Identifying Biological Associations from High-Throughput Datasets

Dissertation
zur Erlangung des Grades des
Doktors der Ingenieurswissenschaften (Dr.-Ing.) der
Naturwissenschaftlich-Technischen Fakultäten
der Universität des Saarlandes

Ahmad Barghash

Saarbrücken
September 2015

Dekan - Dean:

Prof. Dr. Markus Bläser
Saarland University
Saarbrücken, Germany

Kolloquiums - Defense

Datum - Date
September 30, 2015, in Saarbrücken

Vorsitzender - Head of Colloquium:
Prof. Dr. Raimund Seidel

Prüfer - Examiners:

Prof. Dr. Volkhard Helms

Prof. Dr. Tobias Marschall

Akademischer Beisitzer - Scientific Assistant:

Dr. Mazen Ahmad

Acknowledgements
All the praises and thanks be to Allah the "Most Beneficent" and "Most Merciful" for his help, guidance, and all the gifts I was granted.

First, I want to thank my supervisor Prof. Volkhard Helms for his support, ideas, motivation, and fruitful discussions during my thesis. I really can not count what I learned from you during my master and PhD study. Most of my research principles are based on your advices and I even have some of your words quoted and saved in my notebook. I want also to give special thanks for making the Helms group a comfortable and easygoing research environment for us. Although, I will miss being part of the Helms group, I will be happy if I can use what I learned from you to create a similar environment later for my students.

I want also to provide my deep appreciations to Dr. Sonja Keßler and Prof. Alexandra Kiemer for their ideas and friendly collaborations in the highly productive cancer analysis projects.

I owe special thanks to the friends and colleagues I met in the Helms group. Many thanks to Dr. Nadine Schaadt for the fruitful discussions in the early days. Also many thanks to Dr. Siba Shanaq for the wonderful collaboration and for the continuous help. Many thanks to the other members of the Helms group especially Kerstin Gronow-Pudelek for the endless support up from the first days in Saarbrücken.

My biggest thanks go to my family especially my father. My gratitude can never be enough for what you have given me till now and the nights you spent caring for my own benefit. My dear brothers, sisters, and my dear daughter "Hala", this would not have worked without you. Of course I do not forget to send many mercy prayers for the soul of my late mother.

I want to give special thanks to the friends I met in Saarbrücken. Although it is hard to list your names here, still I have a wonderful memory from each of you.

Finally, I want to acknowledge the German Jordanian University (GJU), DFG SFB1027, and the graduate school of computer science in Saarland University for funding part of my PhD study.

Abstract

High-throughput biological datasets are the basis for most modern basic research in the fields of genomics, systems biology, and disease diagnostics. Currently, one sample can contain thousands of measurements in some datasets. The omnipresence of such huge datasets created the urgent need for efficient and robust computational approaches to handle and analyze such database and to identify informative associations.

This thesis deals with different types of large scale datasets and aims to identify with high confidence underlying biological associations. Our computational approach consists of four core parts. In the first part, we analyzed amino acid datasets of membrane transporters from different organisms for the purpose of transferring functional annotations of the transporters across species. Here, we mapped the experimentally validated functions of one protein to another one from a different organism based on their sequence similarity. Sequence similarity results in this work were combinations of similarity decisions of several tools (BLAST, HMMER, MEME). Initially, we defined confidence thresholds and which we then applied for predictions. We found that, up to certain thresholds, membrane transporters sharing high sequence similarity have similar functions or transporting mechanisms even if they exist in different organisms.

Our second computational approach was designed to deal with expression and methylation datasets. We found that expression and methylation datasets often suffer from outliers at gene or sample levels. Performing analyses before dealing with outliers might lead to misleading results. Thus, we present an approach that includes several outlier detection algorithms for detecting sample and gene outliers in expression/methylation datasets. As some outlier algorithms report at least one outlier value even if there is none, we first defined the margin of allowed outlier observations. We tested how many outlier observations are needed to ruin a perfect co-expression and then fixed that threshold for the rest of analyses. Additionally, in this work we considered the distribution underlying the gene expression/methylation before outlier detection. However, outliers might carry useful information. Therefore, we labelled only extreme outliers for removal and marked those possibly carrying useful information for further analysis.

In the next step, we used published expression and methylation datasets from GEO to analyse and confirm possible tumor markers for HCC, liver diseases, and breast cancer. These were later validated in the wetlab through our collaboration with the group of Prof. Kiemer in pharmacy. In addition to their possible roles in the change of survival rates, we also tested the role of several possible markers in tumor initiation and progression.

The final part of this thesis dealt with large scale exon expression, methylation, and chromatin modification datasets for 11 different developmental stages from the Human Epigenome Atlas. Our aim in this genome wide analysis was to identify cases of differential exon usage in different dataset. Our findings suggested a set of strong associations of epigenetic modifications and alternative splicing especially in early human developmental stages.

In summary, the combination of the approaches presented in this thesis may advance the current stages of tumor marker identification. Membrane transporters play key roles in cancer progression. Once their function is defined with the help of similar transporters in other organisms, one may compare their expression and methylation profiles in normal and tumor tissues. The expression/methylation datasets should be cleared first from outliers. Once a tumor marker is defined or confirmed, further analysis is suggested especially for possible different splice variants.

Kurzfassung

Biologische Datensätze aus Hochdurchsatzverfahren sind meist die Basis zeitgemäßer Grundlagenforschung in Genomik, Systembiologie und Krankheitsdiagnstik. Eine Probe kann in manchen Datensätzen momentan tausende Messungen umfassen. Die Allgegenwärtigkeit solch enormer Datenmengen brachte den dringenden Bedarf an effizienten und robusten computergestützten Ansätzen mit sich, die diese Daten verarbeiten und analysieren können und die informative Assoziationen ermitteln.

Diese Arbeit beschäftigt sich mit unterschiedlichen Arten von umfangreichen Datensätzen und beabsichtigt zu Grunde liegende biologische Zusammenhänge mit hoher Zuverlässigkeit zu erkennen. Unsere Methodik besteht aus vier Kernteilen. Im ersten Teil analysierten wir Aminosäure-Daten von Transporterproteinen aus verschiedenen Organismen um funktionelle Annotierungen der Membranproteine speziesübergreifend transferieren zu können. In unserem Fall bildeten wir anhand der Sequenzähnlichkeit die experimentell validierte Funktionen eines Proteins auf ein anderes aus einem anderen Organismus ab. Die Sequenzähnlichkeit in dieser Studie war eine Kombination aus Ähnlichkeitsmaßen verschiedener Softwarewerkzeuge (BLAST, HMMER, MEME). Zuerst definierten wir Vertrauensgrenzwerte (für besagte Werkzeuge) die wir dann für die Vorhersage anwendeten. Wir fanden heraus, dass Membrantransporter mit hoher Sequenzähnlichkeit bis zu gewissen Schwellenwerten sogar dann ähnliche Funktionen oder Transportmechanismen haben wenn sie aus unterschiedlichen Organismen stammen.

Unser zweiter rechnergestützter Ansatz wurde entworfen um Expressions- und Methylierungsdaten zu handhaben. Wir sahen, dass diese Daten oft durch Ausreißer auf Gen- oder Probenebene in Mitleidenschaften gezogen werden. Das Durchführen von Untersuchungen vor einer Bereinigung dieser Ausreißer kann irreführende Ergebnisse zur Folge haben. Daher bieten wir eine Methode die mehrere Ausreißererkennungsalgorithmen beinhaltet um Proben- und Gensonderfälle in Expressions-/Methylierungsdatensätzen zu erkennen. Da einige Ausreißererkennungsmethoden auch dann zumindest einen Ausreißer melden wenn eigentlich keiner vorhanden ist, legten wir zuerst einen Grenzwert für erlaubte Ausnahmefälle fest. Wir prüften wie viele Ausreißerbeobachtungen benötigt wurden um perfekte Koexpression zunichte zu machen und setzten diesen Grenzwert dann für die verbleibende Analyse fest. Zusätzlich haben wir in dieser Arbeit die Verteilung von Genexprimierung/Methylierung vor der Ausreißererkennung bedacht. Dennoch könnten Ausreißer dienliche Information mit sich bringen. Daher markierten wir nur extreme Ausreißer explizit zur Entfernung und solche, die für weitere Untersuchungen potentiell nützliche Information beinhalteten, markierten wir gesondert.

Im nächsten Schritt nutzten wir publizierte Expressions- und Methylierungsdatensätze von GEO um mögliche Tumormarker für HCC, Leberkrankheiten und Brustkrebs zu analysieren und zu bestätigen. Diese wurden später durch unsere pharmazeutischen Kollaborationspartner der Gruppe von Prof. Kiemer im Labor validiert. Zusätzlich zu ihren eventuellen Rollen in der Veränderung von Überlebensraten haben wir auch die Funktion mehrerer möglicher Marker bezüglich Tumorinitiierung- und progression untersucht.

Der letzte Teil dieser Arbeit befasste sich mit umfangreichen Datensätzen für Exonexpression, Methylierung und Chromatinmodifikationen über 11 verschiedenen Entwicklungsstadien aus dem Human Epigenome Atlas. In dieser genomweiten Untersuchung war es unser Ziel Fälle von veränderter Exonnutzung in verschiedenen Datensätzen zu finden. Unsere Resultate legen insbesondere in frühen menschlichen Entwicklungsstadien einige gewichtige Zusammenhänge zwischen epigenetischen Modifikationen und alternativem Spleißen nahe.

Zusammenfassend lässt sich sagen, dass die Kombination der hier präsentierten Ansätze gegenwärtige Stufen der Tumormarkererkennung beschleunigen/verbessern könnte. Membrantransporter haben Schlüsselrollen in der Krebsprogression inne. Sobald ihre Funktion mit der Hilfe ähnlicher Transporter in anderen Lebewesen aufgeklärt ist, könnte man ihre Expressions- und Methylierungsverläufe in gesundem und in Tumorgewebe vergleichen. Die Expressions/Methylierungsdaten sollten hierbei erst von Aureißern bereinigt werden. Sobald ein Tumormarker definiert oder bestätigt ist, ist weitere Untersuchung insbesondere im Hinblick auf verschiedene Spleißvarianten angeraten.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The invention of microscopes in the 17th century was a big step forward to uncover the mystery of what are the basic components of the tissues of living organisms. Scientists were able to see cells with clear membranes for the first time. In the 19th century, scientists like Moritz Traube wondered how permeable cell membranes are. The cell membrane has been a research target ever since. Although genomics is considered a new research field, first DNA was isolated by Friedrich Miescher in 1869 from white blood cells. The first genome sequence was then determined around 100 years later by Fred Sanger and his team. Epigenetics on the other hand appeared first as a side research field linked to evolution studies early in the 20th century. However, it underwent many principal changes till the 1950s. The definitions of modern epigenetics appeared early in 1990s by Riggs, Herrings, and others. Currently, genomics and epigenetics have become standard methods in modern cancer research where membrane transporters are reported to play key roles in cancer initiation and progression.

1.1 Central Dogma of Molecular Biology

The blue print for organisms is passed from one generation to the next one through the hereditary material called deoxyribonucleic acid (DNA). DNA is sometimes called the "book of life" because it is normally presented as a huge list of characters referring to the specific sequence of DNA nucleotide base units that are transcribed and translated later into cellular components. For example, human DNA is about 3.3 x 10^9 bases long. The DNA language is complex although the alphabet is really simple. DNA bases are usually represented by a 4-letter alphabet {A,G,C,T} with reference to the nucleotides Adenine, Cytosine, Guanine, and Thymine. Additionally, DNA typically consists of two strands where each base is in contact with another base in the other strand according to a standard set of pairing relationships {(A-T),(C-G)}.

DNA does not perform much by itself but its coding subsets are copied consistently from its major set into another cellular object called messenger ribonucleic acid (mRNA). The copying process is called transcription and the copied subsets are called genes. In eukaryotic cells, the transcription process is completed in the nucleus and then the mRNA exported to the cytoplasm. Most genes carry the code needed to produce proteins at certain times and locations and thus are called coding genes. Later, each three DNA nucleotide bases in the gene code are decoded -using a cytoplasmic component called ribosome- into an amino acid in a process called translation. The ribosome synthesizes a chain of connected amino acids which is later called a protein. Like in DNA, proteins are presented for simplicity as a set of characters but with an alphabet containing 20 characters referring to the common 20 amino acids. The two-step process of transcription and translation is called the central dogma of molecular biology which is the fundamental basis of gene expression (Figure 1.1).

Gene expression is an important aspect of current genomics due to its sensitive response to clinical conditions, toxic agents, or even in time dependent manner during certain biological processes showing up/down regulation states. Previously, expression analysis was performed in a low-throughput fashion where researchers used northern blotting to analyse one gene at a time or western blotting to analyse one protein at a time. High throughput

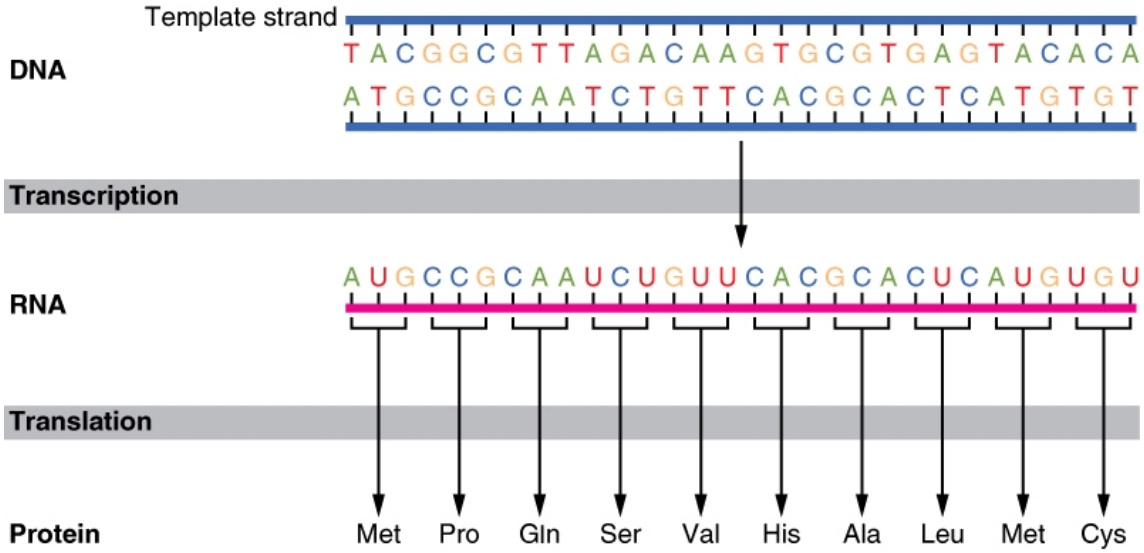

Figure 1.1: Central dogma of molecular biology

technologies started in the early 1990ies by the invention of DNA microarrays nowadays allow to evaluate the expression of thousands of genes at the same time. Since around two decades, gene expression analysis through microarrays has been widely used to identify significant biomarkers for genetic epidemiology [9], to assist the characterization of neurological diseases like Alzheimer's [10], and to open new means for diagnosis and therapeutic interventions for cancer research [11]. The basic idea behind this technology is that the quantification of the mRNA transcripts found can be considered as an approximation for the level of expression of the genes they present. For this purpose, gene probes are hybridized with dye-recognized RNAs on a chip using a separate chamber for each probe. Later, the microarray is illuminated with a laser light causing the labelled molecules to emit fluorescence in proportion corresponding to their abundance. This fluorescence is captured creating an image which is converted later into numbers according to the emitting intensities.

Microarrays are nowadays considered quite affordable and easy to use but they suffer from a fundamental design bias as they only return results for the regions of the design probes. In contrast, the recently invented technology of RNA-seq covers practically all aspects of the transcriptome without any previous knowledge about it allowing the discovery of novel non-coding transcripts, for example. The basic idea is to use the next generation sequencing capabilities to reveal the identity and abundance of most RNAs in a cell [12]. The major disadvantage of RNA-seq is that it is currently more expensive compared to microarrays and thus the microarray technology remains popular.

Regardless of the technology used, gene expression data are normally presented as an array of numerical values where rows correspond to genes and columns correspond to samples. The next standard analysis is to compare gene expression under several conditions like in normal and disease tissues. Computational methods aid the gene expression analysis by providing methods for background correction and normalizing expression as pre-processing steps [13][14], methods and tests to identify differentially expressed genes [15][16], methods to cluster genes according to conditions, or methods for dataset evaluation or testing for possible experimental errors.

1.2 Epigenetic Modifications

Cells in multicellular eukaryotic organisms contain more or less identical copies of DNA although they perform different functions and grow in different ways. The main reason

for this lies in a finely tuned transcriptional program and chromatin organization so that only genes with needed functions in that cell are "on" (expressed) and the others are "off". Some genes are "on" in different cell types as they code for proteins performing general functions needed in such cell types. Moreover, eukaryotic genes might be alternatively spliced so that different parts of it are translated into different proteins in different cells types. Cellular heritable changes not accombined with DNA sequence changes are called "epigenetic changes" and are said to play roles in regulating gene expression [17]. An epigenetic event refers to the structural adaptation of chromosomal regions to register, signal or perpetuate altered activity states [18]. Epigenetic changes occur frequently during normal development but may also have severe consequences. DNA methylation, for example, might alter gene expression in critical cell phases like division or differentiation and thus results in permanent changes. Methylation errors may have severe consequences including diseases. Along the same side, diseases might additionally employ methylation to inactivate genes coding for disease suppressors like in cancer. One example is for this illustrated in figure 1.2.

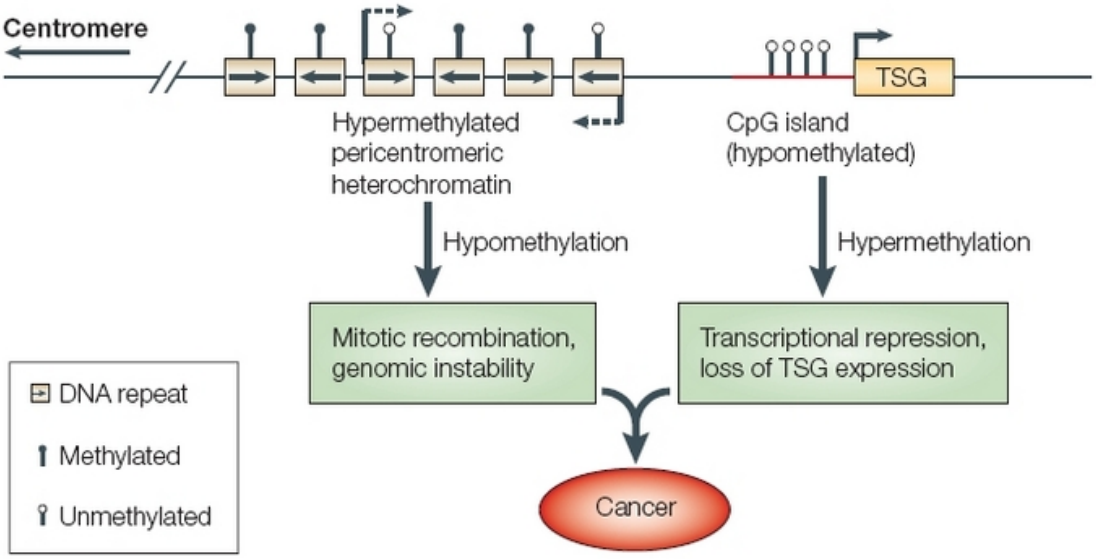

Figure 1.2: In normal cells, the repeat-rich region on the left is hypermethylated whereas the transcribed tumor suppressor gene on the right is hypomethylated. In tumor cells, the methylation is flipped in both regions causing genomic instability and a repression of the tumor suppressor gene. Figure taken from [1]

In mammals and plants, DNA methylation involves the addition of a methyl group on the DNA base cytosine(C). In contrast, bacterial genomes are mostly methylated at adenines. DNA methylation of promoter regions is generally considered as one of the regulation mechanisms associated with gene transcriptional repression. Actively transcribed genes often show a considerable amount of body methylation in their central regions. Furthermore, DNA methylation is required for mammalian development, X inactivation, and genomic imprinting [19]. In the human genome, DNA methylation is mostly restricted to the cytosines followed by DNA base guanine (G) which are called the CpG dinucleotides. DNA regions with high CpG frequency are called CpG islands and they occur frequently in the region that initiates the gene transcription called promoter.

Next generation sequencing is nowadays preferred in DNA methylation analysis over the probe based microarray platforms [20] where possible cross hybridisation forbids the use of any repetitive fraction of genomes in microarrays. This is not the case in NGS-based approaches where the material is directly sequenced and not interrogated by hybridisation [21].

Similar to the gene expression datasets, DNA methylation is provided in datasets of genes where the methylation intensity of each CpG position is presented on a scale from 0 to 1 where 0 means no methylation and 1 means full methylation of this position. The methylation effect is related to the location where it is detected. Therefore, it is standard to provide the exact DNA position for methylation incidents. Several algorithms and software tools were created to aid the methylation analysis starting from the detection of the methylation boundaries, normalizing methylation datasets, checking the differential methylation and statistical significance of methylated regions, and correlating the methylation to gene expression of altered cellular processes like pathways.

Another epigenetic change are the histone modifications. Histones are proteins around which DNA is packed. They might change the binding affinity of specific DNA regions [22] and thus alter the expression of genes in their area [23].

In this thesis, epigenetic modifications have been analyzed in chapters 5 and 6.

1.3 Cellular Differentiation and Carcinogenesis

The structure of a mammalian organism contains many specialized cell types like bones and muscles in human. Cell types are practically identical genetically although they differ in appearance and function. In human, the simple fertilized egg (zygote) gives rise during cell division to about 200 different cell types that form the complex tissues of the human body. Unspecialised primal cells are called stem cells. The process by which a less specialized cell type changes into a more specialized one is called cell differentiation. Stem cells differ according to their differential potential. The most primal type is the totipotent cell formed by the fusion of an egg and a sperm cell that can differentiate into all basic embryonic and the extra-embryonic cell types. Pluripotent cells are descendents of totipotent cells. They can differentiate into cells of any of the three germ layers. Multipotent cells have a narrower differentiation ability. They can produce only cells of related families like the hematopoietic stem cell differentiating into red or white blood cells or other blood cells. The cells with least differential potency are called unipotent because they can produce only one cell type 1.3.

Cellular differentiation analysis is frequently accombined with analysis of epigenetic factors as they play a key role in cell fate determination [24][25]. Some epigenetic modification might be harmful leading to an uncontrolled cell division producing cancer cells, for example, especially if changes cause inactivation of tumor suppressor genes. This process is called carcinogenesis or oncogenesis.

The relationship between oncogenesis and genetic and epigenetic markers is heavily targeted by computational methods since the last decade. This analysis often incorporates up/down regulated pathways due to the changes in activation of the participating components [26]. For this purpose, a variety of computational approaches were used incorporating methods like gene set enrichment, singular value decomposition, and several parametric and non-parametric statistical tests.

1.4 Tumor Markers

Tumor cells or normal cells responding to tumor might produce substances that can be mapped to the current conditions of the cancer. Such substances are called tumor markers and they are mostly proteins found in urine, blood, stool, or certain tissues of the cancer patients. Before accepting the markers for clinical use, they must show high sensitivity and specificity for cancer and fulfill certain criteria [27].

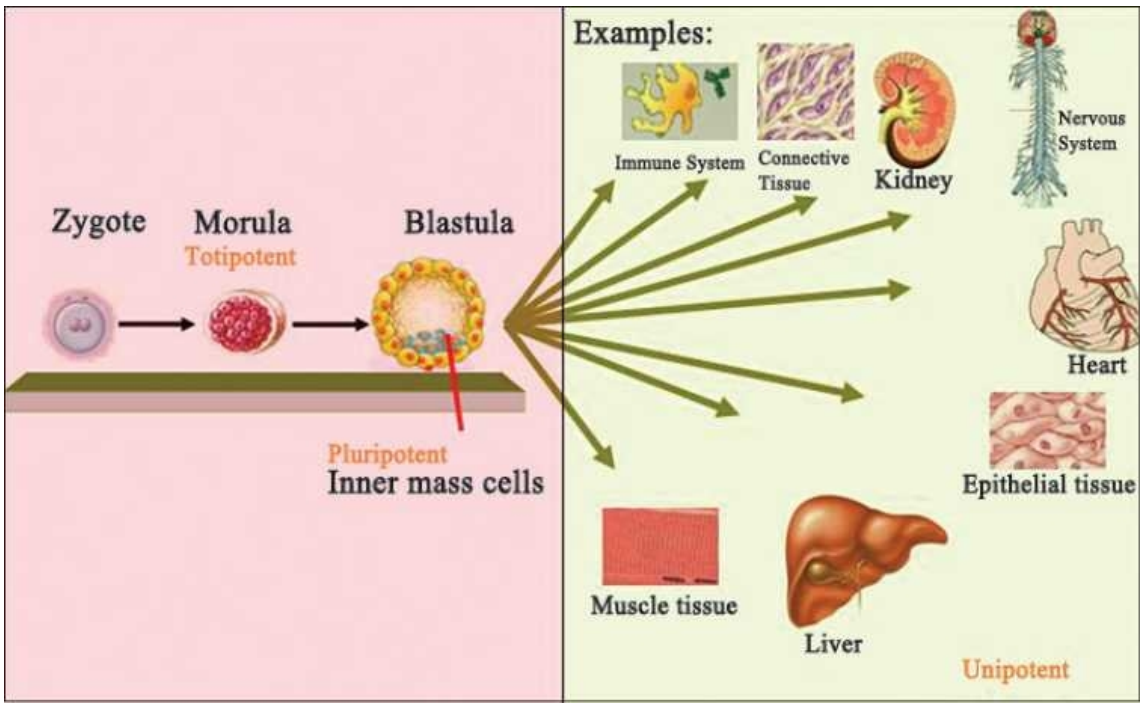

Figure 1.3: Potencies of the stem cells. Figure from [2]

The development of microarrays and the modern sequencing techniques facilitated the discovery of tumor marker genes/proteins. In this work, we tried to validate IMP2 and HAMP genes and their protein products as possible tumor markers using expression analysis and wet-lab work by our collaborators in pharmacy. Although IMP2 was originally identified as an autoantigen in HCC patients, the functional impact of IMP2 on cancer initiation was not described yet in detail sofar. Hepcidine, the protein encoded by the HAMP gene, is strongly correlated with liver diseases. However, its role in HCC is quite unclear.

