. and Wain-Hobson, S. (2009). The 30th anniversary of quasispecies. Meeting on 'Quasispecies: past, present and future'. In *EMBO Rep.*, pages 444–448.

Doolittle, J. M. and Gomez, S. M. (2010). Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens. *Virol. J.*, 7:82.

Doolittle, J. M. and Gomez, S. M. (2011). Mapping protein interactions between Dengue virus and its human and insect hosts. *PLoS Negl Trop Dis*, 5(2):e954.

Doolittle, W. F. W. and Bapteste, E. E. (2007). Pattern pluralism and the Tree of Life hypothesis. *Proc Natl Acad Sci USA*, 104(7):2043–2049.

Dorner, M., Horwitz, J. A., Robbins, J. B., Barry, W. T., Feng, Q., Mu, K., Jones, C. T., Schoggins, J. W., Catanese, M. T., Burton, D. R., Law, M., Rice, C. M., and Ploss, A. (2011). A genetically humanized mouse model for hepatitis C virus infection. *Nature*, 474(7350):208–211.

Dowdall, N. P., Evans, A. D., and Thibeault, C. (2010). Air Travel and TB: An airline perspective. *Travel Medicine and Infectious Disease*, 8(2):96–103.

Dragic, T. T., Trkola, A. A., Thompson, D. A. D., Cormier, E. G. E., Kajumo, F. A. F., Maxwell, E. E., Lin, S. W. S., Ying, W. W., Smith, S. O. S., Sakmar, T. P. T., and Moore, J. P. J. (2000). A binding pocket for a small molecule inhibitor of HIV-1 entry within the transmembrane helices of CCR5. *Proc Natl Acad Sci U S A*, 97(10):5639–5644.

Drake, J. W. (1999). The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes. *Ann. N. Y. Acad. Sci.*, 870:100–107.

Drake, J. W., Charlesworth, B., Charlesworth, D., and Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics*, 148(4):1667–1686.

Dröge, J. and McHardy, A. C. (2012). Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Brief Bioinform*, 13(6):646–655.

Duarte, E., Clarke, D., Moya, A., Domingo, E., and Holland, J. (1992). Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet. *Proc Natl Acad Sci USA*, 89(13):6015–6019.

Duarte, E. A., Clarke, D. K., Moya, A., Elena, S. F., Domingo, E., and Holland, J. (1993). Many-trillionfold amplification of single RNA virus particles fails to overcome the Muller's ratchet effect. *J Virol*, 67(6):3620–3623.

Dubin, G., Socolof, E., Frank, I., and Friedman, H. M. (1991). Herpes simplex virus type 1 Fc receptor protects infected cells from antibody-dependent cellular cytotoxicity. *J Virol*, 65(12):7046–7050.

Duensing, S., Lee, L. Y., Duensing, A., Basile, J., Piboonniyom, S., Gonzalez, S., Crum, C. P., and Münger, K. (2000). The human papillomavirus type 16 E6 and E7 oncoproteins cooperate to induce mitotic defects and genomic instability by uncoupling centrosome duplication from the cell division cycle. *Proc Natl Acad Sci USA*, 97(18):10002–10007.

Duensing, S. and Munger, K. (2002). The human papillomavirus type 16 E6 and E7 oncoproteins independently induce numerical and structural chromosome instability. *Cancer Research*, 62(23):7075–7082.

Duensing, S. and Munger, K. (2003). Human papillomavirus type 16 E7 oncoprotein can induce abnormal centrosome duplication through a mechanism independent of inactivation of retinoblastoma protein family members. *J Virol*, 77(22):12331–12335.

Duesberg, P. H. and Vogt, P. K. (1970). Differences between the ribonucleic acids of transforming and nontransforming avian tumor viruses. *Proc Natl Acad Sci USA*, 67(4):1673–1680.

Duffy, S., Shackelton, L. A., and Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet*, 9(4):267–276.

Duhaime, M. B. M. and Sullivan, M. B. M. (2012). Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology*, 434(2):181–186.

Duncan, C. G., Leary, R. J., Lin, J. C.-H., Cummins, J., Di, C., Schaefer, C. F., Wang, T.-L., Riggins, G. J., Edwards, J., Bigner, D., Kopelovich, L., Vogelstein, B., Kinzler, K. W., Velculescu, V. E., and Yan, H. (2009). Identification of microbial DNA in human cancer. *BMC medical genomics*, 2:22.

Durand, J. D. (1977). Historical Estimates of World Population: an Evaluation. *Population and Development Review*, 3(3):253–296.

Durmuş Tekir, S., Cakir, T., and Ülgen, K. Ö. (2012). Infection Strategies of Bacterial and Viral Pathogens through Pathogen-Human Protein-Protein Interactions. *Front Microbiol*, 3:46.

Durmuş Tekir, S. D. and Ülgen, K. Ö. (2013). Systems biology of pathogen-host interaction: networks of protein-protein interaction within pathogens and pathogen-human interactions in the post-genomic era. *Biotechnol J*, 8(1):85–96.

Dusheiko, G., Main, J., Thomas, H., Reichard, O., Lee, C., Dhillon, A., Rassam, S., Fryden, A., Reesink, H., Bassendine, M., Norkrans, G., Cuypers, T., Lelie, N., Telfer, P., Watson, J., Weegink, C., Sillikens, P., and Weiland, O. (1996). Ribavirin treatment for patients with chronic hepatitis C: results of a placebo-controlled study. *J Hepatol*, 25(5):591–598.

Dyer, M. D., Murali, T. M., and Sobral, B. W. (2007). Computational prediction of host-pathogen protein-protein interactions. *Bioinformatics*, 23(13):i159–66.

Dyer, M. D., Murali, T. M., and Sobral, B. W. (2008). The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathog*, 4(2):1–14.

Dyer, M. D., Murali, T. M., and Sobral, B. W. (2011). Supervised learning and prediction of physical interactions between human and HIV proteins. *Infect. Genet. Evol.*, 11(5):917–923.

Dyson, N., M, H. P., Münger, K., and Harlow, E. (1989). The human papilloma virus-16 E7 oncoprotein is able to bind to the retinoblastoma gene product. *Science*, 243(4893):934–937.

Dziembowski, A. and Séraphin, B. (2004). Recent developments in the analysis of protein complexes. *FEBS Lett*, 556(1-3):1–6.

Eash, S., Manley, K., Gasparovic, M., Querbes, W., and Atwood, W. J. (2006). The human polyomaviruses. *Cell Mol Life Sci*, 63(7-8):865–876.

Eckert, K. A. and Kunkel, T. A. (1991). DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl.*, 1(1):17–24.

Edinger, A. L. A., Amedee, A. A., Miller, K. K., Doranz, B. J. B., Endres, M. M., Sharron, M. M., Samson, M. M., Lu, Z. H. Z., Clements, J. E. J., Murphey-Corb, M. M., Peiper, S. C. S., Parmentier, M. M., Broder, C. C. C., and Doms, R. W. R. (1997). Differential utilization of CCR5 by macrophage and T cell tropic simian immunodeficiency virus strains. *Proc Natl Acad Sci U S A*, 94(8):4005–4010.

Edlin, B. R. (2011). Perspective: test and treat this silent killer. *Nature*, 474(7350):S18–9.

Edwards, R. A. and Rohwer, F. (2005). Opinion: Viral metagenomics. *Nat Rev Microbiol*, 3(6):504–510.

Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10):465–523.

Eigen, M. (1987). New concepts for dealing with the evolution of nucleic acids. *Cold Spring Harbor Symposia on Quantitative Biology*, 52:307–320.

Eigen, M. (1993a). The origin of genetic information: viruses as models. *Gene*, 135(1-2):37–47.

Eigen, M. (1993b). Viral quasispecies. *Sci. Am.*, 269(1):42–49.

Eigen, M. (2000). Natural selection: a phase transition? *Biophys. Chem.*, 85(2-3):101–123.

Eigen, M. and Schuster, P. (1977). The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. *Naturwissenschaften*, 64(11):541–565.

Eisen, J. A. (2007). Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biol*, 5(3):e82.

Eisenberg, D., Marcotte, E. M., Xenarios, I., and Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature*, 405(6788):823–826.

Elgui de Oliveira, D. (2007). DNA viruses in human cancer: an integrated overview on fundamental mechanisms of viral carcinogenesis. *Cancer Lett.*, 247(2):182–196.

Elion, G. B., Furman, P. A., Fyfe, J. A., de Miranda, P., Beauchamp, L., and Schaeffer, H. J. (1977). Selectivity of action of an antiherpetic agent, 9-(2-hydroxyethoxymethyl) guanine. *Proc Natl Acad Sci USA*, 74(12):5716–5720.

Ellerman, V. (1908). *Experimental leukemia in chickens*. Zentralbl. Bakteriol. Parasitenkd. Infectionskr. Hyg. Abt.

Embretson, J. J., Zupancic, M. M., Ribas, J. L. J., Burke, A. A., Racz, P. P., Tenner-Racz, K. K., and Haase, A. T. A. (1993). Massive covert infection of helper T lymphocytes and macrophages by HIV during the incubation period of AIDS. *Nature*, 362(6418):359–362.

Enam, S., Del Valle, L., Lara, C., Gan, D.-D., Ortiz-Hidalgo, C., Palazzo, J. P., and Khalili, K. (2002). Association of human polyomavirus JCV with colon cancer: evidence for interaction of viral T-antigen and beta-catenin. *Cancer Research*, 62(23):7093–7101.

Endres, M. J. M., Garlisi, C. G. C., Xiao, H. H., Shan, L. L., and Hedrick, J. A. J. (1999). The Kaposi's sarcoma-related herpesvirus (KSHV)-encoded chemokine vMIP-I is a specific agonist for the CC chemokine receptor (CCR)8. *J Exp Med*, 189(12):1993–1998.

Engels, E. A., Frisch, M., Goedert, J. J., Biggar, R. J., and Miller, R. W. (2002). Merkel cell carcinoma and HIV infection. *Lancet*, 359(9305):497–498.

Engels, E. A., Pfeiffer, R. M., Goedert, J. J., Virgo, P., McNeel, T. S., Scoppa, S. M., Biggar, R. J., and for the HIV/AIDS Cancer Match Study (2006). Trends in cancer risk among people with AIDS in the United States 1980-2002. *AIDS*, 20(12):1645–1654.

Enomoto, N., Sakuma, I., Asahina, Y., Kurosaki, M., Murakami, T., Yamamoto, C., Ogura, Y., Izumi, N., Marumo, F., and Sato, C. (1996). Mutations in the nonstructural protein 5A gene and response to interferon in patients with chronic hepatitis C virus 1b infection. *New England Journal of Medicine*, 334(2):77–81.

Epstein, M. A., Achong, B. G., and Barr, Y. M. (1964). Virus Particles In Cultured Lymphoblasts From Burkitt's Lymphoma. *Lancet*, 1(7335):702–703.

Erickson, A. L., Kimura, Y., Igarashi, S., Eichelberger, J., Houghton, M., Sidney, J., McKinney, D., Sette, A., Hughes, A. L., and Walker, C. M. (2001). The outcome of hepatitis C virus infection is predicted by escape mutations in epitopes targeted by cytotoxic T lymphocytes. *Immunity*, 15(6):883–895.

Eriksson, B., Helgstrand, E., Johansson, N. G., Larsson, A., Misiorny, A., Norén, J. O., Philipson, L., Stenberg, K., Stening, G., Stridh, S., and Oberg, B. (1977). Inhibition of influenza virus ribonucleic acid polymerase by ribavirin triphosphate. *Antimicrob Agents Chemother*, 11(6):946–951.

Eriksson, N., Pachter, L., Mitsuya, Y., Rhee, S.-Y., Wang, C., Gharizadeh, B., Ronaghi, M., Shafer, R. W., and Beerenwinkel, N. (2008). Viral population estimation using pyrosequencing. *PLoS Comput Biol*, 4(4):e1000074.

Ermolaeva, M. A., Michallet, M.-C., Papadopoulou, N., Utermöhlen, O., Kranidioti, K., Kollias, G., Tschopp, J., and Pasparakis, M. (2008). Function of TRADD in tumor necrosis factor receptor 1 signaling and in TRIF-dependent inflammatory responses. *Nat. Immunol.*, 9(9):1037–1046.

Ertl, R. R., Birzele, F. F., Hildebrandt, T. T., and Klein, D. D. (2011). Viral transcriptome analysis of feline immunodeficiency virus infected cells using second generation sequencing technology. *Vet Immunol Immunopathol*, 143(3-4):314–324.

Escarmís, C., Dávila, M., and Domingo, E. (1999). Multiple molecular pathways for fitness recovery of an RNA virus debilitated by operation of Muller's ratchet. *J Mol Biol*, 285(2):495–505.

Escarmís, C., Lazaro, E., and Manrubia, S. C. (2006). Population bottlenecks in quasispecies dynamics. *Curr. Top. Microbiol. Immunol.*, 299:141–170.

Esquela-Kerscher, A. and Slack, F. J. (2006). Oncomirs - microRNAs with a role in cancer. *Nat Rev Cancer*, 6(4):259–269.

Esteller, M. (2008). Epigenetics in cancer. *N. Engl. J. Med.*, 358(11):1148–1159.

Etkind, P., Du, J., Khan, A., Pillitteri, J., and Wiernik, P. H. (2000). Mouse mammary tumor virus-like ENV gene sequences in human breast tumors and in a lymphoma of a breast cancer patient. *Clin. Cancer Res.*, 6(4):1273–1278.

Evans, A. S. (1976). Causation and disease: the Henle-Koch postulates revisited. *Yale J Biol Med*, 49(2):175–195.

Evans, C., Kadner, S. V., Darroch, L. J., and Wilson, W. H. (2007). The relative significance of viral lysis and microzooplankton grazing as pathways of dimethylsulfoniopropionate (DMSP) cleavage: An Emiliania huxleyi culture study. *Limnol. Oceanogr.*, 52(3):1036–1045.

Evans, P., Dampier, W., Ungar, L., and Tozeren, A. (2009). Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. *BMC medical genomics*, 2:27.

Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genes Dev*, 8(3):186–194.

Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M. D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y. V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J.-P., Duewel, H. S., Stewart, I. I., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S.-L., Moran, M. F., Morin, G. B., Topaloglou, T., and Figeys, D. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol*, 3:89.

Eyckerman, S., Verhee, A., der Heyden, J. V., Lemmens, I., Ostade, X. V., Vandekerckhove, J., and Tavernier, J. (2001). Design and application of a cytokine-receptor-based interaction trap. *Nat Cell Biol*, 3(12):1114–1119.

Fahey, M. E., Bennett, M. J., Mahon, C., Jäger, S., Pache, L., Kumar, D., Shapiro, A., Rao, K., Chanda, S. K., Craik, C. S., Frankel, A. D., and Krogan, N. J. (2011). GPS-Prot: a web-based visualization platform for integrating host-pathogen interaction data. *BMC Bioinformatics*, 12:298.

Falkow, S. (1988). Molecular Koch's postulates applied to microbial pathogenicity. *Rev Infect Dis*, (10(Suppl 2)):S274–S276.

Falkowska, E., Kajumo, F., Garcia, E., Reinus, J., and Dragic, T. (2007). Hepatitis C virus envelope glycoprotein E2 glycans modulate entry, CD81 binding, and neutralization. *J Virol*, 81(15):8072–8079.

Fan, X., Mao, Q., Zhou, D., Lu, Y., Xing, J., Xu, Y., Ray, S. C., and Di Bisceglie, A. M. (2009). High diversity of hepatitis C viral quasispecies is associated with early virological response in patients undergoing antiviral therapy. *Hepatology (Baltimore, Md)*, 50(6):1765–1772.

Fancello, L., Raoult, D., and Desnues, C. (2012). Computational tools for viral metagenomics and their application in clinical research. *Virology*, 434(2):162–174.

Fang, S. H., Hwang, L. H., Chen, D. S., and Chiang, B. L. (2000). Ribavirin enhancement of hepatitis C virus core antigen-specific type 1 T helper cell response correlates with the increased IL-12 level. *J Hepatol*, 33(5):791–798.

Fang, Z. and Cui, X. (2011). Design and validation issues in RNA-seq experiments. *Brief Bioinform*, 12(3):280–287.

Farci, P. (2011). New insights into the HCV quasispecies and compartmentalization. *Semin. Liver Dis.*, 31(4):356–374.

Farci, P., Shimoda, A., Coiana, A., Diaz, G., Peddis, G., Melpolder, J. C., Strazzera, A., Chien, D. Y., Munoz, S. J., Balestrieri, A., Purcell, R. H., and Alter, H. J. (2000). The outcome of acute hepatitis C predicted by the evolution of the viral quasispecies. *Science*, 288(5464):339–344.

Farci, P., Shimoda, A., Wong, D., Cabezon, T., De Gioannis, D., Strazzera, A., Shimizu, Y., Shapiro, M., Alter, H. J., and Purcell, R. H. (1996). Prevention of hepatitis C virus infection in chimpanzees by hyperimmune serum against the hypervariable region 1 of the envelope 2 protein. *Proc Natl Acad Sci USA*, 93(26):15394–15399.

Fares, M. A. M., Ruiz-González, M. X. M., Moya, A. A., Elena, S. F. S., and Barrio, E. E. (2002). Endosymbiotic bacteria: groEL buffers against deleterious mutations. *Nature*, 417(6887):398–398.

Fätkenheuer, G. G., Nelson, M. M., Lazzarin, A. A., Konourina, I. I., Hoepelman, A. I. M. A., Lampiris, H. H., Hirschel, B. B., Tebas, P. P., Raffi, F. F., Trottier, B. B., Bellos, N. N., Saag, M. M., Cooper, D. A. D., Westby, M. M., Tawadrous, M. M., Sullivan, J. F. J., Ridgway, C. C., Dunne, M. W. M., Felstead, S. S., Mayer, H. H., and van der Ryst, E. E. (2008). Subgroup analyses of maraviroc in previously treated R5 HIV-1 infection. *New England Journal of Medicine*, 359(14):1442–1455.

Fauci, A. S. (2001). Infectious Diseases: Considerations for the 21st Century. *Clin Infect Dis*, 32(5):675–685.

Fauquet, C. M. and Fargette, D. (2005). International Committee on Taxonomy of Viruses and the 3,142 unassigned species. *Virol. J.*, 2:64–64.

Fauquet, C. M. C. and Stanley, J. J. (2005). Revising the way we conceive and name viruses below the species level: a review of geminivirus taxonomy calls for new standardized isolate descriptors. *Arch Virol*, 150(10):2151–2179.

Faust, K. and Raes, J. (2012). Microbial interactions: from networks to models. *Nat Rev Microbiol*, 10(8):538–550.

Fay, J. C., Wyckoff, G. J., and Wu, C. I. (2001). Positive and negative selection on the human genome. *Genetics*, 158(3):1227–1234.

Fazilleau, N., Eisenbraun, M. D., Malherbe, L., Ebright, J. N., Pogue-Caley, R. R., McHeyzer-Williams, L. J., and McHeyzer-Williams, M. G. (2007). Lymphoid reservoirs of antigen-specific memory T helper cells. *Nat. Immunol.*, 8(7):753–761.

Feigelstock, D. A., Mihalik, K. B., and Feinstone, S. M. (2011). Selection of hepatitis C virus resistant to ribavirin. *Virol. J.*, 8:402.

Feinberg, A. P., Cui, H., and Ohlsson, R. (2002). DNA methylation and genomic imprinting: insights from cancer into epigenetic mechanisms. *Semin. Cancer Biol.*, 12(5):389–398.

Feld, J. J. (2012). Is there a role for ribavirin in the era of hepatitis C virus direct-acting antivirals? *Gastroenterology*, 142(6):1356–1359.

Feld, J. J. and Hoofnagle, J. H. (2005). Mechanism of action of interferon and ribavirin in treatment of hepatitis C. *Nature*, 436(7053):967–972.

Feld, J. J., Nanda, S., Huang, Y., Chen, W., Cam, M., Pusek, S. N., Schweigler, L. M., Theodore, D., Zacks, S. L., Liang, T. J., and Fried, M. W. (2007). Hepatic gene expression during treatment with peginterferon and ribavirin: Identifying molecular pathways for treatment response. *Hepatology (Baltimore, Md)*, 46(5):1548–1563.

Feldhahn, M., Menzel, M., Weide, B., Bauer, P., Meckbach, D., Garbe, C., Kohlbacher, O., and Bauer, J. (2011). No evidence of viral genomes in whole-transcriptome sequencing of three melanoma metastases. *Exp. Dermatol.*, 20(9):766–768.

Felini, M., Preacely, N., Shah, N., Christopher, A., Sarda, V., Elfaramawi, M., Sall, M., Bangara, S., Gandhi, S., and Johnson, E. S. (2012). A case-cohort study of lung cancer in poultry and control workers: occupational findings. *Occup Environ Med*, 69(3):191–197.

Fellay, J., Thompson, A. J., Ge, D., Gumbs, C. E., Urban, T. J., Shianna, K. V., Little, L. D., Qiu, P., Bertelsen, A. H., Watson, M., Warner, A., Muir, A. J., Brass, C., Albrecht, J., Sulkowski, M., McHutchison, J. G., and Goldstein, D. B. (2010). ITPA gene variants protect against anaemia in patients treated for chronic hepatitis C. *Nature*, 464(7287):405–408.

Feng, H., Shuda, M., Chang, Y., and Moore, P. S. (2008). Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science*, 319(5866):1096–1100.

Feng, H., Taylor, J. L., Benos, P. V., Newton, R., Waddell, K., Lucas, S. B., Chang, Y., and Moore, P. S. (2007). Human transcriptome subtraction by using short sequence tags to search for tumor viruses in conjunctival carcinoma. *J Virol*, 81(20):11332–11340.

Fernandez, A. F. and Esteller, M. (2010). Viral epigenomes in human tumorigenesis. *Oncogene*, 29(10):1405–1420.

Fernandez, A. F., Rosales, C., Lopez-Nieva, P., Graña, O., Ballestar, E., Ropero, S., Espada, J., Melo, S. A., Lujambio, A., Fraga, M. F., Pino, I., Javierre, B., Carmona, F. J., Acquadro, F., Steenbergen, R. D. M., Snijders, P. J. F., Meijer, C. J., Pineau, P., Dejean, A., Lloveras, B., Capella, G., Quer, J., Buti, M., Esteban, J.-I., Allende, H., Rodriguez-Frias, F., Castellsague, X., Minarovits, J., Ponce, J., Capello, D., Gaidano, G., Cigudosa, J. C., Gomez-Lopez, G., Pisano, D. G., Valencia, A., Piris, M. A., Bosch, F. X., Cahir-McFarland, E., Kieff, E., and Esteller, M. (2009). The dynamic DNA methylomes of double-stranded DNA viruses associated with human cancer. *Genome Res*, 19(3):438–451.

Ferrari, R., Berk, A. J., and Kurdistani, S. K. (2009). Viral manipulation of the host epigenome for oncogenic transformation. *Nat Rev Genet*, 10(5):290–294.

Ferrari, R., Pellegrini, M., Horwitz, G. A., Xie, W., Berk, A. J., and Kurdistani, S. K. (2008). Epigenetic reprogramming by adenovirus e1a. *Science*, 321(5892):1086–1088.

Feschotte, C. and Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet*, 13(4):283–296.

Fields, S. (2005). High-throughput two-hybrid analysis. The promise and the peril. *FEBS J.*, 272(21):5391–5399.

Fields, S., Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., and Rothberg, J. M. (2000). A comprehensive analysis of protein | [ndash] | protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770):623–627.

Figlerowicz, M., Alejska, M., Kurzyńska-Kokorniak, A., and Figlerowicz, M. (2003). Genetic variability: the key problem in the prevention and therapy of RNA-based virus infections. *Med Res Rev*, 23(4):488–518.

Filée, J., Forterre, P., and Laurent, J. (2003). The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies. *Res. Microbiol.*, 154(4):237–243.

Finkbeiner, S. R., Allred, A. F., Tarr, P. I., Klein, E. J., Kirkwood, C. D., and Wang, D. (2008). Metagenomic analysis of human diarrhea: viral detection and discovery. *PLoS Pathog*, 4(2):e1000011.

Finlay, C. A., Hinds, P. W., and Levine, A. J. (1989). The p53 proto-oncogene can act as a suppressor of transformation. *Cell*, 57(7):1083–1093.

Finn, R., Marshall, M., and Bateman, A. (2004). iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, 21(3):410–412.

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res*, 39(SUPPL. 2):W29–W37.

Finotello, F., Lavezzo, E., Fontana, P., Peruzzo, D., Albiero, A., Barzon, L., Falda, M., Di Camillo, B., and Toppo, S. (2012). Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data. *Brief Bioinform*, 13(3):269–280.

Firth, C., Kitchen, A., Shapiro, B., Suchard, M. A., Holmes, E. C., and Rambaut, A. (2010). Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Mol Biol Evol*, 27(9):2038–2051.

Fischer, M., Oberthuer, A., Brors, B., Kahlert, Y., Skowron, M., Voth, H., Warnat, P., Ernestus, K., Hero, B., and Berthold, F. (2006). Differential expression of neuronal genes defines subtypes of disseminated neuroblastoma with favorable and unfavorable outcome. *Clin. Cancer Res.*, 12(17):5118–5128.

Flaegstad, T., Andresen, P. A., Johnsen, J. I., Asomani, S. K., Jørgensen, G. E., Vignarajan, S., Kjuul, A., Kogner, P., and Traavik, T. (1999). A possible contributory role of BK virus infection in neuroblastoma development. *Cancer Research*, 59(5):1160–1163.

Flaherty, P., Natsoulis, G., Muralidharan, O., Winters, M., Buenrostro, J., Bell, J., Brown, S., Holodniy, M., Zhang, N., and Ji, H. P. (2012). Ultrasensitive detection of rare mutations using next-generation targeted resequencing. *Nucleic Acids Res*, 40(1):e2.

Flajolet, M., Rotondo, G., Daviet, L., Bergametti, F., Inchauspé, G., Tiollais, P., Transy, C., and Legrain, P. (2000). A genomic approach of the hepatitis C virus generates a protein interaction map. *Gene*, 242(1-2):369–379.

Flanagan, J. M. (2007). Host epigenetic modifications by oncogenic viruses. *Br. J. Cancer*, 96(2):183–188.

Flanegan, J. B., Petterson, R. F., Ambros, V., Hewlett, N. J., and Baltimore, D. (1977). Covalent linkage of a protein to a defined nucleotide sequence at the 5′-terminus of virion and replicative intermediate RNAs of poliovirus. *Proc Natl Acad Sci USA*, 74(3):961–965.

Flexner, C. (2007). HIV drug development: the next 25 years. *Nat Rev Drug Disc*, 6(12):959–966.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A. K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H. S., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sobral, D., Tang, Y. A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X. M., Harrow, J., Herrero, J., Hubbard, T.

J. P., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A., and Searle, S. M. J. (2012). Ensembl 2012. *Nucleic Acids Res*, 40(Database issue):D84–90.

Flores, O., Lee, G., Kessler, J., Miller, M., Schlief, W., Tomassini, J., and Hazuda, D. (1999). Host-cell positive transcription elongation factor b kinase activity is essential and limiting for HIV type 1 replication. *Proc Natl Acad Sci USA*, 96(13):7208–7213.

Fontana, W. W. and Schuster, P. P. (1987). A computer model of evolutionary optimization. *Biophys. Chem.*, 26(2-3):123–147.

Ford, C. E., Faedo, M., Crouch, R., Lawson, J. S., and Rawlinson, W. D. (2004a). Progression from normal breast pathology to breast cancer is associated with increasing prevalence of mouse mammary tumor virus-like sequences in men and women. *Cancer Research*, 64(14):4755–4759.

Ford, C. E., Faedo, M., and Rawlinson, W. D. (2004b). Mouse mammary tumor virus-like RNA transcripts and DNA are found in affected cells of human breast cancer. *Clin. Cancer Res.*, 10(21):7284–7289.

Forst, C. V. (2006). Host-pathogen systems biology. *Drug Discov. Today*, 11(5-6):220–227.

Forterre, P. (2002). The origin of DNA genomes and DNA replication proteins. *Curr Opin Microbiol*, 5(5):525–532.

Forterre, P. (2005). The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. *Biochimie*, 87(9-10):793–803.

Forterre, P. (2006a). The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res*, 117(1):5–16.

Forterre, P. (2006b). Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain. *Proc Natl Acad Sci USA*, 103(10):3669–3674.

Forterre, P. (2011). A new fusion hypothesis for the origin of Eukarya: better than previous ones, but probably also wrong. *Res. Microbiol.*, 162(1):77–91.

Forterre, P. and Philippe, H. (1999). Where is the root of the universal tree of life? *Bioessays*, 21(10):871–879.

Forterre, P. and Prangishvili, D. (2009a). The great billion-year war between ribosome- and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties. *Ann. N. Y. Acad. Sci.*, 1178:65–77.

Forterre, P. and Prangishvili, D. (2009b). The origin of viruses. *Res. Microbiol.*, 160(7):466–472.

Fouchier, R. A. M., Hartwig, N. G., Bestebroer, T. M., Niemeyer, B., de Jong, J. C., Simon, J. H., and Osterhaus, A. D. M. E. (2004). A previously undescribed coronavirus associated with respiratory disease in humans. *Proc Natl Acad Sci USA*, 101(16):6212–6216.

Fox, E. J. and Loeb, L. A. (2010). Lethal mutagenesis: targeting the mutator phenotype in cancer. *Semin. Cancer Biol.*, 20(5):353–359.

Fox, J. L. (2007). Antivirals become a broader enterprise. *Nat Biotechnol*, 25(12):1395–1402.

Frank, C., Werber, D., Cramer, J. P., Askar, M., Faber, M., An der Heiden, M., Bernard, H., Fruth, A., Prager, R., Spode, A., Wadl, M., Zoufaly, A., Jordan, S., Kemper, M. J., Follin, P., Müller, L., King, L. A., Rosner, B., Buchholz, U., Stark, K., and Krause, G. (2011). Epidemic Profile of Shiga-Toxin–Producing Escherichia coliO104:H4 Outbreak in Germany. *N. Engl. J. Med.*, 365(19):1771–1780.

Franzosa, E. A., Garamszegi, S., and Xia, Y. (2012). Toward a three-dimensional view of protein networks between species. *Front Microbiol*, 3:428.

Franzosa, E. A. and Xia, Y. (2011). Structural principles within the human-virus protein-protein interaction network. *Proc Natl Acad Sci USA*, 108(26):10538–10543.

Fredericks, D. N. and Relman, D. A. (1996). Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clin. Microbiol. Rev.*, 9(1):18–33.

Frenkel, N. and Roizman, B. (1972). Ribonucleic acid synthesis in cells infected with herpes simplex virus: controls of transcription and of RNA abundance. *Proc Natl Acad Sci USA*, 69(9):2654–2658.

Friborg, J., Kong, W., Hottiger, M. O., and Nabel, G. J. (1999). p53 inhibition by the LANA protein of KSHV protects against cell death. *Nature*, 402(6764):889–894.

Fried, M. W., Shiffman, M. L., Reddy, K. R., Smith, C., Marinos, G., Gonçales, F. L., Häussinger, D., Diago, M., Carosi, G., Dhumeaux, D., Craxi, A., Lin, A., Hoffman, J., and Yu, J. (2002). Peginterferon alfa-2a plus ribavirin for chronic hepatitis C virus infection. *N. Engl. J. Med.*, 347(13):975–982.

Friedel, C. C. and Haas, J. (2011). Virus-host interactomes and global models of virus-infected cells. *Trends in microbiology*, 19(10):501–508.

Friedel, C. C., Krumsiek, J., and Zimmer, R. (2009). Bootstrapping the interactome: unsupervised identification of protein complexes in yeast. *J Comput Biol*, 16(8):971–987.

Froussard, P. (1992). A random-PCR method (rPCR) to construct whole cDNA library from low amounts of RNA. *Nucleic Acids Res*, 20(11):2900.

Fu, W., Sanders-Beer, B. E., Katz, K. S., Maglott, D. R., Pruitt, K. D., and Ptak, R. G. (2009). Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Res*, 37(Database issue):D417–22.

Futse, J. E., Brayton, K. A., Dark, M. J., Knowles, D. P., and Palmer, G. H. (2008). Superinfection as a driver of genomic diversification in antigenically variant pathogens. *Proc Natl Acad Sci USA*, 105(6):2123–2127.

Gale, M. J., Korth, M. J., Tang, N. M., Tan, S. L., Hopkins, D. A., Dever, T. E., Polyak, S. J., Gretch, D. R., and Katze, M. G. (1997). Evidence that hepatitis C virus resistance to interferon is mediated through repression of the PKR protein kinase by the nonstructural 5A protein. *Virology*, 230(2):217–227.

Gallagher, A., Perry, J., Freeland, J., Alexander, F. E., Carman, W. F., Shield, L., Cartwright, R., and Jarrett, R. F. (2003). Hodgkin lymphoma and Epstein-Barr virus (EBV): no evidence to support hit-and-run mechanism in cases classified as non-EBV-associated. *Int. J. Cancer*, 104(5):624–630.

Gane, E. J., Stedman, C. A., Hyland, R. H., Ding, X., Svarovskaia, E., Symonds, W. T., Hindes, R. G., and Berrey, M. M. (2013). Nucleotide polymerase inhibitor sofosbuvir plus ribavirin for hepatitis C. *N. Engl. J. Med.*, 368(1):34–44.

Gánti, T. (2003). *The principles of life*. Oxford University Press, Oxford.

Gao, M., Nettles, R. E., Belema, M., Snyder, L. B., Nguyen, V. N., Fridell, R. A., Serrano-Wu, M. H., Langley, D. R., Sun, J.-H., O'Boyle, D. R., Lemm, J. A., Wang, C., Knipe, J. O., Chien, C., Colonno, R. J., Grasela, D. M., Meanwell, N. A., and Hamann, L. G. (2010). Chemical genetics strategy identifies an HCV NS5A inhibitor with a potent clinical effect. *Nature*, 465(7294):96–100.

Garbelli, A., Radi, M., Falchi, F., Beermann, S., Zanoli, S., Manetti, F., Dietrich, U., Botta, M., and Maga, G. (2011). Targeting the human DEAD-box polypeptide 3 (DDX3) RNA helicase as a novel strategy to inhibit viral replication. *Curr. Med. Chem.*, 18(20):3015–3027.

Garcia-Alvarez, L., Dawson, S., Cookson, B., and Hawkey, P. (2012). Working across the veterinary and human health sectors. *J. Antimicrob. Chemother.*, 67(Suppl. 1):i37–i49.

García-Arriaza, J., Domingo, E., and Briones, C. (2007). Characterization of minority subpopulations in the mutant spectrum of HIV-1 quasispecies by successive specific amplifications. *Virus Res*, 129(2):123–134.

Garcia-Pichel, F., Belnap, J., Neuer, S., and Schanz, F. (2003). Estimates of global cyanobacterial biomass and its distribution. *Algo Stud*, 109(1):213–227.

Gardner, S. D., Field, A. M., Coleman, D. V., and Hulme, B. (1971). New human papovavirus (B.K.) isolated from urine after renal transplantation. *Lancet*, 1(7712):1253–1257.

Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M.-A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.-M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Küster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636.

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147.

Gavin, A.-C., Maeda, K., and Kühner, S. (2011). Recent advances in charting protein-protein interaction: mass spectrometry-based approaches. *Curr. Opin. Biotechnol.*, 22(1):42–49.

Gavin, A.-C. and Superti-Furga, G. (2003). Protein complexes and proteome organization from yeast to man. *Current opinion in chemical biology*, 7(1):21–27.

Gazdar, A. F., Butel, J. S., and Carbone, M. (2002). SV40 and human tumours: myth, association or causality? *Nat Rev Cancer*, 2(12):957–964.

Ge, D., Fellay, J., Thompson, A. J., Simon, J. S., Shianna, K. V., Urban, T. J., Heinzen, E. L., Qiu, P., Bertelsen, A. H., Muir, A. J., Sulkowski, M., McHutchison, J. G., and Goldstein, D. B. (2009). Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature*, 461(7262):399–401.

Geisbert, T. W. and Jahrling, P. B. (2004). Exotic emerging viral diseases: progress and challenges. *Nat Med*, 10(12 Suppl):S110–21.

Geller, R., Taguwa, S., and Frydman, J. (2012). Broad action of Hsp90 as a host chaperone required for viral replication. *Biochim. Biophys. Acta*, 1823(3):698–706.

Geller, R., Vignuzzi, M., Andino, R., and Frydman, J. (2007). Evolutionary constraints on chaperone-mediated folding provide an antiviral approach refractory to development of drug resistance. *Genes Dev*, 21(2):195–205.

Gerlach, W., Jünemann, S., Tille, F., Goesmann, A., and Stoye, J. (2009). WebCARMA: A web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics*, 10:430.

Gerlach, W. and Stoye, J. (2011). Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Res*, 39(14):e91.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N. Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C. R., Nohadani, M., Eklund, A. C., Spencer-Dene, B., Clark, G., Pickering, L., Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P. A., and Swanton, C. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.*, 366(10):883–892.

Gerrish, P. J. and García-Lerma, J. G. (2003). Mutation rate and the efficacy of antimicrobial drug treatment. *Lancet Infect Dis*, 3(1):28–32.

Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., and Beerenwinkel, N. (2012). Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat Commun*, 3:811.

Geuking, M. B., Weber, J., Dewannieux, M., Gorelik, E., Heidmann, T., Hengartner, H., Zinkernagel, R. M., and Hangartner, L. (2009). Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. *Science*, 323(5912):393–396.

Gewurz, B. E., Gaudet, R., Tortorella, D., Wang, E. W., Ploegh, H. L., and Wiley, D. C. (2001). Antigen presentation subverted: Structure of the human cytomegalovirus protein US2 bound to the class I molecule HLA-A2. *Proc Natl Acad Sci USA*, 98(12):6794–6799.

Ghosh, T. S., Mohammed, M. H., Komanduri, D., and Mande, S. S. (2011). ProViDE: A software tool for accurate estimation of viral diversity in metagenomic samples. *Bioinformation*, 6(2):91–94.

Ghosh, T. S., Monzoorul Haque, M., and Mande, S. S. (2010). DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenomic sequences. *BMC Bioinformatics*, 11 Suppl 7:S14.

Gifford, R. and Tristem, M. (2003). The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes*, 26(3):291–315.

Gilbert, C. and Feschotte, C. (2010). Genomic fossils calibrate the long-term evolution of hepadnaviruses. *PLoS Biol*, 8(9).

Gilles, A., Meglécz, E., Pech, N., Ferreira, S., Malausa, T., and Martin, J.-F. (2011). Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics*, 12:245.

Gillespie, J. J., Wattam, A. R., Cammer, S. A., Gabbard, J. L., Shukla, M. P., Dalay, O., Driscoll, T., Hix, D., Mane, S. P., Mao, C., Nordberg, E. K., Scott, M., Schulman, J. R., Snyder, E. E., Sullivan, D. E., Wang, C., Warren, A., Williams, K. P., Xue, T., Yoo, H. S., Zhang, C., Zhang, Y., Will, R., Kenyon, R. W., and Sobral, B. W. (2011). PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect. Immun.*, 79(11):4286–4298.

Gingras, A.-C., Gstaiger, M., Raught, B., and Aebersold, R. (2007). Analysis of protein complexes using mass spectrometry. *Nat Rev Mol Cell Biol*, 8(8):645–654.

Gingras, A.-C. and Raught, B. (2012). Beyond hairballs: The use of quantitative mass spectrometry data to understand proteinâ€"protein interactions. *FEBS Lett*, 586(17):2723–2731.

Glansdorff, N., Xu, Y., and Labedan, B. (2008). The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biol Direct*, 3:29.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Mol Ecol Resour*, 11(5):759–769.

Glickman, M. S. and Sawyers, C. L. (2012). Converting cancer therapies into cures: lessons from infectious diseases. *Cell*, 148(6):1089–1098.

Goff, S. P. (2008). Knockdown screens to knockout HIV-1. *Cell*, 135(3):417–420.

Golden-Mason, L., Palmer, B. E., Kassam, N., Townshend-Bulson, L., Livingston, S., McMahon, B. J., Castelblanco, N., Kuchroo, V., Gretch, D. R., and Rosen, H. R. (2009). Negative immune regulator Tim-3 is overexpressed on T cells in hepatitis C virus infection and its blockade rescues dysfunctional CD4+ and CD8+ T cells. *J Virol*, 83(18):9122–9130.

Gong, Y., Kakihara, Y., Krogan, N., Greenblatt, J., Emili, A., Zhang, Z., and Houry, W. A. (2009). An atlas of chaperone-protein interactions in Saccharomyces cerevisiae: implications to protein folding pathways in the cell. *Mol Syst Biol*, 5:275.

Gonzalez, M. W. and Kann, M. G. (2012). Chapter 4: Protein interactions and disease. *PLoS Comput Biol*, 8(12):e1002819.

Gonzalez, O., Fontanes, V., Raychaudhuri, S., Loo, R., Loo, J., Arumugaswami, V., Sun, R., Dasgupta, A., and French, S. W. (2009). The heat shock protein inhibitor Quercetin attenuates hepatitis C virus production. *Hepatology (Baltimore, Md)*, 50(6):1756–1764.

González-García, I. I., Solé, R. V. R., and Costa, J. J. (2002). Metapopulation dynamics and spatial heterogeneity in cancer. *Proc Natl Acad Sci U S A*, 99(20):13085–13089.

González-López, C., Gómez-Mariano, G., Escarmís, C., and Domingo, E. (2005). Invariant aphthovirus consensus nucleotide sequence in the transition to error catastrophe. *Infect. Genet. Evol.*, 5(4):366–374.

Gorbalenya, A. E., Koonin, E. V., and Wolf, Y. I. (1990). A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. *FEBS Lett*, 262(1):145–148.

Goren, A., Ozsolak, F., Shoresh, N., Ku, M., Adli, M., Hart, C., Gymrek, M., Zuk, O., Regev, A., Milos, P. M., and Bernstein, B. E. (2010). Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. *Nat Methods*, 7(1):47–49.

Gorry, P. R. P. and Ancuta, P. P. (2011). Coreceptors and HIV-1 pathogenesis. *Curr HIV/AIDS Rep*, 8(1):45–53.

Goswami, B. B., Borek, E., Sharma, O. K., Fujitaki, J., and Smith, R. A. (1979). The broad spectrum antiviral agent ribavirin inhibits capping of mRNA. *Biochem. Biophys. Res. Commun.*, 89(3):830–836.

Goya, R., Sun, M. G. F., Morin, R. D., Leung, G., Ha, G., Wiegand, K. C., Senz, J., Crisan, A., Marra, M. A., Hirst, M., Huntsman, D., Murphy, K. P., Aparicio, S., and Shah, S. P. (2010). SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, 26(6):730–736.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*, 29(7):644–652.

Graci, J. D. and Cameron, C. E. (2002). Quasispecies, error catastrophe, and the antiviral activity of ribavirin. *Virology*.

Graci, J. D. and Cameron, C. E. (2008). Therapeutically targeting RNA viruses via lethal mutagenesis. *Future virology*, 3(6):553–566.

Graci, J. D., Gnädig, N. F., Galarraga, J. E., Castro, C., Vignuzzi, M., and Cameron, C. E. (2012). Mutational robustness of an RNA virus influences sensitivity to lethal mutagenesis. *J Virol*, 86(5):2869–2873.

Grakoui, A., Shoukry, N. H., Woollard, D. J., Han, J.-H., Hanson, H. L., Ghrayeb, J., Murthy, K. K., Rice, C. M., and Walker, C. M. (2003). HCV persistence and immune evasion in the absence of memory T cell help. *Science*, 302(5645):659–662.

Grande-Pérez, A., Lázaro, E., Lowenstein, P., Domingo, E., and Manrubia, S. C. (2005). Suppression of viral infectivity through lethal defection. *Proc Natl Acad Sci USA*, 102(12):4448–4452.

Grande-Pérez, A., Sierra, S., Castro, M. G., Domingo, E., and Lowenstein, P. R. (2002). Molecular indetermination in the transition to error catastrophe: systematic elimination of lymphocytic choriomeningitis virus through mutagenesis does not correlate linearly with large increases in mutant spectrum complexity. *Proc Natl Acad Sci USA*, 99(20):12938–12943.

Gray, M. W. and Lang, B. F. (1998). Transcription in chloroplasts and mitochondria: a tale of two polymerases. *Trends in microbiology*, 6(1):1–3.

Grayson, P., Evilevitch, A., Inamdar, M. M., Purohit, P. K., Gelbart, W. M., Knobler, C. M., and Phillips, R. (2006). The effect of genome length on ejection forces in bacteriophage lambda. *Virology*, 348(2):430–436.

Greenblatt, M. S., Bennett, W. P., Hollstein, M., and Harris, C. C. (1994). Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. *Cancer Research*, 54(18):4855–4878.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E. E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., Clements, J., Cole, J., Dicks, E., Forbes, S., Gray, K., Halliday, K., Harrison, R., Hills, K., Hinton, J., Jenkinson, A., Jones, D., Menzies, A., Mironenko, T., Perry, J., Raine, K., Richardson, D., Shepherd, R., Small, A., Tofts, C., Varian, J., Webb, T., West, S., Widaa, S., Yates, A., Cahill, D. P., Louis, D. N., Goldstraw, P., Nicholson, A. G., Brasseur, F., Looijenga, L., Weber, B. L., Chiew, Y.-E., deFazio, A., Greaves, M. F., Green, A. R., Campbell, P., Birney, E., Easton, D. F., Chenevix-Trench, G., Tan, M.-H., Khoo, S. K., Teh, B. T., Yuen, S. T., Leung, S. Y., Wooster, R., Futreal, P. A., and Stratton, M. R. (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158.

Grégoire, I. P., Richetta, C., Meyniel-Schicklin, L., Borel, S., Pradezynski, F., Diaz, O., Deloire, A., Azocar, O., Baguet, J., Le Breton, M., Mangeot, P. E., Navratil, V., Joubert, P.-E., Flacher, M., Vidalain, P.-O., André, P., Lotteau, V., Biard-Piechaczyk, M., Rabourdin-Combe, C., and Faure, M. (2011). IRGM is a common target of RNA viruses that subvert the autophagy network. *PLoS Pathog*, 7(12):e1002422.

Greninger, A. L., Chen, E. C., Sittler, T., Scheinerman, A., Roubinian, N., Yu, G., Kim, E., Pillai, D. R., Guyard, C., Mazzulli, T., Isa, P., Arias, C. F., Hackett, J., Schochetman, G., Miller, S., Tang, P., and Chiu, C. Y. (2010). A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. *PLoS ONE*, 5(10):e13381.

Gretch, D. R. (1997). Diagnostic tests for hepatitis C. In *Hepatology*, pages 43S–47S. Department of Laboratory Medicine, University of Washington Medical Center, Seattle 98195, USA.

Grimm, T., Schneider, S., Naschberger, E., Huber, J., Guenzi, E., Kieser, A., Reitmeir, P., Schulz, T. F., Morris, C. A., and Stürzl, M. (2005). EBV latent membrane protein-1 protects B cells from apoptosis by inhibition of BAX. *Blood*, 105(8):3263–3269.

Gross, L. (1951). "Spontaneous" leukemia developing in C3H mice following inoculation in infancy, with AK-leukemic extracts, or AK-embrvos. *Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine (New York, N.Y.)*, 76(1):27–32.

Gruener, N. H., Lechner, F., Jung, M. C., Diepolder, H., Gerlach, T., Lauer, G., Walker, B., Sullivan, J., Phillips, R., Pape, G. R., and Klenerman, P. (2001). Sustained dysfunction of antiviral CD8+ T lymphocytes after infection with hepatitis C virus. *J Virol*, 75(12):5550–5558.

Gruhne, B., Kamranvar, S. A., Masucci, M. G., and Sompallae, R. (2009). EBV and genomic instability–a new look at the role of the virus in the pathogenesis of Burkitt's lymphoma. *Semin. Cancer Biol.*, 19(6):394–400.

Grulich, A. E., van Leeuwen, M. T., Falster, M. O., and Vajdic, C. M. (2007). Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: a meta-analysis. *Lancet*, 370(9581):59–67.

Guidotti, L. G. and Chisari, F. V. (1996). To kill or to cure: options in host defense against viral infection. *Curr. Opin. Immunol.*, 8(4):478–483.

Guimaraes, K. S., Jothi, R., Zotenko, E., and Przytycka, T. (2006). Predicting domain-domain interactions using a parsimony approach. *Genome Biol*, 7(11):R104.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, 59(3):307–321.

Gulbahce, N., Yan, H., Dricot, A., Padi, M., Byrdsong, D., Franchi, R., Lee, D.-S., Rozenblatt-Rosen, O., Mar, J. C., Calderwood, M. A., Baldwin, A., Zhao, B., Santhanam, B., Braun, P., Simonis, N., Huh, K.-W., Hellner, K., Grace, M., Chen, A., Rubio, R., Marto, J. A., Christakis, N. A., Kieff, E., Roth, F. P., Roecklein-Canfield, J., DeCaprio, J. A., Cusick, M. E., Quackenbush, J., Hill, D. E., Munger, K., Vidal, M., and Barabási, A.-L. (2012). Viral perturbations of host networks reflect disease etiology. *PLoS Comput Biol*, 8(6):e1002531.

Guldener, U., Münsterkötter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H.-W., and Stümpflen, V. (2006). MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res*, 34(Database issue):D436–41.

Gundry, M. and Vijg, J. (2012). Direct mutation analysis by high-throughput sequencing: from germline to low-abundant, somatic variants. *Mutat. Res.*, 729(1-2):1–15.

Gupta, R. S. (2000). The natural evolutionary relationships among prokaryotes. *Crit. Rev. Microbiol.*, 26(2):111–131.

Guruharsha, K. G., Rual, J.-F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., Beekman, C., Wong, C., Rhee, D. Y., Cenaj, O., McKillip, E., Shah, S., Stapleton, M., Wan, K. H., Yu, C., Parsa, B., Carlson, J. W., Chen, X., Kapadia, B., VijayRaghavan, K., Gygi, S. P., Celniker, S. E., Obar, R. A., and Artavanis-Tsakonas, S. (2011). A protein complex network of Drosophila melanogaster. *Cell*, 147(3):690–703.

Haagmans, B. L. B., Andeweg, A. C. A., and Osterhaus, A. D. M. E. A. (2009). The application of genomics to emerging zoonotic viral diseases. *CORD Conference Proceedings*, 5(10):e1000557–e1000557.

Haaland, R. E., Hawkins, P. A., Salazar-Gonzalez, J., Johnson, A., Tichacek, A., Karita, E., Manigart, O., Mulenga, J., Keele, B. F., Shaw, G. M., Hahn, B. H., Allen, S. A., Derdeyn, C. A., and Hunter, E. (2009). Inflammatory genital infections mitigate a severe genetic bottleneck in heterosexual transmission of subtype A and C HIV-1. *PLoS Pathog*, 5(1):e1000274.

Haasnoot, J., de Vries, W., Geutjes, E.-J., Prins, M., de Haan, P., and Berkhout, B. (2007). The Ebola virus VP35 protein is a suppressor of RNA silencing. *PLoS Pathog*, 3(6):e86.

Habjan, M., Andersson, I., Klingström, J., Schümann, M., Martin, A., Zimmermann, P., Wagner, V., Pichlmair, A., Schneider, U., Mühlberger, E., Mirazimi, A., and Weber, F. (2008). Processing of genome 5′ termini as a strategy of negative-strand RNA viruses to avoid RIG-I-dependent interferon induction. *PLoS ONE*, 3(4):e2032.

Haigh, J. (1978). The accumulation of deleterious genes in a population–Muller's Ratchet. *Theor Popul Biol*, 14(2):251–267.

Haldane, J. B. S. (1929). The origin of life. *The Rationalist Annual*, 148:3–10.

Halfon, P. and Locarnini, S. (2011). Hepatitis C virus resistance to protease inhibitors. *J Hepatol*, 55(1):192–206.

Halfon, P. and Sarrazin, C. (2012). Future treatment of chronic hepatitis C with direct acting antivirals: is resistance important? *Liver Int*, 32 Suppl 1:79–87.

Hamady, M., Walker, J. J., Harris, J. K., Gold, N. J., and Knight, R. (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods*, 5(3):235–237.

Hamilton, A. J. (1999). A Species of Small Antisense RNA in Posttranscriptional Gene Silencing in Plants. *Science*, 286(5441):950–952.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674.

Handel, A., Regoes, R. R., and Antia, R. (2006). The role of compensatory mutations in the emergence of drug resistance. *PLoS Comput Biol*, 2(10):e137–e137.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chem Biol*, 5(10):R245–R249.

Hanekamp, K., Bohnebeck, U., Beszteri, B., and Valentin, K. (2007). PhyloGena - A user-friendly system for automated phylogenetic annotation of unknown sequences. *Bioinformatics*, 23(7):793–801.

Hansen, J. L., Long, A. M., and Schultz, S. C. (1997). Structure of the RNA-dependent RNA polymerase of poliovirus. *Structure*, 5(8):1109–1122.

Hansen, T. H. and Bouvier, M. (2009). MHC class I antigen presentation: learning from viral evasion strategies. *Nat Rev Immunol*, 9(7):503–513.

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., Schork, N. J., Murray, S. S., Topol, E. J., Levy, S., and Frazer, K. A. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol*, 10(3):R32.

Harismendy, O., Schwab, R. B., Bao, L., Olson, J., Rozenzhak, S., Kotsopoulos, S. K., Pond, S., Crain, B., Chee, M. S., Messer, K., Link, D. R., and Frazer, K. A. (2011). Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. *Genome Biol*, 12(12):R124.

Harki, D. A., Graci, J. D., Galarraga, J. E., Chain, W. J., Cameron, C. E., and Peterson, B. R. (2006). Synthesis and antiviral activity of 5-substituted cytidine analogues: identification of a potent inhibitor of viral RNA-dependent RNA polymerases. *J. Med. Chem.*, 49(21):6166–6169.

Harrigan, P. R. and Larder, B. A. (2002). Extent of cross-resistance between agents used to treat human immunodeficiency virus type 1 infection in clinically derived isolates. *Antimicrob Agents Chemother*, 46(3):909–912.

Harris, C. C. (1996). Structure and function of the p53 tumor suppressor gene: clues for rational cancer therapeutic strategies. *J. Natl. Cancer Inst.*, 88(20):1442–1455.

Harris, J. R., Neil, K. P., Behravesh, C. B., Sotir, M. J., and Angulo, F. J. (2010). Recent Multistate Outbreaks of Human Salmonella Infections Acquired from Turtles: A Continuing Public Health Challenge. *Clin Infect Dis*, 50(4):554–559.

Harris, R. S., Kong, Q., and Maizels, N. (1999). Somatic hypermutation and the three R's: repair, replication and recombination. *Mutat. Res.*, 436(2):157–178.

Hart, G. T., Lee, I., and Marcotte, E. R. (2007). A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics*, 8:236.

Hart, G. T., Ramani, A. K., and Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? *Genome Biol*, 7(11):120.

Hatem, A., Bozda, D., Toland, A. E., and Çatalyürek, Ü. V. (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics*, 14(1):184.

Havugimana, P. C., Hart, G. T., Nepusz, T., Yang, H., Turinsky, A. L., Li, Z., Wang, P. I., Boutz, D. R., Fong, V., Phanse, S., Babu, M., Craig, S. A., Hu, P., Wan, C., Vlasblom, J., Dar, V.-u.-N., Bezginov, A., Clark, G. W., Wu, G. C., Wodak, S. J., Tillier, E. R. M., Paccanaro, A., Marcotte, E. M., and Emili, A. (2012). A census of human soluble protein complexes. *Cell*, 150(5):1068–1081.

Hayward, W. S., Neel, B. G., and Astrin, S. M. (1981). Activation of a cellular onc gene by promoter insertion in ALV-induced lymphoid leukosis. *Nature*, 290(5806):475–480.

He, Y., King, M. S., Kempf, D. J., Lu, L., Lim, H. B., Krishnan, P., Kati, W., Middleton, T., and Molla, A. (2008). Relative replication capacity and selective advantage profiles of protease inhibitor-resistant hepatitis C virus (HCV) NS3 protease mutants in the HCV genotype 1b replicon system. *Antimicrob Agents Chemother*, 52(3):1101–1110.

Heath, S. L., Tew, J. G., Tew, J. G., Szakal, A. K., and Burton, G. F. (1995). Follicular dendritic cells and human immunodeficiency virus infectivity. *Nature*, 377(6551):740–744.

Hebner, C., Martin, R., Mo, H. M., and Svarovskaia, E. S. (2011). The effects of combining two direct acting antivirals, ribavirin, and pegylated interferon on the detection of drug resistance mutations early in treatment of HCV. *Hepatology (Baltimore, Md)*, (54):997a.

Heck, J. E., Ritz, B., Hung, R. J., Hashibe, M., and Boffetta, P. (2009). The epidemiology of neuroblastoma: a review. *Paediatr Perinat Epidemiol*, 23(2):125–143.

Hedskog, C., Mild, M., Jernberg, J., Sherwood, E., Bratt, G., Leitner, T., Lundeberg, J., Andersson, B., and Albert, J. (2010). Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. *PLoS ONE*, 5(7):e11345.

Hegde, N. R., Maddur, M. S., Kaveri, S. V., and Bayry, J. (2009). Reasons to include viruses in the tree of life. *Nat Rev Microbiol*, 7(8):615–author reply 615.

Heidmann, O. O., Vernochet, C. C., Dupressoir, A. A., and Heidmann, T. T. (2008). Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: a new syncytin in a third order of mammals. *CORD Conference Proceedings*, 6:107–107.

Hein, J., Boichuk, S., Wu, J., Cheng, Y., Freire, R., Jat, P. S., Roberts, T. M., and Gjoerup, O. V. (2009). Simian virus 40 large T antigen disrupts genome integrity and activates a DNA damage response via Bub1 binding. *J Virol*, 83(1):117–127.

Helle, F., Goffard, A., Morel, V., Duverlie, G., McKeating, J., Keck, Z.-Y., Foung, S., Penin, F., Dubuisson, J., and Voisset, C. (2007). The neutralizing activity of anti-hepatitis C virus antibodies is modulated by specific glycans on the E2 envelope protein. *J Virol*, 81(15):8101–8111.

Henn, M. R., Boutwell, C. L., Charlebois, P., Lennon, N. J., Power, K. A., Macalalad, A. R., Berlin, A. M., Malboeuf, C. M., Ryan, E. M., Gnerre, S., Zody, M. C., Erlich, R. L., Green, L. M., Berical, A., Wang, Y., Casali, M., Streeck, H., Bloom, A. K., Dudek, T., Tully, D., Newman, R., Axten, K. L., Gladden, A. D., Battis, L., Kemper, M., Zeng, Q., Shea, T. P., Gujja, S., Zedlack, C., Gasser, O., Brander, C., Hess, C., Günthard, H. F., Brumme, Z. L., Brumme, C. J., Bazner, S., Rychert, J., Tinsley, J. P., Mayer, K. H., Rosenberg, E., Pereyra, F., Levin, J. Z., Young, S. K., Jessen, H., Altfeld, M., Birren, B. W., Walker, B. D., and Allen, T. M. (2012). Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog*, 8(3):e1002529.

Herfst, S., Schrauwen, E. J. A., Linster, M., Chutinimitkul, S., de Wit, E., Munster, V. J., Sorrell, E. M., Bestebroer, T. M., Burke, D. F., Smith, D. J., Rimmelzwaan, G. F., Osterhaus, A. D. M. E., and Fouchier, R. A. M. (2012). Airborne transmission of influenza A/H5N1 virus between ferrets. *Science*, 336(6088):1534–1541.

Herman, J. G. and Baylin, S. B. (2003). Gene silencing in cancer in association with promoter hypermethylation. *N. Engl. J. Med.*, 349(21):2042–2054.

Hernández, H., Dziembowski, A., Taverner, T., Séraphin, B., and Robinson, C. V. (2006). Subunit architecture of multimeric complexes isolated directly from cells. *EMBO Rep*, 7(6):605–610.

Herrmann, E., Lee, J.-H., Marinos, G., Modi, M., and Zeuzem, S. (2003). Effect of ribavirin on hepatitis C viral kinetics in patients treated with pegylated interferon. *Hepatology (Baltimore, Md)*, 37(6):1351–1358.

Hézode, C., Forestier, N., Dusheiko, G., Ferenci, P., Pol, S., Goeser, T., Bronowicki, J.-P., Bourlière, M., Gharakhanian, S., Bengtsson, L., McNair, L., George, S., Kieffer, T., Kwong, A., Kauffman, R. S., Alam, J., Pawlotsky, J.-M., Zeuzem, S., and PROVE2 Study Team (2009). Telaprevir and peginterferon with or without ribavirin for chronic HCV infection. *N. Engl. J. Med.*, 360(18):1839–1850.

Hiatt, J. B., Patwardhan, R. P., Turner, E. H., Lee, C., and Shendure, J. (2010). Parallel, tag-directed assembly of locally derived short sequence reads. *Nat Methods*, 7(2):119–122.

Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J., and Shendure, J. (2013). Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res*, 23(5):843–854.

Hill, A. B. (1965). The environment and disease: association or causation? In *Proc. R. Soc. Med.*, pages 295–300.

Hinkley, T., Martins, J., Chappey, C., Haddad, M., Stawiski, E., Whitcomb, J. M., Petropoulos, C. J., and Bonhoeffer, S. (2011). A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nat Genet*, 43(5):487–489.

Hiraga, N., Imamura, M., Abe, H., Hayes, C. N., Kono, T., Onishi, M., Tsuge, M., Takahashi, S., Ochi, H., Iwao, E., Kamiya, N., Yamada, I., Tateno, C., Yoshizato, K., Matsui, H., Kanai, A., Inaba, T., Tanaka, S., and Chayama, K. (2011). Rapid emergence of telaprevir resistant hepatitis C virus strain from wildtype clone in vivo. *Hepatology (Baltimore, Md)*, 54(3):781–788.

Hirsch, A. J. A. and Shenk, T. T. (1998). Human cytomegalovirus inhibits transcription of the CC chemokine MCP-1 gene. *J Virol*, 73(1):404–410.

Hiscott, J., Nguyen, T.-L. A., Arguello, M., Nakhaei, P., and Paz, S. (2006). Manipulation of the nuclear factor-kappaB pathway and the innate immune response by viruses. *Oncogene*, 25(51):6844–6867.

Hmwe, S. S., Aizaki, H., Date, T., Murakami, K., Ishii, K., Miyamura, T., Koike, K., Wakita, T., and Suzuki, T. (2010). Identification of hepatitis C virus genotype 2a replicon variants with reduced susceptibility to ribavirin. *Antiviral Res*, 85(3):520–524.

Ho, D. D. (1995). Time to hit HIV, early and hard. *New England Journal of Medicine*, 333(7):450–451.

Hoff, K. J. (2009). The effect of sequencing errors on metagenomic gene prediction. *BMC Genomics*, 10(1):520.

Hoffmann, B. B., Scheuch, M. M., Höper, D. D., Jungblut, R. R., Holsteg, M. M., Schirrmeier, H. H., Eschbaumer, M. M., Goller, K. V. K., Wernike, K. K., Fischer, M. M., Breithaupt, A. A., Mettenleiter, T. C. T., and Beer, M. M. (2012). Novel orthobunyavirus in Cattle, Europe, 2011. *Emerging Infect. Dis.*, 18(3):469–472.

Hofmann, W. P., Herrmann, E., Sarrazin, C., and Zeuzem, S. (2008). Ribavirin mode of action in chronic hepatitis C: from clinical use back to molecular mechanisms. *Liver Int*, 28(10):1332–1343.

Hofmann, W. P., Polta, A., Herrmann, E., Mihm, U., Kronenberger, B., Sonntag, T., Lohmann, V., Schönberger, B., Zeuzem, S., and Sarrazin, C. (2007). Mutagenic effect of ribavirin on hepatitis C nonstructural 5B quasispecies in vitro and during antiviral therapy. *Gastroenterology*, 132(3):921–930.

Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S., and VandePol, S. (1982). Rapid evolution of RNA genomes. *Science*, 215(4540):1577–1585.

Holland, J. J. (2006). Transitions in understanding of RNA viruses: a historical perspective. *Curr. Top. Microbiol. Immunol.*, 299:371–401.

Holland, J. J., Domingo, E., de la Torre, J. C., and Steinhauer, D. A. (1990). Mutation frequencies at defined single codon sites in vesicular stomatitis virus and poliovirus can be increased only slightly by chemical mutagenesis. *J Virol*, 64(8):3960–3962.

Holmes, E. (2003). Error thresholds and the constraints to RNA virus evolution. *Trends in microbiology*, 11(12):543–546.

Holmes, E. C. (2011). The evolution of endogenous viral elements. *Cell Host Microbe*, 10(4):368–377.

Holmes, E. C. and Moya, A. (2002). Is the quasispecies concept relevant to RNA viruses? *J Virol*, 76(1):460–465.

Holmes, R. K., Malim, M. H., and Bishop, K. N. (2007). APOBEC-mediated viral restriction: not simply editing? *Trends Biochem Sci*, 32(3):118–128.

Holt, N., Wang, J., Kim, K., Friedman, G., Wang, X., Taupin, V., Crooks, G. M., Kohn, D. B., Gregory, P. D., Holmes, M. C., and Cannon, P. M. (2010). Human hematopoietic stem/progenitor cells modified by zinc-finger nucleases targeted to CCR5 control HIV-1 in vivo. *Nat Biotechnol*, 28(8):839–847.

Honda, M., Sakai, A., Yamashita, T., Nakamoto, Y., Mizukoshi, E., Sakai, Y., Yamashita, T., Nakamura, M., Shirasaki, T., Horimoto, K., Tanaka, Y., Tokunaga, K., Mizokami, M., Kaneko, S., and Hokuriku Liver Study Group (2010). Hepatic ISG expression is associated with genetic variation in interleukin 28B and the outcome of IFN therapy for chronic hepatitis C. *Gastroenterology*, 139(2):499–509.

Hong, E. L., Balakrishnan, R., Dong, Q., Christie, K. R., Park, J., Binkley, G., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Krieger, C. J., Livstone, M. S., Miyasato, S. R., Nash, R. S., Oughtred, R., Skrzypek, M. S., Weng, S., Wong, E. D., Zhu, K. K., Dolinski, K., Botstein, D., and Cherry, J. M. (2007). Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res*, 36(Database):D577–D581.

Hoofnagle, J. H., Lau, D., Conjeevaram, H., Kleiner, D., and Di Bisceglie, A. M. (1996). Prolonged therapy of chronic hepatitis C with ribavirin. *J Viral Hepat*, 3(5):247–252.

Hopkins, A. L. and Bickerton, G. R. (2010). Drug discovery: Know your chemical space. *Nat. Chem. Biol.*, 6(7):482–483.

Hopkins, S., Scorneaux, B., Huang, Z., Murray, M. G., Wring, S., Smitley, C., Harris, R., Erdmann, F., Fischer, G., and Ribeill, Y. (2010). SCY-635, a novel nonimmunosuppressive analog of cyclosporine that exhibits potent inhibition of hepatitis C virus RNA replication in vitro. *Antimicrob Agents Chemother*, 54(2):660–672.

Horie, M., Honda, T., Suzuki, Y., Kobayashi, Y., Daito, T., Oshida, T., Ikuta, K., Jern, P., Gojobori, T., Coffin, J. M., and Tomonaga, K. (2010). Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature*, 463(7277):84–87.

Hornung, V., Ellegast, J., Kim, S., Brzózka, K., Jung, A., Kato, H., Poeck, H., Akira, S., Conzelmann, K.-K., Schlee, M., Endres, S., and Hartmann, G. (2006). 5'-Triphosphate RNA is the ligand for RIG-I. *Science*, 314(5801):994–997.

Horst, D., Verweij, M. C., Davison, A. J., Ressing, M. E., and Wiertz, E. J. H. J. (2011). Viral evasion of T cell immunity: ancient mechanisms offering new applications. *Curr. Opin. Immunol.*, 23(1):96–103.

Hosseini, P., Sokolow, S. H., Vandegrift, K. J., Kilpatrick, A. M., and Daszak, P. (2010). Predictive Power of Air Travel and Socio-Economic Data for Early Pandemic Spread. *PLoS ONE*, 5(9):e12763.

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., Wu, H., Ye, X., Ye, C., Wu, R., Jian, M., Chen, Y., Xie, W., Zhang, R., Chen, L., Liu, X., Yao, X., Zheng, H., Yu, C., Li, Q., Gong, Z., Mao, M., Yang, X., Yang, L., Li, J., Wang, W., Lu, Z., Gu, N., Laurie, G., Bolund, L., Kristiansen, K., Wang, J., Yang, H., Li, Y., Zhang, X., and Wang, J. (2012). Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell*, 148(5):873–885.

Houben, R., Shuda, M., Weinkam, R., Schrama, D., Feng, H., Chang, Y., Moore, P. S., and Becker, J. C. (2010). Merkel cell polyomavirus-infected Merkel cell carcinoma cells require expression of viral T antigens. *J Virol*, 84(14):7064–7072.

Hsu, T.-H. and Spindler, K. R. (2012). Identifying host factors that regulate viral infection. *PLoS Pathog*, 8(7):e1002772.

Hu, J. and Robinson, J. L. (2010). Treatment of respiratory syncytial virus with palivizumab: a systematic review. *World J Pediatr*, 6(4):296–300.

Huebner, R. J. and Todaro, G. J. (1969). Oncogenes of RNA tumor viruses as determinants of cancer. *Proc Natl Acad Sci USA*, 64(3):1087–1094.

Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biol*, 3(2):0003.1–0003.8.

Huggins, J. W. (1989). Prospects for treatment of viral hemorrhagic fevers with ribavirin, a broad-spectrum antiviral drug. *Rev Infect Dis*, 11 Suppl 4:S750–61.

Hughes, J. F. and Coffin, J. M. (2001). Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. *Nat Genet*, 29(4):487–489.

Hughes, J. P. J., Rees, S. S., Kalindjian, S. B. S., and Philpott, K. L. K. (2011). Principles of early drug discovery. *Br J Pharmacol*, 162(6):1239–1249.

Hung, C. S. C., Vander Heyden, N. N., and Ratner, L. L. (1999). Analysis of the critical domain in the V3 loop of human immunodeficiency virus type 1 gp120 involved in CCR5 utilization. *J Virol*, 73(10):8216–8226.

Hurwitz, B. L., Deng, L., Poulos, B. T., and Sullivan, M. B. (2013). Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ. Microbiol.*, 15(5):1428–1440.

Hurwitz, B. L. B. and Sullivan, M. B. M. (2012). The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology. *PLoS ONE*, 8(2):e57355–e57355.

Huse, S. M., Dethlefsen, L., Huber, J. A., Welch, D. M., Relman, D. A., and Sogin, M. L. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet*, 4(11).

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res*, 17(3):377–386.

Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N., and Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Res*, 21(9):1552–1560.

Huson, D. H., Richter, D. C., Mitra, S., Auch, A. F., and Schuster, S. C. (2009). Methods for comparative metagenomics. *BMC Bioinformatics*, 10 Suppl 1:S12.

Huthoff, H. and Towers, G. J. (2008). Restriction of retroviral replication by APOBEC3G/F and TRIM5alpha. *Trends in microbiology*, 16(12):612–619.

Hutson, C. L. (2007). Monkeypox zoonotic associations: insights from laboratory evaluation of animals associated with the multi-state US outbreak. *Am. J. Trop. Med. Hyg.*, 76:757–768 PB –.

Huynen, M. A. M. (1996). Exploring phenotype space through neutral evolution. *J. Mol. Evol.*, 43(3):165–169.

Huynen, M. A. M., Stadler, P. F. P., and Fontana, W. W. (1996). Smoothness within ruggedness: the role of neutrality in adaptation. *Proc Natl Acad Sci U S A*, 93(1):397–401.

Hyung, S.-J. and Ruotolo, B. T. (2012). Integrating mass spectrometry of intact protein complexes into structural proteomics. *Proteomics*, 12(10):1547–1564.

Ideker, T. and Krogan, N. J. (2012). Differential network biology. *Mol Syst Biol*, 8:565.

Ideker, T. and Sharan, R. (2008). Protein networks in disease. *Genome Res*, 18(4):644–652.

Imai, M. M., Watanabe, T. T., Hatta, M. M., Das, S. C. S., Ozawa, M. M., Shinya, K. K., Zhong, G. G., Hanson, A. A., Katsura, H. H., Watanabe, S. S., Li, C. C., Kawakami, E. E., Yamada, S. S., Kiso, M. M., Suzuki, Y. Y., Maher, E. A. E., Neumann, G. G., and Kawaoka, Y. Y. (2012). Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature*, 486(7403):420–428.

Imperiale, M. J. (2001). Oncogenic transformation by the human polyomaviruses. *Oncogene*, 20(54):7917–7923.

Inda, M.-d.-M., Bonavia, R., Mukasa, A., Narita, Y., Sah, D. W. Y., Vandenberg, S., Brennan, C., Johns, T. G., Bachoo, R., Hadwiger, P., Tan, P., DePinho, R. A., Cavenee, W., and Furnari, F. (2010). Tumor heterogeneity is an active process maintained by a mutant EGFR-induced cytokine circuit in glioblastoma. *Genes Dev*, 24(16):1731–1745.