1.5 Membrane Transporters

Cells import needed compounds and get rid of waste by dealing with the surrounding environment via their biological membrane. Also, membrane-bounded organelles like nucleus and mitochondria exchange compounds in a similar way through their membrane. Transport is essential for every living cell. The research on membrane proteins especially integral transmembrane proteins that span the libid bilayer has undergone an explosive growth in scientific discovery in the last years due to the large-scale genome sequencing projects. Additionally, membrane transporters are a promising target for structural prediction aiming at understanding the molecular mechanisms of fundamental transport processes. When the first atomistic crystal structures of some transporters were revealed, the amount of computational methods specialized in transporter analysis grew rapidly. Additionally, transporters play an important role in cancer analysis. For example, tumor cells exhibit elevated levels of glucose uptake and thus show high expression levels of glucose transporters [28]. Additionally, several transporters have been identified as tumor markers based on changes in their methylation levels [29].

The cell membrane has a hydrophobic nature which prevents hydrophilic substrates and ions from penetration into cells or organelles. However, membrane transporter proteins facilitate such transport and also move molecules against their electrochemical (concentration and electrical) gradients. On the other hand, small uncharged molecules often can passively diffuse though the bilayer membrane.

Membrane proteins either adhere/penetrate the membrane temporarily like peripheral proteins or span it permanently like integral ones (see figure 1.4). Integral proteins might span the whole membrane and are thus called transmembrane proteins or span it partially and are then called monotopic proteins. Transmembrane proteins are labelled alpha helical or beta-barrels depending on the structure of the spanning part. If peripheral proteins penetrate one side of the membrane they can be considered monotopic.

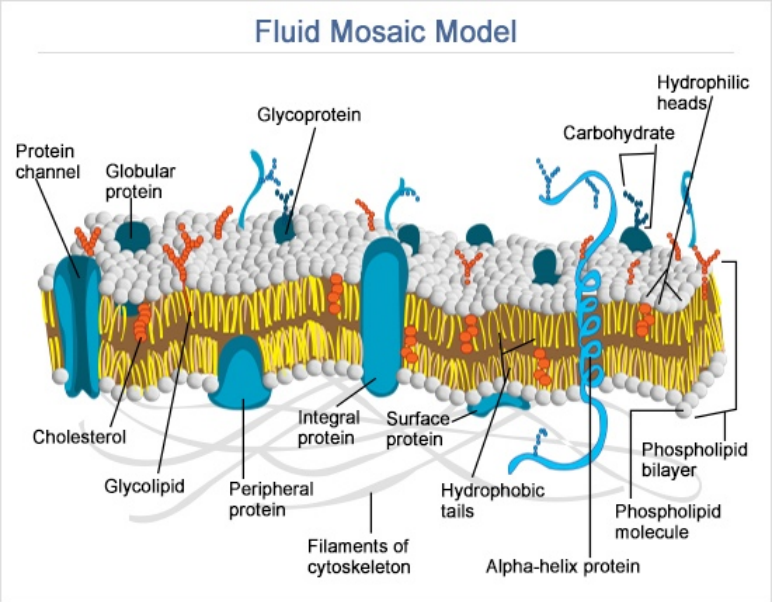

Figure 1.4: Fluid Mosaic Model of the membrane

Considering the transport direction and the amount of transported molecules, transporters might be labelled "Uniporters", "Symporters", or "Antiporters" (see figure 1.5). Uniporters move one solute from one side of the membrane to the other. However, if this co-transport process happens in the same direction then it is performed by a symporter like the glucose-Na^+ symporters frequently found in the kidneys. If the second solute is transferred in the other direction then it is performed by an antiporter like the Na^+-Ca^{2+} exchanger importing Na+ and removing Ca^{2+} from the cells.

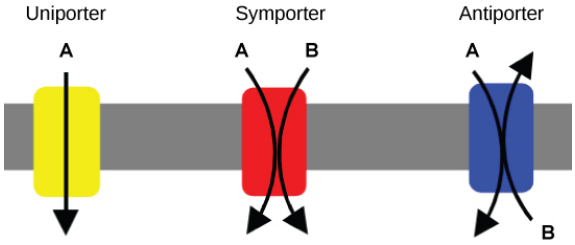

Figure 1.5: Types of transport against the electrochemical gradient. Figure from [3]

However, experimental screening of transporters to determine their function or transporting mechanism is time consuming, costly, and requires sophisticated skills. Therefore, computational methods may make an important contribution by identifying promising candidates for further experimental testing.

In 2003, the International Union of Biochemistry and Molecular Biology (IUBMB) adopted the Transporter Classification (TC) system to categorize transporters [30]. Several databases are based on this classification system such as TCDB [31] and TransportDB [32]. Many databases provide the TC classification as side information such as the Aramemnon database [33]. The TC system hierarchically classifies transporters into classes, subclasses, superfamilies, and families based on their phylogeny, substrate specificity, hydropathy

and transmembrane topology information. For example, the sugar porter (SP) Family 2.A.1.1 belongs to class 2 of Electrochemical Potential-driven Transporters, to the 2.A subclass of Porters (Uniporters, Symporters, Antiporters), and to the 2.A.1 Major Facilitator Superfamily (MFS).

1.6 Goals of this Work

The main goal of this thesis is to help elucidating biological mechanisms by statistical analysis of high-throughput datasets. Membrane transporters play critical roles for cell metabolism by regulating the absorption, distribution, and also the excretion of drugs within the human body. The general goal of chapter 3 is to setup a computational workflow to functionally classify membrane transporters and suggest possible roles of them or some of their splice variants in cancer oncogenesis. As it is costly and time consuming to experimentally define the members of the different functional families, we present a method to computationally classify transporters into their functional families based on text (sequence) similarity and common text pattern (motif) searches.

In the next step, we analyzed expression and methylation profiles of specific membrane transporters in cancer and normal datasets. However, datasets frequently suffer from outlier samples or values that have a severe effect on the proposed correlation or distribution fitting analysis. Therefore, the aim of the method presented in chapter 4 is to filter outliers from expression and methylation datasets before further analysis is performed. This approach was initially implemented in R and later a python tool with GUI was created as the master thesis work of Taner Arslan.

In Chapter 5, we present the results from five projects based on expression, methylation, and aCGH datasets for various tumor types. This project was performed in collaboration with Dr. Sonja Kessler from the group of Prof. Alexandra Kiemer. We analyzed several cancer types and focused the analysis on several sets of genes of interest to the Kiemer group.

In the final project (Chapter 6), we performed a genome-wide analysis of differential exon usage across human developmental stages and how epigenetic modifications are linked to this. This project considered all defined human genes including membrane transporters.

1.7 Publications Resulting From this Thesis

All results chapters of this thesis are based on manuscripts that are either published, submitted, or ready for submission as follows:

- Chapter 3: Barghash A, & Helms V (2013). Transferring functional annotations of membrane transporters on the basis of sequence similarity and sequence motifs. BMC bioinformatics, 14, 343.
- Chapter 4: Barghash A, Arslan T, and Helms, V. A robust approach to detect outlier samples or genes from expression and methylation datasets. (Submitted)
- Chapter 5 section 1: Kessler SM, Laggai S, Barghash A, Schultheiss C, Lederer E, Artl M, Helms V, Haybaeck J, Kiemer A (2015). IMP2/p62 induces genomic instability and an aggressive hepatocellular carcinoma phenotype. Cell Death and Disease. (Just accepted)
  - Chapter 5 section 2: Kessler, SM, Laggai S, Kiemer A, Barghash A, & Helms V

(2015). Hepatic hepcidin expression is decreased in cirrhosis and HCC. Journal of hepatology, 4, 977-979.

- chapter 5 section 3: Barghash A, Helms V, & Kessler SM (2015). Overexpression of IGF2 mRNA-Binding Protein 2 (IMP2/p62) as a feature of basal-like breast cancer correlates with short survival. Scandinavian journal of immunology, 82, 142–143,

- chapter 5 section 4: Kessler, SM, Laggai S, Barghash A, Helms V, & Kiemer A (2014). Lipid Metabolism Signatures in NASH-Associated HCC—Letter. Cancer research, 74, 2903-2904.

- chapter 5 section 5: Kessler SM, Simon Y, Gemperlein K, Gianmoena K, Cadenas C, Zimmer V, Pokorny J, Barghash A, Helms V, van Rooijen N, Bohle RM, Lammert F, Hengstler JG, Müller R, Haybaeck J, & Kiemer AK (2014). Fatty acid elongation in non-alcoholic steatohepatitis and hepatocellular carcinoma. International journal of molecular sciences, 15, 5762-5773.

- Chapter 6: Shanak S, Barghash A, & Helms V. Cross-talk between intragenic epigenetic modifications and exon usage across developmental stages of human cells. (In preparation for submission)

Chapter 2

Fundamentals and Technical Background

In this chapter, we present the computational methods and technical background necessary to understand the bioinformatics contributions presented in this thesis.

2.1 Sequence Analysis

Sequence similarity searches are commonly used as the first analysis of newly determined sequences as sequence similarity provides a hint for functional similarity [34]. For example, that most metagenomic sequences share high similarity with protein sequences in current protein databases. A sequence can be mapped (aligned) to sequence families using computational approaches if they share significant similarities. Such tools might perform either local searches like BLAST [35] and HMMER3 [36], GLocal searches like HMMER2, or global searches like FASTA [37]. It is also known that sequence families might share particular sequence motifs like the case of transmembranes verified motifs [38]. Such motifs can be computationally predicted using tools like MEME [39].

2.1.1 BLAST

The Basic Local Alignment Search Tool (BLAST) is probably the most widely-used bioinformatics tool world-wide (more than 55000 citations to the original publication). BLAST searches for local similarity regions between DNA or protein sequences. The basic idea in sequence alignment is to align a query sequence against another sequence taken from a sequence database (subject sequence) and calculate the statistical significance of the matching. In protein sequences, the matching score is determined, for example, by the blosum62 substitution matrix. All matches with a score exceeding a certain threshold are reported.

BLAST starts by indexing segments of certain size (words) in the query string. Then, each word from the query sequence is aligned against the words from the subject sequences searching for the maximal segment pair MSP. Then, the segment is extended to forward and backward words or toward the next found segment. Later, the extended segment is matched and the score is calculated. This process is repeated as long as the alignment score is increasing and stops when it drops off. Segment pairs whose scores can not be improved by extension or trimming are called high-scoring segment pairs or MSPs. The final statistical significance is reported as expected values indicating the number of times an alignment with an equal or better score than the BLAST alignment can occur based on this database by chance [40].

In brief, given a set of probabilities for the occurrence of individual residues and a set of scores for aligning pairs of residues, the Karlin-Altschul theory provides two parameters λ and K for evaluating the statistical significance of MSP scores. The parameters K and lambda can be thought of simply as natural scales for the search space size and the scoring system, respectively. When two random sequences of lengths m and n are compared, the probability (P-value) of finding a segment pair with a score greater than or equal to S is

given by the equation:

$$P = 1 - e^{-E}$$
$$\text{where the } E\text{-value is given by} \quad E = Kmne^{-\lambda S} \tag{2.1}$$

So, it is expected that the E-values decrease exponentially with the match score. Also, doubling the length of the sequences that are compared would find a double amount of the MSPs resulting in a higher E-value.

In database searches, the E-value is normalized according to the residual size of the database as a last step [41]. If the sequence with length n belongs to a database with total length N in residues, then the E-value for aligning against the n sequence is multiplied by N/n before it is presented to the user.

2.1.2 HMMER

Profile hidden Markov models have provided important assistance to the field of sequence database homology search. Markov models describe memoryless processes where the next state is independent from all previous states except the direct predecessor. Profiles in alignment projects can be considered as set templates or position-specific alignments for sequence families. Hidden Markov models employ a set of hidden states that cannot be observed. They are broadly used in different areas of sequence based bioinformatics. For example, they can with high success match a sequence to another from a sequence database. Once the model λ is created with fixed hidden states (Q=$\{q_1,q_2,..,q_n\}$) and defined probabilities for transitions and output cases (B=$\{b_1,b_2,..,b_n\}$), then it is possible to calculate how probable is the occurrence of any sequence in the space of hidden states (S=$(s_1 s_2 s_4 s_1 s_3$ ...)) and what are the possible transitions through it.

$$Pr[B \wedge Q] = Pr[B \mid Q] \cdot Pr[Q] \tag{2.2}$$

A profile HMM model for sets of sequence alignments considers each column as a hidden state that was produced from hidden state. The standard profile HMM has a deletion and an insertion state at each basic state [42]. The model might start and end with dummy states (Figure 2.1). In the past, profile HMM methods were not as wide spread like BLAST

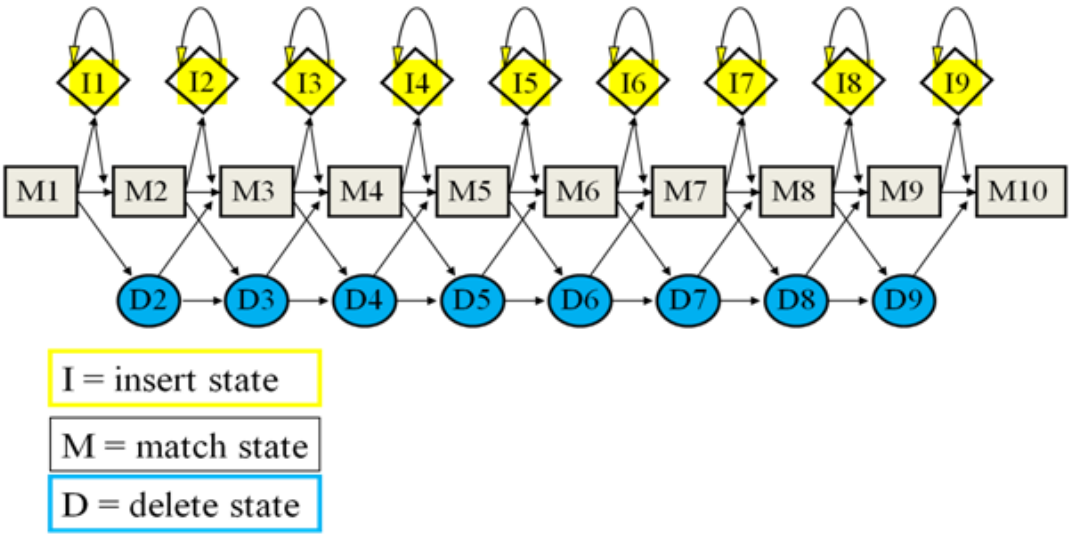

Figure 2.1: A simple profile HMM example. Figure taken from [4]

due to the high computational expense of their software implementations which were slower than BLAST by about 100-fold [43]. Therefore, in current versions of tools based on

profile HMM, acceleration algorithms like the multiple segment Viterbi and sparse rescaling methods are incorporated. With such improvements, HMMER3 became substantially more sensitive and 100-to 1000-fold faster than its preceding version HMMER2 and as fast as BLAST.

The fundamental role of the Viterbi algorithm is to calculate the dynamic programming (DP) matrices for the states of matches, insertions, and deletions [44]. Each element of the DP matrices depends on the value of the previous element as the core principle in HMM. The resulting DP matrices help to find the most probable sequence of hidden states which is called the Viterbi path. This path describes the existing event (match, insertion, deletion). The final result presented to the user is normalized according to the number of found hits.

2.1.3 MEME

In sequence bioinformatics, motifs are pieces of DNA or protein sequence that occur in different DNA or proteins sequences suggesting a possible biological significance. For example, a similar subsequence in the promoters of several genes that appear to be co-regulated might give a hint to a common transcription factor binding there.

Motifs can basically occur in any part of the sequence, at the beginning of one sequence and elsewhere in other sequences. In other words, motif searches need to consider every position from the beginning till the position (sequence length-motif length+1) searching for a subset of special properties with undefined length. Those are the basic properties of the probabilistic model called mixture model where the parameters are unknown with no observed data. Expectation Maximization (EM) is a family of algorithms used for learning such probabilistic models which involve a hidden state. In our case, we used the popular MEME tool [39] for motif searches.

The MEME suite is one of the most widely used tools to discover novel, non-overlapping, approximately matching, and ungapped motifs in a set of unaligned DNA, RNA, or protein sequences using the EM algorithm. The core parts of this method are the calculation of two matrices termed P and Z. Here P_{ck} is the probability of character c to be in position k of the set of sequences and Z_{ij} is the probability that the motif starts at position j in sequence i as motifs might start in different positions in different sequences. When searching for a motif with certain length W initially, the start P should be calculated from the sequence set and Z should be estimated from it in the process called expectation (E-step). In the maximization step (M-step), P is re-estimated from Z. This process is repeated until the change in P is below some given threshold (ϵ). An additional requirement is the probability of characters outside of the motif window which is normally called the background probability (P_{c0}). When P_{c0} is not provided, it is considered $1/4$ for every character in the 4-character DNA alphabet and $1/20$ in the 20-character protein alphabet.

For a hypothetical starting position of the motif:

$$Pr[X_i \mid Z_{ij} = 1, P] = \prod_{k=1}^{j-1} P_{ck,0} \prod_{k=j}^{j+W-1} P_{ck,k-j+1} \prod_{k=j+W}^{L} P_{ck,0} \tag{2.3}$$

Where:

X_i: is the ith sequence

Z_{ij}: is 1 if the motif starts at position j in sequence i, and 0 otherwise

C_k: is the character at position k in sequence i. In the E-step, once the P matrix has

been calculated for the sequence set, the Z matrix is estimated using the equation:

$$Z_{ij}^{(t)} = \frac{Pr[X_i \mid Z_{ij} = 1, P^{(t)}]Pr[Z_{ij} = 1]}{\sum_{k=1}^{L-W+1} Pr[X_i \mid Z_{ik} = 1, P^{(t)}]Pr[Z_{ik} = 1]} \tag{2.4}$$

It is assumed that it is equally likely that the motif can start in any position. Therefore $Pr[Z_{ij}{=}1]$ and $Pr[Z_{ik}{=}1]$ are discarded from the fraction. Later, the Z matrix rows are normalized so that $\sum Z_{ij} = 1$. Next in the M-step, P is re-estimated from the Z matrix using this equation:

$$P_{ck}^{(t+1)} = \frac{n_{c,k} + d_{c,k}}{\sum_b (n_{b,k} + d_{b,k})} \tag{2.5}$$

$$n_{ck} = \begin{cases} \sum_i \sum_{\{j|X_{i,j+k-1}=c\}} Z_{ij} & \text{k} > 0 \\ n_c - \sum_{j=1}^{W} n_{c,j} & \text{k=0} \end{cases}$$

Where:

d: is a pseudo count

n_c: represents how many times this character is found in the dataset

MEME uses the EM algorithm to search for motifs starting at any point allowing multiple motifs to be learned. The predicted motifs are presented as matrix of probabilities. Additionally, sequences can be searched for MEME motifs using another software from the meme package called MAST [45]

2.2 Analysis of Gene Expression Data

In chapter 1 we explained the fundamentals of gene expression and the technologies used. Gene expression data is commonly provided as an array X of $n \times m$ numeric values where n and m correspond to the number of genes and experiments, respectively. Different experiments can be provided for different conditions, patients, tissues, or cell lines. Several pre-processing steps should be applied before the core computational analysis starts. First, poorly expressed genes or genes with constant expression should be filtered out. Genes with missing values can be removed or filled with random numbers. Next, background correction methods are often used to remove possible noise or processing effects. Then, the dataset should be normalized so that unwanted variation is reduced by the normalizing means. RMA and quantile are widely used techniques for dataset normalization since the last decade [46]. Logarithmic transformation is also suggested especially log base 2. It treats up and down regulation equally and uses a continuous mapping space [47]. In this thesis, we either used data that was already RMA-normalized or we applied quantile normalization with the bioconductor packages.

The core statistical analysis can be classified into two major groups of methods: i) methods to find differentially expressed genes, and ii) methods that classify the functional dependency of genes. Methods of the first type aim at identifying genes consistently expressed at different levels under different conditions like healthy and disease. Such a task can be accomplished using classical statistical tests like the ordinary t-test, for example

[48], or the Welch's t-test comparing different-sized samples with unequal variances.

$$t_j = \frac{\overline{X_1} - \overline{X_2}}{s_{\overline{X_1}-\overline{X_2}}} \quad \text{Where:} \quad s_{\overline{X_1}-\overline{X_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{2.6}$$

and s^2 is the unbiased estimator of the variance and n is the number of participants. Additionally at the gene level, a test of co-expression is often requested because genes with similar expression behaviours might have similar functions or participate in the same biological processes. The most commonly used metric to measure the similarity in expression profiles is the Pearson correlation coefficient (PCC)[47].

$$PCC = \frac{\sum_{i=1}^{n}(a_i - \overline{a}) \times (b_i - \overline{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \overline{a})^2} \times \sqrt{\sum_{i=1}^{n}(b_i - \overline{b})^2}} \tag{2.7}$$

Where:

a_i, b_j are the expression values of genes A,B respectively
$\overline{a}$, $\overline{b}$ represent the sample means for the genes
n is the number of samples

The second type of methods aim at identifying common expression patterns to map them later to certain conditions. For example, at the sample level, clustering methods are often applied to group samples of similar conditions or to incorporate new samples from other datasets. Non-hierarchical clustering like the k-means algorithm are less often used as they require pre-knowledge about the dataset. Hierarchical clustering is frequently used with all of its sub-types; Single, Complete, and Average.

$$\begin{aligned} &\text{Single: Distance of the two most similar instances} \\ &\quad dist(c_x, c_y) = min\{dist(a,b) \mid a \in c_x\, b \in c_y\} \\ &\text{Complete: Distance of two least similar instances} \\ &\quad dist(c_x, c_y) = max\{dist(a,b) \mid a \in c_x\, b \in c_y\} \\ &\text{Average: Average distance} \\ &\quad dist(c_x, c_y) = avg\{dist(a,b) \mid a \in c_x\, b \in c_y\} \end{aligned} \tag{2.8}$$

Along the same side, for a group of genes, the gene ontology enrichment analysis is widely used. For sets of co-expressed genes, for example, algorithms like tango [49] or the bioconductor package GOSim identify the enriched GO terms giving a hint at the co-expression.

2.3 Analysis of DNA Methylation Data

In principle, most of the methods used to analyse gene expression datasets can also be used in methylation analysis. However, several new computational methods were created to aid specifically the analysis of DNA methylation data [50]. Particular motifs have been used to predict the methylation status in some DNA sequences [17]. Several studies searched for methylation resistant and methylation prone motifs surrounding apparently methylated

CpGs [51][17] using MEME [39] or pattern searching algorithm. Along the same lines, a tool was presented recently (CpGIMethPred) which employs support vector machines to predict the methylation status in order to speed-up the genome-wide methylation profiling [52].

On the other hand, differential methylation is also an important target of many computational approaches. Bump hunter for example, is a widely used bioconductor package that identifies differentially methylated regions while removing batch effects and labelling regions of statistical uncertainty [53]. Recently, a new computational pipeline (MOABS) was introduced for the analysis of bisulfite sequencing data including the detection of differential methylation [54]. In R-Cran, RnBeads is a newly introduced package developed by the Lengauer and Walter groups for comprehensive analysis of DNA methylation [55].

2.4 Outlier Detection Methods

An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism [Hawkins 1980]. In expression and methylation datasets, outliers can occur at the gene or sample levels. Several studies pointed to the existence and effect of outliers in expression datasets [56][57][58]. By design, outlier samples can result from mislabelling if the dataset contains more than one class like the case of cancer-normal datasets. Outlier values at the gene level might result from experimental or pre-processing errors.

Hierarchical clustering described in section 2.2 can be used to search for outlier samples. Samples not clustered along with other samples under the same conditions might be pure outliers or might correspond to an extreme biological behaviour. Once the clusters are formed, the Silhouette algorithm is used to validate them. This algorithm calculates the average dissimilarity *a(i)* between each point and all other points within the same cluster. So it is a label for how well point i is assigned to its own cluster. Smaller value means better assignment. Then the average dissimilarity between this point and another class is calculated by averaging the distance between this point and all points in that class. If *b(i)* is the lowest average dissimilarity of i to some other cluster which i is not a member of, then this cluster is considered the closest neighbouring cluster. A negative value for Silhouette clustering *S(i)* implies a bad clustering as this point might fit better in the neighbouring cluster.

$$S(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{2.9}$$

To search for outliers at the gene level, several methods can be used. However, some methods require pre-knowledge about the expression/methylation distribution. For example, the generalized extreme studentized deviate (GESD) algorithm is a powerful algorithm for outlier detection only in data approximately following a normal distribution. Therefore, normality tests like Shapiro test should be first used to test the expression/methylation body. In case it does not follow a normal distribution, methods like MAD or boxplot can be used.

GESD

The generalized extreme studentized deviate (GESD) algorithm can be used to detect outliers in data that follow an approximately normal distribution [59]. Studentization is the process of dividing the first degree statistics calculated from the sample by the standard deviation of the sample. Basically, this algorithm assumes that there are r outliers in the dataset and detects them one after the other. In each run, the outlier point is the one that

maximizes the test statistic R_i.

$$R_i = \frac{Max_i \mid x_i - \mu \mid}{\sigma} \tag{2.10}$$

This process is repeated until all r outliers are defined. The number of outliers is determined by finding the count of points fulfilling the condition $R_i > \lambda_i$. Here, λ is the point to determine whether to reject the null hypothesis (critical value) calculated for all points using the percentage points of the t distribution.

$$\lambda_i = \frac{(n-i)t_{p,n-i-1}}{\sqrt{(n-i-1+t^2_{p,n-i-1})(n-i-1)}} \tag{2.11}$$

Simulation studies indicated that for datasets with more than 25 points the critical value estimation is very accurate but it is reasonably accurate for datasets with at least 15 points [59]. It is also reported that the algorithm will detect one outlier even if the data has none.

MAD

The Median Absolute Deviation (MAD) is a robust measure of statistical dispersion. This method has no pre-condition about the data distribution such as GESD. It starts by calculating the absolute value of the deviations of all points from the sample median. Then it calculates the median of the result and labels as outliers the most extreme points away from the new median.

$$MAD_i = median_i(\mid X_i\text{-}median_j(X_j) \mid) \tag{2.12}$$

Boxplot

This initiative algorithm is very popular for conveying the location and variabilities of different datasets. However, based on the type of boxplot used, this algorithm can also be used to detect outliers. The box symbol is basically drawn using the median and the upper and lower data quartiles. The upper and lower whiskers are drawn differently depending on the boxplot type used. They might be drawn at the largest and the smallest point. In that case, no outliers will be assigned. However, in other types they are drawn according the the quartiles or specific standard deviations from the median and the points outside the whiskers are considered outliers.

2.5 Analysis of aCGH Data

Array Comparative Genomic Hybridization (aCGH) is a microarray-based quantitative measure of DNA copy number alterations that can be mapped directly to position and sequence. Tumors might cause severe genetic and epigenetic changes resulting in altered levels of gene expression and thus causing possible modifications to cell growth and survival. These changes might appear as gains/losses in some commonly observed genomic locations in several cancer types. Changes in DNA copy number might have a direct effect on the transcriptional activity of some genes mapped to the altered regions. Therefore, the analysis of DNA copy numbers might be a keypoint toward understanding the major effects caused by diseases like cancer. In Tumor-Normal analysis, the tumor DNA is normally labeled by cy3 and the reference normal DNA is labeled by cy5. To block repetitive sequences, both DNAs are combined with Cot-1 DNA. Then they are denatured and hybridized onto an array containing genomic clones. Then digital images are captured for all fluorescent dyes. The images are later used to calculate the fluoresence intensity for all array targets. To point to the relative DNA copy number between the two hybridized specimen for a certain locus, the ratio of the test to the reference intensities is calculated.

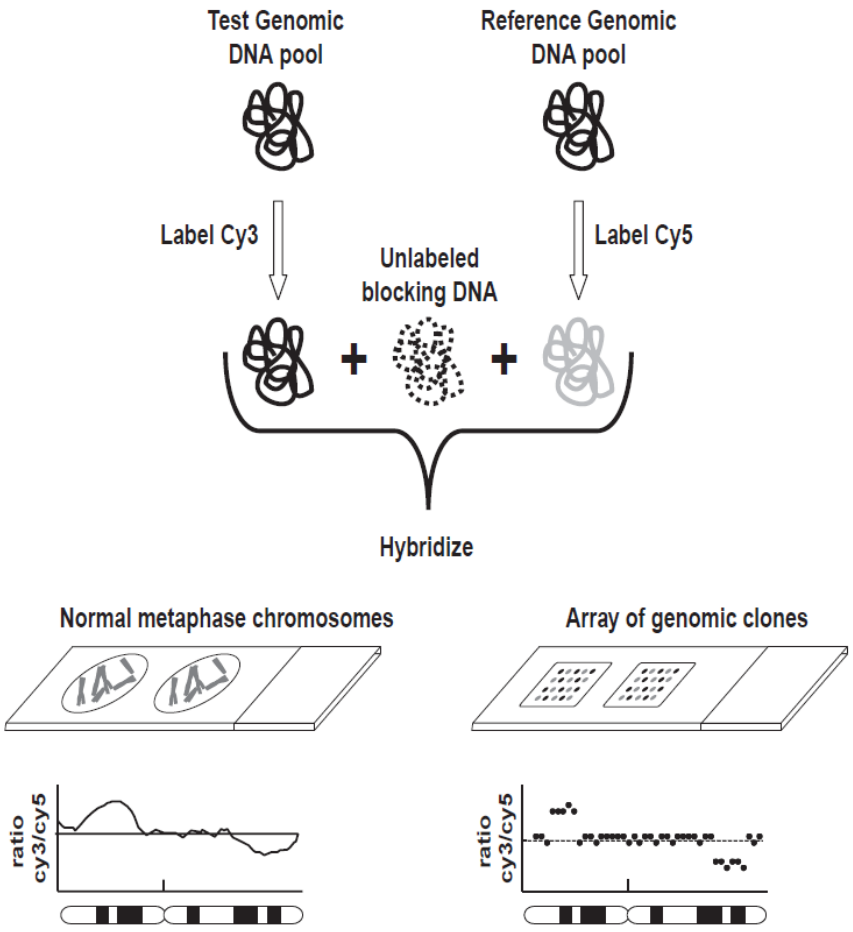

Figure 2.2: Schematic representation of array CGH. Figure taken from [5]

Many algorithms were developed recently to help analyzing aCGH Datasets. Some methods are clone-based and using HMMs to scan through all clones searching for possible gains/losses [5]. However, other algorithms search for segments of gains or losses and not for the individual clones [60].

In this thesis, aCGH data was analyzed in chapter 5.1.

Chapter 3

Transferring Functional Annotations of Membrane Transporters on the Basis of Sequence Similarity and Sequence Motifs

This chapter is based on the following publication:
Barghash A, & Helms V (2013). Transferring functional annotations of membrane transporters on the basis of sequence similarity and sequence motifs. BMC bioinformatics, 14, 343

I conceived this study with Prof. Helms. For this work, I wrote scripts, compiled data-sets, and performed the data analysis. Later we both analyzed the data and jointly wrote the manuscript.

3.1 Abstract

Membrane transporters catalyze the transport of small solute molecules across biological barriers such as lipid bilayer membranes. Experimental identification of the transported substrates is very tedious. Once a particular transport mechanism has been identified in one organism, it is thus highly desirable to transfer this information to related transporter sequences in different organisms based on bioinformatics evidence.

Here, we present a thorough benchmark at which level of sequence identity membrane transporters from Escherichia coli, Saccharomyces cerevisiae, and Arabidopsis thaliana belong to the same families of the Transporter Classification (TC) system, and at what level these membrane transporters mediate the transport of the same substrate. We found that two membrane transporter sequences from different organisms that are aligned with normalized BLAST expectation value better than E-value 1e−8 are highly likely to belong to the same TC family (F-measure around 90%). Enriched sequence motifs identified by MEME at thresholds below 1e−12 support accurate classification into TC families for about two thirds of the sequences (F−measure 80% and higher). For the comparison of transported substrates, we focused on the four largest substrate classes of amino acids, sugars, metal ions, and phosphate. At similar identity thresholds, the nature of the transported substrates was more divergent (F−measure 40 − 75% at the same thresholds) than the TC family membership.

We suggest an acceptable threshold of 1e−8 for BLAST and HMMER where at least three quarters of the sequences are classified according to the TC system with a reasonably high accuracy. Researchers who wish to apply these thresholds in their studies should multiply these thresholds by the size of the database they search against. Our findings should be useful to those who wish to transfer transporter functional annotations across species.

3.2 Background

Prokaryotic and eukaryotic genomes each encode for hundreds of membrane transporter proteins that play essential roles for the cellular import and export of ions and small molecules. Furthermore, transporters mediate signal transduction processes catalyzing the export and uptake of signaling molecules. Therefore, the functional classification of membrane transporters is an important task. The available experimental knowledge about transporter function has been compiled in databases such as TCDB [31], TransportDB [32], SGD [61], and Aramemnon [33]. In these databases, the functional classification is normally done according to the hierarchical transporter classification (TC) system [30] adopted by the International Union of Biochemistry and Molecular Biology (IUBMB).

The TC system categorizes transporter sequences according to their class, subclass, (super) family, and subfamily on the basis of functional or phylogenetic information that is based on sequence similarity. An example for this classification would be the PTS Glucose-Glucoside (Glc) super family 4.A.1 that belongs to class '4' group translocators and subclass 'A' phosphate transfer-driven group translocators. Subfamilies might correspond to transported substrates. A particular transporter sequence in such a family is identified by an extra digit to the right as e.g. 4.A.1.1.1.

A very important detail about each membrane transporter is of course the nature of its transported substrate molecule(s). As an alternative to the TC system, one may also classify transporters into different sets according to their substrates. It is presently unclear how such a substrate-based classification compares with the TC classification system. For example, the Aramemnon database lists members of five different TC families as phosphate transporters in Arabidopsis thaliana. In fact, many databases ignore the fourth digit (subfamily) of the TC system that normally refers to the main substrate. Schaadt and Helms have recently reported that membrane transporters from Arabidopsis thaliana that either transport amino acids, oligopeptides, phosphate, or sugar molecules can be distinguished from each other based on their amino acid composition [62, 63].

An important research question for membrane biology is whether two membrane transporters in organisms X and Y that show a certain sequence similarity will have the same function or not. Previous computational work in this area classified transporters using sequence homology and motif searches [64, 65], amino acid composition [66], and substrate specificity [62]. Interestingly, no study has so far critically analyzed the reliability margins of the individual features. In the general context of protein function, the Pfam repository of protein families has become a quasi-standard. Pfam employs so-called gathering thresholds that are manually curated, family-specific, bit score thresholds that are chosen by Pfam curators at the time a family is built. The threshold used recently corresponds roughly to 'safe' E-value thresholds of $\sim 10^{-2}$[67]. In the TC system, the standard used for establishing homology between two proteins is 9 standard deviations (SDs). This corresponds to a probability of 10^{-19} that the degree of similarity observed arose by chance [68]. Chen and colleagues have recently assessed the performance of different orthology detection strategies for eukaryotic genomes [69].

Here, we have selected the three important model systems Escherichia coli (in the following abbreviated as Ec), Saccharomyces cerevisiae (Sc), and Arabidopsis thaliana (At) that belong arguably to the best characterized species in terms of transport processes. Analyzing homolog databases we found that *Sc* and *At* have more homologs compared to pairs (*Sc*, *Ec*) and (*Ec*, *At*) what reflects the smaller phylogenetic distance between *Sc* and At. According to the InParanoid database [70], 7173 out of the 26207 *At* genes (27.4%) have homologs in *Sc* and 2921 out of the 5884 *Sc* genes (49.6%) have homologs in

At. For comparison, 933 *Sc* genes (15.8%) have homologs in *Ec* and 822 out of 4149 *Ec* genes (19.8%) have homologs in Sc. Finally, only 2778 *At* genes (10.6%) have homologs in *Ec* and 1168 *Ec* genes (28.1%) have homologs in At. Along the same lines, the Arabidopsis sequencing project revealed that a much higher percentage of the proteins in the 12 major functional subsets of the *At* genome had a BLASTP match with E $< 10-30$ to a protein from *Sc* (17−50)% than to a protein from *Ec* (5-32)% [71].

We used three different approaches to transfer transporter functional annotation between the three organisms by relating the level of sequence identity to the functional similarity between the three studied organisms. In this study, we will term this comparison "functional classification". For this, we used the approaches BLAST that generates alignments that optimize a measure of local similarity [35], HMMER that searches for sequence homologs and performs protein sequence alignment using probabilistic methods [36], and MEME that performs motif discovery in protein sequences on the basis of expectation maximization [39]. So far there seem to be no accepted fixed thresholds for the prediction scores of the three tools. Therefore, different studies tend to use their own suitable set of thresholds [11-13,19-21]. Our study establishes a set of thresholds under which the transporter function can safely be transferred between the three model organisms.

3.3 Methods

3.3.1 Overview of the Data

In the training part of this work, we used three sets of membrane transporter sequences from *Ec* (155), *Sc* (177), and *At* (158). In each case, we require that the transporter has been classified in the TC system and that TC/substrate annotations are based on experimental evidence. The sequences and annotations were retrieved from the databases TransportDB [32], SGD [61] and Aramemnon [33], respectively. From TransportDB we downloaded 354 sequences of *Ec* transporters. Among them, 157 have experimentally confirmed annotations about substrate and transporter class. *Sc* transporters were extracted from a list of 6752 ORFs downloaded from SGD. 900 transporters existed in verified ORFs among which 788 had a non-hypothetical function. Only 178 transporters had a clear TC family membership which was obtained by BLASTing SGD extracted transporters against the *Sc* TransportDB by requiring an E-value of 0.0 and a sequence identity of 100%. In Aramemnon, we used the keywords 'transport' and 'carrier' to download 616 transporter sequences from which 159 non-putatives with clear TC classification were extracted. Thereafter, we constructed subsets according to the TC system and according to substrates for later analysis. Obviously, matching a sequence correctly to a particular TC subfamily based on sequence similarity is only possible if this TC subfamily originally contains at least two members (if we take one out for testing, there is at least still one left). Thus, we considered only TC classes with more than one member. Additionally, we also downloaded functional descriptions from the Pfam database [72] for the transporters in the three organisms to assist the substrate information extracted from the individual databases. If substrate information from Pfam conflicted with the original substrate information, the Pfam information was discarded.

The transporters of the three organisms are annotated to 53 (Ec), 29 (At), and 34 (Sc) different TC families. Subclass 2.A (including uniporters, symporters, andantiporters) and subclass 3.A (P-P-bond-hydrolysis-driven transporters) were the most common TC subclasses. In *Sc* and *At*, the Major Facilitator Superfamily 2.A.1 accounts for nearly 40% of all transporters while in *Ec* it is the second largest family after the ATP-binding Cassette (ABC) Superfamily 3.A.1. Shared TC families belong mostly to TC classes Electrochemical Potential Driven Transporters (class 2) and the Primary Active Transporters (class 3).

For the testing part, we created four datasets of experimentally annotated human transporters (Hs). Sugar, amino acid, and metal transporter sets were extracted from the ChEMBL database [73]. Experimentally validated phosphate transporters were obtained from Uniprot [74]. We note that the set of metal transporters contains several proteins that transport several extra substrates besides the metal ion as well.

3.3.2 Prediction Tools

In this work we used the classification approach shown in figure 3.1 (not included in the publication) to transfer the membrane functional annotations of transporters between the organisms *Saccharomyces cerevisiae* (*Sc*), *Escherichia coli* (*Ec*), and *Arabidopsis Thaliana* (*At*) on the basis of sequence homology.

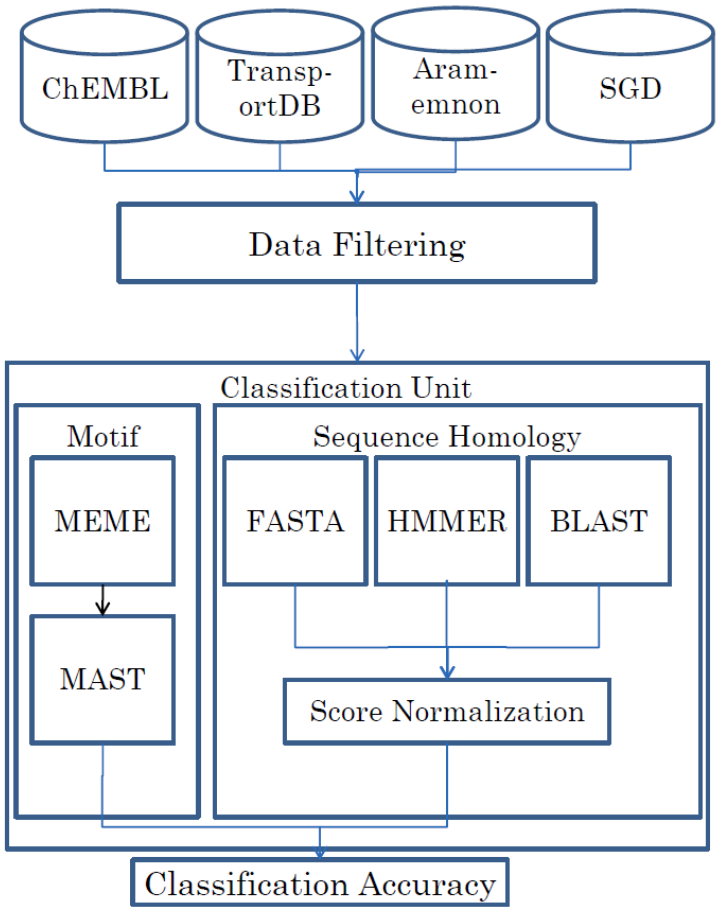

Figure 3.1: A schematic flow diagram of the introduced transporter classification approach

The statistical significance of the sequence similarity between an input sequence and sequences in the input set was determined using the well-known tools BLAST [35] and HMMER [36]. The MEME program suite [39] version 4.6.0 was used to identify enriched sequence motifs in sets of transporter sequences from one organism belonging to the same TC family or that transport the same substrate. Later, the MAST program from the MEME suite provided a score when statistically significant motifs were identified in the sequences from the other organisms. Additionally, we used the tool ggsearch36 from the FASTA suite [37] to test whether sequences transporting the same substrate express not only local but also global sequence similarity.

First, we used NCBI BLAST version 2.2.23 and HMMER version 3.0 for pairwise comparisons of all 90 *Ec* transporters against the 84 *At* transporters that belong to 14 shared TC families. In the MEME analysis, we used only common *At* and *Ec* TC families

with two or more members i.e. 71 *Ec* transporters and 77 *At* transporters belonging to 7 TC families. Next, we aligned the 98 *Ec* transporters belonging to 18 TC families against 131 *Sc* transporters. *Ec* and *Sc* shared 14 TC families that could be searched by MEME involving 87 transporters from *Ec* and 127 from Sc. Finally, we used BLAST and HMMER to compare 157 *Sc* transporters from 23 TC families against 141 *At* transporters. *At* and *Sc* shared 12 TC families involving 130 transporters from *At* and 120 from Sc. Repeatedly, we used sequences from different organisms but belonging to the same TC families as inputs and test sets for the classifiers to test the quality of the prediction. For identifying enriched sequence motifs with MEME, the sequences must be grouped into families that are likely to share motifs. Here, we used MEME to determine up to 3 motifs in each shared TC family between each pair of organisms; 7 such TC families for (*At-Ec*), 14 for (*Sc-Ec*), and 12 for (*At-Sc*). BLAST E-values were normalized by the number of residues in the searched database (see Results section). HMMER E-values were normalized by the number of hits.

In order to identify reliability thresholds at which functional information can be safely transferred between organisms, we tested thresholds (10^{-20}, 10^{-16}, 10^{-12}, 10^{-8} and 10^{-4}) for the E-values and evaluated prediction accordingly. We calculated the accuracy measures precision (positive predictive value), recall (sensitivity) and F-measure (equations 3.1, 3.2, 3.3) at each threshold to evaluate the prediction performance (Tables 3.3, 3.4, 3.5, 3.6, 3.7 and 3.8).

$$Precision = \frac{tp}{tp + fp} \tag{3.1}$$

$$Recall = \frac{tp}{tp + fn} \tag{3.2}$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.3}$$

Precision emphasizes the role of unexpected results whereas recall emphasizes the role of missing classification points. F-measure is a suitable accuracy measure considering precision and recall as we want precision and recall to be evenly weighted. High precision points at a strong prediction boundary while members of other classes rarely match the current class. High recall points at strong similarity within the class members as they rarely match members of other classes. For an actual TC or substrate class, a false negative is a membrane transporter from the class that is predicted to belong to another class, while a false positive is membrane transporter from another class that is predicted to belong to the current class. An example confusion matrix is illustrated in Table 3.1.

		Predicted class	
		3.A.1	Other class
Actual Class	3.A.1	TP	FN
	Other classes	FP	TN

Table 3.1: A confusion matrix corresponding to our method of calculating accuracy measures for the TC and substrate classifications. TPs are members of the actual class correctly classified to the same class from the other organism; Members are considered FNs if they were classified to another class. FPs are members of the other classes that were predicted to belong to the actual. TNs are members of other classes predicted to belong to other classes.

3.4 Results and Discussion

In this work, we perform functional classification of transporter TC families and of transported substrate molecule using datasets from three model organisms. Our aim is to provide a simple guideline to biologists who wish to get a quick information whether available functional information about a transporter in species X may be transferred to

another transporter sequence identified e.g. by BLAST search in species Y. Table 3.2 provides an overview over the main data sets used in this work.

	Ec	*At*	*Sc*
Number of transporters with TC family annotated	156	158	177
Number of transporters with substrate annotation	155	158	848
Number of transporters with TC family and substrate annotation	155	158	177
Metal transporters	10	13	22
Phosphate transporters	5	19	6
Sugar transporters	27	47	24
Amino acid transporters	30	16	27

Table 3.2: Membrane transporters with experimental annotation downloaded from TransportDB, Aramemnon and SGD for *Ec*, *At* and *Sc*, respectively. Only transporters with annotated TC and substrate families were considered in this work.

Figure 3.2 lists common TC families between the three organisms and the distribution of transporters among them.

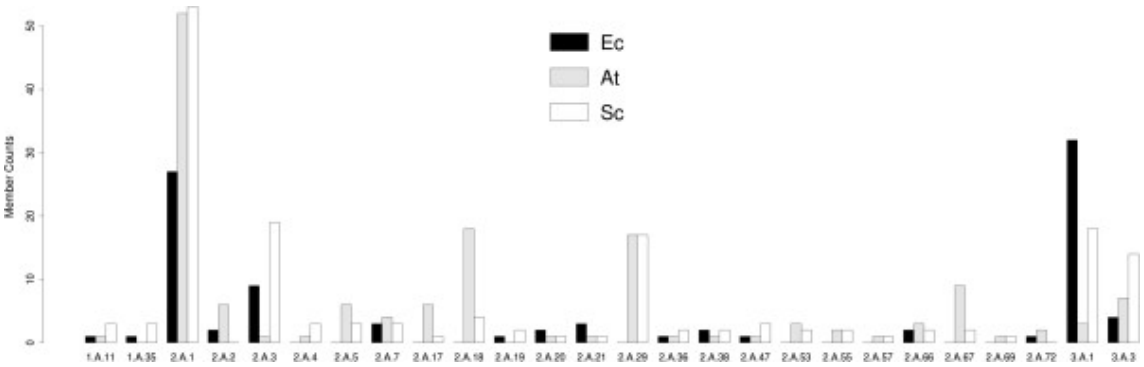

Figure 3.2: Common *Ec*, *At*, and *Sc* TC families with member counts. Most families belong to the Electrochemical Potential Driven Transporters (class 2) and the Primary Active Transporters TC classes (class 3). Shared TC families in the searched organism with more than 2 members were used for MEME motif analysis.

Beside the TC analysis, we also created substrate families of transporters that are annotated to transport the same substrate. For each organism, we collected four large groups of transporters that have been experimentally shown to catalyze the transport of either metal ions, phosphates, sugars, or amino acids. Metal ion transporters account for about 25% of the complete substrate dataset in each organism. *Sc* contains twice as many metal ion transporters as *Ec* and *At* [75]. This can possibly be related with the existence of metallothionein proteins in yeast that function as a metal storage [76]. *At* contains three times as many phosphate transporters as *Ec* and four times as many as in Sc. This is probably due to the essential role of phosphate regulating the *At* root system [77, 78, 79]. Sugar transporters in *At* even account for 50% of the complete substrate dataset which is twice as many as in *Ec* and Sc. One possible explanation for this is that plants need sugar to complete photosynthesis [80]. *Ec* and *Sc* contain twice as many amino acid transporters as At. Figure 3.3 provides an overview to which TC families the members of the created substrate families belong. We noticed that the transporters for these four substrates are spread over many different TC families.

3.4.1 Matching TC families

In this work, we used BLAST for aligning all transporter sequences of one organism against their TC analogues in the two other organisms. Then, we calculated the accuracy

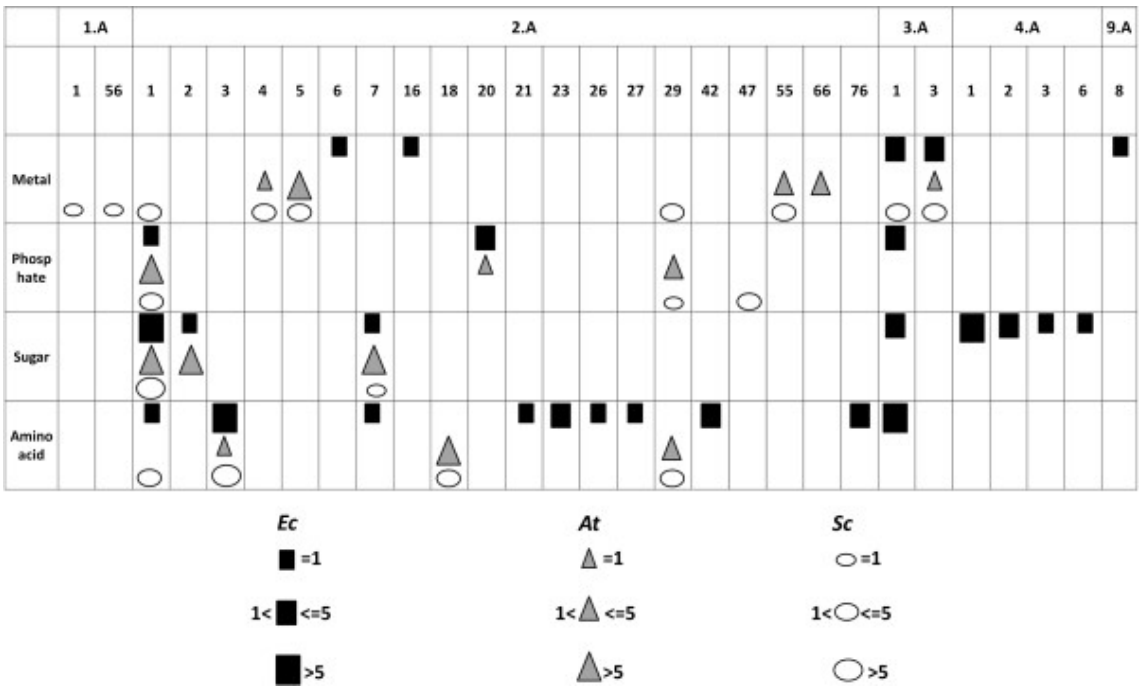

Figure 3.3: Distribution of metal, phosphate, sugar, and amino acid transporters among the different TC families in the three organisms; *Ec* (squares), *At* (triangles) and *Sc* (ovals). The size of the symbols indicates the number of members of this class

measures precision, recall, and F-measure (eq. 3.1–3.3) for various E-value thresholds. BLAST multiplies the significance of a hit by the total number of residues in the database. Thus, to make the obtained results independent from the size of the searched database we divided the E-values by the size of the DB that we were BLASTing against. In this way E-values from searches against different TC sets or substrates sets are comparable to each other. In the following, we will term the normalized BLAST results "normalized E-values". As an example, we BLASTed *Sc* transporter YDR342C either against the *At* dataset (23,567 residues) or against the non-redundant (nr) database of 2011 with 3,877,139,759 residues. Among the matching sequences, we identified the Arabidopsis transporter At3g19940 in both BLAST runs with an E-value of 10^{58} when searching against the *At* dataset and 10^{53} when matching against nr. This difference of reported E-values matches the ratio of the database sizes.

On the other hand, when computing the accuracy measures, we multiplied the results by the member count of each family and then averaged over all TC families considered in order to account for the different member count of each family, see Table 3.3. The last row shows the percentage of transporters that remained unclassified at the given threshold. These are transporters from one organism belonging to the shared TC families that do not share sequence identity better than the given E-value to any transporter in the shared TC family from the other organism.