International Cancer Genome Consortium, Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., Guttmacher, A., Guyer, M., Hemsley, F. M., Jennings, J. L., Kerr, D., Klatt, P., Kolar, P., Kusada, J., Lane, D. P., Laplace, F., Youyong, L., Nettekoven, G., Ozenberger, B., Peterson, J., Rao, T. S., Remacle, J., Schafer, A. J., Shibata, T., Stratton, M. R., Vockley, J. G., Watanabe, K., Yang, H., Yuen, M. M. F., Knoppers, B. M., Bobrow, M., Cambon-Thomsen, A., Dressler, L. G., Dyke, S. O. M., Joly, Y., Kato, K., Kennedy, K. L., Nicolás, P., Parker, M. J., Rial-Sebbag, E., Romeo-Casabona, C. M., Shaw, K. M., Wallace, S., Wiesner, G. L., Zeps, N., Lichter, P., Biankin, A. V., Chabannon, C., Chin, L., Clément, B., de Alava, E., Degos, F., Ferguson, M. L., Geary, P., Hayes, D. N., Hudson, T. J., Johns, A. L., Kasprzyk, A., Nakagawa, H., Penny, R., Piris, M. A., Sarin, R., Scarpa, A., Shibata, T., van de Vijver, M., Futreal, P. A., Aburatani, H., Bayes, M., Botwell, D. D. L., Campbell, P. J., Estivill, X., Gerhard, D. S., Grimmond, S. M., Gut, I., Hirst, M., López-Otín, C., Majumder, P., Marra, M., McPherson, J. D., Nakagawa, H., Ning, Z., Puente, X. S., Ruan, Y., Shibata, T., Stratton, M. R., Stunnenberg, H. G., Swerdlow, H., Velculescu, V. E., Wilson, R. K., Xue, H. H., Yang, L., Spellman, P. T., Bader, G. D., Boutros, P. C., Campbell, P. J., Flicek, P., Getz, G., Guigó, R., Guo, G., Haussler, D., Heath, S., Hubbard, T. J., Jiang, T., Jones, S. M., Li, Q., López-Bigas, N., Luo, R., Muthuswamy, L., Ouellette, B. F. F., Pearson, J. V., Puente, X. S., Quesada, V., Raphael, B. J., Sander, C., Shibata, T., Speed, T. P., Stein, L. D., Stuart, J. M., Teague, J. W., Totoki, Y., Tsunoda, T., Valencia, A., Wheeler, D. A., Wu, H., Zhao, S., Zhou, G., Stein, L. D., Guigó, R., Hubbard, T. J., Joly, Y., Jones, S. M., Kasprzyk, A., Lathrop, M., López-Bigas, N., Ouellette, B. F. F., Spellman, P. T., Teague, J. W., Thomas, G., Valencia, A., Yoshida, T., Kennedy, K. L., Axton, M., Dyke, S. O. M., Futreal, P. A., Gerhard, D. S., Gunter, C., Guyer, M., Hudson, T. J., McPherson, J. D., Miller, L. J., Ozenberger, B., Shaw, K. M., Kasprzyk, A., Stein, L. D., Zhang, J., Haider, S. A., Wang, J., Yung, C. K., Cros, A., Cross, A., Liang, Y., Gnaneshan, S., Guberman, J., Hsu, J., Bobrow, M., Chalmers, D. R. C., Hasel, K. W., Joly, Y., Kaan, T. S. H., Kennedy, K. L., Knoppers, B. M., Lowrance, W. W., Masui, T., Nicolás, P., Rial-Sebbag, E., Rodriguez, L. L., Vergely, C., Yoshida, T., Grimmond, S. M., Biankin, A. V., Bowtell, D. D. L., Cloonan, N., deFazio, A., Eshleman, J. R., Etemadmoghadam, D., Gardiner, B. B., Gardiner, B. A., Kench, J. G., Scarpa, A., Sutherland, R. L., Tempero, M. A., Waddell, N. J., Wilson, P. J., McPherson, J. D., Gallinger, S., Tsao, M.-S., Shaw, P. A., Petersen, G. M., Mukhopadhyay, D., Chin, L., DePinho, R. A., Thayer, S., Muthuswamy, L., Shazand, K., Beck, T., Sam, M., Timms, L., Ballin, V., Lu, Y., Ji, J., Zhang, X., Chen, F., Hu, X., Zhou, G., Yang, Q., Tian, G., Zhang, L., Xing, X., Li, X., Zhu, Z., Yu, Y., Yu, J., Yang, H., Lathrop, M., Tost, J., Brennan, P., Holcatova, I., Zaridze, D., Brazma, A., and Egev... (2010). International network of cancer genome projects. *Nature*, 464(7291):993–998.

International HIV Controllers Study, Pereyra, F., Jia, X., McLaren, P. J., Telenti, A., de Bakker, P. I. W., Walker, B. D., Ripke, S., Brumme, C. J., Pulit, S. L., Carrington, M., Kadie, C. M., Carlson, J. M., Heckerman, D., Graham, R. R., Plenge, R. M., Deeks, S. G., Gianniny, L., Crawford, G., Sullivan, J., Gonzalez, E., Davies, L., Camargo, A., Moore, J. M., Beattie, N., Gupta, S., Crenshaw, A., Burtt, N. P., Guiducci, C., Gupta, N., Gao, X., Qi, Y., Yuki, Y., Piechocka-Trocha, A., Cutrell, E., Rosenberg, R., Moss, K. L., Lemay, P., O'Leary, J., Schaefer, T., Verma, P., Toth, I., Block, B., Baker, B., Rothchild, A., Lian, J., Proudfoot, J., Alvino, D. M. L., Vine, S., Addo, M. M., Allen, T. M., Altfeld, M., Henn, M. R., Le Gall, S., Streeck, H., Haas, D. W., Kuritzkes, D. R., Robbins, G. K., Shafer, R. W., Gulick, R. M., Shikuma, C. M., Haubrich, R., Riddler, S., Sax, P. E., Daar, E. S., Ribaudo, H. J., Agan, B., Agarwal, S., Ahern, R. L., Allen, B. L., Altidor, S., Altschuler, E. L., Ambardar, S., Anastos, K., Anderson, B., Anderson, V., Andrady, U., Antoniskis, D., Bangsberg, D., Barbaro, D., Barrie, W., Bartczak, J., Barton, S., Basden, P., Basgoz, N., Bazner, S., Bellos, N. C., Benson, A. M., Berger, J., Bernard, N. F., Bernard, A. M., Birch, C., Bodner, S. J., Bolan, R. K., Boudreaux, E. T., Bradley, M., Braun, J. F., Brndjar, J. E., Brown, S. J., Brown, K., Brown, S. T., Burack, J., Bush, L. M., Cafaro, V., Campbell, O., Campbell, J., Carlson, R. H., Carmichael, J. K., Casey, K. K., Cavacuiti, C., Celestin, G., Chambers, S. T., Chez, N., Chirch, L. M., Cimoch, P. J., Cohen, D., Cohn, L. E., Conway, B., Cooper, D. A., Cornelson, B., Cox, D. T., Cristofano, M. V., Cuchural, G., Czartoski, J. L., Dahman, J. M., Daly, J. S., Davis, B. T., Davis, K., Davod, S. M., DeJesus, E., Dietz, C. A., Dunham, E., Dunn, M. E., Ellerin, T. B., Eron, J. J., Fangman, J. J. W., Farel, C. E., Ferlazzo, H., Fidler, S., Fleenor-Ford, A., Frankel, R., Freedberg, K. A., French, N. K., Fuchs, J. D., Fuller, J. D., Gaberman, J., Gallant, J. E., Gandhi, R. T., Garcia, E., Garmon, D., Gathe, J. C., Gaultier, C. R., Gebre, W., Gilman, F. D., Gilson, I., Goepfert, P. A., Gottlieb, M. S., Goulston, C., Groger, R. K., Gurley, T. D., Haber, S., Hardwicke, R., Hardy, W. D., Harrigan, P. R., Hawkins, T. N., Heath, S., Hecht, F. M., Henry, W. K., Hladek, M., Hoffman, R. P., Horton, J. M., Hsu, R. K., Huhn, G. D., Hunt, P., Hupert, M. J., Illeman, M. L., Jaeger, H., Jellinger, R. M., John, M., Johnson, J. A., Johnson, K. L., Johnson, H., Johnson, K., Joly, J., Jordan, W. C., Kauffman, C. A., Khanlou, H., Killian, R. K., Kim, A. Y., Kim, D. D., Kinder, C. A., Kirchner, J. T., Kogelman, L., Kojic, E. M., Korthuis, P. T., Kurisu, W., Kwon, D. S., LaMar, M., Lampiris, H., Lanzafame, M., Lederman, M. M., Lee, D. M., Lee, J. M. L., Lee, M. J., Lee, E. T. Y., Lemoine, J., Levy, J. A., Llibre, J. M., Liguori, M. A., Little, S. J., Liu, A. Y., Lopez, A. J., Loutfy, M. R., Loy, D., Mohammed, D. Y., Man, A., Mansour, M. K., Marconi, V. C., Markowitz, M., Marques, R., Martin, J. N., Martin, H. L., Mayer, K. H., McElrath, M. J., McGhee, T. A., McGovern, B. H., McGowan, K., McIntyre, D., Mcleod, G. X., Menezes, P., Mesa, G., Metroka, C. E., Meyer-Olson, D., and Miller, A. (2010). The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science*, 330(6010):1551–1557.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945.

Ioannidis, J. P. J., Havlir, D. V. D., Tebas, P. P., Hirsch, M. S. M., Collier, A. C. A., and Richman, D. D. D. (2000). Dynamics of HIV-1 viral load rebound among patients with previous suppression of viral replication. *AIDS*, 14(11):1481–1488.

Iorio, R. M., Glickman, R. L., Riel, A. M., Sheehan, J. P., and Bratt, M. A. (1989). Functional and neutralization profile of seven overlapping antigenic sites on the HN glycoprotein of Newcastle disease virus: monoclonal antibodies to some sites prevent viral attachment. *Virus Res*, 13(3):245–261.

Iranzo, J. and Manrubia, S. C. (2008). Stochastic extinction of viral infectivity through the action of defectors. *Europhys. Lett.*, 85(1):18001.

Iranzo, J., Perales, C., Domingo, E., and Manrubia, S. C. (2011a). Tempo and mode of inhibitor-mutagen antiviral therapies: a multidisciplinary approach. *Proc Natl Acad Sci USA*, 108(38):16008–16013.

Iranzo, J. J., Perales, C. C., Domingo, E. E., and Manrubia, S. C. S. (2011b). Tempo and mode of inhibitor-mutagen antiviral therapies: a multidisciplinary approach. *Proc Natl Acad Sci U S A*, 108(38):16008–16013.

Isaacson, M. K. M. and Ploegh, H. L. H. (2009). Ubiquitination, Ubiquitin-like Modifiers, and Deubiquitination in Viral Infection. *Cell Host Microbe*, 5(6):12–12.

Isakov, O., Modai, S., and Shomron, N. (2011). Pathogen detection using short-RNA deep sequencing subtraction and assembly. *Bioinformatics*, 27(15):2027–2030.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA*, 98(8):4569–4574.

Ivacik, D., Ely, A., and Arbuthnot, P. (2011). Countering hepatitis B virus infection using RNAi: how far are we from the clinic? *Rev. Med. Virol.*, 21(6):383–396.

Jäger, S., Cimermancic, P., Gulbahce, N., Johnson, J. R., McGovern, K. E., Clarke, S. C., Shales, M., Mercenne, G., Pache, L., Li, K., Hernandez, H., Jang, G. M., Roth, S. L., Akiva, E., Marlett, J., Stephens, M., D'Orso, I., Fernandes, J., Fahey, M., Mahon, C., O'Donoghue, A. J., Todorovic, A., Morris, J. H., Maltby, D. A., Alber, T., Cagney, G., Bushman, F. D., Young, J. A., Chanda, S. K., Sundquist, W. I., Kortemme, T., Hernandez, R. D., Craik, C. S., Burlingame, A., Sali, A., Frankel, A. D., and Krogan, N. J. (2012). Global landscape of HIV-human protein complexes. *Nature*, 481(7381):365–370.

Janoueix-Lerosey, I., Schleiermacher, G., and Delattre, O. (2010). Molecular pathogenesis of peripheral neuroblastic tumors. *Oncogene*, 29(11):1566–1579.

Janoueix-Lerosey, I., Schleiermacher, G., Michels, E., Mosseri, V., Ribeiro, A., Lequin, D., Vermeulen, J., Couturier, J., Peuchmaur, M., Valent, A., Plantaz, D., Rubie, H., Valteau-Couanet, D., Thomas, C., Combaret, V., Rousseau, R., Eggert, A., Michon, J., Speleman, F., and Delattre, O. (2009). Overall genomic pattern is a predictor of outcome in neuroblastoma. *J. Clin. Oncol.*, 27(7):1026–1033.

Javier, R. T. and Butel, J. S. (2008). The History of Tumor Virology. *Cancer Research*, 68(19):7693–7706.

Jeang, K.-T., Giam, C.-z., Majone, F., and Aboud, M. (2004). Life, death, and tax: role of HTLV-I oncoprotein in genetic instability and cellular transformation. *J Biol Chem*, 279(31):31991–31994.

Jenkins, G. M., Worobey, M., Woelk, C. H., and Holmes, E. C. (2001). Evidence for the non-quasispecies evolution of RNA viruses [corrected]. *Mol Biol Evol*, 18(6):987–994.

Jensen, M. A. M. and van 't Wout, A. B. A. (2002). Predicting HIV-1 coreceptor usage with sequence analysis. *AIDS Rev*, 5(2):104–112.

Jern, P. and Coffin, J. M. (2008). Effects of Retroviruses on Host Genome Function. *Annu. Rev. Genet.*, 42(1):709–732.

Jha, A. R., Nixon, D. F., Rosenberg, M. G., Martin, J. N., Deeks, S. G., Hudson, R. R., Garrison, K. E., and Pillai, S. K. (2011). Human endogenous retrovirus K106 (HERV-K106) was infectious after the emergence of anatomically modern humans. *PLoS ONE*, 6(5):e20234.

Jhoti, H. and Leach, A. R. (2007). *Structure-based Drug Discovery*. Springer.

Ji, J. and Loeb, L. A. (1994). Fidelity of HIV-1 reverse transcriptase copying a hypervariable region of the HIV-1 env gene. *Virology*, 199(2):323–330.

Jiang, X., Mu, B., Huang, Z., Zhang, M., Wang, X., and Tao, S. (2010). Impacts of mutation effects and population size on mutation rate in asexual populations: a simulation study. *BMC Evol Biol*, 10:298.

Jin, M. S. and Lee, J.-O. (2008). Structures of TLR-ligand complexes. *Curr. Opin. Immunol.*, 20(4):414–419.

Jinushi, M., Takehara, T., Tatsumi, T., Kanto, T., Miyagi, T., Suzuki, T., Kanazawa, Y., Hiramatsu, N., and Hayashi, N. (2004). Negative regulation of NK cell activities by inhibitory receptor CD94/NKG2A leads to altered NK cell-induced modulation of dendritic cell functions in chronic hepatitis C virus infection. *J. Immunol.*, 173(10):6072–6081.

Jockusch, H., Wiegand, C., Mersch, B., and Rajes, D. (2001). Mutants of tobacco mosaic virus with temperature-sensitive coat proteins induce heat shock response in tobacco leaves. *Mol Plant Microbe Interact*, 14(7):914–917.

Johnson, M. E. and Hummer, G. (2011). Nonspecific binding limits the number of proteins in a cell and shapes their interaction networks. *Proc Natl Acad Sci USA*, 108(2):603–608.

Johnson, T. (1999). The approach to mutation-selection balance in an infinite asexual population, and the evolution of mutation rates. *Proc. Biol. Sci.*, 266(1436):2389–2397.

Johnson, T. T. and Barton, N. H. N. (2002). The effect of deleterious alleles on adaptation in asexual populations. *Genetics*, 162(1):395–411.

Johnson, W. E. and Desrosiers, R. C. (2013). Viral persistence: HIV's strategies of immune system evasion. *Annu. Rev. Med.*, 53:499–518.

Johnston, M. I. and Fauci, A. S. (2008). An HIV vaccine–challenges and prospects. *N. Engl. J. Med.*, 359(9):888–890.

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., and Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature*, 451(7181):990–993.

Jones, P. A. and Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. *Nat Rev Genet*, 3(6):415–428.

Jones, P. L., Korte, T., and Blumenthal, R. (1998). Conformational changes in cell surface HIV-1 envelope glycoproteins are triggered by cooperation between cell surface CD4 and co-receptors. *J Biol Chem*, 273(1):404–409.

Jonjić, S., Mutter, W., Weiland, F., Reddehase, M. J., and Koszinowski, U. H. (1989). Site-restricted persistent cytomegalovirus infection after selective long-term depletion of CD4+ T lymphocytes. *J Exp Med*, 169(4):1199–1212.

Jonsson, C. B. C., Milligan, B. G. B., and Arterburn, J. B. J. (2005). Potential importance of error catastrophe to the development of antiviral strategies for hantaviruses. *Virus Res*, 107(2):11–11.

Jørgensen, G. E., Johnsen, J. I., Ponthan, F., Kogner, P., Flaegstad, T., and Traavik, T. (2000). Human polyomavirus BK (BKV) and neuroblastoma: mechanisms of oncogenic action and possible strategy for novel treatment. *Med. Pediatr. Oncol.*, 35(6):593–596.

Ju, H.-Q., Xiang, Y.-F., Xin, B.-J., Pei, Y., Lu, J.-X., Wang, Q.-L., Xia, M., Qian, C.-W., Ren, Z., Wang, S.-Y., Wang, Y.-F., and Xing, G.-W. (2011). Synthesis and in vitro anti-HSV-1 activity of a novel Hsp90 inhibitor BJ-B11. *Bioorg. Med. Chem. Lett.*, 21(6):1675–1677.

Juang, Y. T., Lowther, W., Kellum, M., Au, W. C., Lin, R., Hiscott, J., and Pitha, P. M. (1998). Primary activation of interferon A and interferon B gene transcription by interferon regulatory factor 3. *Proc Natl Acad Sci USA*, 95(17):9837–9842.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, 110(1-4):462–467.

Kaatsch, P. (2010). Epidemiology of childhood cancer. *Cancer Treat. Rev.*, 36(4):277–285.

Kagan, R. M., Johnson, E. P., Siaw, M., Biswas, P., Chapman, D. S., Su, Z., Platt, J. L., and Pesano, R. L. (2012). A genotypic test for HIV-1 tropism combining Sanger sequencing with ultradeep sequencing predicts virologic response in treatment-experienced patients. *PLoS ONE*, 7(9):e46334.

Kamp, C., Wilke, C. O., Adami, C., and Bornholdt, S. (2003). Viral evolution under the pressure of an adaptive immune system: Optimal mutation rates for viral escape. *Complexity*, 8(2):28–33.

Kamp, C. C. and Bornholdt, S. S. (2002). Coevolution of quasispecies: B-cell mutation rates maximize viral error catastrophes. *Phys Rev Lett*, 88(6):068104–068104.

Kamranvar, S. A., Gruhne, B., Szeles, A., and Masucci, M. G. (2007). Epstein-Barr virus promotes genomic instability in Burkitt's lymphoma. *Oncogene*, 26(35):5115–5123.

Kanagawa, T. T. (2002). Bias and artifacts in multitemplate polymerase chain reactions (PCR). 96(4):7–7.

Kanda, T., Yokosuka, O., Imazeki, F., Tanaka, M., Shino, Y., Shimada, H., Tomonaga, T., Nomura, F., Nagao, K., Ochiai, T., and Saisho, H. (2004). Inhibition of subgenomic hepatitis C virus RNA in Huh-7 cells: ribavirin induces mutagenesis in HCV RNA. *J Viral Hepat*, 11(6):479–487.

Kann, M. G. (2007). Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform*, 8(5):333–346.

Kao, J.-H. and Chen, D.-S. (2002). Global control of hepatitis B virus infection. *Lancet Infect Dis*, 2(7):395–403.

Kapp, E. A., Schütz, F., Connolly, L. M., Chakel, J. A., Meza, J. E., Miller, C. A., Fenyo, D., Eng, J. K., Adkins, J. N., Omenn, G. S., and Simpson, R. J. (2005). An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics*, 5(13):3475–3490.

Karch, H., Denamur, E., Dobrindt, U., Finlay, B. B., Hengge, R., Johannes, L., Ron, E. Z., Tønjum, T., Sansonetti, P. J., and Vicente, M. (2012). The enemy within us: lessons from the 2011 European Escherichia coli O104:H4 outbreak. *EMBO Mol Med*, 4(9):841–848.

Karesh, W. B., Dobson, A., Lloyd-Smith, J. O., Lubroth, J., Dixon, M. A., Bennett, M., Aldrich, S., Harrington, T., Formenty, P., Loh, E. H., Machalaba, C. C., Thomas, M. J., and Heymann, D. L. (2012). Ecology of zoonoses: natural and unnatural histories. *The Lancet*, 380(9857):1936–1945.

Karlas, A., Machuy, N., Shin, Y., Pleissner, K.-P., Artarini, A., Heuer, D., Becker, D., Khalil, H., Ogilvie, L. A., Hess, S., Mäurer, A. P., Müller, E., Wolff, T., Rudel, T., and Meyer, T. F. (2010). Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication. *Nature*, 463(7282):818–822.

Karlin, S., Blaisdell, B. E., and Schachtel, G. A. (1990). Contrasts in codon usage of latent versus productive genes of Epstein-Barr virus: data and hypotheses. *J Virol*, 64(9):4264–4273.

Karlin, S. and Burge, C. (1995). Dinucleotide relative abundance extremes: A genomic signature. *Trends in Genetics*, 11(7):283–290.

Karlin, S., Ladunga, I., and Blaisdell, B. E. (1994). Heterogeneity of genomes: Measures and values. *Proc Natl Acad Sci USA*, 91(26):12837–12841.

Karlin, S., Mrázek, J., and Campbell, A. M. (1997). Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.*, 179(12):3899–3913.

Karp, C. L. (1999). Measles: immunosuppression, interleukin-12, and complement receptors. *Immunol Rev*, 168:91–101.

Kato, N., Hijikata, M., Ootsuyama, Y., Nakagawa, M., Ohkoshi, S., Sugimura, T., and Shimotohno, K. (1990). Molecular cloning of the human hepatitis C virus genome from Japanese patients with non-A, non-B hepatitis. *Proc Natl Acad Sci USA*, 87(24):9524–9528.

Kato, T., Date, T., Miyamoto, M., Sugiyama, M., Tanaka, Y., Orito, E., Ohno, T., Sugihara, K., Hasegawa, I., Fujiwara, K., Ito, K., Ozasa, A., Mizokami, M., and Wakita, T. (2005). Detection of anti-hepatitis C virus effects of interferon and ribavirin by a sensitive replicon system. *J Clin Microbiol*, 43(11):5679–5684.

Kato, T., Date, T., Murayama, A., Morikawa, K., Akazawa, D., and Wakita, T. (2006). Cell culture and infection system for hepatitis C virus. *Nat Protoc*, 1(5):2334–2339.

Katzourakis, A., Gifford, R. J., Tristem, M., Gilbert, M. T. P., and Pybus, O. G. (2009). Macroevolution of complex retroviruses. *Science*, 325(5947):1512.

Kearney, M. F. M., Spindler, J. J., Wiegand, A. A., Shao, W. W., Anderson, E. M. E., Maldarelli, F. F., Ruscetti, F. W. F., Mellors, J. W. J., Hughes, S. H. S., Le Grice, S. F. J. S., and Coffin, J. M. J. (2011). Multiple Sources of Contamination in Samples from Patients Reported to Have XMRV Infection. *PLoS ONE*, 7(2):e30889–e30889.

Keith, C. S., Hoang, D. O., Barrett, B. M., Feigelman, B., Nelson, M. C., Thai, H., and Baysdorfer, C. (1993). Partial sequence analysis of 130 randomly selected maize cDNA clones. *Plant Physiol.*, 101(1):329–332.

Kelly, G. L., Long, H. M., Stylianou, J., Thomas, W. A., Leese, A., Bell, A. I., Bornkamm, G. W., Mautner, J., Rickinson, A. B., and Rowe, M. (2009). An Epstein-Barr virus anti-apoptotic protein constitutively expressed in transformed cells and implicated in burkitt lymphomagenesis: the Wp/BHRF1 link. *PLoS Pathog*, 5(3):e1000341.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R., and Hermjakob, H. (2007). IntAct–open source resource for molecular interaction data. *Nucleic Acids Res*, 35(Database issue):D561–5.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., and Hermjakob, H. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Res*, 40(Database issue):D841–6.

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human Protein Reference Database–2009 update. *Nucleic Acids Res*, 37(Database issue):D767–72.

Khadka, S., Vangeloff, A. D., Zhang, C., Siddavatam, P., Heaton, N. S., Wang, L., Sengupta, R., Sahasrabudhe, S., Randall, G., Gribskov, M., Kuhn, R. J., Perera, R., and LaCount, D. J. (2011). A physical interaction network of dengue virus and human proteins. *Mol Cell Proteomics*, 10(12):M111.012187.

Khakoo, S. I., Thio, C. L., Martin, M. P., Brooks, C. R., Gao, X., Astemborski, J., Cheng, J., Goedert, J. J., Vlahov, D., Hilgartner, M., Cox, S., Little, A.-M., Alexander, G. J., Cramp, M. E., O'Brien, S. J., Rosenberg, W. M. C., Thomas, D. L., and Carrington, M. (2004). HLA and NK cell inhibitory receptor genes in resolving hepatitis C virus infection. *Science*, 305(5685):872–874.

Khalid, S. S., Hamid, S., Siddiqui, A. A., Qureshi, A., and Qureshi, N. (2011). Gene profiling of early and advanced liver disease in chronic hepatitis C patients. *Hepatol Int*, 5(3):782–788.

Khoury, J. D., Tannir, N. M., Williams, M. D., Chen, Y., Yao, H., Zhang, J., Thompson, E. J., TCGA Network, Meric-Bernstam, F., Medeiros, L. J., Weinstein, J. N., and Su, X. (2013). Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq. *J Virol*, 87(16):8916–8926.

Khurana, S., Chearwae, W., Castellino, F., Manischewitz, J., King, L. R., Honorkiewicz, A., Rock, M. T., Edwards, K. M., Del Giudice, G., Rappuoli, R., and Golding, H. (2010). Vaccines with MF59 adjuvant expand the antibody repertoire to target protective sites of pandemic avian H5N1 influenza virus. *Sci Transl Med*, 2(15):15ra5.

Kidwell, M. G. and Lisch, D. R. (2001). Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution*, 55(1):1–24.

Kieffer, T. L., Kwong, A. D., and Picchio, G. R. (2010). Viral resistance to specifically targeted antiviral therapies for hepatitis C (STAT-Cs). *J. Antimicrob. Chemother.*, 65(2):202–212.

Kieffer, T. L., Sarrazin, C., Miller, J. S., Welker, M. W., Forestier, N., Reesink, H. W., Kwong, A. D., and Zeuzem, S. (2007). Telaprevir and pegylated interferon-alpha-2a inhibit wild-type and resistant genotype 1 hepatitis C virus replication in patients. *Hepatology (Baltimore, Md)*, 46(3):631–639.

Kim, K. H. and Bae, J. W. (2011). Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl. Environ. Microbiol.*, 77(21):7663–7668.

Kim, M.-S., Park, E.-J., Roh, S. W., and Bae, J.-W. (2011). Diversity and abundance of single-stranded DNA viruses in human feces. *Appl. Environ. Microbiol.*, 77(22):8062–8070.

Kim, P. M., Lu, L. J., Xia, Y., and Gerstein, M. B. (2006). Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, 314(5807):1938–1941.

Kim, S., Jeong, K., Bhutani, K., Lee, J. H., Patel, A., Scott, E., Nam, H., Lee, H., Gleeson, J. G., and Bafna, V. (2013). Virmid: accurate detection of somatic mutations with sample impurity inference. *Genome Biol*, 14(8):R90.

Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W., and Vogelstein, B. (2011). Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci USA*, 108(23):9530–9535.

Kindt, J., Tzlil, S., Ben-Shaul, A., and Gelbart, W. M. (2001). DNA packaging and ejection forces in bacteriophage. *Proc Natl Acad Sci USA*, 98(24):13671–13674.

King, A. M. Q., Adams, M. J., Lefkowitz, E. J., and Carstens, E. B. (2011). *Virus Taxonomy*. IXth Report of the International Committee on Taxonomy of Viruses. Elsevier Science Limited.

Kirchmair, J., Distinto, S., Liedl, K. R., Markt, P., Rollinger, J. M., Schuster, D., Spitzer, G. M., and Wolber, G. (2011). Development of anti-viral agents using molecular modeling and virtual screening techniques. *Infectious disorders drug targets*, 11(1):64–93.

Kledal, T. N. T., Rosenkilde, M. M. M., Coulin, F. F., Simmons, G. G., Johnsen, A. H. A., Alouani, S. S., Power, C. A. C., Lüttichau, H. R. H., Gerstoft, J. J., Clapham, P. R. P., Clark-Lewis, I. I., Wells, T. N. T., and Schwartz, T. W. T. (1997). A broad-spectrum chemokine antagonist encoded by Kaposi's sarcoma-associated herpesvirus. *Science*, 277(5332):1656–1659.

Klein, E., Kis, L. L., and Klein, G. (2007). Epstein-Barr virus infection in humans: from harmless to life endangering virus-lymphocyte interactions. *Oncogene*, 26(9):1297–1305.

Knipe, D. M. and Cliffe, A. (2008). Chromatin control of herpes simplex virus lytic and latent infection. *Nat Rev Microbiol*, 6(3):211–221.

Knossow, M., Daniels, R. S., Douglas, A. R., Skehel, J. J., and Wiley, D. C. (1984). Three-dimensional structure of an antigenic mutant of the influenza virus haemagglutinin. *Nature*, 311(5987):678–680.

Knudson, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci USA*, 68(4):820–823.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*, 22(3):568–576.

Kochhar, D. M., Penner, J. D., and Knudsen, T. B. (1980). Embryotoxic, teratogenic, and metabolic effects of ribavirin in mice. *Toxicol. Appl. Pharmacol.*, 52(1):99–112.

Kodama, Y., Shumway, M., Leinonen, R., and International Nucleotide Sequence Database Collaboration (2012). The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res*, 40(Database issue):D54–6.

Koev, G. and Kati, W. (2008). The emerging field of HCV drug resistance. *Expert Opin Investig Drugs*, 17(3):303–319.

Komarova, N. L. and Wodarz, D. (2005). Drug resistance in cancer: principles of emergence and prevention. *Proc Natl Acad Sci USA*, 102(27):9714–9719.

König, R., Stertz, S., Zhou, Y., Inoue, A., Hoffmann, H.-H., Bhattacharyya, S., Alamares, J. G., Tscherne, D. M., Ortigoza, M. B., Liang, Y., Gao, Q., Andrews, S. E., Bandyopadhyay, S., De Jesus, P., Tu, B. P., Pache, L., Shih, C., Orth, A., Bonamy, G., Miraglia, L., Ideker, T., García-Sastre, A., Young, J. A. T., Palese, P., Shaw, M. L., and Chanda, S. K. (2010). Human host factors required for influenza virus replication. *Nature*, 463(7282):813–817.

König, R., Zhou, Y., Elleder, D., Diamond, T. L., Bonamy, G. M. C., Irelan, J. T., Chiang, C.-Y., Tu, B. P., De Jesus, P. D., Lilley, C. E., Seidel, S., Opaluch, A. M., Caldwell, J. S., Weitzman, M. D., Kuhen, K. L., Bandyopadhyay, S., Ideker, T., Orth, A. P., Miraglia, L. J., Bushman, F. D., Young, J. A., and Chanda, S. K. (2008). Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell*, 135(1):49–60.

Koonin, E. V. (2009). Darwinian evolution in the light of genomics. *Nucleic Acids Res*, 37(4):1011–1034.

Koonin, E. V. and Dolja, V. V. (2006). Evolution of complexity in the viral world: the dawn of a new vision. *Virus Res*, 117(1):1–4.

Koonin, E. V. and Dolja, V. V. (2013). A virocentric perspective on the evolution of life. *Curr Opin Vir*, 3(5):546–557.

Koonin, E. V., Makarova, K. S., and Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol*, 55:709–742.

Koonin, E. V., Senkevich, T. G., and Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biol Direct*, 1:29.

Koonin, E. V., Senkevich, T. G., and Dolja, V. V. (2009). Compelling reasons why viruses are relevant for the origin of cells. *Nat Rev Microbiol*, 7(8):615–615.

Koonin, E. V. and Wolf, Y. I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res*, 36(21):6688–6719.

Koopmans, M., Wilbrink, B., Conyn, M., Natrop, G., van der Nat, H., Vennema, H., Meijer, A., van Steenbergen, J., Fouchier, R., Osterhaus, A., and Bosman, A. (2004). Transmission of H7N7 avian influenza A virus to human beings during a large outbreak in commercial poultry farms in the Netherlands. *Lancet*, 363(9409):587–593.

Koren, S., Treangen, T. J., and Pop, M. (2011). Bambus 2: Scaffolding metagenomes. *Bioinformatics*, 27(21):2964–2971.

Kostic, A. D., Ojesina, A. I., Pedamallu, C. S., Jung, J., Verhaak, R. G. W., Getz, G., and Meyerson, M. (2011). PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat Biotechnol*, 29(5):393–396.

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., and Turner, D. J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods*, 6(4):291–295.

Koziel, M. J., Dudley, D., Afdhal, N., Choo, Q. L., Houghton, M., Ralston, R., and Walker, B. D. (1993). Hepatitis C virus (HCV)-specific cytotoxic T lymphocytes recognize epitopes in the core and envelope proteins of HCV. *J Virol*, 67(12):7522–7532.

Kragh, H. (1999). *Cosmology and Controversy*. The Historical Development of Two Theories of the Universe. Princeton University Press.

Kreuze, J. F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., and Simon, R. (2009). Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. *Virology*, 388(1):1–7.

Krieger, N., Lohmann, V., and Bartenschlager, R. (2001). Enhancement of hepatitis C virus RNA replication by cell culture-adaptive mutations. *J Virol*, 75(10):4614–4624.

Krilov, L. R. (2001). Respiratory Syncytial Virus: Update on Infection, Treatment, and Prevention. *Curr Infect Dis Rep*, 3(3):242–246.

Kristensen, A. R., Gsponer, J., and Foster, L. J. (2012). A high-throughput approach for measuring temporal changes in the interactome. *Nat Methods*, 9(9):907–909.

Kristensen, D. M., Mushegian, A. R., Dolja, V. V., and Koonin, E. V. (2010). New dimensions of the virus world discovered through metagenomics. *Trends in microbiology*, 18(1):11–19.

Krogan, N. J. N., Cagney, G. G., Yu, H. H., Zhong, G. G., Guo, X. X., Ignatchenko, A. A., Li, J. J., Pu, S. S., Datta, N. N., Tikuisis, A. P. A., Punna, T. T., Peregrín-Alvarez, J. M. J., Shales, M. M., Zhang, X. X., Davey, M. M., Robinson, M. D. M., Paccanaro, A. A., Bray, J. E. J., Sheung, A. A., Beattie, B. B., Richards, D. P. D., Canadien, V. V., Lalev, A. A., Mena, F. F., Wong, P. P., Starostine, A. A., Canete, M. M. M., Vlasblom, J. J., Wu, S. S., Orsi, C. C., Collins, S. R. S., Chandran, S. S., Haw, R. R., Rilstone, J. J. J., Gandi, K. K., Thompson, N. J. N., Musso, G. G., Onge, P. P. S., Ghanny, S. S., Lam, M. H. Y. M., Butland, G. G., Altaf-Ul, A. M. A., Kanaya, S. S., Shilatifard, A. A., O'Shea, E. E., Weissman, J. S. J., Ingles, C. J. C., Hughes, T. R. T., Parkinson, J. J., Gerstein, M. M., Wodak, S. J. S., Emili, A. A., and Greenblatt, J. F. J. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440(7084):637–643.

Krupovic, M. and Bamford, D. H. (2007). Putative prophages related to lytic tailless marine dsDNA phage PM2 are widespread in the genomes of aquatic bacteria. *BMC Genomics*, 8:236.

Krupovic, M. and Bamford, D. H. (2009). Does the evolution of viral polymerases reflect the origin and evolution of viruses? *Nat Rev Microbiol*, 7(3):250–author reply 250.

Krynska, B., Del Valle, L., Croul, S., Gordon, J., Katsetos, C. D., Carbone, M., Giordano, A., and Khalili, K. (1999). Detection of human neurotropic JC virus DNA sequence and expression of the viral oncogenic protein in pediatric medulloblastomas. *Proc Natl Acad Sci USA*, 96(20):11519–11524.

Kuciak, M., Gabus, C., Ivanyi-Nagy, R., Semrad, K., Storchak, R., Chaloin, O., Muller, S., Mély, Y., and Darlix, J.-L. (2008). The HIV-1 transcriptional activator Tat has potent nucleic acid chaperoning activities in vitro. *Nucleic Acids Res*, 36(10):3389–3400.

Kumar, D., Abdulovic, A. L., Viberg, J., Nilsson, A. K., Kunkel, T. A., and Chabes, A. (2011a). Mechanisms of mutagenesis in vivo due to imbalanced dNTP pools. *Nucleic Acids Res*, 39(4):1360–1371.

Kumar, N., Liang, Y., Parslow, T. G., and Liang, Y. (2011b). Receptor tyrosine kinase inhibitors block multiple steps of influenza a virus replication. *J Virol*, 85(6):2818–2827.

Kumar, R. and Nanduri, B. (2010). HPIDB–a unified resource for host-pathogen interactions. *BMC Bioinformatics*, 11 Suppl 6:S16.

Kundrotas, P. J., Zhu, Z., Janin, J., and Vakser, I. A. (2012). Templates are available to model nearly all complexes of structurally characterized proteins. *Proc Natl Acad Sci USA*, 109(24):9438–9441.

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., and Hugenholtz, P. (2008). A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.*, 72(4):557–578.

Kuntz, I. D. (1992). Structure-based strategies for drug design and discovery. *Science*, 257(5073):1078–1082.

Kuo, G., Choo, Q. L., Alter, H. J., Gitnick, G. L., Redeker, A. G., Purcell, R. H., Miyamura, T., Dienstag, J. L., Alter, M. J., and Stevens, C. E. (1989). An assay for circulating antibodies to a major etiologic virus of human non-A, non-B hepatitis. *Science*, 244(4902):362–364.

Kurth, R. and Bannert, N. (2010). Beneficial and detrimental effects of human endogenous retroviruses. *Int. J. Cancer*, 126(2):306–314.

Kwo, P. Y., Lawitz, E. J., McCone, J., Schiff, E. R., Vierling, J. M., Pound, D., Davis, M. N., Galati, J. S., Gordon, S. C., Ravendhran, N., Rossaro, L., Anderson, F. H., Jacobson, I. M., Rubin, R., Koury, K., Pedicone, L. D., Brass, C. A., Chaudhri, E., Albrecht, J. K., and SPRINT-1 investigators (2010). Efficacy of boceprevir, an NS3 protease inhibitor, in combination with peginterferon alfa-2b and ribavirin in treatment-naive patients with genotype 1 hepatitis C infection (SPRINT-1): an open-label, randomised, multicentre phase 2 trial. *Lancet*, 376(9742):705–716.

Kwong, A. D., Najera, I., Bechtel, J., Bowden, S., Fitzgibbon, J., Harrington, P., Kempf, D., Kieffer, T. L., Koletzki, D., Kukolj, G., Lim, S., Pilot-Matias, T., Lin, K., Mani, N., Mo, H., O'Rear, J., Otto, M., Parkin, N., Pawlotsky, J.-M., Petropoulos, C., Picchio, G., Ralston, R., Reeves, J. D., Schooley, R. T., Seiwert, S., Standring, D., Stuyver, L., Sullivan, J., Miller, V., Forum for Collaborative Human Immunodeficiency Virus Research, HCV Drug Development Advisory Group (HCV DRAG), Sequence Analysis Working Group (SAWG), and Phenotype Analysis Working Group (PAWG) (2011). Sequence and phenotypic analysis for resistance monitoring in hepatitis C virus drug development: recommendations from the HCV DRAG. In *Gastroenterology*, pages 755–760. Vertex Pharmaceuticals, Inc, Cambridge, Massachusetts, USA.

Kwong, P. D. (2005). Human immunodeficiency virus: refolding the envelope. *Nature*, 433(7028):815–816.

Kwun, H. J., da Silva, S. R., Shah, I. M., Blake, N., Moore, P. S., and Chang, Y. (2007). Kaposi's sarcoma-associated herpesvirus latency-associated nuclear antigen 1 mimics Epstein-Barr virus EBNA1 immune evasion through central repeat domain effects on protein processing. *J Virol*, 81(15):8225–8235.

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E., and Raoult, D. (2008). The virophage as a unique parasite of the giant mimivirus. *Nature*, 455(7209):100–104.

Lai, B., Ding, R., Li, Y., Duan, L., and Zhu, H. (2012). A de novo metagenomic assembly program for shotgun DNA reads. *Bioinformatics*, 28(11):1455–1462.

Lakadamyali, M., Rust, M. J., Babcock, H. P., and Zhuang, X. (2003). Visualizing infection of individual influenza viruses. *Proc Natl Acad Sci USA*, 100(16):9280–9285.

Lalani, A. S. A., Graham, K. K., Mossman, K. K., Rajarathnam, K. K., Clark-Lewis, I. I., Kelvin, D. D., and McFadden, G. G. (1997). The purified myxoma virus gamma interferon receptor homolog M-T7 interacts with the heparin-binding domains of chemokines. *J Virol*, 71(6):4356–4363.

Lamarre, D., Anderson, P. C., Bailey, M., Beaulieu, P., Bolger, G., Bonneau, P., Bös, M., Cameron, D. R., Cartier, M., Cordingley, M. G., Faucher, A.-M., Goudreau, N., Kawai, S. H., Kukolj, G., Lagacé, L., LaPlante, S. R., Narjes, H., Poupart, M.-A., Rancourt, J., Sentjens, R. E., St George, R., Simoneau, B., Steinmann, G., Thibeault, D., Tsantrizos, Y. S., Weldon, S. M., Yong, C.-L., and Llinàs-Brunet, M. (2003). An NS3 protease inhibitor with antiviral effects in humans infected with hepatitis C virus. *Nature*, 426(6963):186–189.

Lambert, C. G. and Black, L. J. (2012). Learning from our GWAS mistakes: from experimental design to scientific method. *Biostatistics*, 13(2):195–203.

Lambris, J. D., Ricklin, D., and Geisbrecht, B. V. (2008). Complement evasion by human pathogens. *Nat Rev Microbiol*, 6(2):132–142.

Lane, D. P. and Crawford, L. V. (1979). T antigen is bound to a host protein in SV40-transformed cells. *Nature*, 278(5701):261–263.

Lanford, R. E., Chavez, D., Guerra, B., Lau, J. Y., Hong, Z., Brasky, K. M., and Beames, B. (2001). Ribavirin induces error-prone replication of GB virus B in primary tamarin hepatocytes. *J Virol*, 75(17):8074–8081.

Lanford, R. E., Guerra, B., Lee, H., Averett, D. R., Pfeiffer, B., Chavez, D., Notvall, L., and Bigger, C. (2003). Antiviral effect and virus-host interactions in response to alpha interferon, gamma interferon, poly(i)-poly(c), tumor necrosis factor alpha, and ribavirin in hepatitis C virus subgenomic replicons. *J Virol*, 77(2):1092–1104.

Lange, C. M. and Zeuzem, S. (2013). Perspectives and challenges of interferon-free therapy for chronic hepatitis C. *J Hepatol*, 58(3):583–592.

Lanier, L. L. L. (2008). Evolutionary struggles between NK cells and viruses. *Nat Rev Immunol*, 8(4):259–268.

Lapierre, P. and Gogarten, J. P. (2009). Estimating the size of the bacterial pan-genome. *Trends Genet*, 25(3):107–110.

Lappe, M. and Holm, L. (2004). Unraveling protein interaction networks with near-optimal efficiency. *Nat Biotechnol*, 22(1):98–103.

Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., Ley, T. J., Mardis, E. R., Wilson, R. K., and Ding, L. (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317.

Laserson, J., Jojic, V., and Koller, D. (2011). Genovo: De novo assembly for metagenomes. *Journal of Computational Biology*, 18(3):429–443.

Laskus, T., Wilkinson, J., Gallegos-Orozco, J. F., Radkowski, M., Adair, D. M., Nowicki, M., Operskalski, E., Buskell, Z., Seeff, L. B., Vargas, H., and Rakela, J. (2004). Analysis of hepatitis C virus quasispecies transmission and evolution in patients infected through blood transfusion. *Gastroenterology*, 127(3):764–776.

Lauber, C., Goeman, J. J., Parquet, M. d. C., Nga, P. T., Snijder, E. J., Morita, K., and Gorbalenya, A. E. (2013). The footprint of genome architecture in the largest genome expansion in RNA viruses. *PLoS Pathog*, 9(7):e1003500.

Lauer, G. M., Barnes, E., Lucas, M., Timm, J., Ouchi, K., Kim, A. Y., Day, C. L., Robbins, G. K., Casson, D. R., Reiser, M., Dusheiko, G., Allen, T. M., Chung, R. T., Walker, B. D., and Klenerman, P. (2004). High resolution analysis of cellular immune responses in resolved and persistent hepatitis C virus infection. *Gastroenterology*, 127(3):924–936.

Lauring, A. S. and Andino, R. (2010). Quasispecies Theory and the Behavior of RNA Viruses. *PLoS Pathog*, 6(7):e1001005.

Lauring, A. S. and Andino, R. (2011). Exploring the fitness landscape of an RNA virus by using a universal barcode microarray. *J Virol*, 85(8):3780–3791.

Lavallée-Adam, M., Cloutier, P., Coulombe, B., and Blanchette, M. (2011). Modeling contaminants in AP-MS/MS experiments. *J Proteome Res*, 10(2):886–895.

Lawrence, J. G., Hatfull, G. F., and Hendrix, R. W. (2002). Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. *J. Bacteriol.*, 184(17):4891–4905.

Lawson, J. S., Tran, D., and Rawlinson, W. D. (2001). From Bittner to Barr: a viral, diet and hormone breast cancer aetiology hypothesis. *Breast Cancer Res.*, 3(2):81–85.

Lazcano, A., Guerrero, R., Margulis, L., and Oró, J. (1988). The evolutionary transition from RNA to DNA in early cells. *J. Mol. Evol.*, 27(4):283–290.

Le Guen, B., Squadrito, G., Nalpas, B., Berthelot, P., Pol, S., and Bréchot, C. (1997). Hepatitis C virus genome complexity correlates with response to interferon therapy: a study in French patients with chronic hepatitis C. *Hepatology (Baltimore, Md)*, 25(5):1250–1254.

Lechner, F., Wong, D. K., Dunbar, P. R., Chapman, R., Chung, R. T., Dohrenwend, P., Robbins, G., Phillips, R., Klenerman, P., and Walker, B. D. (2000). Analysis of successful immune responses in persons infected with hepatitis C virus. *J Exp Med*, 191(9):1499–1512.

Lecompte, O., Ripp, R., Thierry, J.-C., Moras, D., and Poch, O. (2002). Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res*, 30(24):5382–5390.

Lee, S., Yoon, J., Park, B., Jun, Y., Jin, M., Sung, H. C., Kim, I. H., Kang, S., Choi, E. J., Ahn, B. Y., and Ahn, K. (2000). Structural and functional dissection of human cytomegalovirus US3 in binding major histocompatibility complex class I molecules. *J Virol*, 74(23):11262–11269.

Lee, S.-A., Chan, C.-h., Tsai, C.-H., Lai, J.-M., Wang, F.-S., Kao, C.-Y., and Huang, C.-Y. F. (2008). Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics*, 9 Suppl 12:S11.

Lee, S.-W., Berger, S. J., Martinović, S., Pasa-Tolić, L., Anderson, G. A., Shen, Y., Zhao, R., and Smith, R. D. (2002). Direct mass spectrometric analysis of intact proteins of the yeast large ribosomal subunit using capillary LC/FTICR. *Proc Natl Acad Sci USA*, 99(9):5942–5947.

Lee, W., Jiang, Z., Liu, J., Haverty, P. M., Guan, Y., Stinson, J., Yue, P., Zhang, Y., Pant, K. P., Bhatt, D., Ha, C., Johnson, S., Kennemer, M. I., Mohan, S., Nazarenko, I., Watanabe, C., Sparks, A. B., Shames, D. S., Gentleman, R., de Sauvage, F. J., Stern, H., Pandita, A., Ballinger, D. G., Drmanac, R., Modrusan, Z., Seshagiri, S., and Zhang, Z. (2010). The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature*, 465(7297):473–477.

Lehmann-Grube, F., Ibscher, B., Bugislaus, E., and Kallay, M. (1979). A serological study concerning the role of the golden hamster (Mesocricetus auratus) in transmitting lymphocytic choriomeningitis virus to humans. *Med Microbiol Immunol*, 167(3):205–210.

Leipe, D. D., Aravind, L., and Koonin, E. V. (1999). Did DNA replication evolve twice independently? *Nucleic Acids Res*, 27(17):3389–3401.

Leitner, A., Walzthoeni, T., Kahraman, A., Herzog, F., Rinner, O., Beck, M., and Aebersold, R. (2010). Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Mol Cell Proteomics*, 9(8):1634–1649.

Lengauer, T. (2007). *Bioinformatics - From Genomes to Therapies*. Wiley-VCH.

Lengauer, T., Lemmen, C., Rarey, M., and Zimmermann, M. (2004). Novel technologies for virtual screening. *Drug Discov. Today*, 9(1):27–34.

Lengauer, T., Timmerman, H., Kubinyi, H., and Mannhold, R. (2002). Bioinformatics - from Genomes to Drugs. Wiley-VCH.

Lengauer, T. T. (2011). Bioinformatical assistance of selecting anti-HIV therapies: where do we stand? *Intervirology*, 55(2):108–112.

Lengauer, T. T. and Sing, T. T. (2006). Bioinformatics-assisted anti-HIV therapy. *Nat Rev Microbiol*, 4(10):790–797.

Lengeling, A., Pfeffer, K., and Balling, R. (2001). The battle of two genomes: genetics of bacterial host/pathogen interactions in mice. *Mamm Genome*, 12(4):261–271.

Leroux, M. R. (1999). MtGimC, a novel archaeal chaperone related to the eukaryotic chaperonin cofactor GimC/prefoldin. *EMBO J*, 18(23):6730–6743.

Lesburg, C. A., Cable, M. B., Ferrari, E., Hong, Z., Mannarino, A. F., and Weber, P. C. (1999). Crystal structure of the RNA-dependent RNA polymerase from hepatitis C virus reveals a fully encircled active site. *Nat. Struct. Biol.*, 6(10):937–943.

Lesch, S. M. and Jeske, D. R. (2009). Some Suggestions for Teaching About Normal Approximations to Poisson and Binomial Distribution Functions. *The American Statistician*, 63(3):274–277.

Levi, L. I., Gnädig, N. F., Beaucourt, S., McPherson, M. J., Baron, B., Arnold, J. J., and Vignuzzi, M. (2010). Fidelity variants of RNA dependent RNA polymerases uncover an indirect, mutagenic activity of amiloride compounds. *PLoS Pathog*, 6(10):e1001163.

Levin, H. L. and Moran, J. V. (2011). Dynamic interactions between transposable elements and their hosts. *Nat Rev Genet*, 12(9):615–627.

Levine, A. J. (2009). The common mechanisms of transformation by the small DNA tumor viruses: The inactivation of tumor suppressor gene products: p53. *Virology*, 384(2):285–293.

Levine, A. J. and Enquist, L. W. (2007). History of Virology. In Knipe, D. M. and Howley, P. M., editors, *Fields Virology fifth ed.*, pages 3–24. Lippincott Williams & Wilkins.

Levitskaya, J. J., Coram, M. M., Levitsky, V. V., Imreh, S. S., Steigerwald-Mullen, P. M. P., Klein, G. G., Kurilla, M. G. M., and Masucci, M. G. M. (1995). Inhibition of antigen processing by the internal repeat region of the Epstein-Barr virus nuclear antigen-1. *Nature*, 375(6533):685–688.

Levrero, M. (2006). Viral hepatitis and liver cancer: the case of hepatitis C. *Oncogene*, 25(27):3834–3847.

Lewandowski, G. A., Lo, D., and Bloom, F. E. (1993). Interference with major histocompatibility complex class II-restricted antigen presentation in the brain by herpes simplex virus type 1: a possible mechanism of evasion of the immune response. *Proc Natl Acad Sci USA*, 90(5):2005–2009.

Lewis, K. (2013). Platforms for antibiotic discovery. *Nat Rev Drug Disc*, 12(5):371–387.

Lewthwaite, J., Skinner, A., and Henderson, B. (1998). Are molecular chaperones microbial virulence factors? *Trends in microbiology*, 6(11):426–428.

Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M. D., Chen, K., Dooling, D., Dunford-Shore, B. H., McGrath, S., Hickenbotham, M., Cook, L., Abbott, R., Larson, D. E., Koboldt, D. C., Pohl, C., Smith, S., Hawkins, A., Abbott, S., Locke, D., Hillier, L. W., Miner, T., Fulton, L., Magrini, V., Wylie, T., Glasscock, J., Conyers, J., Sander, N., Shi, X., Osborne, J. R., Minx, P., Gordon, D., Chinwalla, A., Zhao, Y., Ries, R. E., Payton, J. E., Westervelt, P., Tomasson, M. H., Watson, M., Baty, J., Ivanovich, J., Heath, S., Shannon, W. D., Nagarajan, R., Walter, M. J., Link, D. C., Graubert, T. A., Dipersio, J. F., and Wilson, R. K. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, 456(7218):66–72.

Leyssen, P., Balzarini, J., de Clercq, E., and Neyts, J. (2005). The predominant mechanism by which ribavirin exerts its antiviral activity in vitro against flaviviruses and paramyxoviruses is mediated by inhibition of IMP dehydrogenase. *J Virol*, 79(3):1943–1947.

Leyssen, P., de Clercq, E., and Neyts, J. (2006). The anti-yellow fever virus activity of ribavirin is independent of error-prone replication. *Mol. Pharmacol.*, 69(4):1461–1467.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, H. and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform*, 11(5):473–483.

Li, H. and Roossinck, M. J. (2004). Genetic bottlenecks reduce population variation in an experimental RNA virus population. *J Virol*, 78(19):10582–10587.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–1858.

Li, J.-W., Wan, R., Yu, C.-S., Co, N. N., Wong, N., and Chan, T.-F. (2013). ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. *Bioinformatics*, 29(5):649–651.

Li, K., Foy, E., Ferreon, J. C., Nakamura, M., Ferreon, A. C. M., Ikeda, M., Ray, S. C., Gale, M., and Lemon, S. M. (2005). Immune evasion by hepatitis C virus NS3/4A protease-mediated cleavage of the Toll-like receptor 3 adaptor protein TRIF. *Proc Natl Acad Sci USA*, 102(8):2992–2997.

Li, Q., Brass, A. L., Ng, A., Hu, Z., Xavier, R. J., Liang, T. J., and Elledge, S. J. (2009b). A genome-wide genetic screen for host factors required for hepatitis C virus propagation. *Proc Natl Acad Sci USA*, 106(38):16410–16415.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., and Wang, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*, 20(2):265–272.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., and Fan, W. (2012). Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief Funct Genomics*, 11(1):25–37.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A. P., Santonico, E., Castagnoli, L., and Cesareni, G. (2012). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res*, 40(Database issue):D857–61.

Lien, K.-Y., Chuang, Y.-H., Hung, L.-Y., Hsu, K.-F., Lai, W.-W., Ho, C.-L., Chou, C.-Y., and Lee, G.-B. (2010). Rapid isolation and detection of cancer cells by utilizing integrated microfluidic systems. *Lab Chip*, 10(21):2875–2886.

Lilley, C. E., Schwartz, R. A., and Weitzman, M. D. (2007). Using or abusing: viruses and the cellular DNA damage response. *Trends in microbiology*, 15(3):119–126.

Lim, J. K. and Murphy, P. M. (2011). Chemokine control of West Nile virus infection. *Exp. Cell Res.*, 317(5):569–574.

Lim, S.-O., Park, S.-G., Yoo, J.-H., Park, Y.-M., Kim, H.-J., Jang, K.-T., Cho, J.-W., Yoo, B.-C., Jung, G.-H., and Park, C.-K. (2005). Expression of heat shock proteins (HSP27, HSP60, HSP70, HSP90, GRP78, GRP94) in hepatitis B virus-related hepatocellular carcinomas and dysplastic nodules. *World J Gastroenterol*, 11(14):2072–2079.

Lin, C.-c., Philips, L., Xu, C., and Yeh, L.-T. (2004). Pharmacokinetics and safety of viramidine, a prodrug of ribavirin, in healthy volunteers. *J Clin Pharmacol*, 44(3):265–275.

Lindahl, K., Stahle, L., Bruchfeld, A., and Schvarcz, R. (2005). High-dose ribavirin in combination with standard dose peginterferon for treatment of patients with chronic hepatitis C. *Hepatology (Baltimore, Md)*, 41(2):275–279.

Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M., and Chisholm, S. W. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature*, 438(7064):86–89.

Lingappa, V. R., Hurt, C. R., and Garvey, E. (2012). Capsid Assembly as a Point of Intervention for Novel Anti-viral Therapeutics. *Curr Pharm Biotechnol*.

Lipkin, W. I. (2010). Microbe hunting. *Microbiol. Mol. Biol. Rev.*, 74(3):363–377.

Lipkin, W. I. (2013). The changing face of pathogen discovery and surveillance. *Nat Rev Microbiol*, 11(2):133–141.

Lipsitch, M. (2003). Transmission Dynamics and Control of Severe Acute Respiratory Syndrome. *Science*, 300(5627):1966–1970.

Little, R. F. and Yarchoan, R. (2003). Treatment of gammaherpesvirus-related neoplastic disorders in the immunosuppressed host. *Semin. Hematol.*, 40(2):163–171.

Liu, M.-T., Chen, Y.-R., Chen, S.-C., Hu, C.-Y., Lin, C.-S., Chang, Y.-T., Wang, W.-B., and Chen, J.-Y. (2004). Epstein-Barr virus latent membrane protein 1 induces micronucleus formation, represses DNA repair and enhances sensitivity to DNA-damaging agents in human epithelial cells. *Oncogene*, 23(14):2531–2539.

Liu, S. L. S., Rodrigo, A. G. A., Shankarappa, R. R., Learn, G. H. G., Hsu, L. L., Davidov, O. O., Zhao, L. P. L., and Mullins, J. I. J. (1996). HIV quasispecies and resampling. *Science*, 273(5274):415–416.

Ljunggren, H. G. H. and Kärre, K. K. (1990). In search of the 'missing self': MHC molecules and NK cell recognition. *Immunol Today*, 11(7):237–244.

Loeb, L. A., Essigmann, J. M., Kazazi, F., Zhang, J., Rose, K. D., and Mullins, J. I. (1999). Lethal mutagenesis of HIV with mutagenic nucleoside analogs. *Proc Natl Acad Sci USA*, 96(4):1492–1497.

Loeb, L. A. and Mullins, J. I. (2000). Lethal mutagenesis of HIV by mutagenic ribonucleoside analogs. *AIDS Res Hum Retroviruses*, 16(1):1–3.

Lohmann, V., Körner, F., Dobierzewska, A., and Bartenschlager, R. (2001). Mutations in hepatitis C virus RNAs conferring cell culture adaptation. *J Virol*, 75(3):1437–1449.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., and Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol*, 30(5):434–439.

López-Garćia, P. and Moreira, D. (1999). Metabolic symbiosis at the origin of eukaryotes. *Trends Biochem Sci*, 24(3):88–93.

López-Labrador, F. X., Ampurdanès, S., Giménez-Barcons, M., Guilera, M., Costa, J., Jiménez de Anta, M. T., Sánchez-Tapias, J. M., Rodés, J., and Sáiz, J. C. (1999). Relationship of the genomic complexity of hepatitis C virus with liver disease severity and response to interferon in patients with chronic HCV genotype 1b infection [correction of interferon]. *Hepatology (Baltimore, Md)*, 29(3):897–903.

Lorenzi, H. A., Hoover, J., Inman, J., Safford, T., Murphy, S., Kagan, L., and Williamson, S. J. (2011). The viral metagenome annotation pipeline (VMGAP): An automated tool for the functional annotation of viral metagenomic shotgun sequencing data. *Stand. Genomic Sci.*, 4(3):418–429.

Lorenzo-Redondo, R., Bordería, A. V., and López-Galíndez, C. (2011). Dynamics of in vitro fitness recovery of HIV-1. *J Virol*, 85(4):1861–1870.

Lu, X., Yu, H., Liu, S. H., Brodsky, F. M., and Peterlin, B. M. (1998). Interactions between HIV1 Nef and vacuolar ATPase facilitate the internalization of CD4. *Immunity*, 8(5):647–656.

Lubinski, J., Nagashunmugam, T., and Friedman, H. M. (1998). Viral interference with antibody and complement. *Semin. Cell Dev. Biol.*, 9(3):329–337.

Luciani, F. and Alizon, S. (2009). The evolutionary dynamics of a rapidly mutating virus within and between hosts: the case of hepatitis C virus. *PLoS Comput Biol*, 5(11):e1000565.

Lucks, J. B., Nelson, D. R., Kudla, G. R., and Plotkin, J. B. (2008). Genome landscapes and bacteriophage codon usage. *PLoS Comput Biol*, 4(2).

Ludmir, E. B. and Enquist, L. W. (2009). Viral genomes are part of the phylogenetic tree of life. *Nat Rev Microbiol*, 7(8):615–author reply 615.

Luisi, P. L. (1998). About various definitions of life. *Orig Life Evol Biosph*, 28(4-6):613–622.

Lutchman, G., Danehower, S., Song, B.-C., Liang, T. J., Hoofnagle, J. H., Thomson, M., and Ghany, M. G. (2007). Mutation rate of the hepatitis C virus NS5B in patients undergoing treatment with ribavirin monotherapy. *Gastroenterology*, 132(5):1757–1766.

Lwoff, A. (1957). The concept of virus. *Journal of General Microbiology*, 17:239–253.

Lwoff, A., Horne, R., and Tournier, P. (1962). A System of Viruses. *Cold Spring Harbor Symposia on Quantitative Biology*, 27(0):51–55.

Lysholm, F., Wetterbom, A., Lindau, C., Darban, H., Bjerkner, A., Fahlander, K., Lindberg, A. M., Persson, B., Allander, T., and Andersson, B. (2012). Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. *PLoS ONE*, 7(2):e30875.

Ma, M., Huang, Y., Gong, Z., Zhuang, L., Li, C., Yang, H., Tong, Y., Liu, W., and Cao, W. (2011). Discovery of DNA Viruses in Wild-Caught Mosquitoes Using Small RNA High throughput Sequencing. *PLoS ONE*, 6(9):e24758.

Maag, D., Castro, C., Hong, Z., and Cameron, C. E. (2001). Hepatitis C virus RNA-dependent RNA polymerase (NS5B) as a mediator of the antiviral activity of ribavirin. *J Biol Chem*, 276(49):46094–46098.

Macfarlan, T. S. T., Gifford, W. D. W., Driscoll, S. S., Lettieri, K. K., Rowe, H. M. H., Bonanomi, D. D., Firth, A. A., Singer, O. O., Trono, D. D., and Pfaff, S. L. S. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*, 487(7405):57–63.

Machida, K., Cheng, K. T.-H., Lai, C.-K., Jeng, K.-S., Sung, V. M.-H., and Lai, M. M. C. (2006). Hepatitis C virus triggers mitochondrial permeability transition with production of reactive oxygen species, leading to DNA damage and STAT3 activation. *J Virol*, 80(14):7199–7207.

Mackay, J. P., Sunde, M., Lowry, J. A., Crossley, M., and Matthews, J. M. (2007). Protein interactions: is seeing believing? *Trends Biochem Sci*, 32(12):530–531.

MacPherson, J. I., Dickerson, J. E., Pinney, J. W., and Robertson, D. L. (2010). Patterns of HIV-1 protein interaction identify perturbed host-cellular subsystems. *PLoS Comput Biol*, 6(7):e1000863.

Madera, M., Vogel, C., Kummerfeld, S. K., Chothia, C., and Gough, J. (2004). The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res*, 32(Database issue):D235–9.

Maeda, N., Fan, H., and Yoshikai, Y. (2008). Oncogenesis by retroviruses: old and new paradigms. *Rev. Med. Virol.*, 18(6):387–405.

Maggio, E., van den Berg, A., Diepstra, A., Kluiver, J., Visser, L., and Poppema, S. (2002). Chemokines, cytokines and their receptors in Hodgkin's lymphoma cell lines and tissues. *Ann Oncol*, 13 Suppl 1:52–56.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2011). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, 39(Database issue):D52–7.

Maki, D. G. (2006). Don't Eat the Spinach — Controlling Foodborne Infectious Disease. *N. Engl. J. Med.*, 355(19):1952–1955.

Maksakova, I. A., Mager, D. L., and Reiss, D. (2008). Keeping active endogenous retroviral-like elements in check: the epigenetic perspective. *Cell Mol Life Sci*, 65(21):3329–3347.

Maksakova, I. A., Romanish, M. T., Gagnier, L., Dunn, C. A., van de Lagemaat, L. N., and Mager, D. L. (2006). Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. *PLoS Genet*, 2(1):e2.

Maley, C. C., Galipeau, P. C., Finley, J. C., Wongsurawat, V. J., Li, X., Sanchez, C. A., Paulson, T. G., Blount, P. L., Risques, R.-A., Rabinovitch, P. S., and Reid, B. J. (2006). Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat Genet*, 38(4):468–473.

Mallet, F. F., Bouton, O. O., Prudhomme, S. S., Cheynet, V. V., Oriol, G. G., Bonnaud, B. B., Lucotte, G. G., Duret, L. L., and Mandrand, B. B. (2004). The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. *Proc Natl Acad Sci USA*, 101(6):1731–1736.

Malovannaya, A., Li, Y., Bulynko, Y., Jung, S. Y., Wang, Y., Lanz, R. B., O'Malley, B. W., and Qin, J. (2010). Streamlined analysis schema for high-throughput identification of endogenous protein complexes. *Proc Natl Acad Sci USA*, 107(6):2431–2436.

Mande, S. S., Mohammed, M. H., and Ghosh, T. S. (2012). Classification of metagenomic sequences: methods and challenges. *Brief Bioinform*, 13(6):669–681.

Mangeat, B., Turelli, P., Caron, G., Friedli, M., Perrin, L., and Trono, D. (2003). Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. *Nature*, 424(6944):99–103.

Mangeney, M., Renard, M., Schlecht-Louf, G., Bouallaga, I., Heidmann, O., Letzelter, C., Richaud, A., Ducos, B., and Heidmann, T. (2007). Placental syncytins: Genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins. *Proc Natl Acad Sci USA*, 104(51):20534–20539.

Mangili, A. and Gendreau, M. (2005). Transmission of infectious diseases during commercial air travel. *The Lancet*, 365(9463):989–996.

Manns, M. P., Foster, G. R., Rockstroh, J. K., Zeuzem, S., Zoulim, F., and Houghton, M. (2007). The way forward in HCV treatment–finding the right path. *Nat Rev Drug Disc*, 6(12):991–1000.

Manns, M. P., McHutchison, J. G., Gordon, S. C., Rustgi, V. K., Shiffman, M., Reindollar, R., Goodman, Z. D., Koury, K., Ling, M., and Albrecht, J. K. (2001). Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of chronic hepatitis C: a randomised trial. *Lancet*, 358(9286):958–965.

Manns, M. P., Wedemeyer, H., and Cornberg, M. (2006). Treating viral hepatitis C: efficacy, side effects, and complications. *Gut*, 55(9):1350–1359.

Manrubia, S. C., Domingo, E., and Lázaro, E. (2010). Pathways to extinction: beyond the error threshold. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1548):1943–1952.

Manrubia, S. C., Escarmís, C., Domingo, E., and Lázaro, E. (2005). High mutation rates, bottlenecks, and robustness of RNA viral quasispecies. *Gene*, 347(2):273–282.

Mant, C. and Cason, J. (2004). A human murine mammary tumour virus-like agent is an unconvincing aetiological agent for human breast cancer. *Rev. Med. Virol.*, 14(3):169–177.

Mant, C. and Cason, J. (2005). Mouse mammary tumor virus and human breast cancer. *Cancer Research*, 65(3):1112–author reply 1112–3.

Mant, C., Gillett, C., D'Arrigo, C., and Cason, J. (2004a). Human murine mammary tumour virus-like agents are genetically distinct from endogenous retroviruses and are not detectable in breast cancer cell lines or biopsies. *Virology*, 318(1):393–404.

Mant, C., Hodgson, S., Hobday, R., D'Arrigo, C., and Cason, J. (2004b). A viral aetiology for breast cancer: time to re-examine the postulate. *Intervirology*, 47(1):2–13.

Manzin, A., Solforosi, L., Petrelli, E., Macarri, G., Tosone, G., Piazza, M., and Clementi, M. (1998). Evolution of hypervariable region 1 of hepatitis C virus in primary infection. *J Virol*, 72(7):6271–6276.

Mao, Q., Ray, S. C., Laeyendecker, O., Ticehurst, J. R., Strathdee, S. A., Vlahov, D., and Thomas, D. L. (2001). Human immunodeficiency virus seroconversion and evolution of the hepatitis C virus quasispecies. *J Virol*, 75(7):3259–3267.

Marasco, W. A. and Sui, J. (2007). The growth and potential of human antiviral monoclonal antibody therapeutics. *Nat Biotechnol*, 25(12):1421–1434.

Margeridon-Thermet, S. and Shafer, R. W. (2010). Comparison of the Mechanisms of Drug Resistance among HIV, Hepatitis B, and Hepatitis C. *Viruses*, 2(12):2696–2739.

Margeridon-Thermet, S. S., Shulman, N. S. N., Ahmed, A. A., Shahriar, R. R., Liu, T. T., Wang, C. C., Holmes, S. P. S., Babrzadeh, F. F., Gharizadeh, B. B., Hanczaruk, B. B., Simen, B. B. B., Egholm, M. M., and Shafer, R. W. R. (2009). Ultra-deep pyrosequencing of hepatitis B virus quasispecies from nucleoside and nucleotide reverse-transcriptase inhibitor (NRTI)-treated patients and NRTI-naive patients. *J Infect Dis*, 199(9):1275–1285.

Margulies, M. M., Egholm, M. M., Altman, W. E. W., Attiya, S. S., Bader, J. S. J., Bemben, L. A. L., Berka, J. J., Braverman, M. S. M., Chen, Y.-J. Y., Chen, Z. Z., Dewell, S. B. S., Du, L. L., Fierro, J. M. J., Gomes, X. V. X., Godwin, B. C. B., He, W. W., Helgesen, S. S., Ho, C. H. C., Ho, C. H. C., Irzyk, G. P. G., Jando, S. C. S., Alenquer, M. L. I. M., Jarvie, T. P. T., Jirage, K. B. K., Kim, J.-B. J., Knight, J. R. J., Lanza, J. R. J., Leamon, J. H. J., Lefkowitz, S. M. S., Lei, M. M., Li, J. J., Lohman, K. L. K., Lu, H. H., Makhijani, V. B. V., McDade, K. E. K., McKenna, M. P. M., Myers, E. W. E., Nickerson, E. E., Nobile, J. R. J., Plant, R. R., Puc, B. P. B., Ronan, M. T. M., Roth, G. T. G., Sarkis, G. J. G., Simons, J. F. J., Simpson, J. W. J., Srinivasan, M. M., Tartaro, K. R. K., Tomasz, A. A., Vogt, K. A. K., Volkmer, G. A. G., Wang, S. H. S., Wang, Y. Y., Weiner, M. P. M., Yu, P. P., Begley, R. F. R., and Rothberg, J. M. J. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *CORD Conference Proceedings*, 437(7057):376–380.

Maris, J. M., Hogarty, M. D., Bagatell, R., and Cohn, S. L. (2007). Neuroblastoma. *The Lancet*, 369(9579):2106–2120.

Markland, W., McQuaid, T. J., Jain, J., and Kwong, A. D. (2000). Broad-spectrum antiviral activity of the IMP dehydrogenase inhibitor VX-497: a comparison with ribavirin and demonstration of antiviral additivity with alpha interferon. *Antimicrob Agents Chemother*, 44(4):859–866.

Marra, M. A. (2003). The Genome Sequence of the SARS-Associated Coronavirus. *Science*, 300(5624):1399–1404.

Marsh, M. and Helenius, A. (2006). Virus entry: open sesame. *Cell*, 124(4):729–740.

Marsh, S. G. E. (2013). Nomenclature for factors of the HLA system, update June 2013. *Int. J. Immunogenet.*, 40(5):434–437.

Martell, M., Esteban, J. I., Quer, J., Genescà, J., Weiner, A., Esteban, R., Guardia, J., and Gómez, J. (1992). Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution. *J Virol*, 66(5):3225–3229.

Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annual review of biophysics and biomolecular structure*, 29:291–325.

Martin, G. S. (2004). The road to Src. *Oncogene*, 23(48):7910–7917.

Martinez-Picado, J. and Martínez, M. A. (2008). HIV-1 reverse transcriptase inhibitor resistance mutations and fitness: a view from the clinic and ex vivo. *Virus Res*, 134(1-2):104–123.

Más, A., López-Galíndez, C., Cacho, I., Gómez, J., and Martínez, M. A. (2010). Unfinished Stories on Viral Quasispecies and Darwinian Views of Evolution. *J Mol Biol*, 397(4):865–877.