At the strictest threshold of 1e-20, the assignment of TC family has very high confidence but more than 80% of the sequences cannot be assigned for the *Ec-At* comparison and about half in the *Ec-Sc* and *Sc-At* comparisons. When the threshold is made more permissive, the number of correct predictions increased with few false predictions. We found that the precision and recall increased until 1e-8 but at threshold 1e-4 the number of false predictions increased. As expected, the unclassified percentage decreased as the thresholds were made more permissive. Based on this comparison, a rather permissive normalized BLAST threshold of 1e-8 is very acceptable but 1e-4 can still be considered with caution. When using the absolute identity scores of the alignment instead of the extracted E-values,

	Ec – At					*Ec – Sc*					*Sc – At*				
	1e-20	1e-16	1e-12	1e-8	1e-4	1e-20	1e-16	1e-12	1e-8	1e-4	1e-20	1e-16	1e-12	1e-8	1e-4
Precision [%]	83.3	84.4	86.7	90	60.3	78.6	79.6	79.6	87.8	64.1	84.7	85.4	97.5	97.5	54.1
Recall [%]	83.3	84.4	86.7	90	76.2	78.6	79.6	79.6	87.8	65.1	84.7	85.4	97.5	97.5	62.9
F-measure[%]	83.3	84.4	86.7	90	64.7	78.6	79.6	79.6	87.8	63.6	84.7	85.4	97.5	97.5	55.2
Unclassified [%]	82.2	52.2	37.8	25.6	0	56.1	44.9	40.8	29.6	0	48.4	43.3	35	19.1	0

Table 3.3: Accuracy measures of the BLAST prediction results for finding homologous transporter pairs in the *Ec-At*, *Ec-Sc*, and *Sc-At* comparison that belong to the same TC family for various E-value thresholds. The results were normalized by the size of the reference database (see text). Both precision and recall have a peak at thresholds 1e-12 or 1e-8 but showed lower accuracies under other thresholds. The unclassified percentage decreases as the thresholds' values increase.

the results were untrustworthy. The TC family prediction of *Ec* transporters based on *Sc* transporters annotated more sequences than the prediction based on *At* transporters at the strictest thresholds. Additionally, the *Sc-At* analysis resulted in a higher accuracy compared to the *Ec-At* analysis.

We then applied HMMER to the same datasets as for BLAST and calculated the accuracy measures and the unclassified percentage in the same way. Table 3.4 shows the results obtained with HMMER. For the purpose of normalization, the results were divided by the number of found hits in the database that was searched against. Overall, the results are similar to those obtained with BLAST. However, HMMER results are slightly more accurate at loose thresholds and cover a wider annotation fraction at the strictest thresholds with few more false positives. The number of correctly predicted TC family members at the medium-strong thresholds of 1e-16 and 1e-8 is always equal or higher than with BLAST. HMMER also missed fewer points (false negatives) compared to BLAST. This is clearly reflected by the higher recall value calculated most of the times. It should be re-emphasized that the E-values are computed by the three programs used here in different ways and are, thus, not directly comparable. Also, we have applied different normalization procedures - as suggested by the developers - to normalize the results to per-residue or per-sequence levels.

	Ec – At					*Ec – Sc*					*Sc – At*				
	1e-20	1e-16	1e-12	1e-8	1e-4	1e-20	1e-16	1e-12	1e-8	1e-4	1e-20	1e-16	1e-12	1e-8	1e-4
Precision [%]	73.3	85.6	86.7	90	90.8	78.6	78.6	81.6	86.7	92.9	84.7	85.4	85.4	97.5	93.7
Recall [%]	73.3	85.6	86.7	90	92.1	78.6	78.6	81.6	86.7	92.9	84.7	85.4	85.4	97.5	96
F-measure[%]	73.3	85.6	86.7	90	91.4	78.6	78.6	81.6	86.7	92.9	84.7	85.4	85.4	97.5	94.7
Unclassified [%]	76.7	52.2	40	33.3	17.8	21.4	21.4	18.4	13.3	7.1	15.3	14.6	14.6	2.5	2.5

Table 3.4: HMMER prediction results (sequence E-values) under the given E-value confidence thresholds. The results were normalized by the size of the reference database (see text). HMMER gave a better accuracy under loose thresholds compared to BLAST.

Table 3.4. HMMER results for homology between TC families from the three organisms The decisions by HMMER appear similar to BLAST between the three organisms. Apparently, HMMER attained slightly higher precision for almost all thresholds compared to BLAST especially at loose thresholds. Additionally, in the *Ec-Sc* and the *Sc-At* analysis, HMMER made predictions for a larger fraction of the test set with a noticeably higher recall for thresholds till 1e-8 compared to BLAST. For threshold 1e-4, HMMER predicted a slightly smaller fraction of the test set compared to BLAST but HMMER reported much higher prediction accuracy. Hence, we suggest an acceptable HMMER threshold of 1e-4.

The enriched sequence motifs identified by MEME in sequences from one organism were subsequently searched in test sets of sequences from the other two organisms using the MAST program [45] from the MEME suite. Table 3.5 illustrates the results based on using motif searches for family classification of transporters. As can be expected, motif based searches performed better in families with many members such as 2.A.1. For loose thresholds, motif based classification showed lower precision compared to HMMER and

BLAST but a comparable precision at the strictest thresholds of 1e-20 and 1e-16 as in *Ec-Sc* and *Sc-At* analysis. We suggest that motif based methods may be used beneficially in combination with other methods to support transporter classification. At looser thresholds than 1e-8, motif-based searches seem to lead to unreliable results and should be used with high caution.

	Ec – At					*Ec – Sc*					*Sc – At*				
	1e-20	1e-16	1e-12	1e-8	1e-4	1e-20	1e-16	1e-12	1e-8	1e-4	1e-20	1e-16	1e-12	1e-8	1e-4
Precision [%]	45.1	90.1	90.1	68.7	15.8	83.9	83.9	79.1	33.5	13.2	94.2	99.2	100	57.3	9.6
Recall [%]	45.1	90.1	90.1	89.1	51.9	83.9	83.9	83.9	65.3	25.8	94.2	99.2	100	79.3	36.6
F-measure[%]	45.1	90.1	90.1	76.4	21.6	83.9	83.9	81.3	42.6	16.2	94.2	99.2	100	64.3	13
Unclassified [%]	87.3	80.3	45.1	4.2	0	47.1	46	34.5	1.1	0	51.7	40	28.3	4.2	0

Table 3.5: MAST results searching for motifs predicted by MEME in the *Sc* and *At* test sets. Despite the fact that all sequences were classified, prediction accuracy is generally low at loose thresholds and at the strictest threshold in the *Ec-At* analysis.

3.4.2 Matching Substrates Families

In a second step, we used the same three methods to test whether annotations about the transported substrate can be transferred from one organism to the other. For this, we created four subsets of metal ions transporters, phosphate transporters, sugar transporters, and amino acid transporters. These are the four largest known substrate families and comprised 72 *Ec* transporters, 95 *At* transporters, and 79 *Sc* transporters, see Table 3.2.

As shown in Table 3.6, the results were markedly different from the TC family results. Despite the fact that BLAST reported acceptable prediction precision in the *Ec-At* and the *Sc-At* analysis, the program missed classification of many transporters. We noticed that sequences tend to match sequences from their TC families in other substrate families, rather than their analogues in the same substrate family. Thus, the precision for substrate classification is generally lower than for the TC classification, in particular for the *Ec-Sc* comparison. For instance, the metal transporter (YMR301C) from *Sc* was falsely matched to about one third of all *Ec* transporters in the four substrate families irrespective of their substrates since they belong to the same TC family (3.A.1).

	Ec – At					*Ec – Sc*					*Sc – At*				
	1e-20	1e-16	1e-12	1e-8	1e-4	1e-20	1e-16	1e-12	1e-8	1e-4	1e-20	1e-16	1e-12	1e-8	1e-4
Precision [%]	71.6	72.9	66.1	56.8	37.8	57.7	44.1	38.5	39.3	34.9	95.5	79.8	69.9	62.2	37
Recall [%]	93.1	93.1	93.8	90.8	55.6	85.2	84.6	82.7	75.6	51.5	100	100	100	100	100
F-measure[%]	78.9	80.5	71.5	61.3	42.3	64.3	50.6	43.2	43.6	35.7	97.2	87	79	73.6	52.1
Unclassified [%]	90.3	86.1	79.2	72.2	8.3	65.3	56.9	52.8	51.4	1.4	45.7	44.3	37.1	27.1	1.4

Table 3.6: BLAST prediction results for the four created substrate families of metal ion, phosphate, sugar and amino acid transporters. The results were normalized by the size of the reference database (see text). Unlike the TC family prediction, a smaller fraction of transporters was correctly classified and many were misclassified.

Table 3.7 presents the HMMER prediction results for substrate families from the three organisms. Compared to BLAST, HMMER reported higher prediction accuracy in the *Ec-Sc* analysis but slightly lower prediction accuracy in *Ec-At* analysis at the strict thresholds such as in the TC comparisons. In fact, BLAST classified a slightly larger fraction of the test sets than HMMER in almost all runs. HMMER was also affected by transporters tending to match their TC family members in other substrate families rather than their homologues in the same substrate families.

Table 3.8 shows MAST search results for MEME motifs from different substrate families. MEME gave weak predictions in all runs but in the *Sc-At* analysis. However, recall in the medium strict thresholds 1e-16 and 1e-8 in the *Ec-Sc* analysis is generally acceptable but accompanied with many misclassifications. In the *Ec-At* analysis the prediction accuracy

	Ec – At					*Ec – Sc*					*Sc – At*				
	1e-20	1e-16	1e-12	1e-8	1e-4	1e-20	1e-16	1e-12	1e-8	1e-4	1e-20	1e-16	1e-12	1e-8	1e-4
Precision [%]	51.4	58.3	69.1	66.0	57.7	85.2	77.0	72.7	71.3	70.3	99.3	90.4	76.3	71.7	59.9
Recall [%]	51.4	58.3	100.0	93.5	88.7	83.8	82.4	82.3	78.4	74.6	96.2	95.3	93.4	90.1	86.0
F-measure[%]	51.4	58.3	75.9	70.3	61.9	81.9	75.5	73.3	71.1	69.0	97.2	91.4	78.6	75.4	68.2
Unclassified [%]	93.1	88.9	83.3	79.2	65.3	69.4	61.1	55.6	51.4	47.2	45.7	44.3	41.4	34.3	17.1

Table 3.7: HMMER prediction results for substrate families. The results were normalized by the size of the reference database (see text). HMMER gave a slightly higher prediction accuracy than BLAST in the *Ec-Sc* analysis.

was generally low. Here, even the strict threshold of 1e-20 is unreliable because it gave wrong assignments of substrates in two out of three analyses.

	Ec – At					*Ec – Sc*					*Sc – At*				
	1e-20	1e-16	1e-12	1e-8	1e-4	1e-20	1e-16	1e-12	1e-8	1e-4	1e-20	1e-16	1e-12	1e-8	1e-4
Precision [%]	37.5	37.5	52.8	39.8	25.0	34.7	56.3	52.2	34.9	25.1	82.9	81.5	49.4	30.2	25.0
Recall [%]	37.5	37.5	73.2	48.7	44.2	41.7	81.5	85.5	50.4	40.3	96.7	93.0	79.7	39.3	31.7
F-measure[%]	37.5	37.5	61.3	43.8	30.1	37.9	58.5	60.2	40.9	29.4	87.7	85.9	58.8	30.3	27.3
Unclassified [%]	95.8	94.4	80.6	9.7	0.0	90.3	75.0	59.7	9.7	0.0	68.7	55.3	52.0	0.0	0.0

Table 3.8: MAST results searching for up to 3 motifs predicted by MEME in each substrate family from *Sc* and At. Most members of the substrate families were correctly classified for threshold (1e-4) but only with a very low accuracy.

Surprisingly, 22 *Sc* sugar transporters were correctly classified from 3 motifs predicted by MEME in the *At* sugar substrate family. To the best of our knowledge, none of the three motifs have been annotated so far in databases such as [38]. Table 3.9 lists the regular expressions of these three motifs. The motifs were found around positions 420, 150, and 300 of the protein sequences, respectively.

Approximate position	Regular expressions
420	F[AS][WI][GS][WM]GP[LVI][GP]W[LVI][VI]PSEIFPL[ER][IL]R[SGA]A[GA][QG][SA][IL]A[VA][SAL]VN[WM][IFV]F[TNS]F[IL][IV][AGT]Q[SAT]FLS[ML]L[CE][AH]
150	F[LFI]IG[AS][LI][LV][MN][AG]FAPNVA[MV]LI[IV]GR[LI]L[LA]G[FI]G[VI]G[FL][AG][NS][QM]A[VA]P[VL]Y[IL][SA]E[IM][AS]PAKIRG[AG]
300	[GA][VI]G[LI][QP]F[FL]QQ[LF][TS]GIN[AV][VI][ML][FY]Y[AS]P[VT][IL]F[QK][TK]AGF

Table 3.9: Regular expressions of the three motifs predicted in *At* sugar transporters that lead to correct predictions of 22 *Sc* sugar transporters at the second-strictest threshold of 1e-16.

3.4.3 Application of Established Thresholds to Human Datasets

Next, we tested these thresholds on four *Hs* datasets. In comparison to the three model organisms, these datasets are likely much less complete. We used the three tools to align the *Hs* transporters using a set of transporters from *At* and *Sc* and to align *Ec* transporters using *Hs* transporters. The results are in line with the comparisons of the three model organisms. When using BLAST and HMMER, only a small fraction was annotated at strict thresholds but more were classified at more permissive thresholds. Using HMMER, about 50% of the transporters remain not annotated even at the loosest threshold of 1e-4 whereas using BLAST many more were annotated but with a very low prediction accuracy. The reason is that the *Hs* phosphate and metal transporters were not annotated using the *At* and *Sc* sets and even did not help in annotating the *Ec* transporters. However, sugar and amino acid transporters were mostly correctly annotated. Most annotations of *Hs* transporters were based on matching (*Hs*, *Sc*) pairs. In motif searches, two thirds of the *Hs* transporters were annotated at the threshold of 1e-16 but none were annotated at the strictest threshold of 1e-20, see Table 3.10. The complete results of matching (*Hs*, *At*) and (*Ec*, *Hs*) are listed in Appendix table A1.

Additionally, we studied the pairwise global similarity of all organism pairs using the program ggsearch from the FASTA program suite. The results were generally similar to BLAST and HMMER results with a slightly lower accuracy at the loose thresholds and even lower accuracy at the stricter thresholds. Results are listed in Appendix table A2.

	BLAST					HMMER					MEME				
	1e-20	1e-16	1e-12	1e-8	1e-4	1e-20	1e-16	1e-12	1e-8	1e-4	1e-20	1e-16	1e-12	1e-8	1e-4
Precision [%]	66.7	66.7	62.1	59.2	38.6	66.7	66.7	66.7	66.5	65.1	0	66.7	66.7	37.6	25
Recall [%]	66.7	66.7	66.7	66.7	55.4	66.7	66.7	66.7	66.7	64.3	0	66.7	66.7	31.3	34.1
F-measure[%]	66.7	66.7	64.1	62.1	45.1	66.7	66.7	66.7	66.6	64.7	0	66.7	66.7	34.1	27.4
Unclassified [%]	66.7	60	60	60	6.7	66.7	66.7	60	60	53.3	100	33.3	33.3	0	0

Table 3.10: Hs transporters were better annotated using *Sc* transporters compared to At. The results were corrected for the size of the reference database (see text). About half of the transporters remained unannotated in the HMMER runs. Two thirds of the human transporters were annotated using MEME at the threshold of 1e-16.

3.4.4 Prediction of TC Families in Substrate Families

Comparison of the two preceding sections shows that substrate families have less sequence similarity on average compared to TC families. Now, we tested the combination of both properties, see Figure 3.4. We performed this comparison in a systematic way. For this, we named the extracted families in the form "substrate family_TC family". The four substrate families (amino acids, sugars, phosphates, metals) belong to 19 TC families in *Ec*, 13 in *At* and 14 in Sc. 7 families substrate-TC are shared between *Ec* and *At*, 7 also are shared between *Ec* and *Sc* and 11 are shared between *Sc* and At. Some TC families belong to many different substrate families like the family 3.A.1 that contains members of 4 *Ec* substrate families. We used BLAST to analyze the affiliation of test sequences toward their TC or substrate families. Here, only the best match of each substrate_TC family is considered. The heatmap in Figure 3.4 shows the tendency of *Sc* sequences to match their analogues from *At* TC or substrate families. Some *Sc* transporters matched strongly (black rectangles) their actual substrate_TC families from *At* like sugar_2.A.1, phosphate_2.A.1 and metal_2.A.55. However, most sequences from shared TC families had weaker matches to their TC families rather than their substrate families. Similar results were obtained in the *Ec-At* and *Ec-Sc* comparison, see Appendix figure A1 and Appendix figure A2. Thus, we suggest that it is beneficial to apply substrate information as a pre-filter for transporter TC family classification. On the other hand, transporters that transport the same substrate but belong to different TC families generally do not share noticeable sequence similarity. TC information can be the stand alone feature used to classify transporters but a little tuning by substrate information elevates the prediction accuracy. Misclassification will occur in the small substrate_TC families not in the big TC families.

3.4.5 Limitations and Implications

In some way, our analysis presented here is a bit "circular" since we employ tools to identify sequence pairs belonging to the same TC categories while the TC classification itself was established in part based on phylogenetic analysis that is again based on sequence similarity. However, in a practical use case it is far simpler to run a BLAST or FASTA analysis than to establish a complicated phylogeny. Hence, our results reflect to what extent simple sequence similarity captures the structure of the more elaborate TC classification.

When comparing the results of the four methods (BLAST, FASTA, HMMER3, MEME), the reader should not forget that different strategies are employed by each of the methods to derive E-values for the reported results. Hence, the results of different methods are not directly comparable.

Note that datasets to be used for motif discovery are typically cleaned up for sequence redundancy e.g. using BLASTCLUST with a 25% sequence identity threshold [81]. Here, we did not do this because this would significantly decrease the number of families in the TC dataset that can be used for analysis. Hence, the MEME analysis partially rediscovered

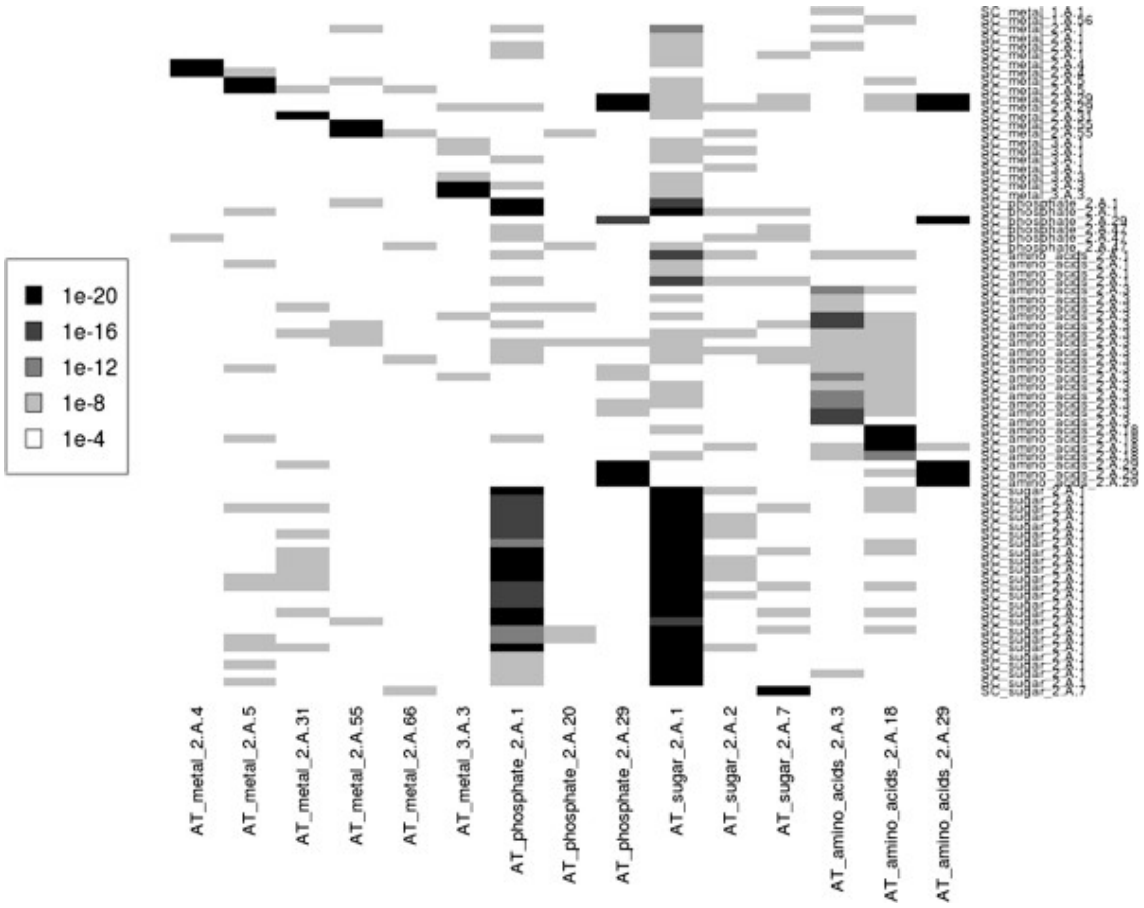

Figure 3.4: BLAST homology search of 69 *Sc* transporters against 84 *At* transporters from 4 substrate families (amino acids, sugars,phosphates, metals) and 13 TC families (Sc) and 12 TC families (At). The grey scale follows a logarithmic scheme where white means no match better than normalized E <1e-04 and black means the best matches better than E <1e-20. Families generally match their `substrate_TC` families. However, they may also match TC families from different `substrate_TC` families

sequence similarities.

This work suggests that the current TC system adopted by IUBMB is a more robust classification feature compared to substrate classification. It is quite likely that phylogenetic inference is a more sensitive indicator of homology than simple sequence similarity or identity. Thus, it appears worthwhile to test the performance of phylogeny-based methods to relate the substrate specificities of membrane transporters.

When trying to completely block the transport of a certain substrate across a particular membrane of an organism it is hard to rely only on the TC information because one substrate can be transported by several transporters from different TC families. One possible explanation in fact is that transporters assigned to different sequence families might actually share a similar 3D structure and the structural similarity might provide an indication about the evolution of the transporter function. Such studies require more sensitive search methods like AlignMe [82].

3.5 Conclusions

We observed that classifying membrane transporters according to TC families gives more accurate results than classifying them according to substrate families. *At* the strictest threshold of 1e-20 for normalized E-values, predictions based on BLAST and HMMER

result generally in high precision, but a huge fraction of the data remains unclassified. We suggest an acceptable threshold of 1e−8 for both programs where at least three quarters of the sequences are classified with a reasonably high accuracy. Researchers who wish to apply these thresholds in their studies should multiply these thresholds by the size of the database they search against. On the other hand, MEME showed unsatisfactory behavior for thresholds below 1e-8. Prediction of TC families split from substrate families showed satisfactory results implying that the application of substrate information as a pre-filter would improve the prediction results. The analysis and suggested thresholds in this study should be useful to those who wish to transfer transporter functional annotations across species without having to build a new phylogeny such as for the TC system. With respect to substrate annotation, the findings of this work may be combined with those of Schaadt et al. [62] who established amino acid composition for substrate annotation of transporters, and with the work of Saier MH Jr. [83].

Chapter 4

A Robust Approach to Detect Outlier Samples or Genes in Expression and Methylation datasets

This chapter is based on the manuscript:
Barghash A, Arslan T, and Helms, V. A robust approach to detect outlier samples or genes from expression and methylation datasets

that has been submitted to the journal of BMC bioinformatics.

I conceived this study with Prof. Volkhard Helms. I wrote scripts, compiled datasets, and performed the data analysis. We both analyzed the data and jointly wrote the manuscript. Taner Arslan created the GUI outlier detection tool in python.

Abstract

Expression and methylation datasets are standard genomic techniques and an increasing number of computational methods are implemented to aid in analyzing the huge and complex amount of generated data. Such datasets often contain a sizeable fraction of outliers that cause misleading results in downstream analysis. Some outliers should be filtered out before starting any analysis while some others must be labeled as they might carry interesting information.

Here, we present a comprehensive approach to detect sample and gene outliers in expression or methylation datasets. We show that the core algorithms detected with high accuracy most outliers that were artificially introduced by us. Sample outliers detected by hierarchical clustering are validated by the Silhouette coefficient. At the gene level, we consider the underlying distribution of a gene expression/methylation dataset and choose a suitable detection algorithm accordingly. The GESD, Boxplot, and MAD algorithms detected with f-measure of at least 83% the simulated outlier genes in non-intersected distributions. We used this approach to detect outliers in publicly available datasets from the TCGA and GEO portals where we found many outliers. However, we frequently found that some functionally similar outliers have outlier observations in common samples. As such cases may be of special interest, they are labeled for further investigations. The presented approach is available as a standalone python tool with GUI via GitHub using this link: https://github.com/TanerArslan/outlier-detection

We suggest that expression and methylation datasets should be checked for outlier points before proceeding with any further analysis. We suggest that 2 outlier observations are enough to label an outlier gene as they are enough to ruin a perfect co-expression. Outliers might also carry useful information and thus functionally similar outliers should be labeled for further investigation. Extremely intersected datasets should be searched for outliers with caution. Our findings should be useful for those using expression or methylation datasets in their research.

4.1 Background

Monitoring gene expression can aid in cancer classification [84] and in identifying clinically-relevant tumor subgroups [85].Additionally, profiling of gene expression is one key approach for finding new biomarkers and therapeutic targets for different cancer types [86]. Several data portals such as the Gene Expression Omnibus (GEO) [87] and The Cancer Genome Atlas (TCGA) now provide convenient access to thousands of normalized expression datasets for most cancer types. However, automatic processing of these data is complicated due to the occasional appearance of outlier samples or outlier genes in such large datasets. In simple words, an outlier is an observation that deviates "too much" from other observations.

Detecting outliers might be important either because the outlier observations are of interest themselves or because they might contaminate the downstream statistical analysis. In the field of gene expression, an outlier can be an abnormal sample that deviates significantly from the other samples in its class. One common reason for this is mislabeling, where accidently a sample of one class might be falsely assigned to another one. Mislabeled samples might then reduce the distinction between true dataset classes. On the other hand, an outlier might also be a gene with abnormal expression values in one or more samples from the same class. In the case of cancer, this may reflect that this patient or his/her disease is a special case. Hence, it is important to identify outliers in expression datasets and, depending on the type of analysis to be performed, to consider whether this data should be removed [5]. Recently, several methods have been proposed for outlier detection in microarray data that used, for example, principle component analysis and estimation of Mahlanobis distances [56], a hybrid evolutionary algorithm [57], cross validation of an SVM classifier [88], a Gene Tissue Index [89], or the OASIS methods [90]. Some studies predicted outliers for the sake of filtering while others predicted them for further analysis. To the best of our knowledge, no approach so far detects sample as well as gene outliers with a set of suitable filters to validate the detection.

In this chapter, we propose and test a simple approach that combines multiple established methods to detect outlier samples or genes in expression and methylation datasets. Average hierarchical clustering is used to detect outlier samples and the clustering is later validated using the Silhouette coefficient. To detect outlier genes we use the three algorithms GESD [59], Boxplot, and MAD [91]. We note that, some outlier genes might carry useful information behind the outlier observations. For this, we introduce functional similarity of abnormal genes as an additional filter for outlier genes. Semantic similarity is analyzed using tool *GOSemSim* [92]. If genes show outlier expression and share high functional similarity with other detected outliers, they are kept for further analysis.