Más, A., Ulloa, E., Bruguera, M., Furcić, I., Garriga, D., Fábregas, S., Andreu, D., Saiz, J. C., and Díez, J. (2004). Hepatitis C virus population analysis of a single-source nosocomial outbreak reveals an inverse correlation between viral load and quasispecies complexity. *J Gen Virol*, 85(Pt 12):3619–3626.

Mascolini, M., Richman, D., Larder, B., Mellors, J., and Boucher, C. A. B. (2008). Clinical implications of resistance to antiretrovirals: new resistance technologies and interpretations. In *Antivir. Ther. (Lond.)*, pages 319–334. VA San Diego Healthcare System, San Diego, USA.

Maslov, S. and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *arXiv*, (5569):910–913.

Masucci, M. G. (2004). Epstein-Barr virus oncogenesis and the ubiquitin-proteasome system. *Oncogene*, 23(11):2107–2115.

Mathers, C. D. and Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med*, 3(11):e442.

Matsen, F. A., Kodner, R. B., and Armbrust, E. V. (2010). pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, page 538.

Matsuoka, M. and Jeang, K.-T. (2007). Human T-cell leukaemia virus type 1 (HTLV-1) infectivity and cellular transformation. *Nat Rev Cancer*, 7(4):270–280.

Matta, H., Sun, Q., Moses, G., and Chaudhary, P. M. (2003). Molecular genetic analysis of human herpes virus 8-encoded viral FLICE inhibitory protein-induced NF-kappaB activation. *J Biol Chem*, 278(52):52406–52411.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., and D'Eustachio, P. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res*, 37(Database issue):D619–22.

Matthews, L., Vaglio, P., Reboul, J., Ge, H., Davis, B., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res*, 11(12):2120–2126.

Mayer, M. P. (2004). Recruitment of Hsp70 chaperones: a crucial part of viral survival strategies. *Rev Physiol Biochem Pharmacol*, 153:1–46.

Mayr, E. (1997). The objects of selection. *Proc Natl Acad Sci USA*, 94(6):2091–2094.

McCloskey, M. L., Stöger, R., Hansen, R. S., and Laird, C. D. (2007). Encoding PCR products with batch-stamps and barcodes. *Biochem. Genet.*, 45(11-12):761–767.

McCormick, J. B., King, I. J., Webb, P. A., Scribner, C. L., Craven, R. B., Johnson, K. M., Elliott, L. H., and Belmont-Williams, R. (1986). Lassa fever. Effective therapy with ribavirin. *New England Journal of Medicine*, 314(1):20–26.

McCown, M. F., Rajyaguru, S., Le Pogam, S., Ali, S., Jiang, W.-R., Kang, H., Symons, J., Cammack, N., and Najera, I. (2008). The hepatitis C virus replicon presents a higher barrier to resistance to nucleoside analogs than to nonnucleoside polymerase or protease inhibitors. *Antimicrob Agents Chemother*, 52(5):1604–1612.

McCraith, S., Holtzman, T., Moss, B., and Fields, S. (2000). Genome-wide analysis of vaccinia virus protein-protein interactions. *Proc Natl Acad Sci USA*, 97(9):4879–4884.

McEwen, S. A. and Fedorka-Cray, P. J. (2002). Antimicrobial use and resistance in animals. *Clin Infect Dis*, 34 Suppl 3:S93–S106.

McFadden, G., Mohamed, M. R., Rahman, M. M., and Bartee, E. (2009). Cytokine determinants of viral tropism. *Nat Rev Immunol*, 9(9):645–655.

McGeoch, D. J. and Gatherer, D. (2005). Integrating reptilian herpesviruses into the family herpesviridae. *J Virol*, 79(2):725–731.

McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P., and Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods*, 4(1):63–72.

McHardy, A. C. and Rigoutsos, I. (2007). What's in the mix: phylogenetic classification of metagenome sequence samples. *Curr Opin Microbiol*, 10(5):499–503.

McHutchison, J. G. and Dev, A. T. (2004). Future trends in managing hepatitis C. *Gastroenterol. Clin. North Am.*, 33(1 Suppl):S51–61.

McHutchison, J. G., Gordon, S. C., Schiff, E. R., Shiffman, M. L., Lee, W. M., Rustgi, V. K., Goodman, Z. D., Ling, M. H., Cort, S., and Albrecht, J. K. (1998). Interferon alfa-2b alone or in combination with ribavirin as initial treatment for chronic hepatitis C. Hepatitis Interventional Therapy Group. *New England Journal of Medicine*, 339(21):1485–1492.

McHutchison, J. G. and Patel, K. (2002). Future therapy of hepatitis C. *Hepatology (Baltimore, Md)*, 36(5 Suppl 1):S245–52.

McHutchison, J. G., Shiffman, M. L., Cheung, R. C., Gordon, S. C., Wright, T. L., Pottage, J. C., McNair, L., Ette, E., Moseley, S., and Alam, J. (2005). A randomized, double-blind, placebo-controlled dose-escalation trial of merimepodib (VX-497) and interferon-alpha in previously untreated patients with chronic hepatitis C. *Antivir Ther (Lond)*, 10(5):635–643.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20(9):1297–1303.

McLaughlin-Drubin, M. E. and Munger, K. (2008). Viruses associated with human cancer. *Biochim. Biophys. Acta*, 1782(3):127–150.

McMullan, L. K., Frace, M., Sammons, S. A., Shoemaker, T., Balinandi, S., Wamala, J. F., Lutwama, J. J., Downing, R. G., Stroeher, U., MacNeil, A., and Nichol, S. T. (2012). Using next generation sequencing to identify yellow fever virus in Uganda. *Virology*, 422(1):1–5.

Meacham, F., Boffelli, D., Dhahbi, J., Martin, D. I. K., Singer, M., and Pachter, L. (2011). Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, 12:451.

Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. *Curr Opin Genet Dev*, 15(6):589–594.

Mehta, S. H., Cox, A., Hoover, D. R., Wang, X.-H., Mao, Q., Ray, S., Strathdee, S. A., Vlahov, D., and Thomas, D. L. (2002). Protection against persistence of hepatitis C. *Lancet*, 359(9316):1478–1483.

Mellmann, A., Harmsen, D., Cummings, C. A., Zentz, E. B., Leopold, S. R., Rico, A., Prior, K., Szczepanowski, R., Ji, Y., Zhang, W., McLaughlin, S. F., Henkhaus, J. K., Leopold, B., Bielaszewska, M., Prager, R., Brzoska, P. M., Moore, R. L., Guenther, S., Rothberg, J. M., and Karch, H. (2011). Prospective Genomic Characterization of the German Enterohemorrhagic Escherichia coli O104:H4 Outbreak by Rapid Next Generation Sequencing Technology. *PLoS ONE*, 6(7):e22751.

Mellors, J. W., Rinaldo, C. R., Gupta, P., White, R. M., Todd, J. A., and Kingsley, L. A. (1996). Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science*, 272(5265):1167–1170.

Melnick, J. L., Crowther, D., and Barrera-Ojo, J. (1961). Rapid development of drug-resistant mutants of poliovirus. *Science*, 134(3478):557.

Menegaux, F., Olshan, A. F., Neglia, J. P., Pollock, B. H., and Bondy, M. L. (2004). Day care, childhood infections, and risk of neuroblastoma. *Am J Epidemiol*, 159(9):843–851.

Menéndez-Arias, L. and Gago, F. (2012). Antiviral agents: structural basis of action and rational design. *Subcell Biochem*, 68:599–630.

Merlo, L. M. F., Pepper, J. W., Reid, B. J., and Maley, C. C. (2006). Cancer as an evolutionary and ecological process. *Nat Rev Cancer*, 6(12):924–935.

Messud-Petit, F. F., Gelfi, J. J., Delverdier, M. M., Amardeilh, M. F. M., Py, R. R., Sutter, G. G., and Bertagnoli, S. S. (1998). Serp2, an inhibitor of the interleukin-1beta-converting enzyme, is critical in the pathobiology of myxoma virus. *J Virol*, 72(10):7830–7839.

Metzenberg, S. (1990). Levels of Epstein-Barr virus DNA in lymphoblastoid cell lines are correlated with frequencies of spontaneous lytic growth but not with levels of expression of EBNA-1, EBNA-2, or latent membrane protein. *J Virol*, 64(1):437–444.

Metzner, K. J., Giulieri, S. G., Knoepfel, S. A., Rauch, P., Burgisser, P., Yerly, S., Günthard, H. F., and Cavassini, M. (2009). Minority quasispecies of drug-resistant HIV-1 that lead to early therapy failure in treatment-naive and -adherent patients. *Clin Infect Dis*, 48(2):239–247.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., and Edwards, R. A. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9:386.

Meyerhans, A., Vartanian, J. P., and Wain-Hobson, S. (1990). DNA recombination during PCR. *Nucleic Acids Res*, 18(7):1687–1691.

Meyniel-Schicklin, L., de Chassey, B., André, P., and Lotteau, V. (2012). Viruses and interactomes in translation. *Mol Cell Proteomics*, 11(7):M111.014738.

Mi, S., Lee, X., Li, X., Veldman, G. M., Finnerty, H., Racie, L., LaVallie, E., Tang, X. Y., Edouard, P., Howes, S., Keith, J. C., and McCoy, J. M. (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, 403(6771):785–789.

Mibayashi, M., Martínez-Sobrido, L., Loo, Y.-M., Cárdenas, W. B., Gale, M., and García-Sastre, A. (2007). Inhibition of retinoic acid-inducible gene I-mediated induction of beta interferon by the NS1 protein of influenza A virus. *J Virol*, 81(2):514–524.

Michallet, M.-C., Meylan, E., Ermolaeva, M. A., Vazquez, J., Rebsamen, M., Curran, J., Poeck, H., Bscheider, M., Hartmann, G., König, M., Kalinke, U., Pasparakis, M., and Tschopp, J. (2008). TRADD protein is an essential component of the RIG-like helicase antiviral pathway. *Immunity*, 28(5):651–661.

Mihm, U., Welker, M.-W., Teuber, G., Wedemeyer, H., Berg, T., Sarrazin, C., Böhm, S., Alshuth, U., Herrmann, E., and Zeuzem, S. (2013). Impact of ribavirin priming on viral kinetics and treatment response in chronic hepatitis C genotype 1 infection. *J Viral Hepat*, pages n/a–n/a.

Mikkers, H. and Berns, A. (2003). Retroviral insertional mutagenesis: tagging cancer pathways. *Adv Cancer Res*, 88:53–99.

Miller, G., Rigsby, M. O., Heston, L., Grogan, E., Sun, R., Metroka, C., Levy, J. A., Gao, S. J., Chang, Y., and Moore, P. (1996). Antibodies to butyrate-inducible antigens of Kaposi's sarcoma-associated herpesvirus in patients with HIV-1 infection. *New England Journal of Medicine*, 334(20):1292–1297.

Miller, R. L., Meng, T.-C., and Tomai, M. A. (2008). The antiviral activity of Toll-like receptor 7 and 7/8 agonists. *Drug News Perspect.*, 21(2):69–87.

Mills, C. E., Robins, J. M., and Lipsitch, M. (2004). Transmissibility of 1918 pandemic influenza. *Nature*, 432(7019):904–906.

Mills, D. R., Peterson, R. L., and Spiegelman, S. (1967). An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proc Natl Acad Sci USA*, 58(1):217–224.

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., and Devine, S. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res*, 16(9):1182–1190.

Miner, B. E., Stöger, R. J., Burden, A. F., Laird, C. D., and Hansen, R. S. (2004). Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. *Nucleic Acids Res*, 32(17):e135.

Minoche, A. E., Dohm, J. C., and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol*, 12(11):R112.

Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D., and Bushman, F. D. (2012). Hypervariable loci in the human gut virome. *Proc Natl Acad Sci USA*, 109(10):3962–3966.

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S. A., Wu, G. D., Lewis, J. D., and Bushman, F. D. (2011). The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Res*, 21(10):1616–1625.

Mira, A., Ochman, H., and Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends Genet*, 17(10):589–596.

Mirarab, S. (2012). SEPP: SATe-Enabled phylogenetic placement. *Pac. Symp. Biocomput.*, 17:247–258.

Misale, S., Yaeger, R., Hobor, S., Scala, E., Janakiraman, M., Liska, D., Valtorta, E., Schiavo, R., Buscarino, M., Siravegna, G., Bencardino, K., Cercek, A., Chen, C.-T., Veronese, S., Zanon, C., Sartore-Bianchi, A., Gambacorta, M., Gallicchio, M., Vakiani, E., Boscaro, V., Medico, E., Weiser, M., Siena, S., Di Nicolantonio, F., Solit, D., and Bardelli, A. (2012). Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature*, 486(7404):532–536.

Missale, G., Bertoni, R., Lamonaca, V., Valli, A., Massari, M., Mori, C., Rumi, M. G., Houghton, M., Fiaccadori, F., and Ferrari, C. (1996). Different clinical behaviors of acute hepatitis C virus infection are associated with different vigor of the anti-viral cell-mediated immune response. *J Clin Invest*, 98(3):706–714.

Mitsuya, Y. Y., Varghese, V. V., Wang, C. C., Liu, T. F. T., Holmes, S. P. S., Jayakumar, P. P., Gharizadeh, B. B., Ronaghi, M. M., Klein, D. D., Fessel, W. J. W., and Shafer, R. W. R. (2008). Minority human immunodeficiency virus type 1 variants in antiretroviral-naive persons with reverse transcriptase codon 215 revertant mutations. *J Virol*, 82(21):10747–10755.

Mokili, J. L., Rohwer, F., and Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. *Curr Opin Vir*, 2(1):63–77.

Mokili, J. L. K., Rogers, M., Carr, J. K., Simmonds, P., Bopopi, J. M., Foley, B. T., Korber, B. T., Birx, D. L., and McCutchan, F. E. (2002). Identification of a novel clade of human immunodeficiency virus type 1 in Democratic Republic of Congo. *AIDS Res Hum Retroviruses*, 18(11):817–823.

Moloney, J. B. (1960). Biological studies on a lymphoid-leukemia virus extracted from sarcoma 37. I. Origin and introductory investigations. *J. Natl. Cancer Inst.*, 24:933–951.

Mondelli, M. U., Cerino, A., Meola, A., and Nicosia, A. (2003). Variability or conservation of hepatitis C virus hypervariable region 1? Implications for immune responses. *J. Biosci.*, 28(3):305–310.

Monzoorul Haque, M., Ghosh, T. S., Komanduri, D., and Mande, S. S. (2009). SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics*, 25(14):1722–1730.

Moore, D. H., Sarkar, N. H., Kelly, C. E., Pillsbury, N., and Charney, J. (1969). Type B particles in human milk. *Tex. Rep. Biol. Med.*, 27(4):1027–1039.

Moore, P. S. and Chang, Y. (1995). Detection of herpesvirus-like DNA sequences in Kaposi's sarcoma in patients with and without HIV infection. *New England Journal of Medicine*, 332(18):1181–1185.

Moore, P. S. and Chang, Y. (2003). Kaposi's sarcoma-associated herpesvirus immunoevasion and tumorigenesis: two sides of the same coin? *Annu Rev Microbiol*, 57:609–639.

Moore, P. S. and Chang, Y. (2010). Why do viruses cause cancer? Highlights of the first century of human tumour virology. *Nat Rev Cancer*, 10(12):878–889.

Moore, R. A., Warren, R. L., Freeman, J. D., Gustavsen, J. A., Chénard, C., Friedman, J. M., Suttle, C. A., Zhao, Y., and Holt, R. A. (2011). The sensitivity of massively parallel sequencing for detecting candidate infectious agents associated with human tissue. *PLoS ONE*, 6(5):e19838.

Moreira, D. (2000). Multiple independent horizontal transfers of informational genes from bacteria to plasmids and phages: implications for the origin of bacterial replication machinery. *Mol. Microbiol.*, 35(1):1–5.

Moreira, D. and López-García, P. (2009). Ten reasons to exclude viruses from the tree of life. *Nat Rev Microbiol*, 7(4):306–311.

Morello, J., Cuenca, L., Soriano, V., Medrano, J., Madejon, A., Vispo, E., Barreiro, P., Labarga, P., Jiménez-Nácher, I., and Rodríguez-Nóvoa, S. (2010). Influence of a single nucleotide polymorphism at the main ribavirin transporter gene on the rapid virological response to pegylated interferon-ribavirin therapy in patients with chronic hepatitis C virus infection. *J Infect Dis*, 202(8):1185–1191.

Moreno, H., Gallego, I., Sevilla, N., de la Torre, J. C., Domingo, E., and Martín, V. (2011). Ribavirin can be mutagenic for arenaviruses. *J Virol*, 85(14):7246–7255.

Moreno, H., Grande-Pérez, A., Domingo, E., and Martín, V. (2012). Arenaviruses and lethal mutagenesis. Prospects for new ribavirin-based interventions. *Viruses*, 4(11):2786–2805.

Morgan, J. L., Darling, A. E., and Eisen, J. A. (2010). Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS ONE*, 5(4):e10209.

Mori, K., Ikeda, M., Ariumi, Y., Dansako, H., Wakita, T., and Kato, N. (2011). Mechanism of action of ribavirin in a novel hepatitis C virus replication cell system. *Virus Res*, 157(1):61–70.

Morice, Y., Ratinier, M., Miladi, A., Chevaliez, S., Germanidis, G., Wedemeyer, H., Laperche, S., Lavergne, J.-P., and Pawlotsky, J.-M. (2009). Seroconversion to hepatitis C virus alternate reading frame protein during acute infection. *Hepatology (Baltimore, Md)*, 49(5):1449–1459.

Morishima, C., Polyak, S. J., Ray, R., Doherty, M. C., Di Bisceglie, A. M., Malet, P. F., Bonkovsky, H. L., Sullivan, D. G., Gretch, D. R., Rothman, A. L., Koziel, M. J., Lindsay, K. L., and Hepatitis C Antiviral Long-Term Treatment Against Cirrhosis Trial Group (2006). Hepatitis C virus-specific immune responses and quasi-species variability at baseline are associated with nonresponse to antiviral therapy during advanced hepatitis C. *J Infect Dis*, 193(7):931–940.

Morse, S. S. (1992). Evolving views of viral evolution: towards an evolutionary biology of viruses. *Hist Philos Life Sci*, 14(2):215–248.

Morse, S. S. (1995). *Factors in the Emergence of Infectious Diseases*. National Emergency Training Center.

Morse, S. S. (2004). Factors and determinants of disease emergence. *Rev. - Off. Int. Epizoot.*, 23(2):443–451.

Morse, S. S., Mazet, J. A. K., Woolhouse, M., Parrish, C. R., Carroll, D., Karesh, W. B., Zambrana-Torrelio, C., Lipkin, W. I., and Daszak, P. (2012). Prediction and prevention of the next pandemic zoonosis. *Lancet*, 380(9857):1956–1965.

Mosca, R., Pons, C., Fernández-Recio, J., and Aloy, P. (2009). Pushing structural information into the yeast interactome by high-throughput protein docking experiments. *PLoS Comput Biol*, 5(8):e1000490.

Mosier, D. E. (2009). How HIV changes its tropism: evolution and adaptation? *Curr Opin HIV AIDS*, 4(2):125–130.

Moya, A., Elena, S. F., Bracho, A., Miralles, R., and Barrio, E. (2000). The evolution of RNA viruses: A population genetics view. *Proc Natl Acad Sci USA*, 97(13):6967–6973.

Moyer, S. E., Lewis, P. W., and Botchan, M. R. (2006). Isolation of the Cdc45/Mcm2-7/GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase. *Proc Natl Acad Sci USA*, 103(27):10236–10241.

Muchowski, P. J. and Wacker, J. L. (2005). Modulation of neurodegeneration by molecular chaperones. *Nat. Rev. Neurosci.*, 6(1):11–22.

Mueller, Y. M., Do, D. H., Altork, S. R., Artlett, C. M., Gracely, E. J., Katsetos, C. D., Legido, A., Villinger, F., Altman, J. D., Brown, C. R., Lewis, M. G., and Katsikis, P. D. (2008). IL-15 treatment during acute simian immunodeficiency virus (SIV) infection increases viral set point and accelerates disease progression despite the induction of stronger SIV-specific CD8+ T cell responses. *J. Immunol.*, 180(1):350–360.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., Courcelle, E., Das, U., Daugherty, L., Dibley, M., Finn, R., Fleischmann, W., Gough, J., Haft, D., Hulo, N., Hunter, S., Kahn, D., Kanapin, A., Kejariwal, A., Labarga, A., Langendijk-Genevaux, P. S., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Nikolskaya, A. N., Orchard, S., Orengo, C., Petryszak, R., Selengut, J. D., Sigrist, C. J. A., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C. (2007). New developments in the InterPro database. *Nucleic Acids Res*, 35(Database):D224–D228.

Muller, H. J. (1929). 4th International Congress of Plant Science. pages 917–918.

Müller, W. E., Maidhof, A., Taschner, H., and Zahn, R. K. (1977). Virazole (1-beta-D-ribofuranosyl-1,2,4-triazole-3-carboxamide; a cytostatic agent. *Biochem. Pharmacol.*, 26(11):1071–1075.

Mullins, J. I., Heath, L., Hughes, J. P., Kicha, J., Styrchak, S., Wong, K. G., Rao, U., Hansen, A., Harris, K. S., Laurent, J.-P., Li, D., Simpson, J. H., Essigmann, J. M., Loeb, L. A., and Parkins, J. (2011). Mutation of HIV-1 genomes in a clinical population treated with the mutagenic nucleoside KP1461. *PLoS ONE*, 6(1):e15135.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, 51 Pt 1:263–273.

Münter, S., Way, M., and Frischknecht, F. (2006). Signaling during pathogen infection. *Sci. STKE*, 2006(335):re5.

Münz, C., Lünemann, J. D., Getts, M. T., and Miller, S. D. (2009). Antiviral immune responses: triggers of or triggered by autoimmunity? *Nat Rev Immunol*, 9(4):246–258.

Murali, T. M., Dyer, M. D., Badger, D., Tyler, B. M., and Katze, M. G. (2011). Network-based prediction and analysis of HIV dependency factors. *PLoS Comput Biol*, 7(9):e1002164.

Murray, A. G. (1992). Viral dynamics: A model of the effects size, shape, motion and abundance of single-celled planktonic organisms and other particles. *Marine ecology progress series Oldendorf*, (89):103–116.

Murray, C. J. and Lopez, A. D. (1997a). Alternative projections of mortality and disability by cause 1990-2020: Global Burden of Disease Study. *Lancet*, 349(9064):1498–1504.

Murray, C. J. and Lopez, A. D. (1997b). Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study. *Lancet*, 349(9063):1436–1442.

Murray, C. J. and Lopez, A. D. (1997c). Mortality by cause for eight regions of the world: Global Burden of Disease Study. *Lancet*, 349(9061):1269–1276.

Mushegian, A. R. and Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci USA*, 93(19):10268–10273.

Musso, F. (2012). On the relation between the Eigen model and the asexual Wright-Fisher model. *Bull. Math. Biol.*, 74(1):103–115.

Myers, E. W. (2005). The fragment assembly string graph. *Bioinformatics*, 21 Suppl 2:ii79–85.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., Anson, E. L., Bolanos, R. A., Chou, H. H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., Dunn, P. J., Lai, Z., Liang, Y., Nusskern, D. R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G. M., Adams, M. D., and Venter, J. C. (2000). A whole-genome assembly of Drosophila. *Science*, 287(5461):2196–2204.

Mykhalovskiy, E. and Weir, L. (2006). The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health. *Can. J. Public Health*, 97:42–44 PB –.

Nabirochkin, S. D., Gabitova, L., Ossokina, M. A., Soldatov, A. V., Gazaryan, T. G., and Gazaryan, K. G. (1998). Oncoviral DNAs induce transposition of endogenous mobile elements in the genome of Drosophila melanogaster. *Mutat. Res.*, 403(1-2):127–136.

Nagarajan, N. and Pop, M. (2013). Sequence assembly demystified. *Nat Rev Genet*, 14(3):157–167.

Nagata, S. (1997). Apoptosis by Death Factor. *Cell*, 88(3):11–11.

Naggie, S., Patel, K., and McHutchison, J. (2010). Hepatitis C virus directly acting antivirals: current developments with NS3/4A HCV serine protease inhibitors. *J. Antimicrob. Chemother.*, 65(10):2063–2069.

Nájera, I., Holguín, A., Quiñones-Mateu, M. E., Muñoz-Fernández, M. A., Nájera, R., López-Galíndez, C., and Domingo, E. (1995). Pol gene quasispecies of human immunodeficiency virus: mutations associated with drug resistance in virus from patients undergoing no drug therapy. *J Virol*, 69(1):23–31.

Nájera, I., Richman, D. D., Olivares, I., Rojas, J. M., Peinado, M. A., Perucho, M., Nájera, R., and López-Galíndez, C. (1994). Natural occurrence of drug resistance mutations in the reverse transcriptase of human immunodeficiency virus type 1 isolates. *AIDS Res Hum Retroviruses*, 10(11):1479–1488.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M., Ogasawara, N., and Kanaya, S. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res*, 39(13):e90.

Nakamura, M., Saito, H., Ikeda, M., Tada, S., Kumagai, N., Kato, N., Shimotohno, K., and Hibi, T. (2008). Possible molecular mechanism of the relationship between NS5B polymorphisms and early clearance of hepatitis C virus during interferon plus ribavirin treatment. *J Med Virol*, 80(4):632–639.

Nakamura, S., Yang, C.-S., Sakon, N., Ueda, M., Tougan, T., Yamashita, A., Goto, N., Takahashi, K., Yasunaga, T., Ikuta, K., Mizutani, T., Okamoto, Y., Tagami, M., Morita, R., Maeda, N., Kawai, J., Hayashizaki, Y., Nagai, Y., Horii, T., Iida, T., and Nakaya, T. (2009). Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS ONE*, 4(1):e4219.

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res*, 40(20):e155.

National Institutes of Health (2002). NIH Consensus Statement on Management of Hepatitis C: 2002. In *NIH Consens State Sci Statements*, pages 1–46.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W. R., Hicks, J., and Wigler, M. (2011). Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94.

Navratil, V., de Chassey, B., Combe, C. R., and Lotteau, V. (2011). When the human viral infectome and diseasome networks collide: towards a systems biology platform for the aetiology of human diseases. *BMC systems biology*, 5(1):13.

Navratil, V., de Chassey, B., Meyniel, L., Delmotte, S., Gautier, C., André, P., Lotteau, V., and Rabourdin-Combe, C. (2009). VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. *Nucleic Acids Res*, 37(Database issue):D661–D668.

Navratil, V., de Chassey, B., Meyniel, L., Pradezynski, F., André, P., Rabourdin-Combe, C., and Lotteau, V. (2010). System-level comparison of protein-protein interactions between viruses and the human type I interferon system network. *J Proteome Res*, 9(7):3527–3536.

Neckers, L. and Tatu, U. (2008). Molecular chaperones in pathogen virulence: emerging new targets for therapy. *Cell Host Microbe*, 4(6):519–527.

Neel, B. G., Hayward, W. S., Robinson, H. L., Fang, J., and Astrin, S. M. (1981). Avian leukosis virus-induced tumors have common proviral integration sites and synthesize discrete new RNAs: oncogenesis by promoter insertion. *Cell*, 23(2):323–334.

Negrini, M., Ferracin, M., Sabbioni, S., and Croce, C. M. (2007). MicroRNAs in human cancer: from research to therapy. *J. Cell. Sci.*, 120(Pt 11):1833–1840.

Nesvizhskii, A. I. (2012). Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments. *Proteomics*, 12(10):1639–1655.

Neumann, A. U., Lam, N. P., Dahari, H., Gretch, D. R., Wiley, T. E., Layden, T. J., and Perelson, A. S. (1998). Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-alpha therapy. *Science*, 282(5386):103–107.

Nevins, J. R. and Vogt, N. J. R. (1996). Cell transformation by viruses. In Fields, B. N., M, K. D., and M, H. P., editors, *Field's Virology*, pages 2111–2148. Lippincott, Philadelphia.

Ng, S., Zhang, Z., Tan, S., and Lin, K. (2003). InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res*, 31(1):251–254.

Ni, M. and Lee, A. S. (2007). ER chaperones in mammalian development and human diseases. *FEBS Lett*, 581(19):3641–3651.

Nicolson, G. L. G. (1987). Tumor cell instability, diversification, and progression to the metastatic phenotype: from oncogene to oncofetal expression. *Cancer Research*, 47(6):1473–1487.

Nijhuis, M., Schuurman, R., de Jong, D., Erickson, J., Gustchina, E., Albert, J., Schipper, P., Gulnik, S., and Boucher, C. A. (1999). Increased fitness of drug resistant HIV-1 protease as a result of acquisition of compensatory mutations during suboptimal therapy. *AIDS*, 13(17):2349–2359.

Ning, Z., Cox, A. J., and Mullikin, J. C. (2001). SSAHA: a fast search method for large DNA databases. *Genome Res*, 11(10):1725–1729.

Nishizawa, T., Okamoto, H., Konishi, K., Yoshizawa, H., Miyakawa, Y., and Mayumi, M. (1997). A novel DNA virus (TTV) associated with elevated transaminase levels in post-transfusion hepatitis of unknown etiology. *Biochem. Biophys. Res. Commun.*, 241(1):92–97.

Niu, B., Fu, L., Sun, S., and Li, W. (2010). Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics*, 11:187.

Noguchi, H., Park, J., and Takagi, T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res*, 34(19):5623–5630.

Nora Dickson, M., Tsinberg, P., Tang, Z., Bischoff, F. Z., Wilson, T., and Leonard, E. F. (2011). Efficient capture of circulating tumor cells with a novel immunocytochemical microfluidic device. *Biomicrofluidics*, 5(3):34119–3411915.

Novella, I. S. (2003). Contributions of vesicular stomatitis virus to the understanding of RNA virus evolution. *Curr Opin Microbiol*, 6(4):399–405.

Novella, I. S., Borrego, B., Mateu, M. G., Domingo, E., Giralt, E., and Andreu, D. (1993). Use of substituted and tandem-repeated peptides to probe the relevance of the highly conserved RGD tripeptide in the immune response against foot-and-mouth disease virus. *FEBS Lett*, 330(3):253–259.

Novella, I. S., Duarte, E. A., Elena, S. F., Moya, A., Domingo, E., and Holland, J. J. (1995). Exponential increases of RNA virus fitness during large population transmissions. *Proc Natl Acad Sci USA*, 92(13):5841–5844.

Novella, I. S., Dutta, R. N., and Wilke, C. O. (2008). A linear relationship between fitness and the logarithm of the critical bottleneck size in vesicular stomatitis virus populations. *J Virol*, 82(24):12589–12590.

Novick, D., Cohen, B., and Rubinstein, M. (1992). Soluble interferon-alpha receptor molecules are present in body fluids. *FEBS Lett*, 314(3):445–448.

Nowak, M. A. M. (1992). What is a quasispecies? *Trends Ecol. Evol. (Amst.)*, 7(4):118–121.

Nowak, M. M. and Schuster, P. P. (1989). Error thresholds of replication in finite populations mutation frequencies and the onset of Muller's ratchet. *J. Theor. Biol.*, 137(4):375–395.

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28.

Nyborg, J. K., Egan, D., and Sharma, N. (2010). The HTLV-1 Tax protein: revealing mechanisms of transcriptional activation through histone acetylation and nucleosome disassembly. *Biochim. Biophys. Acta*, 1799(3-4):266–274.

Nyfeler, B., Michnick, S. W., and Hauri, H.-P. (2005). Capturing protein interactions in the secretory pathway of living cells. *Proc Natl Acad Sci USA*, 102(18):6350–6355.

Ochoa, G. G. (2005). Error thresholds in genetic algorithms. *Evol Comput*, 14(2):157–182.

Odeberg, J., Yun, Z., Sönnerborg, A., Bjøro, K., Uhlén, M., and Lundeberg, J. (1997). Variation of hepatitis C virus hypervariable region 1 in immunocompromised patients. *J Infect Dis*, 175(4):938–943.

Ogert, R. A., Hou, Y., Ba, L., Wojcik, L., Qiu, P., Murgolo, N., Duca, J., Dunkle, L. M., Ralston, R., and Howe, J. A. (2010). Clinical resistance to vicriviroc through adaptive V3 loop mutations in HIV-1 subtype D gp120 that alter interactions with the N-terminus and ECL2 of CCR5. *Virology*, 400(1):11–11.

Ojosnegros, S., Agudo, R., Sierra, M., Briones, C., Sierra, S., González-López, C., Domingo, E., and Cristina, J. (2008). Topology of evolving, mutagenized viral populations: quasispecies expansion, compression, and operation of negative selection. *BMC Evol Biol*, 8:207.

Ojosnegros, S., Perales, C., Más, A., and Domingo, E. (2011). Quasispecies as a matter of fact: viruses and beyond. *Virus Res*, 162(1-2):203–215.

Oliver, G. R. (2012). Considerations for clinical read alignment and mutational profiling using next-generation sequencing. *F1000Res*.

O'Neill, L. A. J. (2008). When signaling pathways collide: positive and negative regulation of toll-like receptor signal transduction. *Immunity*, 29(1):12–20.

O'Neill, L. A. J. and Bowie, A. G. (2007). The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. *Nat Rev Immunol*, 7(5):353–364.

Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*, 1(5):376–386.

Ono, M., Yasunaga, T., Miyata, T., and Ushikubo, H. (1986). Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome. *J Virol*, 60(2):589–598.

Oparin, A. I. (1961). *Life: its Nature, Origin and Development*. Academic Press.

Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., Bridge, A., Briganti, L., Brinkman, F. S. L., Brinkman, F., Cesareni, G., Chatr-aryamontri, A., Chautard, E., Chen, C., Dumousseau, M., Goll, J., Hancock, R. E. W., Hancock, R., Hannick, L. I., Jurisica, I., Khadake, J., Lynn, D. J., Mahadevan, U., Perfetto, L., Raghunath, A., Ricard-Blum, S., Roechert, B., Salwinski, L., Stümpflen, V., Tyers, M., Uetz, P., Xenarios, I., and Hermjakob, H. (2012). Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods*, 9(4):345–350.

Orr, H. A. (2009). Fitness and its role in evolutionary genetics. *Nat Rev Genet*, 10(8):531–539.

Orr, H. A. H. (2000). The rate of adaptation in asexuals. *Genetics*, 155(2):961–968.

Ortmann, B., Copeman, J., Lehner, P. J., Sadasivan, B., Herberg, J. A., Grandea, A. G., Riddell, S. R., Tampé, R., Spies, T., Trowsdale, J., and Cresswell, P. (1997). A critical role for tapasin in the assembly and function of multimeric MHC class I-TAP complexes. *Science*, 277(5330):1306–1309.

Otsuka, M., Kato, N., Moriyama, M., Taniguchi, H., Wang, Y., Dharel, N., Kawabe, T., and Omata, M. (2005). Interaction between the HCV NS3 protein and the host TBK1 protein leads to inhibition of cellular antiviral responses. *Hepatology (Baltimore, Md)*, 41(5):1004–1012.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., and Trajanoski, Z. (2013). A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform*.

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, 276(5313):734–740.

Padgett, B. L., Walker, D. L., ZuRhein, G. M., Eckroade, R. J., and Dessel, B. H. (1971). Cultivation of papova-like virus from human brain with progressive multifocal leucoencephalopathy. *Lancet*, 1(7712):1257–1260.

Padian, K. (2008). *Darwin's enduring legacy.*, volume 451. Museum of Paleontology, University of California, Berkeley, USA.

Paeshuyse, J., Dallmeier, K., and Neyts, J. (2011). Ribavirin for the treatment of chronic hepatitis C virus infection: a review of the proposed mechanisms of action. *Curr Opin Vir*, 1(6):590–598.

Pagano, J. S., Blaser, M., Buendia, M. A., Damania, B., Khalili, K., Raab-Traub, N., and Roizman, B. (2004). Infectious agents and cancer: Criteria for a causal relation. *Semin. Cancer Biol.*, 14(6):453–471.

Palacios, G., Druce, J., Du, L., Tran, T., Birch, C., Briese, T., Conlan, S., Quan, P.-L., Hui, J., Marshall, J., Simons, J. F., Egholm, M., Paddock, C. D., Shieh, W.-J., Goldsmith, C. S., Zaki, S. R., Catton, M., and Lipkin, W. I. (2008). A new arenavirus in a cluster of fatal transplant-associated diseases. *N. Engl. J. Med.*, 358(10):991–998.

Palella, F. J., Delaney, K. M., Moorman, A. C., Loveless, M. O., Fuhrer, J., Satten, G. A., Aschman, D. J., and Holmberg, S. D. (1998). Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators. *New England Journal of Medicine*, 338(13):853–860.

Palmer, D. C. and Restifo, N. P. (2009). Suppressors of cytokine signaling (SOCS) in T cell differentiation, maturation, and function. *Trends Immunol.*, 30(12):592–602.