4.2 Methods

Here, we introduce a hybrid technique based on established algorithms to detect outlier samples and genes in expression datasets. Samples are denoted as outliers if they deviate more than a certain threshold in Euclidean distance from other samples in the same class (tumor/normal). The threshold is not fixed but dataset-dependent. To find outlier samples, we used average hierarchical clustering based on Euclidean distance (AHC-ED). Subsequently, we use the Silhouette measurement to validate the quality of the clustering. On the other hand, genes are labeled as outliers if their reported expression values contain outlier observations that pass a suggested threshold and if they share no significant functional similarity with other detected outlier genes. If the expression of one gene follows a normal distribution, we use the Generalized Extreme Studentized Deviate algorithm (GESD) (see chapter 2.4) [59]. If the gene expression data does not follow a normal distribution, then

we apply the two distribution-free algorithms Boxplot and Median Absolute Deviations about the median (MAD) [91]. We additionally test functional similarity within outlier genes using *GOSemSim* [92]. If such gene pairs are found, we check whether their outlier observations are detected in common samples. Functionally dissimilar outlier genes are later marked for removal. The pipeline is illustrated in Figure 4.1.

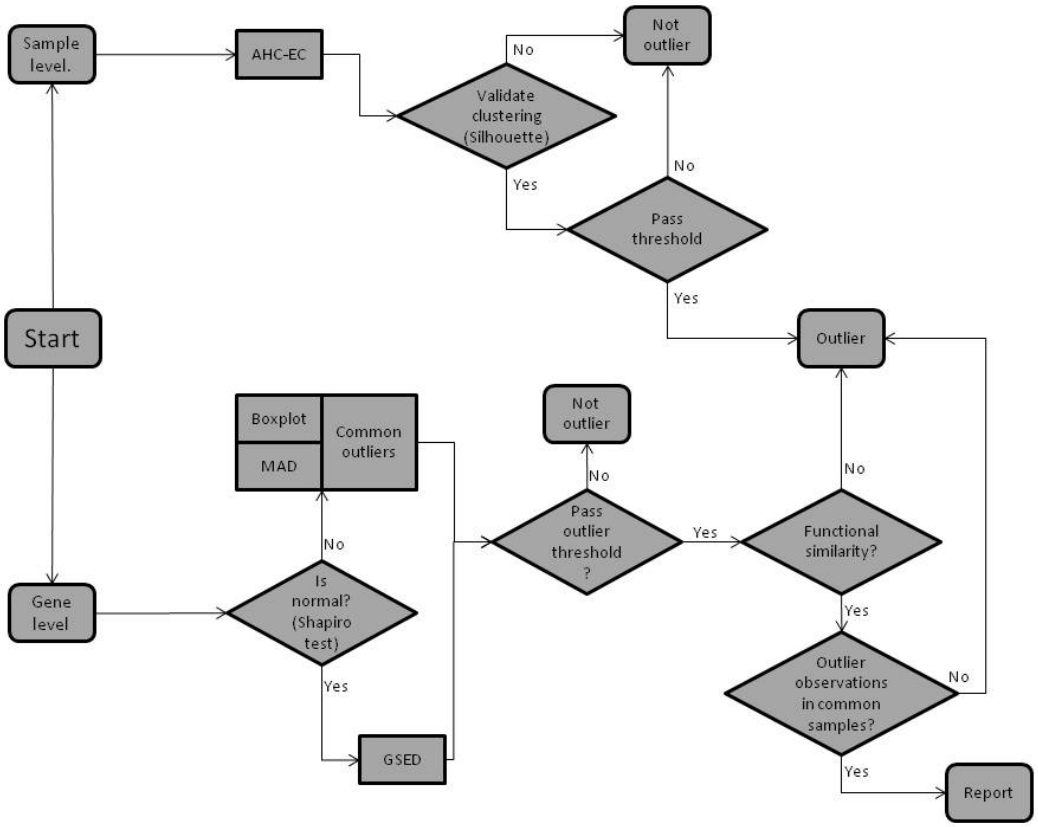

Figure 4.1: Entity relationship model for the outlier detection approach

4.2.1 Datasets

To test the hybrid approach just introduced, we generated four simulated expression datasets with known outliers at the gene and sample levels. Additionally, we tested the workflow on a public colon cancer dataset with known outliers published by [93]. Subsequently, we applied our approach to predict outliers in public datasets of colon cancer, glioblastoma multiforme (GBM), ovarian cancer (OV), and liver cancer obtained from The Cancer Genome Atlas (TCGA) and the Gene omnibus (GEO) databases.

Data with Known Outliers

Initially, we generated four simulated datasets with known outlier samples or genes in a scenario that resembles a typical cancer dataset. Each dataset contains two clearly distinguishable classes of samples. Thus outlier samples either do not match the majority of samples in either of the two classes or are simply mislabeled. On a different manner, a gene is considered an outlier if it presents a clear uneven simulated behavior within either class. In the literature, the overall shape of the distribution of gene expression levels is typically not explicitly mentioned. Several studies apply tests for normality to check whether the data follows a Gaussian distribution [94] [95] [96]. We speculated that in rare cases, the distribution of gene expression might also follow a Poisson distribution. Thus, we created two simulated datasets that obey either a Gaussian distribution or a Poisson distribution.

The simulated datasets contained 100 samples distributed equally to two classes and 1000 genes each. The first 50 samples belonged to class 1 (C1) and the other 50 to class 2 (C2). The form of the first two datasets (SD1/2) is the same and they were both used for identification of sample outliers. At first, the first 900 rows are drawn from the same distribution for both classes but the remaining 100 were drawn from different distributions. In SD1, 900 rows were drawn from the normal distribution N(0,2^2) (see equation 4.1 with $\mu = 0$ and $\sigma = 2$) but the remaining 100 were drawn either from N(10,1^2) or N(20,1^2) for samples of classes C1 and C2, respectively.

$$N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{4.1}$$

In SD2, the first 900 rows were drawn from the same distribution like in SD1 but the remaining 100 were drawn from distributions N(10,2^2) and N(15,1^2) for samples of classes C1 and C2, respectively. SD2 represents clearly overlapping classes. Later, samples 10, 15, and 20 from class 1 were switched with samples 60, 65, and 70 from class 2 as a set of mislabeled samples in both datasets. Additionally, the last sample from each class was replaced by one drawn either from N(25,1^2) or N(30,1^2) in classes 1 and 2, respectively, to create clear outlier samples (see Figure 4.2).

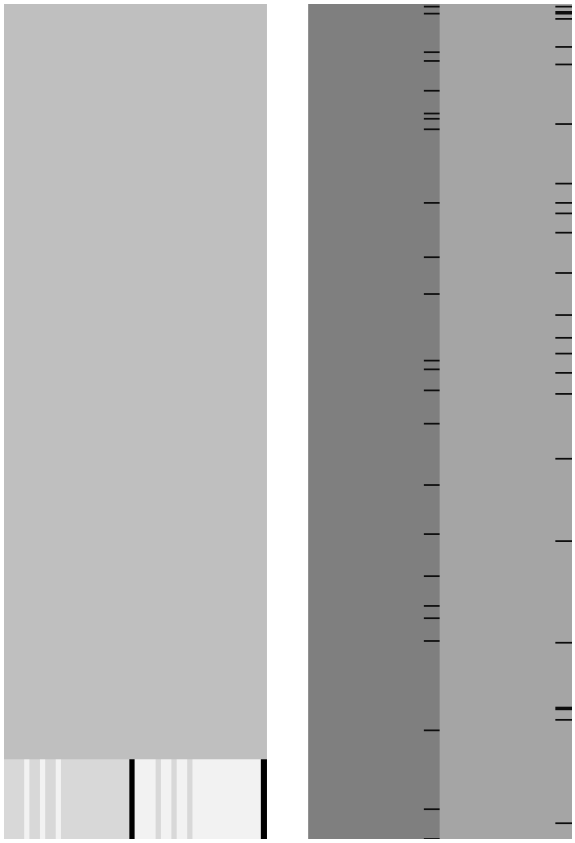

Figure 4.2: Datasets of simulated gene expression. Different gray levels represent different classes. Outlier cases are in black. SD1/2 (left) has two known outliers and 3 known switched samples. SD3/4 (right) Contain 50 outlier each. SD1-3 follow Gaussian distributions while SD4 follows a Poisson distribution

The third and fourth datasets (SD3, SD4) were used for identification of outlier genes. Each had 50 known outlier genes with outlier values at the same positions in classes C1 and C2. In SD3, the 950 non-outlier genes were filled from Gaussian distributions N(0,2^2) and N(15,3^2) for classes C1 and C2, respectively. Regarding the outlier genes, 45 points followed the class rules and the other five were drawn from N(12,1^2) and N(2,1^2) for classes 1 and 2, respectively. To overcome the randomness in the created distributions, we generated 100 arrays in the form of SD3 and passed them later to the outlier detection algorithms. All

normal distributions for non-outlier points were controlled by Shapiro tests with p-value threshold of 0.9. The 950 non-outlier genes in SD4 were filled from a Poisson distribution with λ equal to 2 or 3 for classes 1 and 2, respectively. To simulate outliers in the remaining 50 genes, we filled 45 out of 50 points in each gene with values from the class distribution like before but the remaining five points were filled from Poisson distributions with λ equal to 3 or 0.5 for classes 1 and 2, respectively. Here, we used minimum chi-square estimation [97] to fit the generated distributions and accepted those with an upper p-value threshold of 0.0001. As a further test on an experimental dataset, we considered an extensively studied experimental dataset with documented outlier samples in colon cancer [93]. This dataset has 22 normal and 40 tumor samples. Several classification algorithms were previously applied to this dataset and suggested many outliers and misclassified samples between tumor and normal [56] [88]. Overall, nine samples can be considered as confirmed outliers (T2, T30, T33, T36, T37, N8, N12, N34, N36) and were used here to test our outlier detection approach.

Application to Public Datasets

After validating the workflow shown in Figure 4.1 on the test datasets with known outliers, we applied this hybrid technique to detect unknown outliers in public cancer datasets downloaded from TCGA for colon, GBM, and OV cancers and from GEO for liver cancer (Table 4.1). In GEO, a sample description is included in the main dataset page which is not the case with TCGA. In TCGA datasets, normal and tumor samples can be distinguished by their barcodes. The barcode has several parts separated by hyphens. The third part -with two digit number and a character- describes the sample. Numbers from 0-9 label cancer samples while numbers from 11-19 label normal samples.

Dataset	Raw data type	Normal samples	Tumor samples	Download data	# Genes	# Genes obeying normal distribution
COAD Expression	Agilent	7	143	08.Feb.2013	11687	5971
GBM expression	Agilent	10	594	04.Apr.2013	17430	2820
OV expression	Agilent	7	591	07.Apr.2013	17436	4112
Liver expression (GSE14520)	Affymetrix	239	247	01.July.2013	12701	N:1144 T:1791
COAD Methylation	Illumina Infinium HumanMethylation27	0	129	28.Apr.2013	11633	1082
GBM Methylation	Illumina Infinium HumanMethylation27	0	294	28.Apr.2013	10256	98
OV Methylation	Illumina Infinium HumanMethylation27	8	597	28.Apr.2013	7876	14

Table 4.1: Dataset description.

4.2.2 Detection Algorithms

To detect outlier samples, we cluster samples using the average hierarchical clustering based on Euclidean distance. Subsequently, we use the Silhouette measurement as a measure of the quality of clustering. Based on the clustering vector and the set of distances, the algorithm calculates the average dissimilarity of a point to its current class a(i) and the lowest dissimilarity of the point to other classes b(i). The combination of dissimilarity according to equation 4.2 measures how well elements fit into their clusters. S(i) ranges between (-1,1) where 1 indicates a better fit to the current cluster and -1 means that the

point actually belongs to the other class or a so called neighboring cluster.

$$Silhouette \quad clustering \quad S(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{4.2}$$

The Silhouette coefficient for the objects of one cluster is defined as the arithmetic mean of the Silhouette values of all objects.

To detect outliers at the gene level, we use the 3 algorithms GESD, Boxplot, and MAD. GESD was developed to detect one or more outliers in a dataset assuming that the body of its data points comes from a normal distribution [59]. Precisely, this algorithm calculates the deviation from the mean for every point (see equation 4.3),

$$R(i) = \frac{Max_i \mid x_i - \mu \mid}{SD} \tag{4.3}$$

and then removes the point with the maximum deviation at each iteration. This process is repeated until all outliers that fulfill the condition $R_i > \lambda_i$ are identified where λ is the critical value calculated for all points using the percentage points of the t distribution (see equation 4.4).

$$\lambda_i = \frac{(n-i)t_{p,n-i-1}}{\sqrt{(n-i-1+t^2_{p,n-i-1})(n-i+1)}} \tag{4.4}$$

GESD and its predecessor ESD will always mark at least one data point as outlier [59] ven when there are in fact no outliers present. Therefore, using GESD to detect outliers in microarray data must be accompanied with a threshold of outlier allowance where a certain amount of outliers are detected before marking a gene as an outlier. The GESD method is said to perform best for datasets with more than 25 points [59]. Additionally, the algorithm requires the suspected amount of outliers as an input. The default in this work is half of the tested size.

Besides GESD, we additionally use the well-known Boxplot method which is also a non-parametric algorithm but can detect outliers without pre-assumption about the underlying statistical distribution. Boxplot calculates five key points for plotting; two extremes (whiskers), upper and lower hinges (quartiles), and the median. Data points outside the hinges are labeled as possible outliers. As the quartiles and whiskers are not distribution-driven (related), Boxplot normally suggests many points as outliers and thus datasets might extremely shrink [98]. Therefore, we use this algorithm for gene expression data sets that failed the normality test and we suggest an allowed margin of outliers. The last algorithm we apply is the MAD algorithm. This algorithm does not rely on the variance or standard deviation and thus it assumes no special statistical distribution of the data similar to Boxplot. Here, first the raw median for each gene is calculated over all samples. Then the median absolute deviation (MAD2) of data points from the raw median is calculated as in equation 2.12 where data points with maximum MAD are labeled as possible outliers.

$$MAD_i = median(\mid X_i - median_j(X_j)) \tag{4.5}$$

Hereafter, in this manuscript, we will label as outliers those genes with at least two outlier values (see below). We will use the GESD algorithm only if the gene expression follows a normal distribution and expression data is available from at least 25 samples. For

other genes we use MAD and Boxplot to detect outliers and we accept decisions if they match for at least 2 of the outlier observations.

The analysis in this work was completed in R-cran mainly using the *parody* package. To make it publicly available, The master student Taner Arslan implemented the same workflow as a GUI Python tool for outlier detection under my supervision. This tool offers special implementations of the algorithms mentioned in this work and some other features. AHC-ED followed by Silhouette are used to identify outliers at the sample level while GESD, Modified z-score (MAD) [99], adjusted Boxplot [100], and the median rule [101] are used at the gene level. Once the outliers are detected, the tool offers to group outliers on the basis of their co-expression, functional similarity, or their KEGG pathway participation. The user is asked almost at every step to input his confidence thresholds. The tool provides dataset statistics, detection statistics, and outlier similarity statistics while allowing the user to export the findings at the different stages. Related figures are generated and saved to the disk automatically where needed. The tool is available at GitHub via the link: https://github.com/TanerArslan/outlier-detection

4.3 Results

As a start we illustrate the effect of two outlier data points on co-expression analysis.

4.3.1 Effect of Two Introduced Outlier Points

Co-expression analysis is important for suggesting functional gene-gene interactions. Thus, one may wonder how many outliers are needed to ruin a known co-expression. To test this, we randomly picked one gene each from the 4 public cancer expression datasets studied in this work and introduced two outliers to it. Then we compared the correlation of expression between its raw expression and its modified one. The magnitude of their deviation from the mean was measured in multiples of the standard deviation (SD). Perturbations ranged from 2SD to 12SD. Figure 4.3 illustrates the effect on genes with different numbers of samples.

Figure 4.3 illustrates that introducing only 2 outlier data points with 2 standard deviations from the mean in samples with 143 to 594 data points decreases the auto-correlation of the data from 1 to 0.76-0.94 depending on the size of the dataset. Hence, already few undetected outliers may have a large effect on the biological interpretation of the data. Based on this result, and knowing that some outlier detection algorithms have a marginal error of one outlier, we consider in the following genes as outlier genes if they have at least 2 outlier values.

4.3.2 Detecting Outliers in Data with Known Outliers

Next, we tested the outlier detection approach illustrated in Figure 4.1 using four datasets of simulated expression. SDS1/2 were generated to have two classes to simulate cancer and normal classes. Each class contained a pure outlier sample and three mislabeled samples. In SDS3/4, 50 outliers were distributed among the two classes with five outlier points out of 50 in each class.

Detecting Known Outlier Samples in Simulated Datasets

Here, we first tested the sensitivity of the clustering algorithms using simulated expression data. The first 900 rows in the two classes were filled from the same distribution and the remaining 100 rows were filled from different distributions for the two classes. The outlier

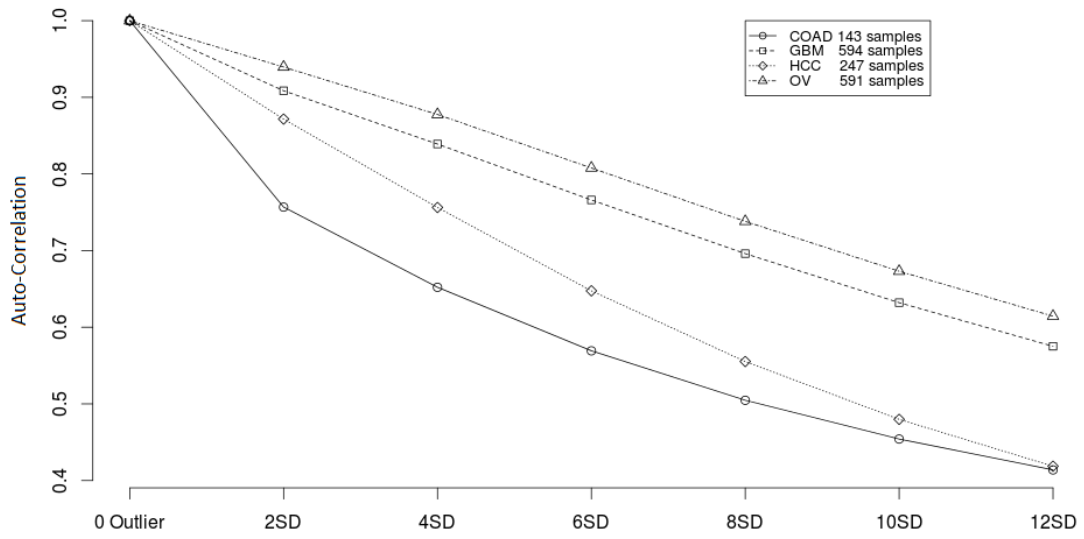

Figure 4.3: Effect of two introduced outlier points on co-expression analysis of a gene with itself. The x-axis illustrates the magnitude of perturbations applied as multiples of standard deviations (SD)

sample detection module successfully classified samples into the two main classes even when only 10% of these rows are different between classes C1 and C2. Additionally, the module detected the two pure outlier samples and labeled them as a third class away from the other two. Finally, the module successfully managed to detect the mislabeled samples 10, 15, 20 from the first class and 60, 65, 70 from the second class and mapped them to the correct classes.

Then, we tested the quality of clustering using the Silhouette method. We found that the two clusters are well separated with an average distance of 0.36 within the SDS1 clusters and 0.14 in SDS2 with semi-nested classes, see Figure 4.4.

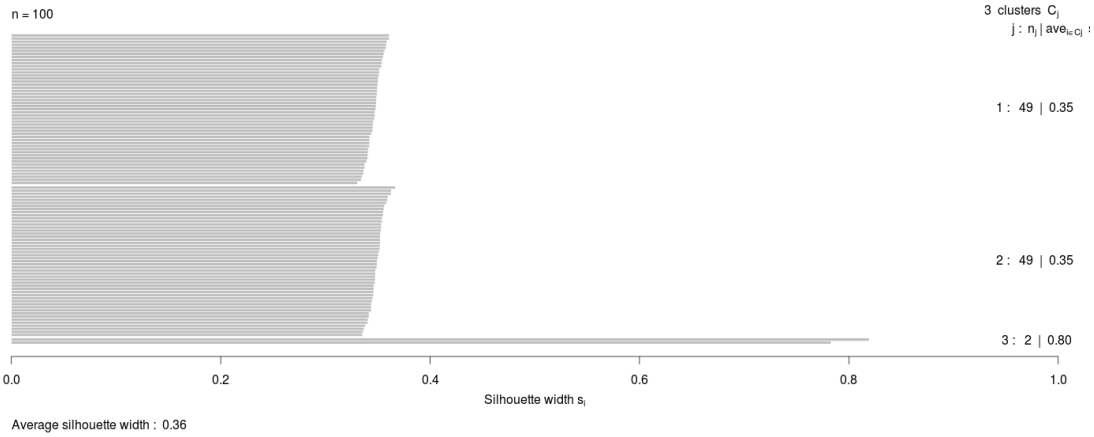

Figure 4.4: Silhouette validation of the AHC-ED clustering of SDS1. The average distance of 0.36 indicates that AHC-ED succeeded in clustering SDS1

Since this first test was very satisfactory, we then tested the stability boundaries of this detection method. First, we varied the proportion of the SDS1 dataset that is being filled from the same distributions. Here we performed 3 runs filling 950, 975, or 990 rows from the raw distribution and filling the remaining rows from the class specific distributions

as before. Then we clustered the samples using AHC-ED and tested the clustering using Silhouette coefficients. In all runs, AHC-ED successfully clustered the samples pointing to the outliers and to the mislabeled ones. Silhouette confirmed the clustering result but with a continuously decreasing average width S(i) of 0.23, 0.14, and 0.07 on average.

As a final test, we filled the differing parts from distributions that have a larger overlap: $N(0, 1^2)$ as raw distribution and $N(8, 1^2)$ and $N(9, 1^2)$ for classes C1 and C2, respectively. Again we tested the four class proportions (900/100, 950/50, 975/25, 990/10) as in the first analysis. Now, Silhouette did not validate the clustering up from the second run, returning negative $S(i)$ width, because of mislabeled samples. Generally, the average Silhouette width was lower than in the first test.

Detecting Known Outlier Genes in Simulated Datasets

For testing the outlier gene detection module, we used the three algorithms GESD, MAD, and Boxplot to identify simulated outliers in 100 generated datasets in the form of SDS3. Each outlier gene was modeled to have 5 known outlier values out of 50 points. We observed that the GESD algorithm was able to detect at least four out of five outlier values in 46 out of 50 outlier genes on average. In contrast, MAD and Boxplot on average detected four out of five outlier points in only 33 and 34 genes, respectively, and some outlier points of the other outlier genes. On average, 31 outlier genes were commonly detected by all algorithms as listed in Table 4.2.

	GSED	Boxplot	MAD
GESD	46±		
Boxplot	33±	34±	
MAD	33±	31±	33±

Table 4.2: Average of commonly detected outliers by GESD, Boxplot, and MAD algorithms in 100 simulated datasets of the SDS3 form. An outlier is considered as correctly detected if four out of five outlier values are detected from the other 50. DS3/4 have in total 50 outlier genes out of 1000.

To test the stability of the detection module, we then performed 3 runs filling the dataset with more intersected distributions each time. In each run we created 100 datasets with 50 outliers each and calculated the average detection of the different algorithms. We found that the GESD detection was more stable than Boxplot and MAD but still failed in the last case showing strong overlap. Table 4.3 lists the distributions used in each run and the detection results.

Approximate Intersection	Class Distributions	Outlier distribution	Detection Result
1SD	C1: N(0,2^2) C2: N(5,1^2)	C1: N(10,2^2) C2: N(11,1^2)	GESD: 45± Boxplot: 37± MAD: 36±
2SD	C1: N(0,2^2) C2: N(5,1^2)	C1: N(8,2^2) C2: N(10,1^2)	GESD: 30± Boxplot: 18± MAD: 17±
3SD	C1: N(0,2^2) C2: N(5,1^2)	C1: N(6,2^2) C2: N(9,1^2)	GESD: 10± Boxplot: 4± MAD: 4±

Table 4.3: Lists of all distributions used in different runs to create simulated expression datasets.

In the datasets following a normal distribution, all three algorithms detected the outliers with good accuracy unless the distributions overlapped to a major extent. To describe the

accuracy, we calculated precision 3.1, recall 3.2 and f−measure 3.3 accuracy measures. As explained by [102], accuracy measures in prediction and classification approaches emphasize the role of unexpected predictions (precision) or the role of missing predictions (recall). Along the same side, F−measure is frequently calculated to merge the precision and recall decisions. In this sense, we consider the known outliers correctly predicted by the algorithms as "True positives (TP)" and the missed known outliers as "False negatives (FN)". Hence, recall for the first runs of the disjoint distributions was calculated as 90%, 74%, and 72% for the GESD, Boxplot, and MAD results, respectively. On the other hand, the algorithms detected at most one additional outlier observation in non−outlier genes (which we did not introduce). Such cases could be considered "False positives (FP)". However, no gene contains two such outlier observations which suggest perfect precision. The F−measure calculated for GESD, Boxplot, and MAD was 94%, 85%, and 83%, respectively.

However, the algorithm detected only few outliers in SDS4 following a Poisson distribution what is rarely the case in gene expression datasets. In that case, GESD detected on average 46% of the outlier points in 16 out of 50 genes and failed to detect any outlier point in the rest. MAD detected 46% of the outlier points in only 3 out of the 50 outlier genes. Boxplot detected only 23% of the outlier points in only 6 out of the 50 outlier genes. This indicates that the algorithms are most robust to detect outliers in expression datasets following more or less a normal distribution.

We now summarize the main decisions taken when establishing the workflow of Figure 4.1 that is implemented in the provided software package. Even in apparently "well behaved" distributed normal distributions, all algorithms detected some less significant outliers (on average one for each gene). More of such insignificant outlier values can be found in real datasets (data not shown). Therefore, we suggest that only genes with at least two outlier observations should be labeled as outliers. We experienced in our analysis that GESD is powerful in detecting outliers in data sets following Gaussian distribution. We also found that Boxplot is a quite restrictive algorithm and places many points outside of the whiskers. Therefore we suggest to implement the GESD decision in data following a normal distribution (Shapiro test p-value >0.1) and accept the decision of Boxplot and MAD for other genes only if they match the positions of at least two outlier observations.

Detect Outlier Samples in Public Datasets with Known Outliers

Next, we tested the outlier sample detection module using a public dataset for colon cancer with known outlier samples in normal and tumor classes [93]. Normal and tumor classes were treated separately. Average hierarchical clustering found 8 out of the 9 reported outlier samples and placed them on the far left in the dendrograms, see Figure 4.5.

4.3.3 Detect Outliers in Public Data Sources

Then, we applied the established workflow to detect outliers in datasets from the public sources TCGA and GEO. At the gene level we checked the normality using Shapiro test as a precondition. Genes with outlier behavior might actually carry useful information behind the outlier values. Therefore, as a last filter, we tested whether the genes with outlier behavior belong to a functional group by analyzing Gene Ontology (GO) annotations using the package *GOSemSim* [92]. We postulate that if two or more outlier genes show a certain degree of functional similarity and have outlier points in the same samples, then the causative outlier behavior of this functional group might be interesting to analyze and thus genes should not be discarded right away. Hence, we first needed to establish a cut-off threshold for meaningful semantic similarity. To this aim, we computed the semantic similarity between all pairs of around 11000 human genes, see Figure 4.6. Based on the

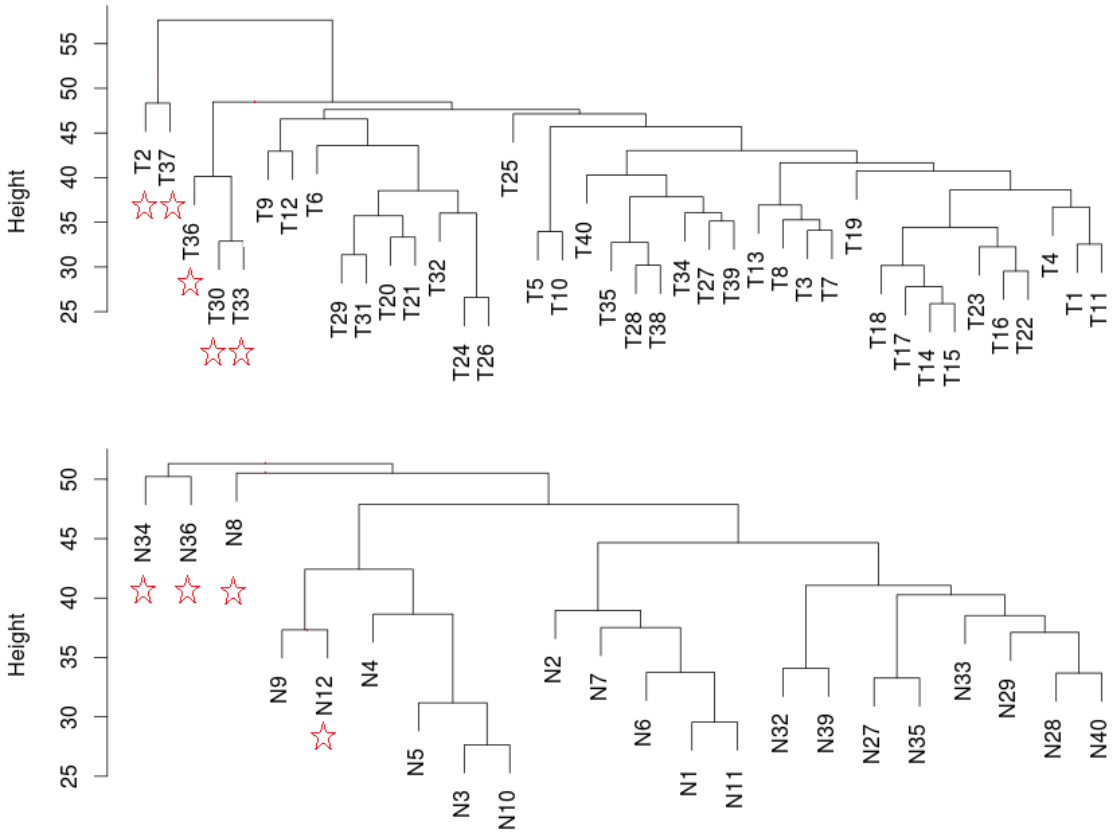

Figure 4.5: Average hierarchical clustering based on Euclidean distances of a public colon cancer dataset with known outliers marked by asterisks

data shown, we suggest that 0.85 is a reasonable cut-off threshold for meaningful functional similarity.

Detection of Outliers in TCGA Datasets

In the colon dataset, AHC-ED clustered the normal samples into one cluster distanced away from most tumor sub-clusters without detecting any clear outlier or mislabeled samples, see Figure 4.7.

The Silhouette coefficient validated this clustering with an overall average width of 0.22 (Figure 4.8). As TCGA datasets sofar contain only few normal samples for most cancer types, we analyzed only the tumor samples for outlier genes.

The gene expression of TCGA datasets frequently followed a normal distribution. Among these genes, GESD detected only four outlier genes with at least 2 outlier values (EIF3G, GLUD1, GSG1L, STARD6). The results of MAD and Boxplot on these genes mostly supported the GESD findings. Among the non-Gaussian genes, Boxplot detected 1692 and MAD detected 1840 outliers. 1586 genes had common outlier observations in at least two samples reported by Boxplot and MAD. Interestingly, 1163 of these outlier genes were also detected by GESD applied to the non-Gaussian expression. When searching for functionally similar outliers using GOSemSim, we found that 400 outlier genes show high pairwise functional similarity to other outliers among these 400 genes.

In the GBM dataset, AHC-ED grouped the normal samples as one of the outer clusters like for the colon dataset. Additionally, several tumor samples were clustered away from the

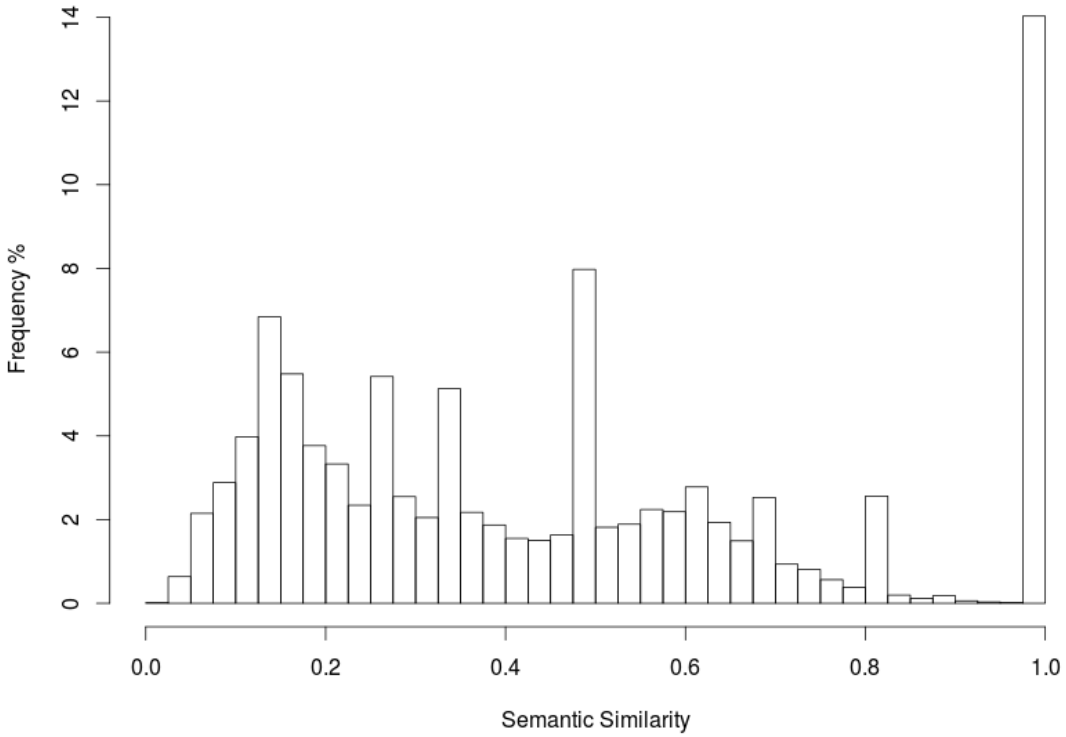

Figure 4.6: Histogram of semantic similarity between all pairs of 11000 genes. 85% of all gene pairs have functional similarity of 0.85 or less according to *GOSemSim*. Those pairs with larger values than 0.85 are considered as functionally similar here

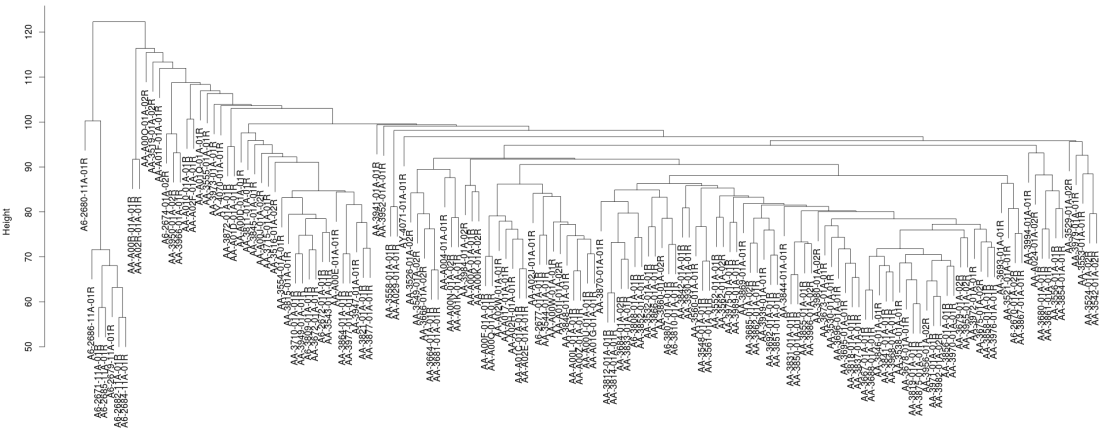

Figure 4.7: Detected clusters in public colon cancer dataset from TCGA. All 7 normal samples with barcode 11A were clustered together on the left side of the dendrogram away from tumor samples with barcode 01A

core clusters and thus they can be labeled as outliers (Data not shown). Overall clustering was validated using Silhouette with overall an average width of 0.22 (Figure 4.9). Here we suggest that further downstream analysis will be slightly improved after removing these outlier samples.

At the gene level, the expression values of 2820 out of 17430 genes followed a Gaussian distribution according to Shapiro test and GESD detected 6 outlier genes among these (C6orf151, DOCK2, EIF2S2, NPR2, PLEKHA8, SH3GL1). Among the genes with non-

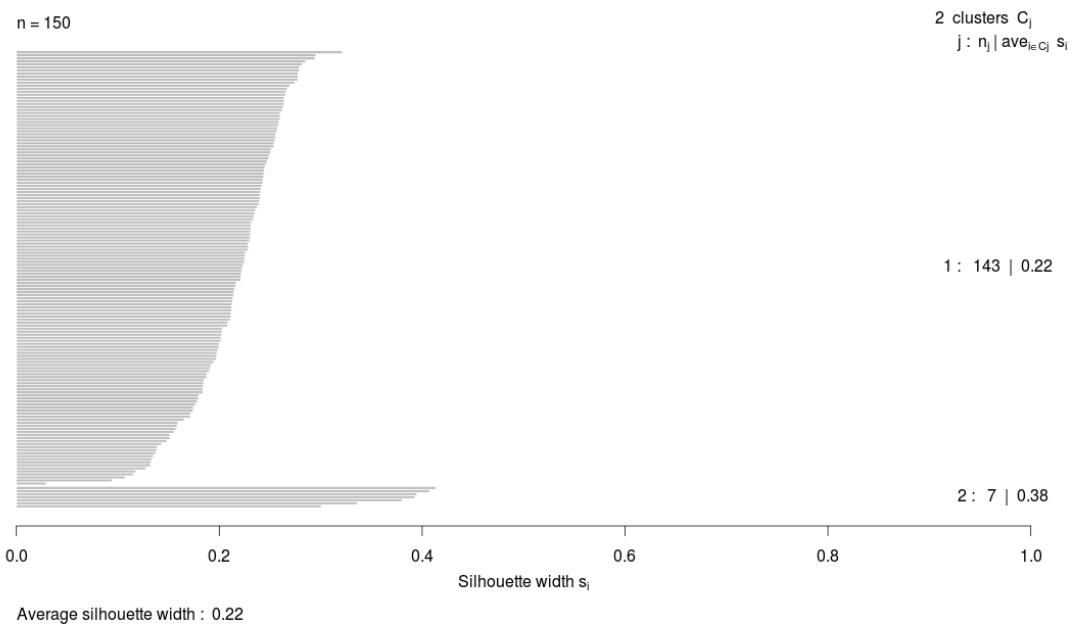

Figure 4.8: Silhouette validation of clustering the TCGA COAD dataset

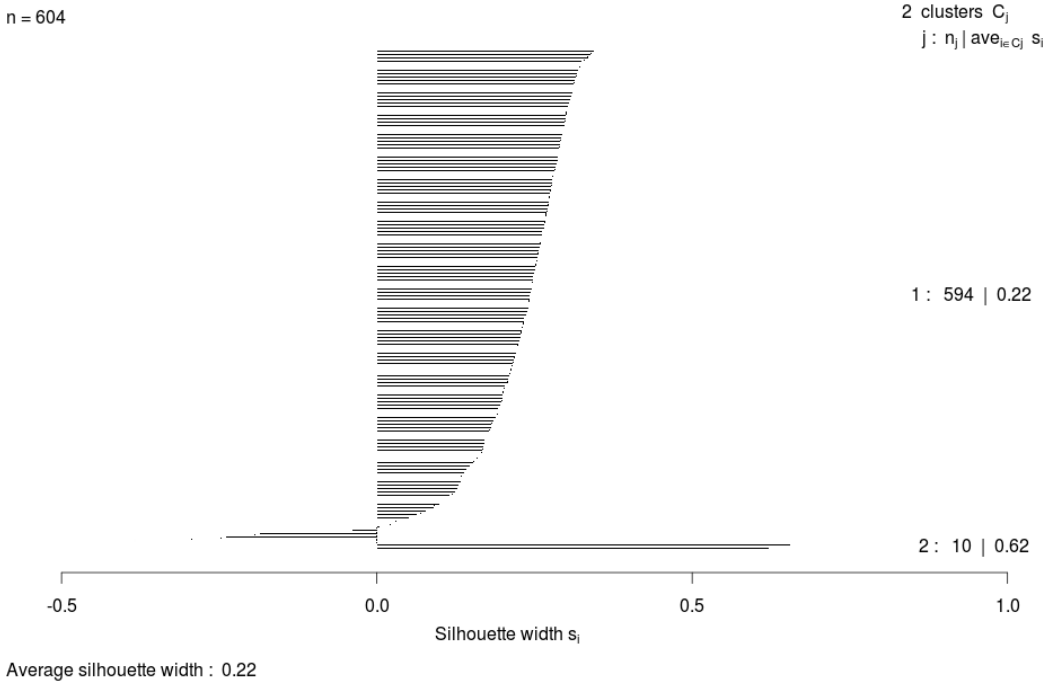

Figure 4.9: Silhouette validation of the GBM dataset clustering.

Gaussian body, Boxplot and MAD detected 6788 and 7130 outlier genes, respectively. Both algorithms detected that 6671 outliers had at least two outlier points in common samples. Additionally, the detection of 5032 of these genes was supported by GESD. 2325 of the 6671 outlier genes shared high functional similarities and outlier observations in at least two common samples.

In the OV dataset, normal samples were clustered together but not on the outer sides. For tumor samples, clustering resulted in many small clusters which indicates weak relations between the samples (Data not shown). Silhouette validated this clustering with average

widths of 0.47 and 0.05 in normal and tumor samples, respectively (Figure 4.10). The removal of the outermost 10 samples improved the clustering only slightly.

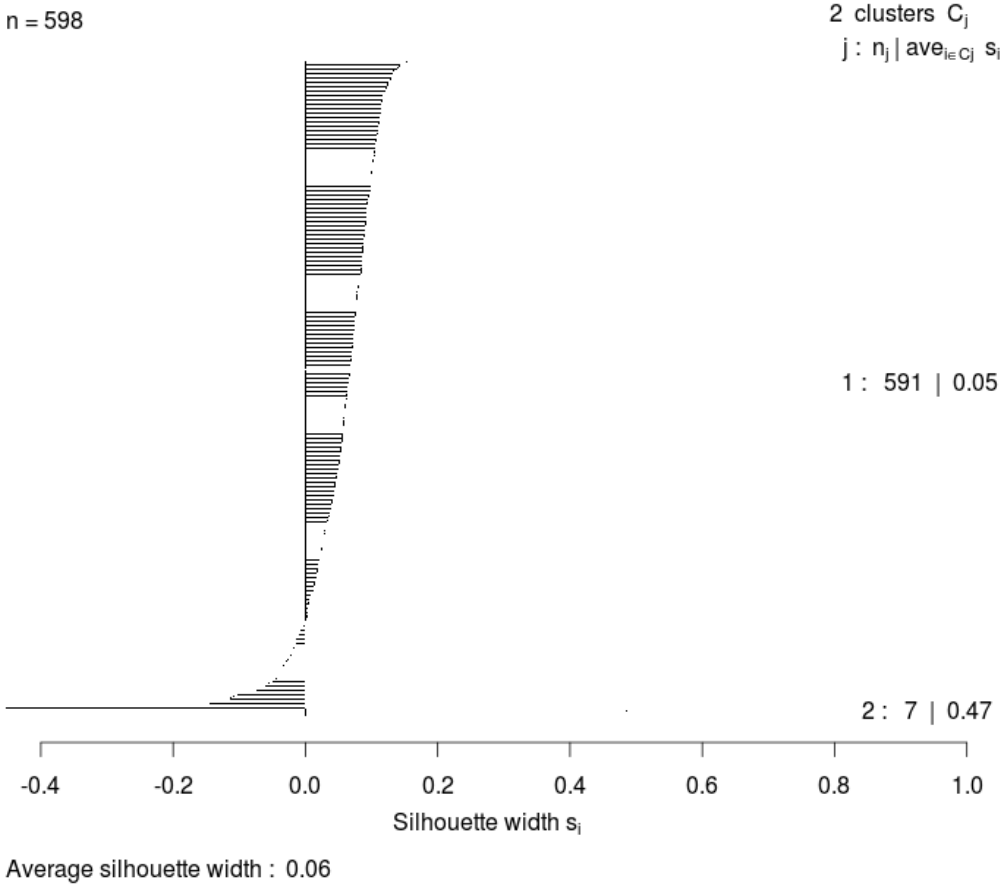

Figure 4.10: Silhouette validation of the OV dataset clustering

At the gene level, the expression of 4112 out of 17436 genes follows a Gaussian distribution. GESD found 8 outlier genes among the non-Gaussian ones. Boxplot and MAD found 5757 and 6067 outlier genes, respectively, of which 5659 have outlier observations in common samples. GESD supported the detection of 786 of the outlier genes. 1665 outliers shared high functional similarity and outlier observations in common samples.

Detect outliers in GEO Datasets

NCBI GEO provides more cancer related datasets than TCGA. Also, GEO datasets normally contain a balanced amount of normal samples. Here we applied our hybrid approach to a liver cancer dataset with 486 samples; 239 normal and 247 tumor. Normally, samples were mostly clustered into one core cluster. However, clustering tumor samples presented at least two clear tumor clusters as shown in Figure 4.11.

Silhouette validated these findings with an average width of 0.4 for normal and 0.03 for tumor samples (Figure 4.12). Here, we suggest removing only the outliers among normal samples clustered outside the core cluster. Also, for this case, we suggest that performing further analysis to tumor clusters separately might achieve clearer results. In this dataset where only 14% of the genes had a Gaussian expression body, we found many outlier genes in this dataset as listed in table 4.4. Boxplot and MAD matched at least 2 outlier positions in 7742 and 6128 outlier genes in normal and tumor samples, respectively. We found 4541 outliers in common between normal and tumor samples. However, 4716 and 3208 outlier

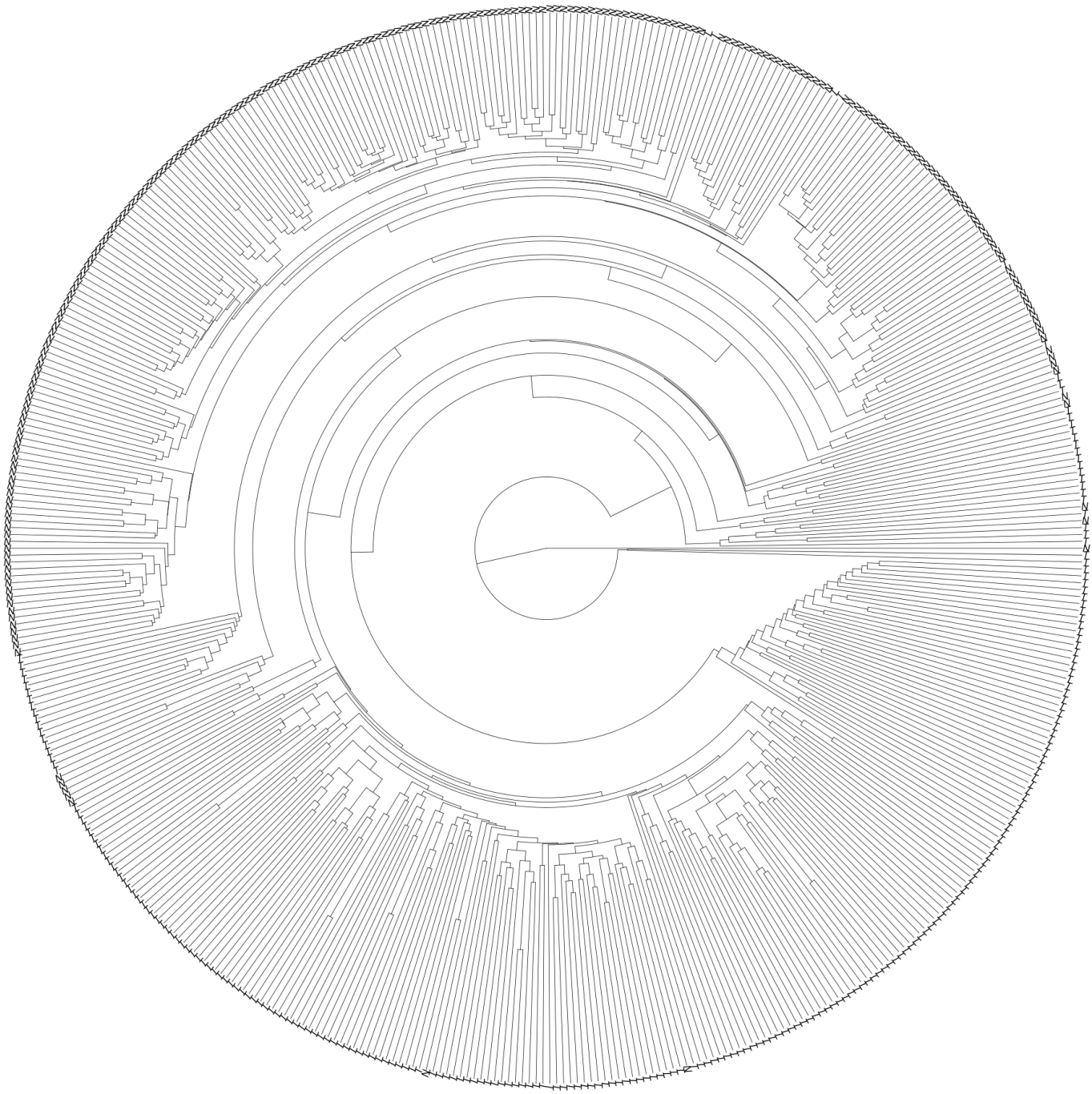

Figure 4.11: Hierarchical clustering of the GEO liver cancer dataset. Sample names are replaced by N for normal and T for tumor

genes shared high functional similarity in normal and tumor samples and they had outlier observations commonly in at least 2 samples.

	GSED	Boxplot	MAD
GESD	7	2	0
Boxplot	6215	7846	4636
MAD	6668	8174	5071

Table 4.4: Statistics of outlier detection in GEO HCC dataset.

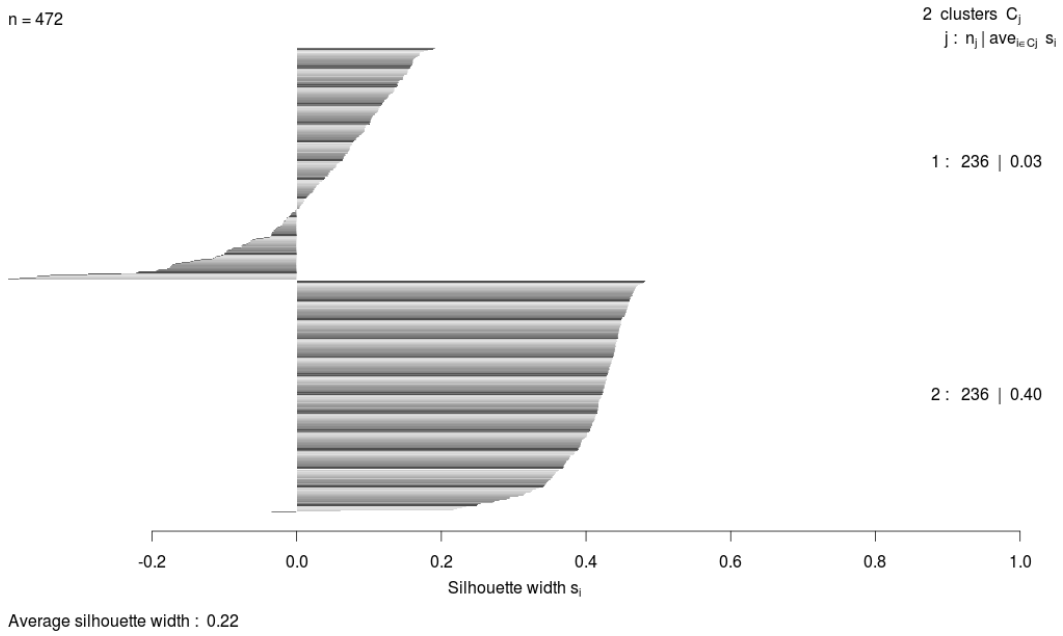

Figure 4.12: Silhouette validation of the clustering on the GEO liver cancer dataset

4.3.4 Detecting Outliers in Methylation Datasets

Finally, we tested the outlier detection approach to identify outliers in 3 methylation datasets downloaded from TCGA for colon, GBM, and OV cancers. Only the OV dataset had normal and tumor samples. Out of these, only the normal samples were clustered together as validated by Silhouette (Figure 4.13).