Palmer, S., Kearney, M., Maldarelli, F., Halvas, E. K., Bixby, C. J., Bazmi, H., Rock, D., Falloon, J., Davey, R. T., Dewar, R. L., Metcalf, J. A., Hammer, S., Mellors, J. W., and Coffin, J. M. (2004). Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *J Clin Microbiol*, 43(1):406–413.

Palmer, S., Maldarelli, F., Wiegand, A., Bernstein, B., Hanna, G. J., Brun, S. C., Kempf, D. J., Mellors, J. W., Coffin, J. M., and King, M. S. (2008). Low-level viremia persists for at least 7 years in patients on suppressive antiretroviral therapy. *Proc Natl Acad Sci USA*, 105(10):3879–3884.

Pan, J., Peng, X., Gao, Y., Li, Z., Lu, X., Chen, Y., Ishaq, M., Liu, D., Dediego, M. L., Enjuanes, L., and Guo, D. (2008). Genome-wide analysis of protein-protein interactions and involvement of viral proteins in SARS-CoV replication. *PLoS ONE*, 3(10):e3299.

Pantry, S. N. and Medveczky, P. G. (2009). Epigenetic regulation of Kaposi's sarcoma-associated herpesvirus replication. *Semin. Cancer Biol.*, 19(3):153–157.

Paprotka, T., Delviks-Frankenberry, K. A., Cingöz, O., Martinez, A., Kung, H.-J., Tepper, C. G., Hu, W.-S., Fivash, M. J., Coffin, J. M., and Pathak, V. K. (2011). Recombinant origin of the retrovirus XMRV. *Science*, 333(6038):97–101.

Parada, L. F., Tabin, C. J., Shih, C., and Weinberg, R. A. (1982). Human EJ bladder carcinoma oncogene is homologue of Harvey sarcoma virus ras gene. *Nature*, 297(5866):474–478.

Pardo, M. and Choudhary, J. S. (2012). Assignment of protein interactions from affinity purification/mass spectrometry data. *J Proteome Res*, 11(3):1462–1474.

Parham, P. and Ohta, T. (1996). Population biology of antigen presentation by MHC class I molecules. *Science*, 272(5258):67–74.

Pariente, N., Sierra, S., Lowenstein, P. R., and Domingo, E. (2001). Efficient virus extinction by combinations of a mutagen and antiviral inhibitors. *J Virol*, 75(20):9723–9730.

Pariente, N. N., Airaksinen, A. A., and Domingo, E. E. (2003). Mutagenesis versus inhibition in the efficiency of extinction of foot-and-mouth disease virus. *J Virol*, 77(12):7131–7138.

Park, B., Kim, Y., Shin, J., Lee, S., Cho, K., Früh, K., Lee, S., and Ahn, K. (2004). Human cytomegalovirus inhibits tapasin-dependent peptide loading and optimization of the MHC class I peptide cargo for immune evasion. *Immunity*, 20(1):71–85.

Park, J.-M., Muñoz, E., and Deem, M. W. (2010). Quasispecies theory for finite populations. *Physical review E, Statistical, nonlinear, and soft matter physics*, 81(1 Pt 1):011902.

Parkin, D. M. (2006). The global health burden of infection-associated cancers in the year 2002. *Int. J. Cancer*, 118(12):3030–3044.

Parkin, D. M., Bray, F., Ferlay, J., and Pisani, P. (2005). Global cancer statistics, 2002. *CA Cancer J Clin*, 55(2):74–108.

Parra, J., Portilla, J., Pulido, F., la Rosa, R. S.-d., Alonso-Villaverde, C., Berenguer, J., Blanco, J. L., Domingo, P., Dronda, F., Galera, C., Gutiérrez, F., Kindelán, J. M., Knobel, H., Leal, M., López-Aldeguer, J., Mariño, A., Miralles, C., Moltó, J., Ortega, E., and Oteo, J. A. (2010). Clinical utility of maraviroc. *Clin Drug Investig*, 31(8):527–542.

Parrish, C. R. and Kawaoka, Y. (2005). THE ORIGINS OF NEW PANDEMIC VIRUSES: The Acquisition of New Host Ranges by Canine Parvovirus and Influenza A Viruses. *Annu Rev Microbiol*, 59(1):553–586.

Paszkiewicz, K. and Studholme, D. J. (2010). De novo assembly of short sequence reads. *Brief Bioinform*, 11(5):457–472.

Pathak, V. K. and Temin, H. M. (1990). Broad spectrum of in vivo forward mutations, hypermutations, and mutational hotspots in a retroviral shuttle vector after a single replication cycle: substitutions, frameshifts, and hypermutations. *Proc Natl Acad Sci USA*, 87(16):6019–6023.

Patil, K. R., Haider, P., Pope, P. B., Turnbaugh, P. J., Morrison, M., Scheffer, T., and McHardy, A. C. (2011). Taxonomic metagenome sequence assignment with structured output models. *Nat Methods*, 8(3):191–192.

Patowary, A., Chauhan, R. K., Singh, M., Kv, S., Periwal, V., Kp, K., Sapkal, G. N., Bondre, V. P., Gore, M. M., Sivasubbu, S., and Scaria, V. (2012). De novo identification of viral pathogens from cell culture hologenomes. *BMC Res Notes*, 5:11.

Paul, J. H., Rose, J. B., Jiang, S. C., Kellogg, C. A., and Dickson, L. (1993). Distribution of viral abundance in the reef environment of Key Largo, Florida. *Appl. Environ. Microbiol.*, 59(3):718–724.

Pauwels, R. R. (2006). Aspects of successful drug discovery and development. *Antiviral Res*, 71(2-3):13–13.

Pawlotsky, J.-M. (2002). Use and interpretation of virological tests for hepatitis C. *Hepatology (Baltimore, Md)*, 36(5 Suppl 1):S65–73.

Pawlotsky, J.-M. (2005). Current and future concepts in hepatitis C therapy. *Semin. Liver Dis.*, 25(1):72–83.

Pawlotsky, J.-M., Dahari, H., Neumann, A. U., Hézode, C., Germanidis, G., Lonjon, I., Castera, L., and Dhumeaux, D. (2004). Antiviral action of ribavirin in chronic hepatitis C. *Gastroenterology*, 126(3):703–714.

Pawlotsky, J.-M., Germanidis, G., Neumann, A. U., Pellerin, M., Frainais, P. O., and Dhumeaux, D. (1998). Interferon resistance of hepatitis C virus genotype 1b: relationship to nonstructural 5A gene quasispecies mutations. *J Virol*, 72(4):2795–2805.

Pawlotsky, J. M. J., Bouvier-Alias, M. M., Hezode, C. C., Darthuy, F. F., Remire, J. J., and Dhumeaux, D. D. (2000). Standardization of Hepatitis C Virus RNA Quantification. *Hepatology (Baltimore, Md)*, 32(3):6–6.

Payer, B. and Lee, J. T. (2008). X chromosome dosage compensation: how mammals keep the balance. *Annu. Rev. Genet.*, 42:733–772.

Peaper, D. R. and Cresswell, P. (2008). Regulation of MHC class I assembly and peptide binding. *Annu. Rev. Cell Dev. Biol.*, 24:343–368.

Pearlman, B. L. (2004). Hepatitis C treatment update. *Am. J. Med.*, 117(5):344–352.

Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N. R., Brucker, W., Kumar, V., Kandasamy, J., Keenan, L., Bardarov, S., Kriakov, J., Lawrence, J. G., Jacobs, W. R., Hendrix, R. W., and Hatfull, G. F. (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell*, 113(2):171–182.

Peiris, J. S. M., Poon, L. L. M., and Guan, Y. (2009). Emergence of a novel swine-origin influenza A virus (S-OIV) H1N1 virus in humans. *J Clin Virol*, 45(3):169–173.

Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428.

Perales, C., Agudo, R., Manrubia, S. C., and Domingo, E. (2011a). Influence of mutagenesis and viral load on the sustained low-level replication of an RNA virus. *J Mol Biol*, 407(1):60–78.

Perales, C., Agudo, R., Tejero, H., Manrubia, S. C., and Domingo, E. (2009a). Potential benefits of sequential inhibitor-mutagen treatments of RNA virus infections. *PLoS Pathog*, 5(11):e1000658.

Perales, C., Henry, M., Domingo, E., Wain-Hobson, S., and Vartanian, J.-P. (2011b). Lethal mutagenesis of foot-and-mouth disease virus involves shifts in sequence space. *J Virol*, 85(23):12227–12240.

Perales, C. C., Agudo, R. R., Tejero, H. H., Manrubia, S. C. S., and Domingo, E. E. (2009b). Potential benefits of sequential inhibitor-mutagen treatments of RNA virus infections. *PLoS Pathog*, 5(11):e1000658–e1000658.

Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T. K. B., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H. N., Rashmi, B. P., Ramya, M. A., Zhao, Z., Chandrika, K. N., Padma, N., Harsha, H. C., Yatish, A. J., Kavitha, M. P., Menezes, M., Choudhury, D. R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S. K., Madavan, V., Joseph, A., Wong, G. W., Schiemann, W. P., Constantinescu, S. N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G. C., Dang, C. V., Garcia, J. G. N., Pevsner, J., Jensen, O. N., Roepstorff, P., Deshpande, K. S., Chinnaiyan, A. M., Hamosh, A., Chakravarti, A., and Pandey, A. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 13(10):2363–2371.

Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567.

Perron, H. and Lang, A. (2010). The human endogenous retrovirus link between genes and environment in multiple sclerosis and in multifactorial diseases associating neuroinflammation. *Clin Rev Allergy Immunol*, 39(1):51–61.

Perry, S. C. and Beiko, R. G. (2010). Distinguishing microbial genome fragments based on their composition: Evolutionary and comparative genomic perspectives. *Genome Biology and Evolution*, 2(1):117–131.

Pessoa, M. G., Bzowej, N., Berenguer, M., Phung, Y., Kim, M., Ferrell, L., Hassoba, H., and Wright, T. L. (1999). Evolution of hepatitis C virus quasispecies in patients with severe cholestatic hepatitis after liver transplantation. *Hepatology (Baltimore, Md)*, 30(6):1513–1520.

Peterson, D. A., Frank, D. N., Pace, N. R., and Gordon, J. I. (2008). Metagenomic Approaches for Defining the Pathogenesis of Inflammatory Bowel Diseases. *Cell Host Microbe*, 3(6):417–427.

Petrosino, J. F., Highlander, S., Luna, R. A., Gibbs, R. A., and Versalovic, J. (2009). Metagenomic pyrosequencing and microbial identification. *Clin. Chem.*, 55(5):856–866.

Pfeiffer, J. K. and Kirkegaard, K. (2003). A single mutation in poliovirus RNA-dependent RNA polymerase confers resistance to mutagenic nucleotide analogs via increased fidelity. *Proc Natl Acad Sci USA*, 100(12):7289–7294.

Pfeiffer, J. K. and Kirkegaard, K. (2005a). Increased fidelity reduces poliovirus fitness and virulence under selective pressure in mice. *PLoS Pathog*, 1(2):e11.

Pfeiffer, J. K. and Kirkegaard, K. (2005b). Ribavirin resistance in hepatitis C virus replicon-containing cell lines conferred by changes in the cell line or mutations in the replicon RNA. *J Virol*, 79(4):2346–2355.

Pichlmair, A., Kandasamy, K., Alvisi, G., Mulhern, O., Sacco, R., Habjan, M., Binder, M., Stefanovic, A., Eberle, C.-A., Goncalves, A., Bürckstümmer, T., Müller, A. C., Fauster, A., Holze, C., Lindsten, K., Goodbourn, S., Kochs, G., Weber, F., Bartenschlager, R., Bowie, A. G., Bennett, K. L., Colinge, J., and Superti-Furga, G. (2012). Viral immune modulators perturb the human molecular network by common and unique strategies. *Nature*, 487(7408):486–490.

Pichlmair, A., Schulz, O., Tan, C. P., Näslund, T. I., Liljeström, P., Weber, F., and Reis e Sousa, C. (2006). RIG-I-mediated antiviral responses to single-stranded RNA bearing 5′-phosphates. *Science*, 314(5801):997–1001.

Pieters, J. (1997). MHC class II restricted antigen presentation. *Curr. Opin. Immunol.*, 9(1):89–96.

Plotch, S. J., Bouloy, M., Ulmanen, I., and Krug, R. M. (1981). A unique cap(m7GpppXm)-dependent influenza virion endonuclease cleaves capped RNAs to generate the primers that initiate viral RNA transcription. *Cell*, 23(3):847–858.

Pobezinskaya, Y. L., Kim, Y.-S., Choksi, S., Morgan, M. J., Li, T., Liu, C., and Liu, Z. (2008). The function of TRADD in signaling through tumor necrosis factor receptor 1 and TRIF-dependent Toll-like receptors. *Nat. Immunol.*, 9(9):1047–1054.

Pockley, A. G., Muthana, M., and Calderwood, S. K. (2008). The dual immunoregulatory roles of stress proteins. *Trends Biochem Sci*, 33(2):71–79.

Podolsky, S. (1996). The role of the virus in origin-of-life theorizing. *J. Hist. Biol.*, 29:79–126.

Pokrovskii, M. V., Bush, C. O., Beran, R. K. F., Robinson, M. F., Cheng, G., Tirunagari, N., Fenaux, M., Greenstein, A. E., Zhong, W., Delaney, W. E., and Paulson, M. S. (2011). Novel mutations in a tissue culture-adapted hepatitis C virus strain improve infectious-virus stability and markedly enhance infection kinetics. *J Virol*, 85(8):3978–3985.

Polanowska, J., Martin, J. S., Fisher, R., Scopa, T., Rae, I., and Boulton, S. J. (2004). Tandem immunoaffinity purification of protein complexes from Caenorhabditis elegans. *BioTechniques*, 36(5):778–80– 782.

Poordad, F., Lawitz, E., Kowdley, K. V., Cohen, D. E., Podsadecki, T., Siggelkow, S., Heckaman, M., Larsen, L., Menon, R., Koev, G., Tripathi, R., Pilot-Matias, T., and Bernstein, B. (2013). Exploratory study of oral combination antiviral therapy for hepatitis C. *N. Engl. J. Med.*, 368(1):45–53.

Poorvin, L., Rinta-Kanto, J. M., Hutchins, D. A., and Wilhelm, S. W. (2004). Viral release of iron and its bioavailability to marine plankton. *Limnol. Oceangr.*, 49(5):1734–1741.

Pop, M. and Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends Genet*, 24(3):142–149.

Popper, K. (2002). The Logic of Scientific Discovery.

Powdrill, M. H., Bernatchez, J. A., and Götte, M. (2010). Inhibitors of the Hepatitis C Virus RNA-Dependent RNA Polymerase NS5B. *Viruses*, 2(10):2169–2195.

Powers, C. J. and Früh, K. (2008). Signal peptide-dependent inhibition of MHC class I heavy chain translation by rhesus cytomegalovirus. *PLoS Pathog*, 4(10):e1000150.

Poynard, T., Marcellin, P., Lee, S. S., Niederau, C., Minuk, G. S., Ideo, G., Bain, V., Heathcote, J., Zeuzem, S., Trepo, C., and Albrecht, J. (1998). Randomised trial of interferon alpha2b plus ribavirin for 48 weeks or for 24 weeks versus interferon alpha2b plus placebo for 48 weeks for treatment of chronic infection with hepatitis C virus. International Hepatitis Interventional Therapy Group (IHIT). *Lancet*, 352(9138):1426–1432.

Price, B. D. (1996). Complete replication of an animal virus and maintenance of expression vectors derived from it in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 93(18):9465–9470.

Pride, D. T., Salzman, J., Haynes, M., Rohwer, F., Davis-Long, C., White, R. A., Loomer, P., Armitage, G. C., and Relman, D. A. (2012). Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *ISME J*, 6(5):915–926.

Prieto, C. and De Las Rivas, J. (2006). APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Res*, 34(Web Server issue):W298–302.

Prieto, C. and De Las Rivas, J. (2010). Structural domain-domain interactions: assessment and comparison with protein-protein interaction data to improve the interactome. *Proteins*, 78(1):109–117.

Pruitt, K. D., Tatusova, T., Klimke, W., and Maglott, D. R. (2009). NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res*, 37(Database issue):D32–6.

Przytycka, T. M., Singh, M., and Slonim, D. K. (2010). Toward the dynamic interactome: it's about time. *Brief Bioinform*, 11(1):15–29.

Pu, S., Wong, J., Turner, B., Cho, E., and Wodak, S. J. (2009). Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res*, 37(3):825–831.

Pugach, P., Marozsan, A. J., Ketas, T. J., Landes, E. L., Moore, J. P., and Kuhmann, S. E. (2007). HIV-1 clones resistant to a small molecule CCR5 inhibitor use the inhibitor-bound form of CCR5 for entry. *Virology*, 361(1):17–17.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Seraphin, B. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods*, 24(3):218–229.

Pulliam, J. R. C., Epstein, J. H., Dushoff, J., Rahman, S. A., Bunning, M., Jamaluddin, A. A., Hyatt, A. D., Field, H. E., Dobson, A. P., Daszak, P., and the Henipavirus Ecology Research Group (HERG) (2011). Agricultural intensification, priming for persistence and the emergence of Nipah virus: a lethal bat-borne zoonosis. *Journal of The Royal Society Interface*, 9(66):89–101.

Purcell, A. W. A. and Elliott, T. T. (2008). Molecular machinations of the MHC-I peptide loading complex. *Curr. Opin. Immunol.*, 20(1):7–7.

Qi, Y., Tastan, O., Carbonell, J. G., Klein-Seetharaman, J., and Weston, J. (2010). Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. *Bioinformatics*, 26(18):i645–52.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13:341.

Quan, P.-L., Palacios, G., Jabado, O. J., Conlan, S., Hirschberg, D. L., Pozo, F., Jack, P. J. M., Cisterna, D., Renwick, N., Hui, J., Drysdale, A., Amos-Ritchie, R., Baumeister, E., Savy, V., Lager, K. M., Richt, J. A., Boyle, D. B., García-Sastre, A., Casas, I., Perez-Breña, P., Briese, T., and Lipkin, W. I. (2007). Detection of respiratory viruses and subtype identification of influenza A viruses by GreeneChipResp oligonucleotide microarray. *J Clin Microbiol*, 45(8):2359–2364.

Quan, P. L., Wagner, T. A., Briese, T., Torgerson, T. R., Hornig, M., Tashmukhamedova, A., Firth, C., Palacios, G., Baisre-de Leon, A., Paddock, C. D., Hutchison, S. K., Egholm, M., Zaki, S. R., Goldman, J. E., Ochs, H. D., and Lipkin, W. I. (2010). Astrovirus encephalitis in boy with X-linked agammaglobulinemia. *Emerging Infect. Dis.*, 16(6):918–925.

Quer, J., Esteban, J.-I., Cos, J., Sauleda, S., Ocaña, L., Martell, M., Otero, T., Cubero, M., Palou, E., Murillo, P., Esteban, R., and Guardia, J. (2005). Effect of bottlenecking on evolution of the nonstructural protein 3 gene of hepatitis C virus during sexually transmitted acute resolving infection. *J Virol*, 79(24):15131–15141.

Quiñones-Mateu, M. E. and Arts, E. J. (2006). Virus fitness: concept, quantification, and application to HIV population dynamics. *Curr. Top. Microbiol. Immunol.*, 299:83–140.

Radziewicz, H., Ibegbu, C. C., Hon, H., Osborn, M. K., Obideen, K., Wehbi, M., Freeman, G. J., Lennox, J. L., Workowski, K. A., Hanson, H. L., and Grakoui, A. (2008). Impaired hepatitis C virus (HCV)-specific effector CD8+ T cells undergo massive apoptosis in the peripheral blood during acute HCV infection and in the liver during the chronic phase of infection. *J Virol*, 82(20):9808–9822.

Raes, J., Foerstner, K. U., and Bork, P. (2007). Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr Opin Microbiol*, 10(5):490–498.

Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schächter, V., Chemama, Y., Labigne, A., and Legrain, P. (2001). The protein-protein interaction map of Helicobacter pylori. *Nature*, 409(6817):211–215.

Rajagopala, S. V., Hughes, K. T., and Uetz, P. (2009). Benchmarking yeast two-hybrid systems using the interactions of bacterial motility proteins. *Proteomics*, 9(23):5296–5302.

Ramachandran, N., Raphael, J. V., Hainsworth, E., Demirkan, G., Fuentes, M. G., Rolfs, A., Hu, Y., and LaBaer, J. (2008). Next-generation high-density self-assembling functional protein arrays. *Nat Methods*, 5(6):535–538.

Ramratnam, B., Bonhoeffer, S., Binley, J., Hurley, A., Zhang, L., Mittler, J. E., Markowitz, M., Moore, J. P., Perelson, A. S., and Ho, D. D. (1999). Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. *Lancet*, 354(9192):1782–1785.

Ranea, J. A. G., Sillero, A., Thornton, J. M., and Orengo, C. A. (2006). Protein superfamily evolution and the last universal common ancestor (LUCA). *J. Mol. Evol.*, 63(4):513–525.

Ranish, J. A., Hahn, S., Lu, Y., Yi, E. C., Li, X.-j., Eng, J., and Aebersold, R. (2004). Identification of TFB5, a new component of general transcription and DNA repair factor IIH. *Nat Genet*, 36(7):707–713.

Ranish, J. A., Yi, E. C., Leslie, D. M., Purvine, S. O., Goodlett, D. R., Eng, J., and Aebersold, R. (2003). The study of macromolecular complexes by quantitative proteomics. *Nat Genet*, 33(3):349–355.

Raoult, D. (2009). There is no such thing as a tree of life (and of course viruses are out!). *Nat Rev Microbiol*, 7(8):615–author reply 615.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., and Claverie, J.-M. (2004). The 1.2-megabase genome sequence of Mimivirus. *Science*, 306(5700):1344–1350.

Raoult, D. and Forterre, P. (2008). Redefining viruses: lessons from Mimivirus. *Nat Rev Microbiol*, 6(4):315–319.

Rappé, M. S. and Giovannoni, S. J. (2003). The uncultured microbial majority. *Annu Rev Microbiol*, 57:369–394.

Rappoport, N. and Linial, M. (2012). Viral proteins acquired from a host converge to simplified domain architectures. *PLoS Comput Biol*, 8(2):e1002364.

Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S. K., McCulle, S. L., Karlebach, S., Gorle, R., Russell, J., Tacket, C. O., Brotman, R. M., Davis, C. C., Ault, K., Peralta, L., and Forney, L. J. (2011). Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci USA*, 108 Suppl 1:4680–4687.

Ray, S. C., Mao, Q., Lanford, R. E., Bassett, S., Laeyendecker, O., Wang, Y. M., and Thomas, D. L. (2000). Hypervariable region 1 sequence stability during hepatitis C virus replication in chimpanzees. *J Virol*, 74(7):3058–3066.

Razick, S., Magklaras, G., and Donaldson, I. M. (2008). iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*, 9:405.

Real, E. E., Rain, J.-C. J., Battaglia, V. V., Jallet, C. C., Perrin, P. P., Tordo, N. N., Chrisment, P. P., D'Alayer, J. J., Legrain, P. P., and Jacob, Y. Y. (2004). Antiviral drug discovery strategy using combinatorial libraries of structurally constrained peptides. *J Virol*, 78(14):7410–7417.

Reesink, H. W., Zeuzem, S., Weegink, C. J., Forestier, N., van Vliet, A., van de Wetering de Rooij, J., McNair, L., Purdy, S., Kauffman, R., Alam, J., and Jansen, P. L. M. (2006). Rapid decline of viral RNA in hepatitis C patients treated with VX-950: a phase Ib, placebo-controlled, randomized study. *Gastroenterology*, 131(4):997–1002.

Reeves, P. M., Bommarius, B., Lebeis, S., McNulty, S., Christensen, J., Swimm, A., Chahroudi, A., Chavan, R., Feinberg, M. B., Veach, D., Bornmann, W., Sherman, M., and Kalman, D. (2005). Disabling poxvirus pathogenesis by inhibition of Abl-family tyrosine kinases. *Nat Med*, 11(7):731–739.

Regoes, R. R. and Bonhoeffer, S. (2005). The HIV coreceptor switch: a population dynamical perspective. *Trends in microbiology*, 13(6):9–9.

Rehermann, B. (2009). Hepatitis C virus versus innate and adaptive immune responses: a tale of coevolution and coexistence. *J Clin Invest*, 119(7):1745–1754.

Relman, D. A. (1999a). Chronic host-parasite interactions. *Microbes and Malignancy: Infection As a Cause of Human Cancers Oxford University Press, Oxford, Uk*, pages 19–34.

Relman, D. A. (1999b). The search for unrecognized pathogens. *Science*, 284(5418):1308–1310.

Ressing, M. E., Luteijn, R. D., Horst, D., and Wiertz, E. J. (2013). Viral interference with antigen presentation: trapping TAP. *Mol. Immunol.*, 55(2):139–142.

Reumers, J., De Rijk, P., Zhao, H., Liekens, A., Smeets, D., Cleary, J., Van Loo, P., Van Den Bossche, M., Catthoor, K., Sabbe, B., Despierre, E., Vergote, I., Hilbush, B., Lambrechts, D., and Del-Favero, J. (2012). Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nat Biotechnol*, 30(1):61–68.

Reyes, A., Haynes, M., Hanson, N., Angly, F. E., Heath, A. C., Rohwer, F., and Gordon, J. I. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*, 466(7304):334–338.

Reyes, G. R. and Kim, J. P. (1991). Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. *Mol. Cell. Probes*, 5(6):473–481.

Rice, G., Stedman, K., Snyder, J., Wiedenheft, B., Willits, D., Brumfield, S., McDermott, T., and Young, M. J. (2001). Viruses from extreme thermal environments. *Proc Natl Acad Sci USA*, 98(23):13341–13345.

Rice, G., Tang, L., Stedman, K., Roberto, F., Spuhler, J., Gillitzer, E., Johnson, J. E., Douglas, T., and Young, M. (2004). The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life. *Proc Natl Acad Sci USA*, 101(20):7716–7720.

Richman, D. D., Margolis, D. M., Delaney, M., Greene, W. C., Hazuda, D., and Pomerantz, R. J. (2009). The challenge of finding a cure for HIV infection. *Science*, 323(5919):1304–1307.

Rider, T., Zook, C., Boettcher, T., and Wick, S. (2011). Broad-Spectrum Antiviral Therapeutics. *PLoS ONE*.

Rieber, N., Zapatka, M., Lasitschka, B., Jones, D., Northcott, P., Hutter, B., Jäger, N., Kool, M., Taylor, M., Lichter, P., Pfister, S., Wolf, S., Brors, B., and Eils, R. (2013). Coverage Bias and Sensitivity of Variant Calling for Four Whole-genome Sequencing Technologies. *PLoS ONE*, 8(6):e66621.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Seraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, 17(10):1030–1032.

Riley, K. J. K., Rabinowitz, G. S. G., Yario, T. A. T., Luna, J. M. J., Darnell, R. B. R., and Steitz, J. A. J. (2012). EBV and human microRNAs co-target oncogenic and apoptotic viral and human genes during latency. *EMBO Journal*, 31(9):2207–2221.

Riley, S. (2003). Transmission Dynamics of the Etiological Agent of SARS in Hong Kong: Impact of Public Health Interventions. *Science*, 300(5627):1961–1966.

Rivers, T. M. (1937). Viruses and Koch's postulates. *J. Bacteriol.*, (33):1–12.

Rizzo, P., Carbone, M., Fisher, S. G., Matker, C., Swinnen, L. J., Powers, A., Di Resta, I., Alkan, S., Pass, H. I., and Fisher, R. I. (1999). Simian virus 40 is present in most United States human mesotheliomas, but it is rarely present in non-Hodgkin's lymphoma. *Chest*, 116(6 Suppl):470S–473S.

Roach, J. C., Glusman, G., Smit, A. F. A., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M., Shendure, J., Drmanac, R., Jorde, L. B., Hood, L., and Galas, D. J. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328(5978):636–639.

Roberts, J. D., Bebenek, K., and Kunkel, T. A. (1988). The accuracy of reverse transcriptase from HIV-1. *Science*, 242(4882):1171–1173.

Robertson, K. D. and Ambinder, R. F. (1997). Methylation of the Epstein-Barr virus genome in normal lymphocytes. *Blood*, 90(11):4480–4484.

Robinson, C. V., Sali, A., and Baumeister, W. (2007). The molecular sociology of the cell. *Nature*, 450(7172):973–982.

Robinson, M., Tian, Y., Delaney, W. E., and Greenstein, A. E. (2011). Preexisting drug-resistance mutations reveal unique barriers to resistance for distinct antivirals. *Proc Natl Acad Sci USA*, 108(25):10290–10295.

Roehr, B. (2012). Texas records worst outbreak of West Nile virus on record. *BMJ*, 345(sep06 3):e6019–e6019.

Rohwer, F. (2003). Global Phage Diversity. *Cell*, 113(2):141–141.

Rohwer, F. and Edwards, R. (2002). The Phage Proteomic Tree: a genome-based taxonomy for phage. *J. Bacteriol.*, 184(16):4529–4535.

Rohwer, F. and Thurber, R. V. (2009). Viruses manipulate the marine environment. *Nature*, 459(7244):207–212.

Rohwer, F. and Youle, M. (2011). Consider something viral in your search. *Nat Rev Microbiol*, 9(5):308–309.

Roman, E., Simpson, J., Ansell, P., Kinsey, S., Mitchell, C. D., McKinney, P. A., Birch, J. M., Greaves, M., Eden, T., and United Kingdom Childhood Cancer Study Investigators (2007). Childhood acute lymphoblastic leukemia and infections in the first year of life: a report from the United Kingdom Childhood Cancer Study. *Am J Epidemiol*, 165(5):496–504.

Rong, L., Dahari, H., Ribeiro, R. M., and Perelson, A. S. (2010). Rapid emergence of protease inhibitor resistance in hepatitis C virus. *Sci Transl Med*, 2(30):30ra32.

Roomp, K., Beerenwinkel, N., and Sing, T. (2006). Arevir: A secure platform for designing personalized antiretroviral therapies against HIV. *Data Integration in the . . . .*

Rosario, K. and Breitbart, M. (2011). Exploring the viral world through metagenomics. *Curr Opin Vir*, 1(4):289–297.

Rose, T. M., Schultz, E. R., Henikoff, J. G., Pietrokovski, S., McCallum, C. M., and Henikoff, S. (1998). Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Res*, 26(7):1628–1635.

Rosen, G. L., Reichenberger, E. R., and Rosenfeld, A. M. (2011). NBC: The naïve Bayes classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics*, 27(1):127–129.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., and Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biol*, 14(5):R51.

Rota, P. A. (2003). Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. *Science*, 300(5624):1394–1399.

Rothman, A. L., Morishima, C., Bonkovsky, H. L., Polyak, S. J., Ray, R., Di Bisceglie, A. M., Lindsay, K. L., Malet, P. F., Chang, M., Gretch, D. R., Sullivan, D. G., Bhan, A. K., Wright, E. C., Koziel, M. J., and HALT-C Trial Group (2005). Associations among clinical, immunological, and viral quasispecies measurements in advanced chronic hepatitis C. *Hepatology (Baltimore, Md)*, 41(3):617–625.

Rotman, Y., Noureddin, M., Feld, J. J., Guedj, J., Witthaus, M., Han, H., Park, Y. J., Park, S.-H., Heller, T., Ghany, M. G., Doo, E., Koh, C., Abdalla, A., Gara, N., Sarkar, S., Thomas, E., Ahlenstiel, G., Edlich, B., Titerence, R., Hogdal, L., Rehermann, B., Dahari, H., Perelson, A. S., Hoofnagle, J. H., and Liang, T. J. (2013). Effect of ribavirin on viral kinetics and liver gene expression in chronic hepatitis C. *Gut*.

Roulston, A., Marcellus, R. C., and Branton, P. E. (1999). Viruses and apoptosis. *Annu Rev Microbiol*, 53:577–628.

Rous, P. (1973). Transmission of a malignant new growth by means of a cell-free filtrate. *Conn Med*, 37(10):526.

Rous, P. and Beard, J. W. (1935). The progression to carcinoma of virus-induced rabbit papillomas (Shope). *J Exp Med*, 62:523–548.

Roux, S. S., Enault, F. F., Robin, A. A., Ravet, V. V., Personnic, S. S., Theil, S. S., Colombet, J. J., Sime-Ngando, T. T., and Debroas, D. D. (2011). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS ONE*, 7(3):e33641–e33641.

Rowe, J. M., DeBruyn, J. M., Poorvin, L., LeCleir, G. R., Johnson, Z. I., Zinser, E. R., and Wilhelm, S. W. (2012). Viral and bacterial abundance and production in the Western Pacific Ocean and the relation to other oceanic realms. *FEMS Microbiol. Ecol.*, 79(2):359–370.

Rowe, M. M., Glaunsinger, B. B., van Leeuwen, D. D., Zuo, J. J., Sweetman, D. D., Ganem, D. D., Middeldorp, J. J., Wiertz, E. J. H. J. E., and Ressing, M. E. M. (2007). Host shutoff during productive Epstein-Barr virus infection is mediated by BGLF5 and may contribute to immune evasion. *Proc Natl Acad Sci USA*, 104(9):3366–3371.

Rozenblatt-Rosen, O., Deo, R. C., Padi, M., Adelmant, G., Calderwood, M. A., Rolland, T., Grace, M., Dricot, A., Askenazi, M., Tavares, M., Pevzner, S. J., Abderazzaq, F., Byrdsong, D., Carvunis, A.-R., Chen, A. A., Cheng, J., Correll, M., Duarte, M., Fan, C., Feltkamp, M. C., Ficarro, S. B., Franchi, R., Garg, B. K., Gulbahce, N., Hao, T., Holthaus, A. M., James, R., Korkhin, A., Litovchick, L., Mar, J. C., Pak, T. R., Rabello, S., Rubio, R., Shen, Y., Singh, S., Spangle, J. M., Tasan, M., Wanamaker, S., Webber, J. T., Roecklein-Canfield, J., Johannsen, E., Barabási, A.-L., Beroukhim, R., Kieff, E., Cusick, M. E., Hill, D. E., Munger, K., Marto, J. A., Quackenbush, J., Roth, F. P., DeCaprio, J. A., and Vidal, M. (2012). Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. *Nature*, 487(7408):491–495.

Ruiz-Jarabo, C. M., Arias, A., Baranowski, E., Escarmís, C., and Domingo, E. (2000). Memory in viral quasispecies. *J Virol*, 74(8):3543–3547.

Ruprecht, K., Ferreira, H., Flockerzi, A., Wahl, S., Sauter, M., Mayer, J., and Mueller-Lantzsch, N. (2008). Human endogenous retrovirus family HERV-K(HML-2) RNA transcripts are selectively packaged into retroviral particles produced by the human germ cell tumor line Tera-1 and originate mainly from a provirus on chromosome 22q11.21. *J Virol*, 82(20):10008–10016.

Russell, C. A., Fonville, J. M., Brown, A. E. X., Burke, D. F., Smith, D. L., James, S. L., Herfst, S., van Boheemen, S., Linster, M., Schrauwen, E. J., Katzelnick, L., Mosterín, A., Kuiken, T., Maher, E., Neumann, G., Osterhaus, A. D. M. E., Kawaoka, Y., Fouchier, R. A. M., and Smith, D. J. (2012). The potential for respiratory droplet-transmissible A/H5N1 influenza virus to evolve in a mammalian host. *Science*, 336(6088):1541–1547.

Russell, J. and Cohn, R. (2012). Palivizumab. Tbilisi State University.

Russell, R. B., Alber, F., Aloy, P., Davis, F. P., Korkin, D., Pichaud, M., Topf, M., and Sali, A. (2004). A structural perspective on protein-protein interactions. *Curr Opin Struct Biol*, 14(3):313–324.

Rux, J. J. and Burnett, R. M. (1998). Spherical viruses. *Curr Opin Struct Biol*, 8(2):142–149.

Saakian, D. B., Biebricher, C. K., and Hu, C.-K. (2009). Phase diagram for the Eigen quasispecies theory with a truncated fitness landscape. *Physical review E, Statistical, nonlinear, and soft matter physics*, 79(4 Pt 1):041905.

Saakian, D. B. and Hu, C.-K. (2006). Exact solution of the Eigen model with general fitness functions and degradation rates. *Proc Natl Acad Sci USA*, 103(13):4935–4939.

Sabin, A. B., Hennessen, W. A., and Winsser, J. (1954). Studies on variants of poliomyelitis virus. I. Experimental segregation and properties of avirulent variants of three immunologic types. *J Exp Med*, 99(6):551–576.

Sadasivan, B., Lehner, P. J., Ortmann, B., Spies, T., and Cresswell, P. (1996). Roles for Calreticulin and a Novel Glycoprotein, Tapasin, in the Interaction of MHC Class I Molecules with TAP. *Immunity*, 5(2):12–12.

Sadegh-Nasseri, S. S., Chen, M. M., Narayan, K. K., and Bouvier, M. M. (2008). The convergent roles of tapasin and HLA-DM in antigen presentation. *Trends Immunol.*, 29(3):7–7.

Sadler, A. J. and Williams, B. R. G. (2008). Interferon-inducible antiviral effectors. *Nat Rev Immunol*, 8(7):559–568.

Saito, T. and Gale, M. (2008). Differential recognition of double-stranded RNA by RIG-I-like receptors in antiviral immunity. *J Exp Med*, 205(7):1523–1527.

Salama, I. and Quade, D. (1982). A nonparametric comparison of two multiple regressions by means of a weighted measure of correlation. *Communications in Statistics - Theory and Methods*, 11(11):1185 — 1195.

Salazar-Gonzalez, J. F., Bailes, E., Pham, K. T., Salazar, M. G., Guffey, M. B., Keele, B. F., Derdeyn, C. A., Farmer, P., Hunter, E., Allen, S., Manigart, O., Mulenga, J., Anderson, J. A., Swanstrom, R., Haynes, B. F., Athreya, G. S., Korber, B. T. M., Sharp, P. M., Shaw, G. M., and Hahn, B. H. (2008). Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J Virol*, 82(8):3952–3970.

Salwínski, L. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, 32(90001):449D–451.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*, 74(12):5463–5467.

Sanger, H. L., Klotz, G., Riesner, D., Gross, H. J., and Kleinschmidt, A. K. (1976). Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures. *Proc Natl Acad Sci USA*, 73(11):3852–3856.

Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M., and Belshaw, R. (2010). Viral mutation rates. *J Virol*, 84(19):9733–9748.

Sanjuán, R. R., Moya, A. A., and Elena, S. F. S. (2004). The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc Natl Acad Sci U S A*, 101(22):8396–8401.

Sano, E., Carlson, S., Wegley, L., and Rohwer, F. (2004). Movement of viruses between biomes. *Appl. Environ. Microbiol.*, 70(10):5842–5846.

Sardanyés, J., Elena, S. F., and Solé, R. V. (2008). Simple quasispecies models for the survival-of-the-flattest effect: The role of space. *J. Theor. Biol.*, 250(3):560–568.

Sardanyés, J., Solé, R. V., and Elena, S. F. (2009). Replication mode and landscape topology differentially affect RNA virus mutational load and robustness. *J Virol*, 83(23):12579–12589.

Sardiu, M. E., Cai, Y., Jin, J., Swanson, S. K., Conaway, R. C., Conaway, J. W., Florens, L., and Washburn, M. P. (2008). Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proc Natl Acad Sci USA*, 105(5):1454–1459.

Sarid, R. and Gao, S.-J. (2011). Viruses and human cancer: from detection to causality. *Cancer Lett.*, 305(2):218–227.

Sarid, R., Olsen, S. J., and Moore, P. S. (1999). Kaposi's sarcoma-associated herpesvirus: epidemiology, virology, and molecular biology. *Adv Virus Res*, 52:139–232.

Sarmady, M., Dampier, W., and Tozeren, A. (2011). HIV protein sequence hotspots for crosstalk with host hub proteins. *PLoS ONE*, 6(8):e23293.

Sarrazin, C., Kieffer, T. L., Bartels, D., Hanzelka, B., Müh, U., Welker, M., Wincheringer, D., Zhou, Y., Chu, H.-M., Lin, C., Weegink, C., Reesink, H., Zeuzem, S., and Kwong, A. D. (2007). Dynamic hepatitis C virus genotypic and phenotypic changes in patients treated with the protease inhibitor telaprevir. *Gastroenterology*, 132(5):1767–1777.

Sarrazin, C. and Zeuzem, S. (2010). Resistance to Direct Antiviral Agents in Patients With Hepatitis C Virus Infection. *Gastroenterology*, 138(2):447–462.

Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., and Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28(14):1811–1817.

Sawyer, L. A. L. (2000). Antibodies for the prevention and treatment of viral diseases. *Antiviral Res*, 47(2):57–77.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Federhen, S., Feolo, M., Fingerman, I. M., Geer, L. Y., Helmberg, W., Kapustin, Y., Krasnov, S., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Karsch-Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., Wilbur, W. J., Yaschenko, E., and Ye, J. (2012). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 40(Database issue):D13–25.

Scheffner, M., Werness, B. A., Huibregtse, J. M., Levine, A. J., and M, H. P. (1990). The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. *Cell*, 63(6):1129–1136.

Scheffner, M. and Whitaker, N. J. (2003). Human papillomavirus-induced carcinogenesis and the ubiquitin-proteasome system. *Semin. Cancer Biol.*, 13(1):59–67.

Schelhorn, S.-E., Fischer, M., Tolosi, L., Altmüller, J., Nürnberg, P., Pfister, H., Lengauer, T., and Berthold, F. (2013). Sensitive Detection of Viral Transcripts in Human Tumor Transcriptomes. *PLoS Comput Biol*, 9(10):e1003228.

Schelhorn, S.-E., Lengauer, T., and Albrecht, Mario (2008). An integrative approach for predicting interactions of protein regions. *Bioinformatics*, 24(16):i35–41.

Schelhorn, S.-E., Mestre, J., Albrecht, M., and Zotenko, E. (2011). Inferring physical protein contacts from large-scale purification data of protein complexes. *Mol Cell Proteomics*, 10(6):M110.004929.

Schelhorn, U. (1993). *Untersuchungen zur Charakterisierung der zellulären Ribosomen gleichenden Partikel in den Viria des Virus der Lymphozytären Choriomeningitis*. PhD thesis.

Schiller, J. T. and Lowy, D. R. (2010). Vaccines to prevent infections by oncoviruses. *Annu Rev Microbiol*, 64:23–41.

Schinkel, J., de Jong, M. D., Bruning, B., van Hoek, B., Spaan, W. J. M., and Kroes, A. C. M. (2003). The potentiating effect of ribavirin on interferon in the treatment of hepatitis C: lack of evidence for ribavirin-induced viral mutagenesis. *Antivir Ther (Lond)*, 8(6):535–540.

Schlom, J., Spiegelman, S., and Moore, D. (1971). RNA-dependent DNA polymerase activity in virus-like particles isolated from human milk. *Nature*, 231(5298):97–100.

Schmitt, M. W., Kennedy, S. R., Salk, J. J., Fox, E. J., Hiatt, J. B., and Loeb, L. A. (2012). Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci USA*, 109(36):14508–14513.

Schneider-Schaulies, J. J. (2000). Cellular receptors for viruses: links to tropism and pathogenesis. *J Gen Virol*, 81(Pt 6):1413–1429.

Schoenhals, G. J. G., Krishna, R. M. R., Grandea, A. G. A., Spies, T. T., Peterson, P. A. P., Yang, Y. Y., and Früh, K. K. (1999). Retention of empty MHC class I molecules by tapasin is essential to reconstitute antigen presentation in invertebrate cells. *EMBO Journal*, 18(3):743–753.

Schreiber, M. M., Sedger, L. L., and McFadden, G. G. (1997). Distinct domains of M-T2, the myxoma virus tumor necrosis factor (TNF) receptor homolog, mediate extracellular TNF binding and intracellular apoptosis inhibition. *J Virol*, 71(3):2171–2181.

Schröder, J., Schröder, H., Puglisi, S. J., Sinha, R., and Schmidt, B. (2009). SHREC: a short-read error correction method. *Bioinformatics*, 25(17):2157–2163.

Schuitemaker, H., van 't Wout, A. B., and Lusso, P. (2010). Clinical significance of HIV-1 coreceptor usage. *J Transl Med*, 9 Suppl 1:S5–S5.

Schulman, J. L. and Kilbourne, E. D. (1969). Independent variation in nature of hemagglutinin and neuraminidase antigens of influenza virus: distinctiveness of hemagglutinin antigen of Hong Kong-68 virus. *Proc Natl Acad Sci USA*, 63(2):326–333.

Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092.

Schulz, T. F. (2009). Cancer and viral infections in immunocompromised individuals. *Int. J. Cancer*, 125(8):1755–1763.

Schuster, P. P. and Swetina, J. J. (1987). Stationary mutant distributions and evolutionary optimization. *Bull. Math. Biol.*, 50(6):635–660.

Schwartz, A. S., Yu, J., Gardenour, K. R., Finley, R. L., and Ideker, T. (2009). Cost-effective strategies for completing the interactome. *Nat Methods*, 6(1):55–61.

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. *Genome Res*, 13(1):103–107.

Sears, C. L. (2005). A dynamic partnership: celebrating our gut flora. *Anaerobe*, 11(5):247–251.

Seeger, C. and Mason, W. S. (2000). Hepatitis B virus biology. *Microbiology and Molecular Biology Reviews*, 64(1):51–68.

Selling, B. H., Allison, R. F., and Kaesberg, P. (1990). Genomic RNA of an insect virus directs synthesis of infectious virions in plants. *Proc Natl Acad Sci USA*, 87(1):434–438.

Senkevich, T. G. T., Bugert, J. J. J., Sisler, J. R. J., Koonin, E. V. E., Darai, G. G., and Moss, B. B. (1996). Genome sequence of a human tumorigenic poxvirus: prediction of specific host response-evasion genes. *Science*, 273(5276):813–816.

Shah, K. V. (2007). SV40 and human cancer: a review of recent data. *Int. J. Cancer*, 120(2):215–223.

Shah, K. V., Daniel, R. W., and Warszawski, R. M. (1973). High prevalence of antibodies to BK virus, an SV40-related papovavirus, in residents of Maryland. *J Infect Dis*, 128(6):784–787.

Shamay, M., Krithivas, A., Zhang, J., and Hayward, S. D. (2006). Recruitment of the de novo DNA methyltransferase Dnmt3a by Kaposi's sarcoma-associated herpesvirus LANA. *Proc Natl Acad Sci USA*, 103(39):14554–14559.

Shan, G. (2010). RNA interference as a gene knockdown technique. *Int. J. Biochem. Cell Biol.*, 42(8):1243–1251.

Shaner, L. (2005). The Yeast Hsp110 Sse1 Functionally Interacts with the Hsp70 Chaperones Ssa and Ssb. *J Biol Chem*, 280(50):41262–41269.

Shapira, S. D., Gat-Viks, I., Shum, B. O. V., Dricot, A., de Grace, M. M., Wu, L., Gupta, P. B., Hao, T., Silver, S. J., Root, D. E., Hill, D. E., Regev, A., and Hacohen, N. (2009). A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell*, 139(7):1255–1267.

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nat Biotechnol*, 26(10):1135–1145.

Sheridan, I., Pybus, O. G., Holmes, E. C., and Klenerman, P. (2004). High-resolution phylogenetic analysis of hepatitis C virus adaptation and its relationship to disease progression. *J Virol*, 78(7):3447–3454.

Shimakami, T., Lanford, R. E., and Lemon, S. M. (2009). Hepatitis C: recent successes and continuing challenges in the development of improved treatment modalities. *Curr Opin Pharmacol*, 9(5):537–544.

Shiraishi, Y., Sato, Y., Chiba, K., Okuno, Y., Nagata, Y., Yoshida, K., Shiba, N., Hayashi, Y., Kume, H., Homma, Y., Sanada, M., Ogawa, S., and Miyano, S. (2013). An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res*, 41(7):e89.

Shisler, J. J., Yang, C. C., Walter, B. B., Ware, C. F. C., and Gooding, L. R. L. (1997). The adenovirus E3-10.4K/14.5K complex mediates loss of cell surface Fas (CD95) and resistance to Fas-induced apoptosis. *J Virol*, 71(11):8299–8306.

Shivapurkar, N., Wiethege, T., Wistuba, I. I., Salomon, E., Milchgrub, S., Muller, K. M., Churg, A., Pass, H., and Gazdar, A. F. (1999). Presence of simian virus 40 sequences in malignant mesotheliomas and mesothelial cell proliferations. *J. Cell. Biochem.*, 76(2):181–188.

Shresta, S., Pham, C. T., Thomas, D. A., Graubert, T. A., and Ley, T. J. (1998). How do cytotoxic lymphocytes kill their targets? *Curr. Opin. Immunol.*, 10(5):581–587.

Shuangshoti, S., Shuangshoti, S., Nuchprayoon, I., Kanjanapongkul, S., Marrano, P., Irwin, M. S., and Thorner, P. S. (2012). Natural course of low risk neuroblastoma. *Pediatr Blood Cancer*, 58(5):690–694.

Shuman, E. K. (2010). Global Climate Change and Infectious Diseases. *N. Engl. J. Med.*, 362(12):1061–1063.

Shutt, T. E. and Gray, M. W. (2006). Bacteriophage origins of mitochondrial replication and transcription proteins. *Trends Genet*, 22(2):90–95.

Si, H. and Robertson, E. S. (2006). Kaposi's sarcoma-associated herpesvirus-encoded latency-associated nuclear antigen induces chromosomal instability through inhibition of p53 function. *J Virol*, 80(2):697–709.

Sidwell, R. W., Huffman, J. H., Khare, G. P., Allen, L. B., Witkowski, J. T., and Robins, R. K. (1972). Broad-spectrum antiviral activity of Virazole: 1-beta-D-ribofuranosyl-1,2,4-triazole-3-carboxamide. *Science*, 177(4050):705–706.

Sierra, M., Airaksinen, A., González-López, C., Agudo, R., Arias, A., and Domingo, E. (2007). Foot-and-mouth disease virus mutant with decreased sensitivity to ribavirin: implications for error catastrophe. *J Virol*, 81(4):2012–2024.

Sierra, S., Dávila, M., Lowenstein, P. R., and Domingo, E. (2000). Response of foot-and-mouth disease virus to increased mutagenesis: influence of viral load and fitness in loss of infectivity. *J Virol*, 74(18):8316–8323.

Simbiri, K. O., Murakami, M., Feldman, M., Steenhoff, A. P., Nkomazana, O., Bisson, G., and Robertson, E. S. (2010). Multiple oncogenic viruses identified in Ocular surface squamous neoplasia in HIV-1 patients. *Infect. Agents Cancer*, 5:6.

Simmonds, P. (2004). Genetic diversity and evolution of hepatitis C virus–15 years on. *J Gen Virol*, 85(Pt 11):3173–3188.

Simmonds, P., Bukh, J., Combet, C., Deleage, G., Enomoto, N., Feinstone, S., Halfon, P., Inchauspé, G., Kuiken, C., Maertens, G., Mizokami, M., Murphy, D. G., Okamoto, H., Pawlotsky, J.-M., Penin, F., Sablon, E., Shin-I, T., Stuyver, L. J., Thiel, H.-J., Viazov, S., Weiner, A. J., and Widell, A. (2005). Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. *Hepatology (Baltimore, Md)*, 42(4):962–973.

Simmonds, P. P. (2001). Reconstructing the origins of human hepatitis viruses. *Philos Trans R Soc Lond, B, Biol Sci*, 356(1411):1013–1026.

Simon, C. and Daniel, R. (2011). Metagenomic analyses: Past and future trends. *Appl. Environ. Microbiol.*, 77(4):1153–1161.

Simon, C. E. (1923). The Filterable Viruses. *Physiological Reviews*, 3(4):483–508.

Simpson, J. T. and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Res*, 22(3):549–556.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Res*, 19(6):1117–1123.

Sing, T., Beerenwinkel, N., and Kaiser, R. (2005). Geno2pheno [coreceptor]: a tool for predicting coreceptor usage from genotype and for monitoring coreceptor-associated sequence alterations. *3rd European HIV ....*

Singh, I., Tastan, O., and Klein-Seetharaman, J. (2010). Comparison of virus interactions with human signal transduction pathways. In *BCB '10*, page 17, New York, New York, USA. ACM Press.

Sinz, A. (2010). Investigation of protein-protein interactions in living cells by chemical crosslinking and mass spectrometry. *Anal Bioanal Chem*, 397(8):3433–3440.

Small, M. B., Gluzman, Y., and Ozer, H. L. (1982). Enhanced transformation of human fibroblasts by origin-defective simian virus 40. *Nature*, 296(5858):671–672.

Smith, G. L. (1994). Virus strategies for evasion of the host response to infection. *Trends in microbiology*, 2(3):81–88.

Smith, K. M., Anthony, S. J., Switzer, W. M., Epstein, J. H., Seimon, T., Jia, H., Sanchez, M. D., Huynh, T. T., Galland, G. G., Shapiro, S. E., Sleeman, J. M., McAloose, D., Stuchin, M., Amato, G., Kolokotronis, S.-O., Lipkin, W. I., Karesh, W. B., Daszak, P., and Marano, N. (2012). Zoonotic Viruses Associated with Illegally Imported Wildlife Products. *PLoS ONE*, 7(1):e29505.

Smith-Tsurkan, S. D., Wilke, C. O., and Novella, I. S. (2010). Incongruent fitness landscapes, not tradeoffs, dominate the adaptation of vesicular stomatitis virus to novel host types. *J Gen Virol*, 91(Pt 6):1484–1493.

Sniegowski, P. D. P., Gerrish, P. J. P., Johnson, T. T., and Shaver, A. A. (2000). The evolution of mutation rates: separating causes from consequences. *Bioessays*, 22(12):1057–1066.

Snijder, B. and Pelkmans, L. (2011). Origins of regulated cell-to-cell variability. *Nat Rev Mol Cell Biol*, 12(2):119–125.

Snuderl, M., Fazlollahi, L., Le, L. P., Nitta, M., Zhelyazkova, B. H., Davidson, C. J., Akhavanfard, S., Cahill, D. P., Aldape, K. D., Betensky, R. A., Louis, D. N., and Iafrate, A. J. (2011). Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma. *Cancer Cell*, 20(6):810–817.

Söderholm, J., Ahlén, G., Kaul, A., Frelin, L., Alheim, M., Barnfield, C., Liljeström, P., Weiland, O., Milich, D. R., Bartenschlager, R., and Sällberg, M. (2006). Relation between viral fitness and immune escape within the hepatitis C virus protease. *Gut*, 55(2):266–274.

Solé, R. V. R., Rodríguez-Caso, C. C., Deisboeck, T. S. T., and Saldaña, J. J. (2008). Cancer stem cells as the engine of unstable tumor progression. *J. Theor. Biol.*, 253(4):9–9.

Solmone, M., Vincenti, D., Prosperi, M. C. F., Bruselles, A., Ippolito, G., and Capobianchi, M. R. (2009). Use of massively parallel ultradeep pyrosequencing to characterize the genetic diversity of hepatitis B virus in drug-resistant and drug-naive patients and to detect minor variants in reverse transcriptase and hepatitis B S antigen. *J Virol*, 83(4):1718–1726.

Soriano, V., Perelson, A. S., and Zoulim, F. (2008). Why are there different dynamics in the selection of drug resistance in HIV and hepatitis B and C viruses? *J. Antimicrob. Chemother.*, 62(1):1–4.

Sowa, M. E., Bennett, E. J., Gygi, S. P., and Harper, J. W. (2009). Defining the human deubiquitinating enzyme interaction landscape. *Cell*, 138(2):389–403.

Spence, S. L. and Pipas, J. M. (1994). SV40 large T antigen functions at two distinct steps in virion assembly. *Virology*, 204(1):200–209.

Spitz, R., Hero, B., Ernestus, K., and Berthold, F. (2003). FISH analyses for alterations in chromosomes 1, 2, 3, and 11 define high-risk groups in neuroblastoma. *Med. Pediatr. Oncol.*, 41(1):30–35.

Spriggs, M. K. M. (1995). One step ahead of the game: viral immunomodulatory molecules. *Immunology*, 14:101–130.

Sprinzak, E. and Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, 311(4):681–692.

Sprinzak, E., Sattath, S., and Margalit, H. (2003). How reliable are experimental protein-protein interaction data? *J Mol Biol*, 327(5):919–923.

Spyrakis, F., BidonChanal, A., Barril, X., and Luque, F. J. (2011). Protein flexibility and ligand recognition: challenges for molecular modeling. *Curr Top Med Chem*, 11(2):192–210.

Srikiatkhachorn, A. and Green, S. (2010). Markers of dengue disease severity. *Curr. Top. Microbiol. Immunol.*, 338:67–82.

Staal, F. J. F., Ela, S. W. S., Roederer, M. M., Anderson, M. T. M., Herzenberg, L. A. L., and Herzenberg, L. A. L. (1992). Glutathione deficiency and human immunodeficiency virus infection. *Lancet*, 339(8798):909–912.

Stack, J., Haga, I. R., Schröder, M., Bartlett, N. W., Maloney, G., Reading, P. C., Fitzgerald, K. A., Smith, G. L., and Bowie, A. G. (2005). Vaccinia virus protein A46R targets multiple Toll-like-interleukin-1 receptor adaptors and contributes to virulence. *J Exp Med*, 201(6):1007–1018.

Staheli, J. P., Boyce, R., Kovarik, D., and Rose, T. M. (2011). CODEHOP PCR and CODEHOP PCR primer design. *Methods in molecular biology (Clifton, N.J.)*, 687:57–73.

Stark, M., Berger, S. A., Stamatakis, A., and Von Mering, C. (2010). MLTreeMap - accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics*, 11(1).

Stebbins, C. E. (2005). Structural microbiology at the pathogen-host interface. *Cell Microbiol*, 7(9):1227–1236.

Stehelin, D., Varmus, H. E., Bishop, J. M., and Vogt, P. K. (1976). DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*, 260(5547):170–173.

Stein, A., Ceol, A., and Aloy, P. (2011). 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res*, 39(Database issue):D718–23.

Stein, A., Panjkovich, A., and Aloy, P. (2009). 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Res*, 37(Database issue):D300–4.

Stein, A., Russell, R. B., and Aloy, P. (2005). 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res*, 33(Database issue):D413–7.

Stelzl, U. and Wanker, E. E. (2006). The value of high quality protein-protein interaction networks for systems biology. *Current opinion in chemical biology*, 10(6):551–558.

Stengel, F., Aebersold, R., and Robinson, C. V. (2012). Joining forces: integrating proteomics and cross-linking with the mass spectrometry of intact complexes. *Mol Cell Proteomics*, 11(3):R111.014027.

Stewart, T. H., Sage, R. D., Stewart, A. F., and Cameron, D. W. (2000). Breast cancer incidence highest in the range of one species of house mouse, Mus domesticus. *Br. J. Cancer*, 82(2):446–451.

Stolt, A., Kjellin, M., Sasnauskas, K., Luostarinen, T., Koskela, P., Lehtinen, M., and Dillner, J. (2005). Maternal human polyomavirus infection and risk of neuroblastoma in the child. *Int. J. Cancer*, 113(3):393–396.

Stoye, J. P., Moroni, C., and Coffin, J. M. (1991). Virological events leading to spontaneous AKR thymomas. *J Virol*, 65(3):1273–1285.

Straight, S. W., Hinkle, P. M., Jewers, R. J., and McCance, D. J. (1993). The E5 oncoprotein of human papillomavirus type 16 transforms fibroblasts and effects the downregulation of the epidermal growth factor receptor in keratinocytes. *J Virol*, 67(8):4521–4532.

Streeter, D. G., Witkowski, J. T., Khare, G. P., Sidwell, R. W., Bauer, R. J., Robins, R. K., and Simon, L. N. (1973). Mechanism of action of 1- -D-ribofuranosyl-1,2,4-triazole-3-carboxamide (Virazole), a new broad-spectrum antiviral agent. *Proc Natl Acad Sci USA*, 70(4):1174–1178.

Stumpf, M. P. H., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. *Proc Natl Acad Sci USA*, 105(19):6959–6964.

Su, C. H., Hsu, M. T., Wang, T. Y., Chiang, S., Cheng, J. H., Weng, F. C., Kao, C. Y., Wang, D., and Tsai, H. K. (2011). MetaABC - An integrated metagenomics platform for data adjustment, binning and clustering. *Bioinformatics*, 27(16):2298–2299.

Sugimoto, M., Tahara, H., Okubo, M., Kobayashi, T., Goto, M., Ide, T., and Furuichi, Y. (2004). WRN gene and other genetic factors affecting immortalization of human B-lymphoblastoid cell lines transformed by Epstein-Barr virus. *Cancer Genet. Cytogenet.*, 152(2):95–100.

Sukumar, S., Notario, V., Martin Zanca, D., and Barbacid, M. (1983). Induction of mammary carcinomas in rats by nitrosomethylurea involves malignant activation of H-ras-1 by single point mutations. *Nature*, 306(5944):658–661.

Sulkowski, M. S. (2003). Anemia in the treatment of hepatitis C virus infection. *Clin Infect Dis*, 37 Suppl 4:S315–22.

Sullivan, C. S. and Pipas, J. M. (2001). The virus-chaperone connection. *Virology*, 287(1):1–8.

Sullivan, P. F., Allander, T., Lysholm, F., Goh, S., Persson, B., Jacks, A., Evengård, B., Pedersen, N. L., and Andersson, B. (2011). An unbiased metagenomic search for infectious agents using monozygotic twins discordant for chronic fatigue. *BMC Microbiol*, 11:2.

Sun, C., Skaletsky, H., Rozen, S., Gromoll, J., Nieschlag, E., Oates, R., and Page, D. C. (2000). Deletion of azoospermia factor a (AZFa) region of human Y chromosome caused by recombination between HERV15 proviruses. *Hum. Mol. Genet.*, 9(15):2291–2296.

Sun, Q., Matta, H., and Chaudhary, P. M. (2005). Kaposi's sarcoma associated herpes virus-encoded viral FLICE inhibitory protein activates transcription from HIV-1 Long Terminal Repeat via the classical NF-kappaB pathway and functionally cooperates with Tat. *Retrovirology*, 2:9.

Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S., Stocks, K., Allen, E. E., Ellisman, M., Grethe, J., and Wooley, J. (2011). Community cyberinfrastructure for advanced microbial ecology research and analysis: The CAMERA resource. *Nucleic Acids Res*, 39(SUPPL. 1):D546–D551.

Sun, S., Schiller, J. H., and Gazdar, A. F. (2007). Lung cancer in never smokers–a different disease. *Nat Rev Cancer*, 7(10):778–790.

Sung, W.-K., Zheng, H., Li, S., Chen, R., Liu, X., Li, Y., Lee, N. P., Lee, W. H., Ariyaratne, P. N., Tennakoon, C., Mulawadi, F. H., Wong, K. F., Liu, A. M., Poon, R. T., Fan, S. T., Chan, K. L., Gong, Z., Hu, Y., Lin, Z., Wang, G., Zhang, Q., Barber, T. D., Chou, W.-C., Aggarwal, A., Hao, K., Zhou, W., Zhang, C., Hardwick, J., Buser, C., Xu, J., Kan, Z., Dai, H., Mao, M., Reinhard, C., Wang, J., and Luk, J. M. (2012). Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat Genet*, 44(7):765–769.

Suratanee, A., Rebhan, I., Matula, P., Kumar, A., Kaderali, L., Rohr, K., Bartenschlager, R., Eils, R., and König, R. (2010). Detecting host factors involved in virus infection by observing the clustering of infected cells in siRNA screening images. *Bioinformatics*, 26(18):i653–8.

Suttle, C. A. (1994). The significance of viruses to mortality in aquatic microbial communities. *Microbial Ecology*, (28):237–243.

Suttle, C. A. (2005). Viruses in the sea. *Nature*, 437(7057):356–361.

Suttle, C. A. (2007). Marine viruses–major players in the global ecosystem. *Nat Rev Microbiol*, 5(10):801–812.

Suzuki, S., Ono, N., Furusawa, C., Ying, B.-W., and Yomo, T. (2011). Comparison of sequence reads obtained from three next-generation sequencing platforms. *PLoS ONE*, 6(5):e19534.

Szabo, S., Haislip, A. M., and Garry, R. F. (2005). Of mice, cats, and men: is human breast cancer a zoonosis? *Microsc. Res. Tech.*, 68(3-4):197–208.

Szakács, G., Paterson, J. K., Ludwig, J. A., Booth-Genthe, C., and Gottesman, M. M. (2006). Targeting multidrug resistance in cancer. *Nat Rev Drug Disc*, 5(3):219–234.

Taddei, F. F., Radman, M. M., Maynard-Smith, J. J., Toupance, B. B., Gouyon, P. H. P., and Godelle, B. B. (1997). Role of mutator alleles in adaptive evolution. *Nature*, 387(6634):700–702.

Taira, A. V., Neukermans, C. P., and Sanders, G. D. (2004). Evaluating human papillomavirus vaccination programs. *Emerging Infect. Dis.*, 10(11):1915–1923.

Takahashi, I. and Marmur, J. (1963). Replacement of thymidylic acid by deoxyuridylic acid in the deoxyribonucleic acid of a transducing phage for Bacillus subtilis. *Nature*, 197:794–795.

Takaki, A., Wiese, M., Maertens, G., Depla, E., Seifert, U., Liebetrau, A., Miller, J. L., Manns, M. P., and Rehermann, B. (2000). Cellular immune responses persist and humoral responses decrease two decades after recovery from a single-source outbreak of hepatitis C. *Nat Med*, 6(5):578–582.

Takeuchi, N. and Hogeweg, P. (2007). Error-threshold exists in fitness landscapes with lethal mutants. *BMC Evol Biol*, 7(1):15.

Takeuchi, O. and Akira, S. (2008). MDA5/RIG-I and virus recognition. *Curr. Opin. Immunol.*, 20(1):17–22.

Tan, S.-L., Ganji, G., Paeper, B., Proll, S., and Katze, M. G. (2007). Systems biology and the host response to viral infection. *Nat Biotechnol*, 25(12):1383–1389.

Tanabe, Y., Sakamoto, N., Enomoto, N., Kurosaki, M., Ueda, E., Maekawa, S., Yamashiro, T., Nakagawa, M., Chen, C.-H., Kanazawa, N., Kakinuma, S., and Watanabe, M. (2004). Synergistic inhibition of intracellular hepatitis C virus replication by combination of ribavirin and interferon- alpha. *J Infect Dis*, 189(7):1129–1139.

Tang, K.-W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M., and Larsson, E. (2013). The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat Commun*, 4:2513.

Tannenbaum, E. E. and Shakhnovich, E. I. E. (2004). Solution of the quasispecies model for an arbitrary gene network. *Physical review E, Statistical, nonlinear, and soft matter physics*, 70(2 Pt 1):021903–021903.

Tapia, N., Fernàndez, G., Parera, M., Gómez-Mariano, G., Clotet, B., Quiñones-Mateu, M., Domingo, E., and Martínez, M. A. (2005). Combination of a mutagenic agent with a reverse transcriptase inhibitor results in systematic inhibition of HIV-1 infection. *Virology*, 338(1):1–8.

Tarassov, K., Messier, V., Landry, C. R., Radinovic, S., Molina, M. M. S., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., and Michnick, S. W. (2008). An in Vivo Map of the Yeast Protein Interactome. *Science*, 320(5882):1465–1470.

Tastan, O., Qi, Y., Carbonell, J. G., and Klein-Seetharaman, J. (2009). Prediction of interactions between HIV-1 and human proteins by information integration. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 516–527.

Taubenberger, J. K. and Morens, D. M. (2006). 1918 Influenza: the mother of all pandemics. *Emerging Infect. Dis.*, 12(1):15–22.

Tavernier, J., Eyckerman, S., Lemmens, I., Van der Heyden, J., Vandekerckhove, J., and Van Ostade, X. (2002). MAPPIT: a cytokine receptor-based two-hybrid method in mammalian cells. *Clin. Exp. Allergy*, 32(10):1397–1404.

Taylor, L. H., Latham, S. M., and Woolhouse, M. E. (2001). Risk factors for human disease emergence. *Philos Trans R Soc Lond, B, Biol Sci*, 356(1411):983–989.

Taylor, W. R., Chelliah, V., Hollup, S. M., MacDonald, J. T., and Jonassen, I. (2009). Probing the "dark matter" of protein fold space. *Structure*, 17(9):1244–1252.

Teeling, H., Meyerdierks, A., Bauer, M., Amann, R., and Glöckner, F. O. (2004). Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.*, 6(9):938–947.

Temin, H. M. (1964). Nature of the provirus of Rous sarcoma. *Natl Cancer Inst Monogr*, 17:557–570.

Temin, H. M. and Mizutani, S. (1970). Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature*, 226(5252):1211–1213.

Temin, H. M. and Rubin, H. (1958). Characteristics of an assay for Rous sarcoma virus and Rous sarcoma cells in tissue culture. *Virology*, 6(3):669–688.

Teodoro, J. G. and Branton, P. E. (1997). Regulation of apoptosis by viral gene products. *J Virol*, 71(3):1739–1746.

Thielen, A. and Lengauer, T. (2012). Geno2pheno[454]: a Web server for the prediction of HIV-1 coreceptor usage from next-generation sequencing data. *Intervirology*, 55(2):113–117.

Thielen, A. A., Lengauer, T. T., Swenson, L. C. L., Dong, W. W Y. W., McGovern, R. A. R., Lewis, M. M., James, I. I., Heera, J. J., Valdez, H. H., and Harrigan, P. R. P. (2010). Mutations in gp41 are correlated with coreceptor tropism but do not improve prediction methods substantially. *Antivir Ther (Lond)*, 16(3):319–328.

Thieu, T., Joshi, S., Warren, S., and Korkin, D. (2012). Literature mining of host-pathogen interactions: comparing feature-based supervised learning and language-based approaches. *Bioinformatics*, 28(6):867–875.

Thimme, R., Binder, M., and Bartenschlager, R. (2012). Failure of innate and adaptive immune responses in controlling hepatitis C virus infection. *FEMS Microbiology Reviews*, 36(3):663–683.

Thiry, E., Bublot, M., Dubuisson, J., Van Bressem, M. F., Lequarre, A. S., Lomonte, P., Vanderplasschen, A., and Pastoret, P. P. (1992). Molecular biology of bovine herpesvirus type 4. *Vet. Microbiol.*, 33(1-4):79–92.

Thomas, E., Feld, J. J., Li, Q., Hu, Z., Fried, M. W., and Liang, T. J. (2011). Ribavirin potentiates interferon action by augmenting interferon-stimulated gene induction in hepatitis C virus cell culture models. *Hepatology (Baltimore, Md)*, 53(1):32–41.

Thomas, J. A. and Gorelick, R. J. (2008). Nucleocapsid protein function in early infection processes. *Virus Res*, 134(1-2):39–63.

Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp*, 2(1):3.

Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L., and Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes. *Nat Protoc*, 4(4):470–483.

Timm, J., Li, B., Daniels, M. G., Bhattacharya, T., Reyor, L. L., Allgaier, R., Kuntzen, T., Fischer, W., Nolan, B. E., Duncan, J., Schulze Zur Wiesch, J., Kim, A. Y., Frahm, N., Brander, C., Chung, R. T., Lauer, G. M., Korber, B. T., and Allen, T. M. (2007). Human leukocyte antigen-associated sequence polymorphisms in hepatitis C virus reveal reproducible immune responses and constraints on viral evolution. *Hepatology (Baltimore, Md)*, 46(2):339–349.

Tomlins, S. A., Laxman, B., Dhanasekaran, S. M., Helgeson, B. E., Cao, X., Morris, D. S., Menon, A., Jing, X., Cao, Q., Han, B., Yu, J., Wang, L., Montie, J. E., Rubin, M. A., Pienta, K. J., Roulston, D., Shah, R. B., Varambally, S., Mehra, R., and Chinnaiyan, A. M. (2007). Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature*, 448(7153):595–599.

Tortorella, D., Gewurz, B. E., Furman, M. H., Schust, D. J., and Ploegh, H. L. (2000). Viral subversion of the immune system. *Annu Rev Immunol*, 18:861–926.

Trifonov, V. and Rabadan, R. (2010). Frequency analysis techniques for identification of viral genetic data. *MBio*, 1(3).

Trivedi, P., Takazawa, K., Zompetta, C., Cuomo, L., Anastasiadou, E., Carbone, A., Uccini, S., Belardelli, F., Takada, K., Frati, L., and Faggioni, A. (2004). Infection of HHV-8+ primary effusion lymphoma cells with a recombinant Epstein-Barr virus leads to restricted EBV latency, altered phenotype, and increased tumorigenicity without affecting TCL1 expression. *Blood*, 103(1):313–316.

Tsai, W.-L. and Chung, R. T. (2010). Viral hepatocarcinogenesis. *Oncogene*, 29(16):2309–2324.

Tsibris, A. M. N., Korber, B., Arnaout, R., Russ, C., Lo, C.-C., Leitner, T., Gaschen, B., Theiler, J., Paredes, R., Su, Z., Hughes, M. D., Gulick, R. M., Greaves, W., Coakley, E., Flexner, C., Nusbaum, C., and Kuritzkes, D. R. (2009). Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. *PLoS ONE*, 4(5):e5683.

Tulip, W. R., Varghese, J. N., Webster, R. G., Laver, W. G., and Colman, P. M. (1992). Crystal structures of two mutant neuraminidase-antibody complexes with amino acid substitutions in the interface. *J Mol Biol*, 227(1):149–159.