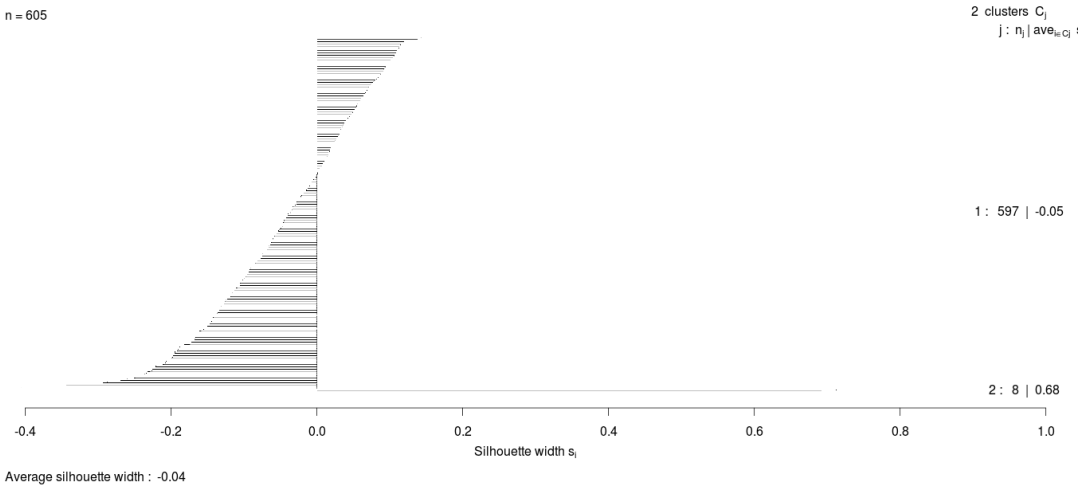

Figure 4.13: Silhouette validation of clustering OV methylation dataset

At the gene level, fewer genes had a Gaussian methylation body compared to expression datasets. However, most outliers found shared high functional similarity with other detected outlier genes and thus were not removed except the case of outliers detected by MAD in the COAD dataset. Interestingly, we noticed that the 3 algorithms matched at least two outlier positions in most of the detected outliers although only few had a Gaussian body. Additionally, at least 50% of the commonly detected outliers shared high functional similarity. The fraction of outliers detected and returned by the three algorithms is shown

in Figure 4.14.

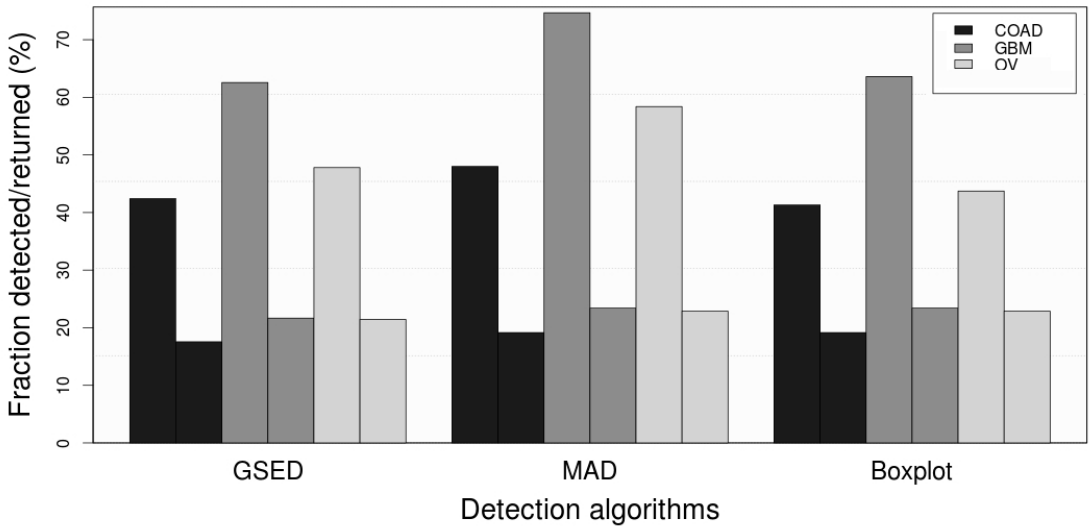

Figure 4.14: Percentage of detected and returned outliers -due to functional similarity and common positions- in the TCGA methylation datasets COAD, GBM and OV. The left column in each group refers to the detected and the right column refers to the returned.

4.4 Discussion

Here, we presented a new robust strategy for detecting outlier samples and genes from gene expression and methylation datasets. As outliers might carry useful information we set filters to remove only the extreme outliers while labeling interesting outliers for further analysis. We presented two modules for outlier detection working at the sample and gene levels. The outlier sample detection module consists of AHC-ED to define outlier samples and the Silhouette coefficient to validate the clustering. In the outlier gene detection module we observed that the underlying distributions of the expression or methylation play a key role in the detection process. The underlying distributions are frequently Gaussian und thus the GESD algorithm would fit for detecting outliers. This module includes two other methods (Boxplot, MAD) that detect outliers regardless of the underlying distribution found.

To validate this approach, we created several expression simulated datasets with introduced sample and gene outliers and searched them using the proposed methods. Simulation datasets were filled either from disjoint or intersected distributions. AHC-ED clustered successfully samples into two classes even in the case where less than 10% of the class rows were generated from two disjoint distributions while the rest came from the same distribution. On the other hand, the more intersected the classes are the less they can be distinguished on the basis of clustering dendrograms. AHC-ED successfully clustered samples filled from intersected distributions but with a less strong Silhouette validation compared to the completely disjoint ones. In simulated datasets, we also introduced 3 mislabeled samples and the clustering mapped them to their original classes. Two additionally introduced pure outlier samples were successfully clustered far most from other classes. Later we tested the outlier sample detection module using one colon cancer public dataset that has a set of known outlier samples. Here the module detected 8 out of the 9 known outlier samples.

We used a similar method to test the outlier gene detection module. We created expression simulated datasets and introduced outlier points for a set of genes. The datasets were filled from several normal distributions. The GSED algorithm detected 90% of the outliers coming from disjoint distributions where Boxplot and MAD detected around 70%. On the other hand, the three algorithms performed less well when the outliers were drawn from a distribution intersecting with the original distribution.

The amount of outlier observations defining an outlier gene remains an open question. In this work we found that two outlier observations can ruin a known co-expression and thus was used as a threshold. Once the outliers are defined, we tested how functionally similar they can be. It is an interesting research topic to study functionally similar outliers that have outlier observations in the same samples. Therefore, outliers fulfilling these conditions were not removed but labeled for further analysis.

This approach was used later to detect outliers in expression and methylation datasets downloaded from public sources TCGA and GEO.

In this approach, it is not possible to automate the removal of sample outliers as it is impossible to fix a threshold for the cuts. The tool generates a dendrogram for the basic clustering and lets the user decide what the tool shall remove.

In summary, we have demonstrated the dramatic effect how a few outlier points may contaminate gene expression or methylation data for further downstream analysis. We make available a convienient tool that implemented established algorithms for detecting outliers. We presented a clear workflow that chooses the most appropriate algorithms depending on the form of the data and on the type of analysis to be presented.

Chapter 5

p62, Hepcidine, and ELOVL6 as Possible Tumor Markers in NASH, Hepatocellular Carcinoma, or Breast Cancer

This chapter presents the results published in two full papers and three letters

- Section 5.1: Kessler SM, Laggai S, Barghash A, Schultheiss C, Lederer E, Artl M, Helms V, Haybaeck J, Kiemer A (2015). IMP2/p62 induces genomic instability and an aggressive hepatocellular carcinoma phenotype. Cell Death and Disease.

- Section 5.2: Kessler, SM, Laggai S, Kiemer A, Barghash A, & Helms V (2015). Hepatic hepcidin expression is decreased in cirrhosis and HCC. Journal of hepatology, 4, 977-979.

- Section 5.3: Barghash A, Helms V, & Kessler SM (2015). Overexpression of IGF2 mRNA-Binding Protein 2 (IMP2/p62) as a feature of basal-like breast cancer correlates with short survival. Scandinavian journal of immunology,82,142–143,

- Section 5.4: Kessle, SM, Laggai S, Barghash A, Helms V, & Kiemer A (2014). Lipid Metabolism Signatures in NASH-Associated HCC—Letter. Cancer research, 74, 2903-2904.

- Section 5.5: Kessler SM, Simon Y, Gemperlein K, Gianmoena K, Cadenas C, Zimmer V, Pokorny J, Barghash A, Helms V, van Rooijen N, Bohle RM, Lammert F, Hengstler JG, Müller R, Haybaeck J, & Kiemer AK (2014). Fatty acid elongation in non-alcoholic steatohepatitis and hepatocellular carcinoma. International journal of molecular sciences, 15, 5762-5773.

resulting from collaboration projects with Dr. Sonja Kessler from the group of Prof. Alexandra Kiemer. For the bioinformatics part, we used the approach displayed in figure 5.1 or some parts of it to analyze complete datasets or parts related to genes of interest for different cancer types. The first section and the letter in the second section show a comprehensive analysis of the IMP2 roles in Hepatocellular carcinoma (HCC) and breast cancer while the letter in third section concentrates on the roles of hepcidine in HCC and liver diseases. The letter in the fourth section and the manuscript in the fifth section discuss the correlations between ELOVL6 expression and liver diseases.

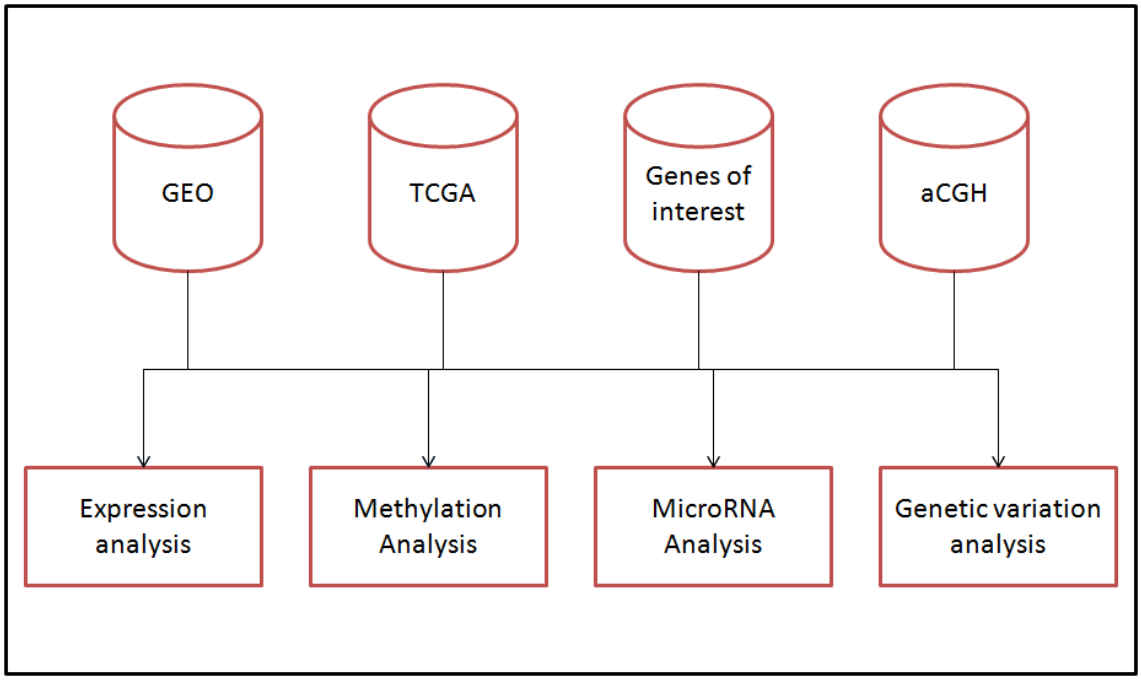

Figure 5.1: Schematic approach used used to process data from tumor samples

5.1 IMP2/p62 Induces Genomic Instability and an Aggressive Hepatocellular Carcinoma Phenotype

Abstract Hepatocellular carcinoma (HCC) represents the second leading cause of cancer−related deaths and commonly develops in inflammatory environments. The *IGF2* mRNA binding protein *IMP2−2/IGF2BP2−2/p62* was originally identified as an autoantigen in hepatocellular carcinoma (HCC). Aim of this study was to investigate a potential pathophysiological role of *p62* in hepatocarcinogenesis. Human HCC tissue showed overexpression of *IMP2*, which strongly correlated with the fetal markers *AFP* and *DLK1/Pref−1/FA−1*, and was particularly elevated in tumors with stem−like features and hypervascularization. Molecular classification of *IMP2* overexpressing tumors revealed an aggressive phenotype. Livers of mice overexpressing the *IMP2* splice variant *p62* highly expressed the stem cell marker *DLK1* and secreted *DLK1* into blood. *p62* was oncogenic: diethylnitrosamine (DEN)−treated *p62* transgenic mice exhibited a higher tumor incidence and multiplicity than wild−types. Tumors of transgenics showed a more aggressive and stem−like phenotype and displayed more oncogenic chromosomal aberrations determined by aCGH analysis. DEN−treated *p62* transgenic mice exhibited distinct signs of inflammation, such as inflammatory cytokine expression and oxidative stress markers, i.e. TBARS levels. Reactive oxygen species (ROS) production was elevated in HepG2 cells, which either overexpressed *p62* or were treated with *DLK1*. *p62* induced this ROS production by a DLK1−dependent induction and activation of the small Rho−GTPase *RAC1*, activating NADPH oxidase and being overexpressed in human HCC. Our data indicate that *p62/IMP2* promotes hepatocarcinogenesis by an amplification of inflammation.

5.1.1 Introduction

Hepatocellular carcinoma (HCC) is the second leading cause of cancer related death [103]. In most cases HCC develops based on an inflammatory etiology, namely chronic hepatitis provoked by either viruses, or alcoholic, and non−alcoholic steatohepatitis. Elevated reactive oxygen species (ROS) generation represents a hallmark of inflammation and promotes carcinogenesis [104]. The insulin−like growth factor 2 (IGF2) mRNA binding protein p62/IMP2−2/IGF2BP2−2 represents a shortened splice variant of IMP2, but harboring the identical mRNA binding domain [105]. Although *p62* was originally identified as an autoantigen in an HCC patient [106], a functional impact of *p62* or IMP2 on hepatocarcinogenesis has not been described as yet. Still, other members of the IMP family, i.e. IMP1 and IMP3, were reported to promote HCC [107][108] and other tumors [109][110]. p62 transgenic mice expressing the transgene exclusively in the liver develop steatosis [111][112] and are more prone to develop steatohepatitis [113]. The animals express elevated levels of the imprinted genes H19 and Igf2 [112], suggesting an effect of *p62* on a specific cluster of imprinted genes [114]. IGF2 displays a key regulator in mammalian growth through metabolic and growth−promoting effects. While *p62* was recently reported to exert its lipogenic actions via IGF2 [111], its anti−apoptotic actions are independent of IGF2 [6]. Also IMP2 was suggested to promote HCC cell survival[115]. Employing transgenic animals and hepatoma cells we here show that *p62* induces an aggressive HCC phenotype, which is linked to inflammatory and oxidant actions of p62. Analyses of publicly available human HCC gene expression data further support *p62* as a marker of human HCC with poor prognosis.

5.1.2 Materials and Methods

The experiments described below were carried out in the Kiemer group. They have been added here for completeness.

Animals

All animal procedures were performed in accordance with the local animal welfare committee. Mice were kept under controlled conditions regarding temperature, humidity, 12 h day/night rhythm, and food access. *p62* transgenic mice expressing the transgene exclusively in the liver were established as previously described [112]. For the short−term experiment *p62* transgenic (p62 tg) mice and matched wild−type (wt) littermates were treated with 100 mg/kg BW DEN i.p. at the age of 2.5 or 5 weeks and sacrifized 48 h later [116][117]. For tumor induction *p62* transgenic mice and wild−type littermates were injected with 5 mg/kg BW at the age of 2 weeks. In long−term experiments mice were sacrifized at an age of 6 and 8 months modelling an early (tumor initiation) and late (tumor progression) tumor stage, respectively [118]. Metastases were investigated in animals older than 10 months (metastatic phase; wt:n=21, tg:n=18).

Real−time Quantitative Polymerase Chain Reaction

Experiments and quantification were performed as described in detail previously [111].

ELISA

Serum levels of IL6 (#m6000b, R&D Systems), TNFα (#mta00b, R&D Systems), and DLK1 (#CSB−EL006945MO, Cusabio Biotech) were performed by ELISA according to manufacturers' instructions.

Immunohistochemistry (IHC)

Primary antibodies used were specific to glutamine synthetase (GS), Golgi membrane protein 73 (Gp73), β−catenin, and *p62* [6]. Samples were examined by two independent investigators blinded to experimental conditions.

Western Blot

Western blot analysis of *p62* protein levels was performed according to [6].

Quantification of Thiobarbituric Acid Reactive Substances

(TBARS) Products of lipid peroxidation were measured as previously described [119].

Cell Culture

Knockdown and overexpression experiments for *p62* in HepG2 were performed as previously described[6]. *p62* sense and antisense constructs are available at Addgene (#42174, #42175). Recombinant DLK1 was used for treatement (#1144−PR−025, R&D Systems).

Caspase-3-like activity assay

Caspase-3-like activity assay was perfomed as previously described [6]. The extraction buffer was slightly modified: 25 mM HEPES pH 7.5, 5 mM MgCl2, 1 mM EGTA, pepstatin, leupeptin, aprotinin (1 μg/ml each).

ROS Assay

ROS assay was performed as previously described [120]. HepG2 cells were loaded with either 20 μM 2',7'−dichlorodihydrofluorescein diacetate (DCF−DA) alone or with the RAC1 inhibitor NSC23766 (#2161/10, R&D Systems) in PBS 60 minutes prior to DLK1 treatment for 48 h after transfection and 50 minutes before measurement, respectively. Combined DLK1 and NSC32766 treatment was done for 5 min. DLK1/H_2O_2 (positive control) treatment over time (5−30 min) was performed in quintuplicates. Combined DLK1 and NSC23766 treatment was done for 5 min.

RAC1 pulldown−assay

Activated RAC1 levels were measured by pulldown assay as previously described [120]. The affinity precipitation assay detects binding of active RAC1 to a fusion protein consisting of the RAC1 target p21−activated kinase 1 and glutathione S−transferase (GST). GST-PBD was expressed in Escherichia coli, purified, and bound to glutathione Sepharose beads (#17-0756-01, GE Healthcare Life Sciences, Germany). For RAC1 pull−down assays HepG2 cells were treated with 1 μg/ml recombinant DLK1 (#1144-PR-025, R&D Systems, Germany), cells were washed with ice-cold PBS, and lysed with PBD-buffer (Tris pH 8.0 25 mM, DTT 1 mM, MgCl2 20mM, NaCl 100 mM, EDTA 0.5 mM, Triton X-100 1%, Aproptinin 0.1%, Leupeptin 0.1%, and PMSF 0.1%). As a positive control one sample was lysed with GTPγ S−PBD−buffer (Tris pH 8.0 25 mM, DTT 1 mM, MgCl2 5 mM, NaCl 100 mM, EDTA 1 mM, Triton X−100 1%, Aprotinin 0.1%, Leupeptin 0.1%, and PMSF 0.1%). After scraping cells off, cells were incubated for complete lysis for 15 min at 4 °C under vigorous shaking. The positive control was incubated for 10 min with GTPγS (10 mM), leading to an exchange of RAC−GDP to RAC−GTP which was stopped by adding MgCl2 (1 M). After centrifugation the supernatants of cell lysates and and positive control were incubated with 30 μl GST−PBD−beads for 2 h at 4 °C under vigorous shaking. After centrifugation and one wash step with PBD−/GTPγS−PBD−buffer, the pellet was frozen at −80 °C.

Pull−down supernatants and pellets with loading buffer were boiled for 10 min. Subsequently, the samples were separated by SDS–PAGE on 12% gels and transferred onto Immobilon−FLPVDF membranes (Rockland, Gilbertsville, PA, USA). The membranes were blocked and incubated with primary antibody overnight at 4 °C, followed by incubation with IRDye conjugated secondary antibody. After washing, blots were scanned with an Odyssey Infrared Imaging System and signal intensities were determined using the Odyssey software.

aCGH analysis

Paraffin−embedded liver tumors were micro−dissected and hybridized against three month−old wild−type liver tissues. Labeling was performed following the BioPrime aCGH Genomic Labeling Module protocol (Invitrogen). The samples were hybridized on an 8x60k CGH Array under the conditions of the Agilent protocol (Version 7.2). The arrays were analyzed with an Agilent DNA Microarray Scanner G2505C and the extraction software Agilent Feature Extraction 11.0.1.1. All data analysis described below was performed by the author of this thesis. The data were analyzed by the statistical software R Bioconductor packages aCGH [121] and CGHcall [60]. In order to compute the similarity of aberrations in the primary tumor and the corresponding metastasis permutation tests were used to calculate the pair−wise statistical significance similar to the method described in [122]. Aberrations were labelled using the bioconductor package aCGH with standard log ratio threshold of 0.25 [121]. The number of matching positions was calculated in the two samples. The

aberration positions of the sample containing fewer aberrations were randomly re−ordered, matched to a random set of aberration positions of the other sample, and the new number of matching positions (ri) was calculated. This step was repeated n=100,000 times and the number of times r, which showed a higher number of matching aberrations of the randomly reshuffled samples compared to the original samples was counted as r = sum (ri> o). The p−value for the statistical significance of matching positions of gains or losses was estimated as p=r/n. Locations of aberrations specifically observed in the *p62* transgenic animals were detected by Golden Helix software: analysis was done using SNP & Variation Suite v8. These loci were compared to the aberrant loci of human HCC samples on www.progenetix.org. GOSim was used to identify enriched Gene Ontology terms [123]. The mutation data were obtained from the Sanger Institute COSMIC web site, http://www.sanger.ac.uk/cosmic. Additionally, the CGHcall package [60] was used to search for significant alterations. CGHcall employs DNAcopy methods [124] to normalize and smoothen the data and defines equal copy number segments for further analysis.

Human GEO datasets

For differential gene expression between tumor (n=247) and non−tumor (n=239) samples the log2 of an RMA normalized dataset (GSE14520) [125] of an AffymetrixGeneChip HG−U133A 2.0 was analyzed. Similarly, differential gene expression was analysed in dataset GSE5975 between positive (n=95) and negative (n=143) EpCAM samples and in dataset GSE20238 between vascular invasive (n=45) and non−invasive (n=34) HCC samples. Outlier data points were removed using part of the methodology described in chapter 4. Differential expression analysis was based on the Kolmogorov–Smirnov test. Pearson correlation was applied to detect correlations between genes of interest.

Identification of common molecular HCC subclasses

Complete hierarchical clustering of dataset GSE14520 [125] was performed using the marker genes presented by Hoshida et al. and Chiang et al [7] [8]. The cluster dendograms are provided below Fig. A3, A4. To test the affiliation of genes with HCC subtypes, the signal−noise−ratio (SNR) was calculated for each marker gene as described in [7][126].

Methylation analysis using a TCGA dataset

TCGA analysis of DNA methylation in HCC was performed using an Illumina Infinium Human Methylation 450 platform. The dataset contains 50 normal and 109 tumor samples. We considered methylation only in the promoter regions (defined within 2,000 bp from the transcription start site provided in the EPD promoter DB [127]. Averages were considered for regions covered by multiple probes.

Statistical analysis

Data analysis and statistics of experimental data were performed using Origin software (OriginPro 8.1G; OriginLabs). All data are displayed either as columns with mean values ± SD or as individual values and boxplots ± interquartile range with mean and median. Statistical differences were estimated by independent two−sample t−test or Wilcoxon−rank−sum test depending on normal distribution, which was tested by the Shapiro−Wilk method, or Fisher−exact−test for categorical data. Normally distributed data comparing multiple groups were analyzed by ANOVA combined with Bonferroni posthoc test. All tests are two−sided and differences were considered statistically significant when p values were less than 0.05.

5.1.3 Results

p62 expression correlates with the stem cell marker DLK1 and promotes hepatocarcinogenesis We investigated IMP2 expression in a large patient cohort (GSE14520) of almost 250 predominantly HBV−positive HCC cases. IMP2 was distinctly overexpressed in tumor (Figures 5.2) compared to normal tissue.

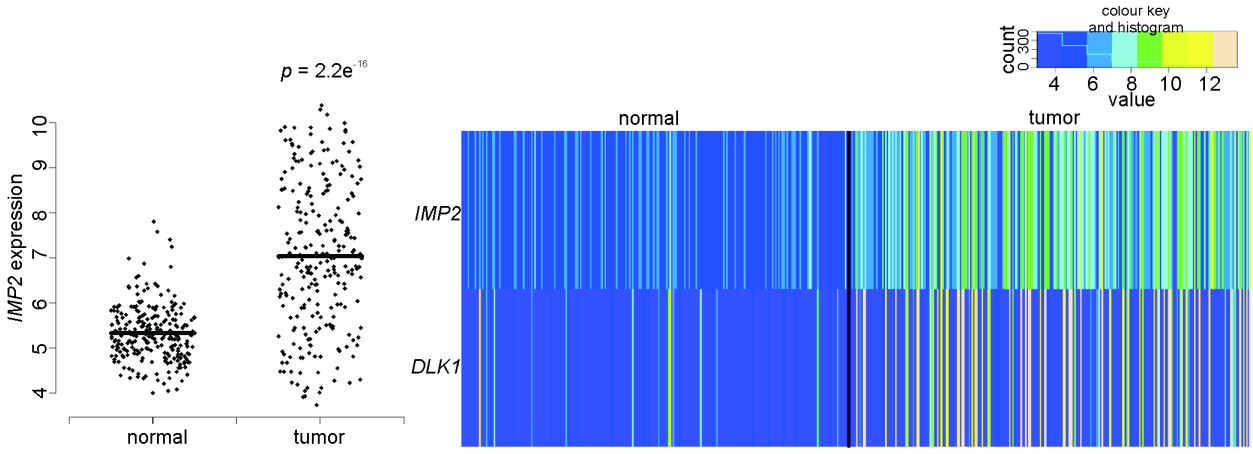

Figure 5.2: Expression analysis of IMP2 in human HCC tumor (n=247) and normal liver (n=239) samples (GSE14520)

IMP2 strongly correlated with α−fetoprotein (AFP) as a marker of poor prognosis (R2=0.63; p<2.2e−16), which was also differentially expressed compared to normal tissue (p<2.2e−16). *p62* was previously shown to induce the expression of the imprinted gene IGF2 [112][6]. Another gene of the same imprinted gene cluster [114], DLK1, represents a marker of hepatic stem cells [128]. DLK1 was significantly overexpressed (p=1.3e−7) (figure 5.2(right)) and its promoter was hypomethylated (p=1.3e−13) (Figure 5.3) in human tumor tissue compared to normal samples. Additionally, its expression is strongly correlated with IMP2 (R2=0.548; p<2.2e−16) and AFP (R2=0.535; p<2.2e−16).