Turinsky, A. L., Razick, S., Turner, B., Donaldson, I. M., and Wodak, S. J. (2010). Literature curation of protein interactions: measuring agreement across major public databases. *Database (Oxford)*, 2010:baq026.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The Human Microbiome Project. *Nature*, 449(7164):804–810.

Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., Morrison, K., Donaldson, I. M., and Wodak, S. J. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database (Oxford)*, 2010:baq023.

Tzeng, S.-R. and Kalodimos, C. G. (2011). Protein dynamics and allostery: an NMR view. *Curr Opin Struct Biol*, 21(1):62–67.

Uetz, P., Dong, Y.-A., Zeretzke, C., Atzler, C., Baiker, A., Berger, B., Rajagopala, S. V., Roupelieva, M., Rose, D., Fossum, E., and Haas, J. (2006). Herpesviral protein networks and their interaction with the human proteome. *Science*, 311(5758):239–242.

Vajdic, C. M. and van Leeuwen, M. T. (2009). Cancer incidence and risk factors after solid organ transplantation. *Int. J. Cancer*, 125(8):1747–1754.

Vallejo, A., Molina-Pinelo, S., Abad, M. A., Gómez, G., Leal, M., Sánchez-Quijano, A., and Lissen, E. (2004). Analysis of quasispecies in the viral 5′ untranslated region of hepatitis C virus to evaluate ribavirin mutagenic effect in patients receiving ribavirin and interferon-alfa. *Eur. J. Clin. Microbiol. Infect. Dis.*, 23(12):923–926.

van Nood, E., Vrieze, A., Nieuwdorp, M., Fuentes, S., Zoetendal, E. G., de Vos, W. M., Visser, C. E., Kuijper, E. J., Bartelsman, J. F. W. M., Tijssen, J. G. P., Speelman, P., Dijkgraaf, M. G. W., and Keller, J. J. (2013). Duodenal infusion of donor feces for recurrent Clostridium difficile. *N. Engl. J. Med.*, 368(5):407–415.

Van Regenmortel, M. H. M. (1989). Virus species, a much overlooked but essential concept in virus classification. *Intervirology*, 31(5):241–254.

Van Regenmortel, M. H. M., Bishop, D. H. D., Fauquet, C. M. C., Mayo, M. A. M., Maniloff, J. J., and Calisher, C. H. C. (1996). Guidelines to the demarcation of virus species. *Arch Virol*, 142(7):1505–1518.

van Regenmortel, M. H. V. (2000). 7th Report of the International Committee on Taxonomy of Viruses. pages 3–16.

van Regenmortel, M. H. V. (2008). Encyclopedia of Virology. pages 398–402.

Van Regenmortel, M. H. V. M. (2003). Viruses are real, virus species are man-made, taxonomic constructions. *Arch Virol*, 148(12):2481–2488.

Varela, F. G., Maturana, H. R., and Uribe, R. (1974). Autopoiesis: the organization of living systems, its characterization and a model. *Curr Mod Biol*, 5(4):187–196.

Varela, I., Tarpey, P., Raine, K., Huang, D., Ong, C. K., Stephens, P., Davies, H., Jones, D., Lin, M.-L., Teague, J., Bignell, G., Butler, A., Cho, J., Dalgliesh, G. L., Galappaththige, D., Greenman, C., Hardy, C., Jia, M., Latimer, C., Lau, K. W., Marshall, J., McLaren, S., Menzies, A., Mudie, L., Stebbings, L., Largaespada, D. A., Wessels, L. F. A., Richard, S., Kahnoski, R. J., Anema, J., Tuveson, D. A., Perez-Mancera, P. A., Mustonen, V., Fischer, A., Adams, D. J., Rust, A., Chanon, W., Subimerb, C., Dykema, K., Furge, K., Campbell, P. J., Teh, B. T., Stratton, M. R., and Futreal, P. A. (2011). Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature*, 469(7331):539–542.

Varghese, J. N., Webster, R. G., Laver, W. G., and Colman, P. M. (1988). Structure of an escape mutant of glycoprotein N2 neuraminidase of influenza virus A/Tokyo/3/67 at 3 A. *J Mol Biol*, 200(1):201–203.

Varghese, V., Shahriar, R., Rhee, S.-Y., Liu, T., Simen, B. B., Egholm, M., Hanczaruk, B., Blake, L. A., Gharizadeh, B., Babrzadeh, F., Bachmann, M. H., Fessel, W. J., and Shafer, R. W. (2009). Minority variants associated with transmitted and acquired HIV-1 nonnucleoside reverse transcriptase inhibitor resistance: implications for the use of second-generation nonnucleoside reverse transcriptase inhibitors. *J. Acquir. Immune Defic. Syndr.*, 52(3):309–315.

Vasileva, A. A. and Jessberger, R. R. (2005). Precise hit: adeno-associated virus in gene targeting. *Nat Rev Microbiol*, 3(11):837–847.

Vella, S. (1994). HIV therapy advances. Update on a proteinase inhibitor. *AIDS*, 8 Suppl 3:S25–9.

Vembar, S. S. and Brodsky, J. L. (2008). One step at a time: endoplasmic reticulum-associated degradation. *Nat Rev Mol Cell Biol*, 9(12):944–957.

Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.-I., Yildirim, M. A., Simonis, N., Heinzmann, K., Gebreab, F., Sahalie, J. M., Cevik, S., Simon, C., de Smet, A.-S., Dann, E., Smolyar, A., Vinayagam, A., Yu, H., Szeto, D., Borick, H., Dricot, A., Klitgord, N., Murray, R. R., Lin, C., Lalowski, M., Timm, J., Rau, K., Boone, C., Braun, P., Cusick, M. E., Roth, F. P., Hill, D. E., Tavernier, J., Wanker, E. E., Barabási, A.-L., and Vidal, M. (2009). An empirical framework for binary interactome mapping. *Nat Methods*, 6(1):83–90.

Veraksa, A., Bauer, A., and Artavanis-Tsakonas, S. (2005). Analyzing protein complexes in Drosophila with tandem affinity purification-mass spectrometry. *Dev. Dyn.*, 232(3):827–834.

Verbinnen, T., Van Marck, H., Vandenbroucke, I., Vijgen, L., Claes, M., Lin, T.-I., Simmen, K., Neyts, J., Fanning, G., and Lenz, O. (2010). Tracking the evolution of multiple in vitro hepatitis C virus replicon variants under protease inhibitor selection pressure by 454 deep sequencing. *J Virol*, 84(21):11124–11133.

Verma, S. C., Borah, S., and Robertson, E. S. (2004). Latency-associated nuclear antigen of Kaposi's sarcoma-associated herpesvirus up-regulates transcription of human telomerase reverse transcriptase promoter through interaction with transcription factor Sp1. *J Virol*, 78(19):10348–10359.

Verma, S. C., Choudhuri, T., and Robertson, E. S. (2007). The minimal replicator element of the Kaposi's sarcoma-associated herpesvirus terminal repeat supports replication in a semiconservative and cell-cycle-dependent manner. *J Virol*, 81(7):3402–3413.

Vetsigian, K., Woese, C., and Goldenfeld, N. (2006). Collective evolution and the genetic code. *Proc Natl Acad Sci USA*, 103(28):10696–10701.

Victoria, J. G., Kapoor, A., Li, L., Blinkova, O., Slikas, B., Wang, C., Naeem, A., Zaidi, S., and Delwart, E. (2009). Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *J Virol*, 83(9):4642–4651.

Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E., and Andino, R. (2006). Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature*, 439(7074):344–348.

Vignuzzi, M., Wendt, E., and Andino, R. (2008). Engineering attenuated virus vaccines by controlling replication fidelity. *Nat Med*, 14(2):154–161.

Vilchez, R. A. and Butel, J. S. (2004). Emergent human pathogen simian virus 40 and its role in cancer. *Clin. Microbiol. Rev.*, 17(3):495–508.

Virgin, H. W. and Walker, B. D. (2010). Immunology and the elusive AIDS vaccine. *Nature*, 464(7286):224–231.

Virgin, H. W., Wherry, E. J., and Ahmed, R. (2009). Redefining Chronic Viral Infection. *Cell*, 138(1):21–21.

Vo, N. V., Young, K.-C., and Lai, M. M. C. (2003). Mutagenic and inhibitory effects of ribavirin on hepatitis C virus RNA polymerase. *Biochemistry*, 42(35):10462–10471.

Vogelstein, B., Fearon, E. R., Kern, S. E., Hamilton, S. R., Preisinger, A. C., Nakamura, Y., and White, R. (1989). Allelotype of colorectal carcinomas. *Science*, 244(4901):207–211.

Voisset, C., Weiss, R. A., and Griffiths, D. J. (2008). Human RNA "rumor" viruses: the search for novel human retroviruses in chronic disease. *Microbiol. Mol. Biol. Rev.*, 72(1):157–196.

Vollmer, J., Rankin, R., Hartmann, H., Jurk, M., Samulowitz, U., Wader, T., Janosch, A., Schetter, C., and Krieg, A. M. (2004). Immunopharmacology of CpG oligodeoxynucleotides and ribavirin. *Antimicrob Agents Chemother*, 48(6):2314–2317.

von Brunn, A., Teepe, C., Simpson, J. C., Pepperkok, R., Friedel, C. C., Zimmer, R., Roberts, R., Baric, R., and Haas, J. (2007). Analysis of intraviral protein-protein interactions of the SARS coronavirus ORFeome. *PLoS ONE*, 2(5):e459.

von Hahn, T., Yoon, J. C., Alter, H., Rice, C. M., Rehermann, B., Balfe, P., and McKeating, J. A. (2007). Hepatitis C virus continuously escapes from neutralizing antibody and T-cell responses during chronic infection in vivo. *Gastroenterology*, 132(2):667–678.

von Neumann, J. (1966). Theory of Self-Reproducing Automata. In Burks, A. W., editor, *Essays on Cellular Automata*. University of Illinois Press, Urbana.

von Schwedler, U. K., Stuchell, M., Müller, B., Ward, D. M., Chung, H.-Y., Morita, E., Wang, H. E., Davis, T., He, G.-P., Cimbora, D. M., Scott, A., Kräusslich, H.-G., Kaplan, J., Morham, S. G., and Sundquist, W. I. (2003). The protein network of HIV budding. *Cell*, 114(6):701–713.

Wainberg, M. A. (2008). Perspectives on antiviral drug development. *Antiviral Res*, 81(1):5–5.

Walker, C. M. (2010). Adaptive immunity to the hepatitis C virus. *Adv Virus Res*, 78:43–86.

Walker, L. M., Phogat, S. K., Chan-Hui, P.-Y., Wagner, D., Phung, P., Goss, J. L., Wrin, T., Simek, M. D., Fling, S., Mitcham, J. L., Lehrman, J. K., Priddy, F. H., Olsen, O. A., Frey, S. M., Hammond, P. W., Protocol G Principal Investigators, Kaminsky, S., Zamb, T., Moyle, M., Koff, W. C., Poignard, P., and Burton, D. R. (2009). Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science*, 326(5950):285–289.

Wang, C., Mitsuya, Y., Gharizadeh, B., Ronaghi, M., and Shafer, R. W. (2007). Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res*, 17(8):1195–1201.

Wang, D., Urisman, A., Liu, Y.-T., Springer, M., Ksiazek, T. G., Erdman, D. D., Mardis, E. R., Hickenbotham, M., Magrini, V., Eldred, J., Latreille, J. P., Wilson, R. K., Ganem, D., and DeRisi, J. L. (2003). Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol*, 1(2):E2.

Wang, L.-F. (2011). Discovering novel zoonotic viruses. *N S W Public Health Bull*, 22(5-6):113–117.

Wang, R. Y. and Li, K. (2012). Host factors in the replication of positive-strand RNA viruses. *Chang Gung Med J*, 35(2):111–124.

Wang-Johanning, F., Frost, A. R., Jian, B., Epp, L., Lu, D. W., and Johanning, G. L. (2003). Quantitation of HERV-K env gene expression and splicing in human breast cancer. *Oncogene*, 22(10):1528–1535.

Wang-Johanning, F., Radvanyi, L., Rycaj, K., Plummer, J. B., Yan, P., Sastry, K. J., Piyathilake, C. J., Hunt, K. K., and Johanning, G. L. (2008). Human endogenous retrovirus K triggers an antigen-specific immune response in breast cancer patients. *Cancer Research*, 68(14):5869–5877.

Ward, C. L., Dev, A., Rigby, S., Symonds, W. T., Patel, K., Zekry, A., Pawlotsky, J.-M., and McHutchison, J. G. (2008). Interferon and ribavirin therapy does not select for resistance mutations in hepatitis C virus polymerase. *J Viral Hepat*, 15(8):571–577.

Wargo, A. R., Huijben, S., de Roode, J. C., Shepherd, J., and Read, A. F. (2007). Competitive release and facilitation of drug-resistant parasites after therapeutic chemotherapy in a rodent malaria model. *Proc Natl Acad Sci USA*, 104(50):19914–19919.

Warren, R. A. (1980). Modified bases in bacteriophage DNAs. *Annu Rev Microbiol*, 34:137–158.

Watkins, W. J., Ray, A. S., and Chong, L. S. (2010). HCV NS5B polymerase inhibitors. *Curr Opin Drug Discov Devel*, 13(4):441–465.

Watson, J. D. (1954). The structure of tobacco mosaic virus. I. X-ray evidence of a helical arrangement of sub-units around the longitudinal axis. *Biochim. Biophys. Acta*, 13(1):10–19.

Weaver, S. C. S., Brault, A. C. A., Kang, W. W., and Holland, J. J. J. (1999). Genetic and fitness changes accompanying adaptation of an arbovirus to vertebrate and invertebrate cells. *J Virol*, 73(5):4316–4326.

Weber, G., Shendure, J., Tanenbaum, D. M., Church, G. M., and Meyerson, M. (2002). Identification of foreign gene sequences by transcript filtering against the human genome. *Nat Genet*, 30(2):141–142.

Weber, T. and Major, E. O. (1997). Progressive multifocal leukoencephalopathy: molecular biology, pathogenesis and clinical impact. *Intervirology*, 40(2-3):98–111.

Webster, D. R., Hekele, A. G., Lauring, A. S., Fischer, K. F., Li, H., Andino, R., and DeRisi, J. L. (2009). An enhanced single base extension technique for the analysis of complex viral populations. *PLoS ONE*, 4(10):e7453.

Wedemeyer, H., He, X.-S., Nascimbeni, M., Davis, A. R., Greenberg, H. B., Hoofnagle, J. H., Liang, T. J., Alter, H., and Rehermann, B. (2002). Impaired effector function of hepatitis C virus-specific CD8+ T cells in chronic hepatitis C virus infection. *J. Immunol.*, 169(6):3447–3458.

Wei, Z., Wang, W., Hu, P., Lyon, G. J., and Hakonarson, H. (2011). SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res*, 39(19):e132–e132.

Weinberg, R. A. (1989). Oncogenes, antioncogenes, and the molecular bases of multistep carcinogenesis. *Cancer Research*, 49(14):3713–3721.

Weitzman, M. D., Lilley, C. E., and Chaurushiya, M. S. (2010). Genomes in conflict: maintaining genome integrity during virus infection. *Annu Rev Microbiol*, 64:61–81.

Westby, M., Lewis, M., Whitcomb, J., Youle, M., Pozniak, A. L., James, I. T., Jenkins, T. M., Perros, M., and van der Ryst, E. (2006). Emergence of CXCR4-using human immunodeficiency virus type 1 (HIV-1) variants in a minority of HIV-1-infected patients following treatment with the CCR5 antagonist maraviroc is from a pretreatment CXCR4-using virus reservoir. *J Virol*, 80(10):4909–4920.

Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA*, 95(12):6578–6583.

WHO (2003). Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. WHO Library.

Wiertz, E. J. E., Devlin, R. R., Collins, H. L. H., and Ressing, M. E. M. (2007). Herpesvirus interference with major histocompatibility complex class II-restricted T-cell activation. *J Virol*, 81(9):4389–4396.

Wilhelm, S. W. (1999). Viruses and nutrient cycles in the sea. *Bioscience*, (49):781–788.

Wilke, C. O. (2005). Quasispecies theory in the context of population genetics. *BMC Evol Biol*, 5:44.

Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E., and Adami, C. (2001). Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333.

Williams, T. A., Embley, T. M., and Heinz, E. (2010). Informational gene phylogenies do not support a fourth domain of life for nucleocytoplasmic large DNA viruses. *PLoS ONE*, 6(6):e21080–e21080.

Williamson, S. J., Allen, L. Z., Lorenzi, H. A., Fadrosh, D. W., Brami, D., Thiagarajan, M., McCrow, J. P., Tovchigrechko, A., Yooseph, S., and Venter, J. C. (2012). Metagenomic Exploration of Viruses throughout the Indian Ocean. *PLoS ONE*, 7(10):e42047.

Williamson, S. J., Rusch, D. B., Yooseph, S., Halpern, A. L., Heidelberg, K. B., Glass, J. I., Andrews-Pfannkoch, C., Fadrosh, D., Miller, C. S., Sutton, G., Frazier, M., and Venter, J. C. (2008). The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE*, 3(1):e1456.

Willner, D., Furlan, M., Haynes, M., Schmieder, R., Angly, F. E., Silva, J., Tammadoni, S., Nosrat, B., Conrad, D., and Rohwer, F. (2009). Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE*, 4(10):e7370.

Willner, D., Furlan, M., Schmieder, R., Grasis, J. A., Pride, D. T., Relman, D. A., Angly, F. E., McDole, T., Mariella, R. P., Rohwer, F., and Haynes, M. (2011). Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. *Proc Natl Acad Sci USA*, 108 Suppl 1:4547–4553.

Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., Khor, C. C., Petric, R., Hibberd, M. L., and Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res*, 40(22):11189–11201.

Winnenburg, R., Urban, M., Beacham, A., Baldwin, T. K., Holland, S., Lindeberg, M., Hansen, H., Rawlings, C., Hammond-Kosack, K. E., and Köhler, J. (2008). PHI-base update: additions to the pathogen host interaction database. *Nucleic Acids Res*, 36(Database issue):D572–6.

Wintersberger, U. and Wintersberger, E. (1987). RNA makes DNA: a speculative view of the evolution of DNA replication mechanisms. *Trends in Genetics*, 3:198–202.

Wodak, S. J., Pu, S., Vlasblom, J., and Séraphin, B. (2009). Challenges and rewards of interaction proteomics. *Mol Cell Proteomics*, 8(1):3–18.

Woese, C. R. (1987). Bacterial evolution. *Microbiol. Rev.*, 51(2):221–271.

Woese, C. R. (2000). Interpreting the universal phylogenetic tree. *Proc Natl Acad Sci USA*, 97(15):8392–8396.

Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA*, 74(11):5088–5090.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA*, 87(12):4576–4579.

Wommack, K. E., Bhavsar, J., Polson, S. W., Chen, J., Dumas, M., Srinivasiah, S., Furman, M., Jamindar, S., and Nasko, D. J. (2012). VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand. Genomic Sci.*, 6(3):427–439.

Woodhouse, S. D. S., Narayan, R. R., Latham, S. S., Lee, S. S., Antrobus, R. R., Gangadharan, B. B., Luo, S. S., Schroth, G. P. G., Klenerman, P. P., and Zitzmann, N. N. (2010). Transcriptome sequencing, microarray, and proteomic analyses reveal cellular and metabolic impact of hepatitis C virus infection in vitro. *Hepatology (Baltimore, Md)*, 52(2):443–453.

Woolhouse, M., Scott, F., Hudson, Z., Howey, R., and Chase-Topping, M. (2012). Human viruses: discovery and emergence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1604):2864–2871.

Woolhouse, M. E. J. (2001). Population Biology of Multihost Pathogens. *Science*, 292(5519):1109–1112.

Woolhouse, M. E. J., Haydon, D. T., and Antia, R. (2005). Emerging pathogens: the epidemiology and evolution of species jumps. *Trends Ecol. Evol. (Amst.)*, 20(5):238–244.

Woolhouse, M. E. J., Howey, R., Gaunt, E., Reilly, L., Chase-Topping, M., and Savill, N. (2008). Temporal trends in the discovery of human viruses. *Proc. Biol. Sci.*, 275(1647):2111–2115.

Woolhouse, M. E. J. M. and Gowtage-Sequeria, S. S. (2005). Host range and emerging and reemerging pathogens. *Emerging Infect. Dis.*, 11(12):1842–1847.

Woolhouse, M. E. J. M., Webster, J. P. J., Domingo, E. E., Charlesworth, B. B., and Levin, B. R. B. (2002). Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat Genet*, 32(4):569–577.

Woolhouse, M. M. and Gaunt, E. E. (2006). Ecological origins of novel human pathogens. *Crit. Rev. Microbiol.*, 33(4):231–242.

Wright, C. F., Morelli, M. J., Thébaud, G., Knowles, N. J., Herzyk, P., Paton, D. J., Haydon, D. T., and King, D. P. (2011). Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. *J Virol*, 85(5):2266–2275.

Wu, J. Z., Walker, H., Lau, J. Y. N., and Hong, Z. (2003). Activation and deactivation of a broad-spectrum antiviral drug by a single enzyme: adenosine deaminase catalyzes two consecutive deamination reactions. *Antimicrob Agents Chemother*, 47(1):426–431.

Wu, M. and Scott, A. J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics*, 28(7):1033–1034.

Wu, Q., Luo, Y., Lu, R., Lau, N., Lai, E. C., Li, W.-X., and Ding, S.-W. (2010). Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. *Proc Natl Acad Sci USA*, 107(4):1606–1611.

Wuchty, S. (2011). Computational prediction of host-parasite protein interactions between P. falciparum and H. sapiens. *PLoS ONE*, 6(11):e26960.

Wyles, D. L. D. (2013). Antiviral resistance and the future landscape of hepatitis C virus infection therapy. *J Infect Dis*, 207 Suppl 1:S33–S39.

Wylie, K. M., Weinstock, G. M., and Storch, G. A. (2013). Virome genomics: a tool for defining the human virome. *Curr Opin Microbiol*, 16(4):479–484.

Xia, J., Chen, X., Zhou, C. Z., Li, Y. G., and Peng, Z. H. (2011). Development of a low-cost magnetic microfluidic chip for circulating tumour cell capture. *IET Nanobiotechnol*, 5(4):114–120.

Xiang, Z., Tian, Y., and He, Y. (2007). PHIDIAS: a pathogen-host interaction data integration and analysis system. *Genome Biol*, 8(7):R150.

Xiao, A., Wong, J., and Luo, H. (2010). Viral interaction with molecular chaperones: role in regulating viral infection. *Arch Virol*, 155(7):1021–1031.

Xie, G., Chain, P. S. G., Lo, C. C., Liu, K. L., Gans, J., Merritt, J., and Qi, F. (2010). Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing. *Molecular Oral Microbiology*, 25(6):391–405.

Xiong, Y. and Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO Journal*, 9(10):3353–3362.

Xu, F., Wang, W., Wang, P., Jun Li, M., Chung Sham, P., and Wang, J. (2012a). A fast and accurate SNP detection algorithm for next-generation sequencing data. *Nat Commun*, 3:1258.

Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., He, W., Zeng, L., Xing, M., Wu, R., Jiang, H., Liu, X., Cao, D., Guo, G., Hu, X., Gui, Y., Li, Z., Xie, W., Sun, X., Shi, M., Cai, Z., Wang, B., Zhong, M., Li, J., Lu, Z., Gu, N., Zhang, X., Goodman, L., Bolund, L., Wang, J., Yang, H., Kristiansen, K., Dean, M., Li, Y., and Wang, J. (2012b). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, 148(5):886–895.

Xu, Y., Stange-Thomann, N., Weber, G., Bo, R., Dodge, S., David, R. G., Foley, K., Beheshti, J., Harris, N. L., Birren, B., Lander, E. S., and Meyerson, M. (2003). Pathogen discovery from human tissue by sequence-based computational subtraction. *Genomics*, 81(3):329–335.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R. H., Eshleman, J. R., Nowak, M. A., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., and Iacobuzio-Donahue, C. A. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*, 467(7319):1114–1117.

Yalcin, B., Wong, K., Agam, A., Goodson, M., Keane, T. M., Gan, X., Nellåker, C., Goodstadt, L., Nicod, J., Bhomra, A., Hernandez-Pliego, P., Whitley, H., Cleak, J., Dutton, R., Janowitz, D., Mott, R., Adams, D. J., and Flint, J. (2011). Sequence-based characterization of structural variation in the mouse genome. *Nature*, 477(7364):326–329.

Yam, A. Y. W. (2005). Hsp110 Cooperates with Different Cytosolic HSP70 Systems in a Pathway for de Novo Folding. *J Biol Chem*, 280(50):41252–41261.

Yang, C.-W. (2012). A comparative study of short linear motif compositions of the influenza A virus ribonucleoproteins. *PLoS ONE*, 7(6):e38637.

Yang, J., Yang, F., Ren, L., Xiong, Z., Wu, Z., Dong, J., Sun, L., Zhang, T., Hu, Y., Du, J., Wang, J., and Jin, Q. (2011a). Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *J Clin Microbiol*, 49(10):3463–3469.

Yang, X., Charlebois, P., Gnerre, S., Coole, M. G., Lennon, N. J., Levin, J. Z., Qu, J., Ryan, E. M., Zody, M. C., and Henn, M. R. (2012). De novo assembly of highly diverse viral populations. *BMC Genomics*, 13(1):1–13.

Yang, X., Dorman, K. S., and Aluru, S. (2010a). Reptile: representative tiling for short read error correction. *Bioinformatics*, 26(20):2526–2533.

Yang, Z. Z., Bruno, D. P. D., Martens, C. A. C., Porcella, S. F. S., and Moss, B. B. (2010b). Simultaneous high-resolution analysis of vaccinia virus and host cell transcriptomes by deep RNA sequencing. *Proc Natl Acad Sci USA*, 107(25):11513–11518.

Yang, Z. Z., Reynolds, S. E. S., Martens, C. A. C., Bruno, D. P. D., Porcella, S. F. S., and Moss, B. B. (2011b). Expression profiling of the intermediate and late stages of poxvirus replication. *J Virol*, 85(19):9899–9908.

Yap, T. A., Omlin, A., and de Bono, J. S. (2013). Development of therapeutic combinations targeting major cancer signaling pathways. *J. Clin. Oncol.*, 31(12):1592–1605.

Yasunaga, J.-i. and Matsuoka, M. (2007). Leukaemogenic mechanism of human T-cell leukaemia virus type I. *Rev. Med. Virol.*, 17(5):301–311.

Ye, C., Cannon, C. H., Ma, Z. S., Yu, D. W., and Pop, M. (2011). SparseAssembler2: Sparse k-mer Graph for Memory Efficient Genome Assembly. *arXiv*.

Ye, Y. and Tang, H. (2009). An orfome assembly approach to metagenomics sequences analysis. *J Bioinform Comput Biol*, 7(3):455–471.

Yedavalli, V. S. R. K., Neuveut, C., Chi, Y.-H., Kleiman, L., and Jeang, K.-T. (2004). Requirement of DDX3 DEAD box RNA helicase for HIV-1 Rev-RRE export function. *Cell*, 119(3):381–392.

Yellaboina, S., Tasneem, A., Zaykin, D. V., Raghavachari, B., and Jothi, R. (2011). DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res*, 39(Database issue):D730–5.

Yeung, M. L., Houzet, L., Yedavalli, V. S. R. K., and Jeang, K.-T. (2009). A genome-wide short hairpin RNA screening of jurkat T-cells for human proteins contributing to productive HIV-1 replication. *J Biol Chem*, 284(29):19463–19473.

Yin, Y. and Fischer, D. (2008). Identification and investigation of ORFans in the viral world. *BMC Genomics*, 9:24.

Yoder, J. A., Walsh, C. P., and Bestor, T. H. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet*, 13(8):335–340.

Yost, S. E., Alakus, H., Matsui, H., Schwab, R. B., Jepsen, K., Frazer, K. A., and Harismendy, O. (2013). Mutascope: sensitive detection of somatic mutations from deep amplicon sequencing. *Bioinformatics*, 29(15):1908–1909.

Young, K.-C., Lindsay, K. L., Lee, K.-J., Liu, W.-C., He, J.-W., Milstein, S. L., and Lai, M. M. C. (2003). Identification of a ribavirin-resistant NS5B mutation of hepatitis C virus during ribavirin monotherapy. *Hepatology (Baltimore, Md)*, 38(4):869–878.

Young, L. S. and Rickinson, A. B. (2004). Epstein-Barr virus: 40 years on. *Nat Rev Cancer*, 4(10):757–768.

Yozwiak, N. L., Skewes-Cox, P., Stenglein, M. D., Balmaseda, A., Harris, E., and DeRisi, J. L. (2012). Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Negl Trop Dis*, 6(2):e1485.

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabási, A.-L., Tavernier, J., Hill, D. E., and Vidal, M. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110.

Yu, X., Ivanic, J., Memisević, V., Wallqvist, A., and Reifman, J. (2011). Categorizing biases in high-confidence high-throughput protein-protein interaction data sets. *Mol Cell Proteomics*, 10(12):M111.012500.

Yu, X., Ivanic, J., Wallqvist, A., and Reifman, J. (2009). A novel scoring approach for protein co-purification data reveals high interaction specificity. *PLoS Comput Biol*, 5(9):e1000515.

Yuste, E., Sánchez-Palomino, S., Casado, C., Domingo, E., and López-Galíndez, C. (1999). Drastic fitness loss in human immunodeficiency virus type 1 upon serial bottleneck events. *J Virol*, 73(4):2745–2751.

Zagordi, O., Klein, R., Däumer, M., and Beerenwinkel, N. (2010). Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res*, 38(21):7400–7409.

Zaldumbide, A., Ossevoort, M., Wiertz, E. J. H. J., and Hoeben, R. C. (2007). In cis inhibition of antigen processing by the latency-associated nuclear antigen I of Kaposi sarcoma herpes virus. *Mol. Immunol.*, 44(6):1352–1360.

Zampieri, C. A., Sullivan, N. J., and Nabel, G. J. (2007). Immunopathology of highly virulent pathogens: insights from Ebola virus. *Nat. Immunol.*, 8(11):1159–1164.

Zandi, R. R., Reguera, D. D., Bruinsma, R. F. R., Gelbart, W. M. W., and Rudnick, J. J. (2004). Origin of icosahedral symmetry in viruses. *Proc Natl Acad Sci U S A*, 101(44):15556–15560.

Zeller, S. J. and Kumar, P. (2011). RNA-based gene therapy for the treatment and prevention of HIV: from bench to bedside. *Yale J Biol Med*, 84(3):301–309.

Zerbino, D. R. (2009). *Genome assembly and comparison using de Bruijn graphs*. PhD thesis.

Zerbino, D. R. and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, 18(5):821–829.

Zeuzem, S., Buggisch, P., Agarwal, K., Marcellin, P., Sereni, D., Klinker, H., Moreno, C., Zarski, J.-P., Horsmans, Y., Mo, H., Arterburn, S., Knox, S., Oldach, D., McHutchison, J. G., Manns, M. P., and Foster, G. R. (2012). The protease inhibitor, GS-9256, and non-nucleoside polymerase inhibitor tegobuvir alone, with ribavirin, or pegylated interferon plus ribavirin in hepatitis C. *Hepatology (Baltimore, Md)*, 55(3):749–758.

Zeuzem, S., Schmidt, J. M., Lee, J. H., von Wagner, M., Teuber, G., and Roth, W. K. (1998). Hepatitis C virus dynamics in vivo: effect of ribavirin and interferon alfa on viral turnover. *Hepatology (Baltimore, Md)*, 28(1):245–252.

Zeyaullah, M. M., Patro, M. M., Ahmad, I. I., Ibraheem, K. K., Sultan, P. P., Nehal, M. M., and Ali, A. A. (2012). Oncolytic viruses in the treatment of cancer: a review of current strategies. *Pathol. Oncol. Res.*, 18(4):771–781.

Zhang, L. Q., Simmonds, P., Ludlam, C. A., and Brown, A. J. (1991). Detection, quantification and sequencing of HIV-1 from the plasma of seropositive individuals and from factor VIII concentrates. *AIDS*, 5(6):675–681.

Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., Maniatis, T., Califano, A., and Honig, B. (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, 490(7421):556–560.

Zhang, T., Breitbart, M., Lee, W. H., Run, J.-Q., Wei, C. L., Soh, S. W. L., Hibberd, M. L., Liu, E. T., Rohwer, F., and Ruan, Y. (2006). RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol*, 4(1):e3.

Zhang, Y. and Sun, Y. (2011). HMM-FRAME: accurate protein domain classification for metagenomic sequences containing frameshift errors. *BMC Bioinformatics*, 12:198.

Zhang, Y., Zheng, N., and Zhong, Y. (2007). Computational characterization and design of SARS coronavirus receptor recognition and antibody neutralization. *Comput Biol Chem*, 31(2):5–5.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *J Comput Biol*, 7(1-2):203–214.

Zhao, Q.-Y., Wang, Y., Kong, Y.-M., Luo, D., Li, X., and Hao, P. (2011a). Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics*, 12 Suppl 14:S2.

Zhao, Z., Xia, J., Tastan, O., Singh, I., Kshirsagar, M., Carbonell, J., and Klein-Seetharaman, J. (2011b). Virus interactions with human signal transduction pathways. *Int J Comput Biol Drug Des*, 4(1):83–105.

Zhaxybayeva, O. and Doolittle, W. F. (2011). Lateral gene transfer. *Curr. Biol.*, 21(7):R242–6.

Zheng, W., Chung, L. M., and Zhao, H. (2011). Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics*, 12(1):290.

Zhou, H., Xu, M., Huang, Q., Gates, A. T., Zhang, X. D., Castle, J. C., Stec, E., Ferrer, M., Strulovici, B., Hazuda, D. J., and Espeseth, A. S. (2008a). Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host Microbe*, 4(5):495–504.

Zhou, S., Liu, R., Baroudy, B. M., Malcolm, B. A., and Reyes, G. R. (2003). The effect of ribavirin and IMPDH inhibitors on hepatitis C virus subgenomic replicon RNA. *Virology*, 310(2):333–342.

Zhou, Y., Bartels, D. J., Hanzelka, B. L., Müh, U., Wei, Y., Chu, H.-M., Tigges, A. M., Brennan, D. L., Rao, B. G., Swenson, L., Kwong, A. D., and Lin, C. (2008b). Phenotypic characterization of resistant Val36 variants of hepatitis C virus NS3-4A serine protease. *Antimicrob Agents Chemother*, 52(1):110–120.

Zhou, Y., Müh, U., Hanzelka, B. L., Bartels, D. J., Wei, Y., Rao, B. G., Brennan, D. L., Tigges, A. M., Swenson, L., Kwong, A. D., and Lin, C. (2007). Phenotypic and structural analyses of hepatitis C virus NS3 protease Arg155 variants: sensitivity to telaprevir (VX-950) and interferon alpha. *J Biol Chem*, 282(31):22619–22628.

Zhu, T., Mo, H., Wang, N., Nam, D. S., Cao, Y., Koup, R. A., and Ho, D. D. (1993). Genotypic and phenotypic characterization of HIV-1 patients with primary infection. *Science*, 261(5125):1179–1181.

Zhu, X.-D., Li, C.-L., Lang, Z.-W., Gao, G. F., and Tien, P. (2004). Significant correlation between expression level of HSP gp96 and progression of hepatitis B virus induced diseases. *World J Gastroenterol*, 10(8):1141–1145.

Zinger, L., Amaral-Zettler, L. A., Fuhrman, J. A., Horner-Devine, M. C., Huse, S. M., Welch, D. B. M., Martiny, J. B. H., Sogin, M., Boetius, A., and Ramette, A. (2011). Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS ONE*, 6(9):e24570.

Zou, P. P., Isegawa, Y. Y., Nakano, K. K., Haque, M. M., Horiguchi, Y. Y., and Yamanishi, K. K. (1999). Human herpesvirus 6 open reading frame U83 encodes a functional chemokine. *J Virol*, 73(7):5926–5933.

zur Hausen, H. (2001). Oncogenic DNA viruses. *Oncogene*, 20(54):7820–7823.

zur Hausen, H. (2002). Papillomaviruses and cancer: from basic studies to clinical application. *Nat Rev Cancer*, 2(5):342–350.

zur Hausen, H. (2006). *Infections Causing Human Cancer*. Wiley-VCH, Weinheim.

zur Hausen, H. (2009a). Childhood leukemias and other hematopoietic malignancies: interdependence between an infectious event and chromosomal modifications. *Int. J. Cancer*, 125(8):1764–1770.

zur Hausen, H. (2009b). The search for infectious causes of human cancers: where and why. *Virology*, 392(1):1–10.

zur Hausen, H. (2012). Red meat consumption and cancer: reasons to suspect involvement of bovine infectious factors in colorectal cancer. *Int. J. Cancer*, 130(11):2475–2483.

zur Hausen, H. and de Villiers, E. M. (2009). TT viruses: oncogenic or tumor-suppressive properties? *Curr. Top. Microbiol. Immunol.*, 331:109–116.