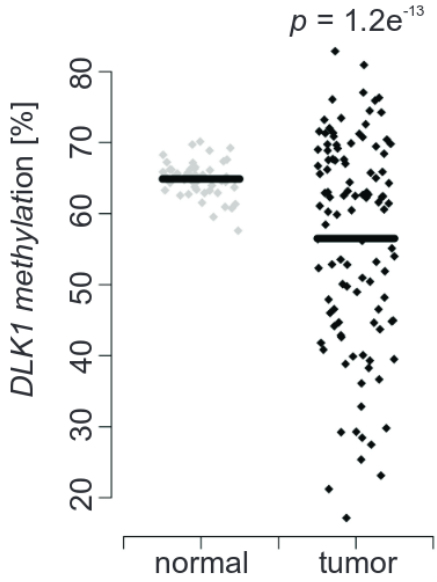

Figure 5.3: DLK1 promoter methylation in human HCC tumor (n=109) and normal liver (n=50) samples (TCGA)

In order to connect IMP2, AFP, and DLK1 overexpression to already known molecular subsets of HCC [129] we performed hierarchical clustering of dataset GSE14520 according to the marker genes identified by Hoshida et al. for three HCC subtypes in mouse [7] and

by Chiang et al. for 5 HCC subtypes in human (Appendix Figures A3,A4) [8]. Subsequent signal−to−noise ratio (SNR) analysis revealed that IMP2, DLK1, and AFP show similar expression patterns as marker genes of class 1, in which 84% of A3 and 65% of A4 marker genes were found. Class 2 can be described by subclass S3 presented by Hoshida et al. (see Fig. 5.4(left)).Clustering by Chiang's marker genes resulted in three major classes (Fig. A4). Here, class 1, which included IMP2, DLK1, and AFP, was well related to Chiang's proliferation class. Class 2a can be described by elevated CTNNB1, Interferon, and Poly7 subclasses. Class 2b, however, was not related to any of Chiang's subclasses (Fig. 5.4 (right)).

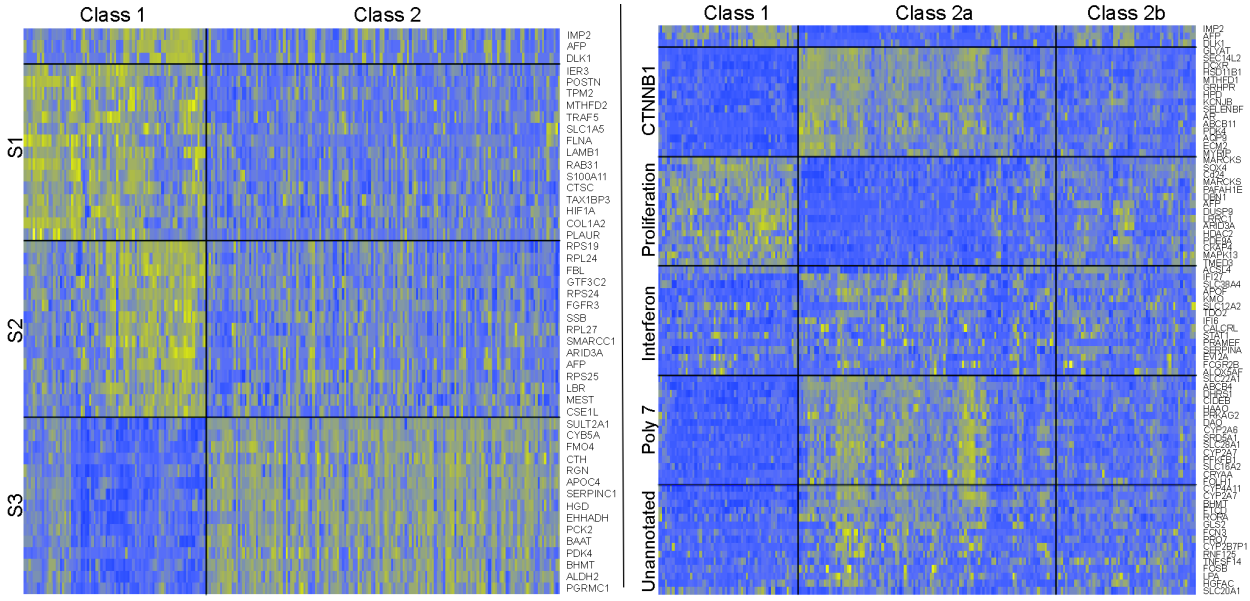

Figure 5.4: Heatmaps of clustering analysis according to Hoshida's (left) and Chiang's (right) HCC subsets

A causal link of DLK1 expression to IMP2 was given by the fact that DLK1 mRNA and protein were increased in livers overexpressing the IMP2 splice variant *p62* compared to wild−types (Figures 5.5).

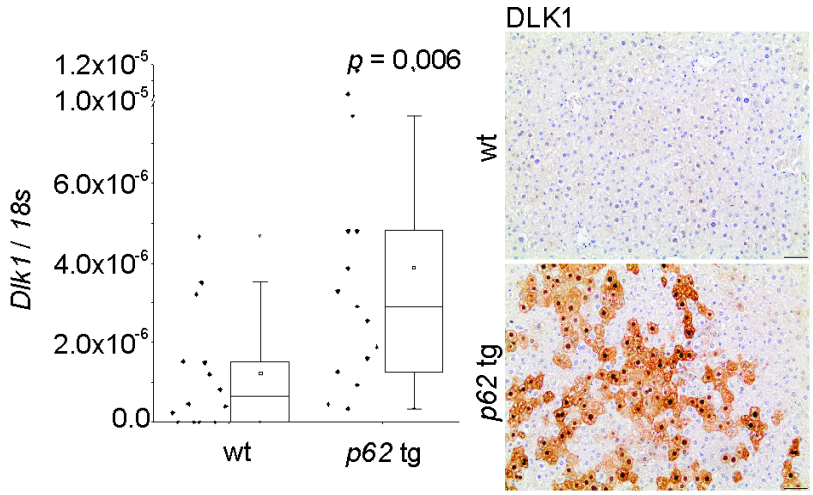

Figure 5.5: Left: DLK1 mRNA levels in livers of untreated animals 5 weeks of age: wild-type (wt) (n=14), *p62* transgenic (p62 tg) (n=15). Error bars show the interquartile range
Right: Representative immunohistochemical staining for DLK1 in untreated 5 week-old mice. Scale bars: 50 μm

Interestingly, also secreted DLK1 was elevated in serum of *p62* transgenic animals (Figure 5.6). As *p62* induced the stem cell marker DLK1 we aimed to investigate the role of

p62 in hepatocarcinogenesis employing the DEN model. Both tumor incidence and tumor multiplicity were increased in DEN−treated *p62* transgenic animals during the early and late stage of tumor development (Figure 5.7).

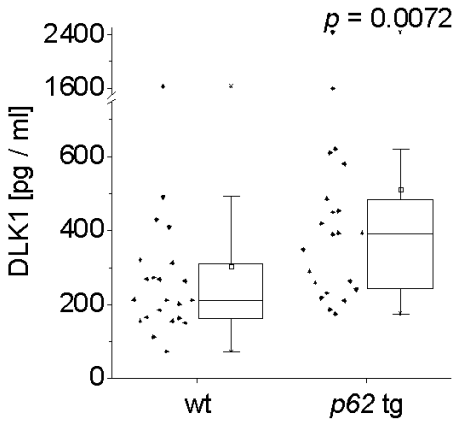

Figure 5.6: Serum DLK1 protein levels in 5 week-old wt (n=22) and *p62* tg (n=22) mice. Error bars show the interquartile range

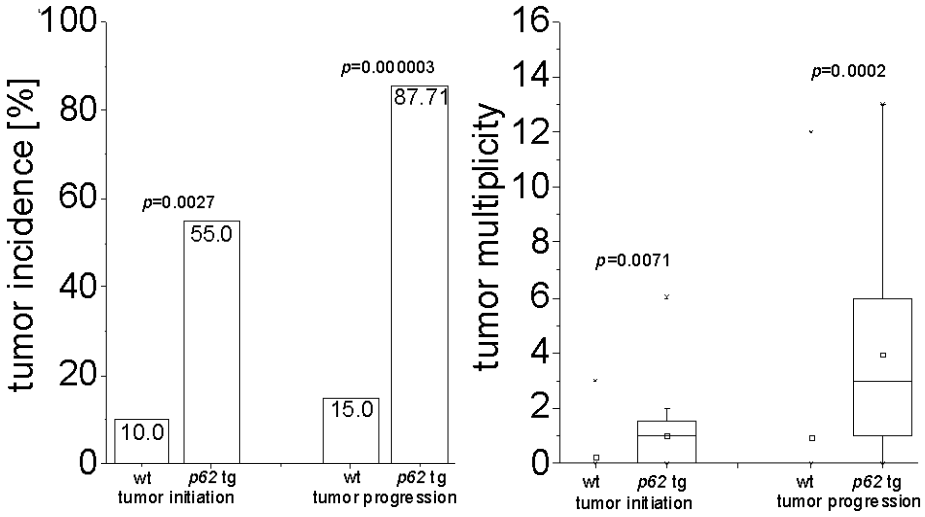

Figure 5.7: Tumor incidence (left) and tumor multiplicity (right) in early stage (6 months: wt: n=20; *p62* tg: n=20) (tumor initiation) and late stage (8 months: wt: n=20; *p62* tg: n=20) (tumor progression) of DEN-treated mice. Error bars show the interquartile range

After 48 h of DEN treatment, which models early liver cell damage [117], *p62* transgenic mice revealed a more pronounced inflammatory response as shown by increased lobular lymphocytic as well as granulocytic infiltrations (Figure 5.8) and by elevated serum levels of the inflammatory cytokines IL6 and TNFα (Figure 5.9). Neither AST nor ALT levels were different in *p62* transgenic animals compared to DEN−treated wild−type mice. Still, apoptosis was reduced in DEN-treated p62 transgenic animals (Figure 5.10)

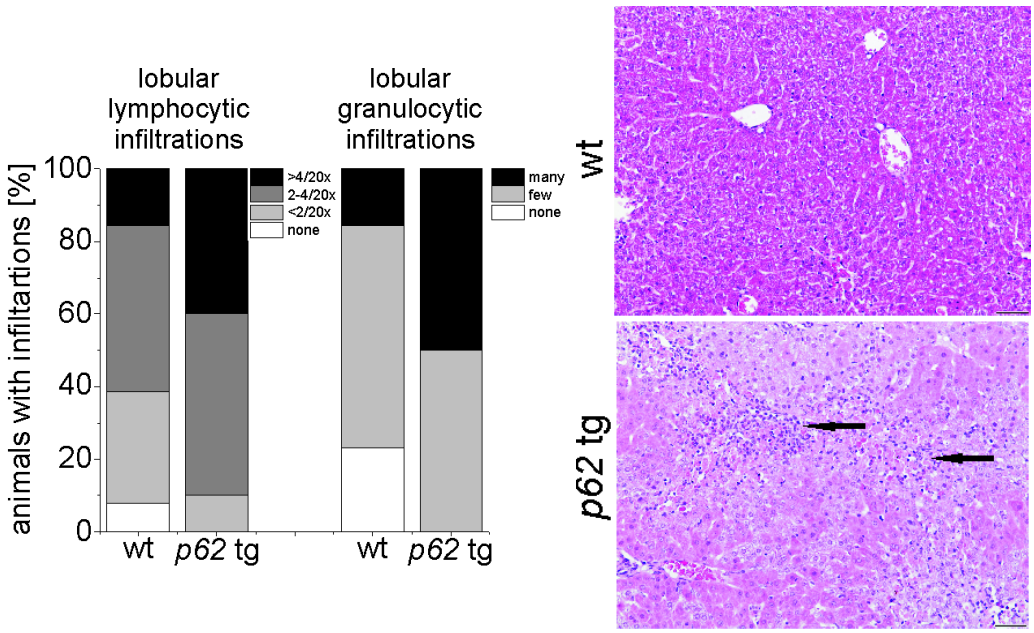

Figure 5.8: Histological scoring of HE stainings and representative picture for lobular lymphocytic and granulocytic infiltrations 48 h after DEN application. Arrows denote mixed lymphocytic and granulocytic infiltrations. Scale bar: 50 μm

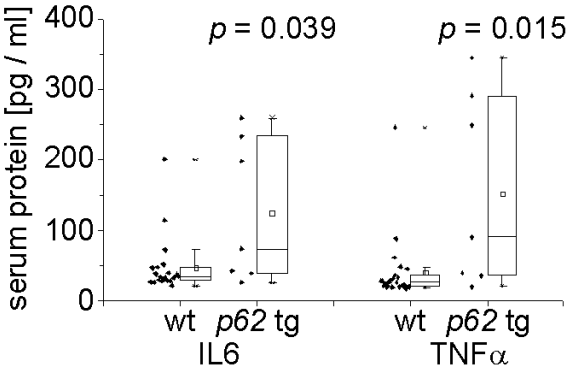

Figure 5.9: Serum protein levels of IL6 (left) and TNFα (right) of 5 week-old wt (n=22) and *p62* tg (n=7) mice 48 h after DEN application. Error bars show the interquartile range

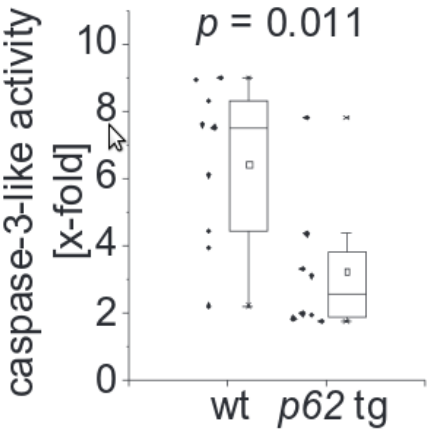

Figure 5.10: Caspase-3-like activity in DEN-treated 5 week-old wt (n=9) and p62 tg (n=8) mice 48 h after DEN injection normalized to untreated wt

Tumors of *p62* transgenic mice show a more aggressive phenotype

In order to characterize the DEN−induced tumors, paraffin sections were stained for the tumor markers Golgi membrane protein 73 (Gp73) and glutamine synthetase (GS). All wild−type tumors were Gp73 positive, whereas in transgenics only 70.31% were Gp73 positive. Interestingly, while none of the wild−type tumors stained positive for GS, 29.69% of p62−tumors were GS positive and half of them were positive for both markers (Figure 5.11).

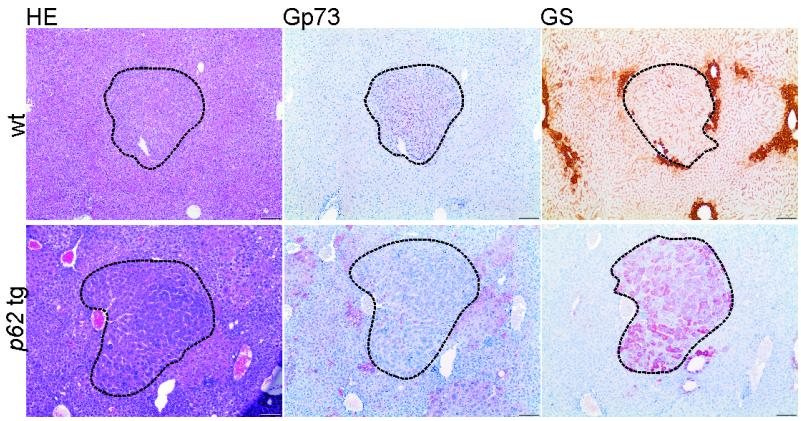

Figure 5.11: Representative HE and immunostainings against Golgi membrane protein 73 (Gp73) and glutamine synthetase (GS) in wt and *p62* tg mice in late stage tumors

GS positivity is regarded as a marker of β−catenin activation [130], which can be regulated by activation of the canonical wingless−int (WNT) pathway. Concordantly, β−catenin staining confirmed its activation by nuclear and cytoplasmatic localization in tumor tissue, while normal tissue showed a membranous pattern (Figure 5.12).

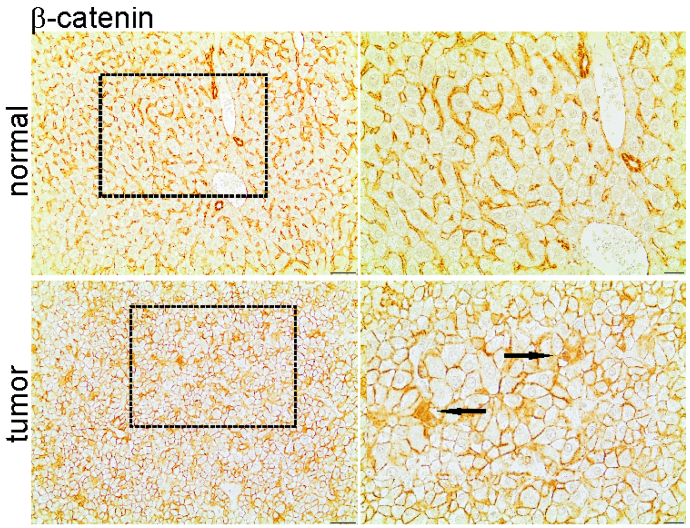

Figure 5.12: Representative β-catenin immunostaining in adjacent normal and tumor tissue of livers bearing GS-positive tumors. Scale bars: 50 μm (left), inset (right): 20 μm

WNT10B, a canonical WNT pathway member, which is highly expressed in fetal, but shut down in adult liver, was increased in *p62* transgenic animals (Figure 5.13). Tumors of transgenic animals were more mitotically active (p=0.0477) by irregular mitosis (Figure 5.14) and were rather pleomorphic (0% in wt versus 15.6% in tg, p=0.014). mRNA levels of the pro-proliferation growth factor *igf2* tended to be increased in *p62* transgenic animals (figure 5.15). CK19−positive oval cell compartments were solely observed in tumors of transgenics (Figure 5.16).

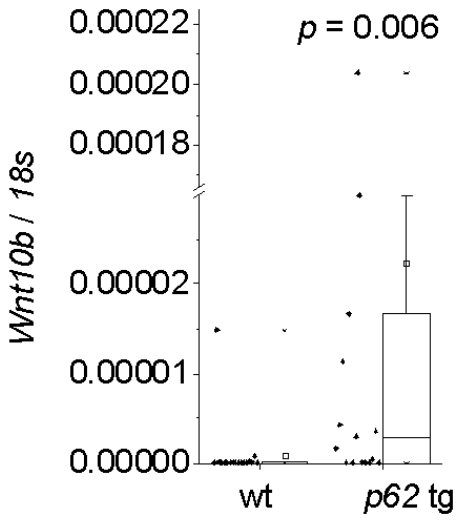

Figure 5.13: *WNT10B* mRNA levels in wt (n=18) and *p62* tg (n=18) in the late tumor stage. Error bars show the interquartile range

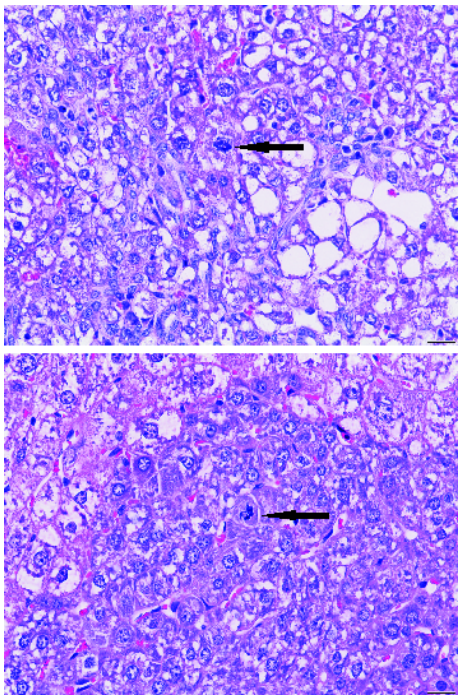

Figure 5.14: Arrows show irregular mitosis in representative HE stainings in tumors of *p62* tg mice. Scale bars: 20 μm.

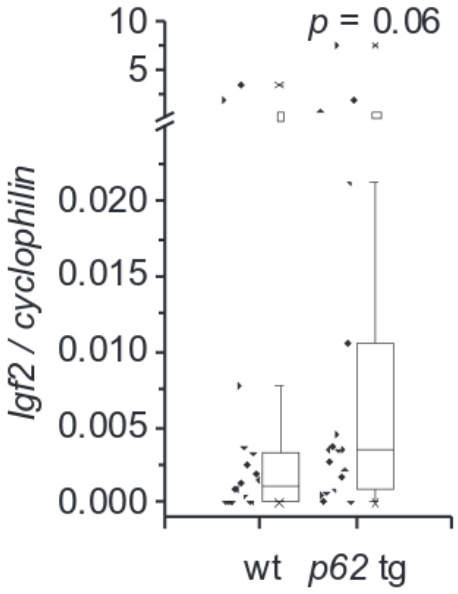

Figure 5.15: *Igf2* mRNA levels in wt (n=18) and p62 tg (n=18) in the late tumor stage. Error bars show the interquartile range

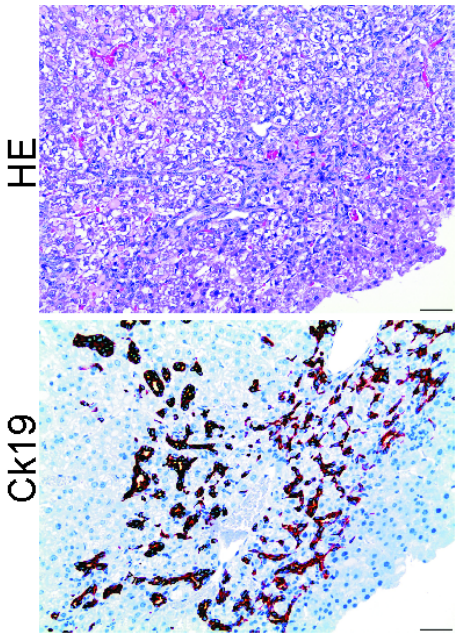

Figure 5.16: Representative HE and corresponding oval cell marker CK19 immunostaining in *p62* tg tumors. Scale bars: 50 μm

Concordantly, human HCCs positive for the oval cell marker EpCAM exhibited higher expression levels of IMP2 compared to EpCAM−negative HCCs in an HBV−positive HCC cohort (238 samples; GSE5975) (Figure 5.17). Vascular invasion as well as lung metastases developed in both wild−type as well as in transgenic animals(Figure 5.18).

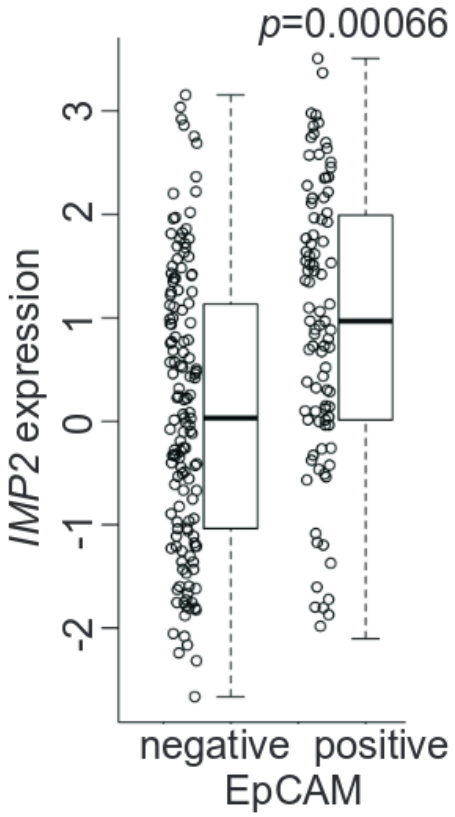

Figure 5.17: *IMP2* expression in human HCCs grouped into EpCAM-positive and -negative tumors (238 samples; GSE5975)

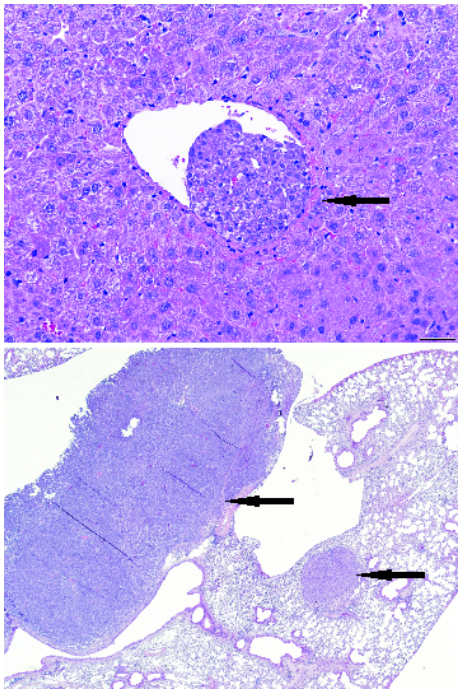

Figure 5.18: Representative HE sections showing vascular invasion (upper panel, scale bar: 50 μm) and lung metastases (lower panel, original magnification: 40x). Arrows designate metastatic foci

Analysis of a GEO dataset of 226 predominantly viral hepatitis related HCC cases (GSE20238) categorized by the presence or absence of vascular invasion revealed increased IMP2 expression in patients with vascular invasion (Figure 5.19). Murine lung metastases showed the same staining pattern for the HCC markers GS and Gp73 as the primary liver

tumors of wild−type and *p62* transgenic mice (Figure 5.20).

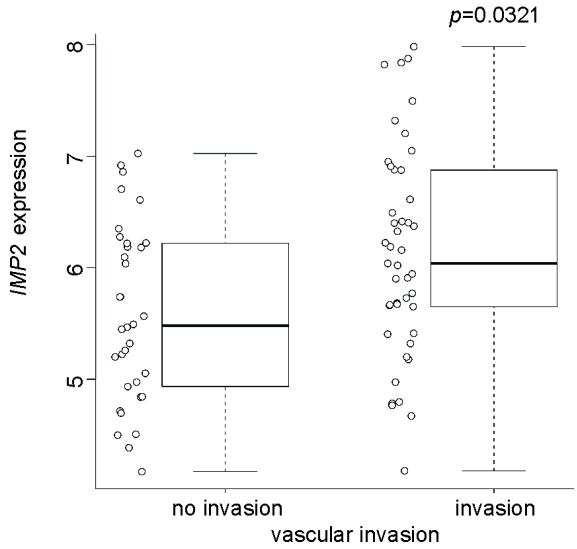

Figure 5.19: *IMP2* expression in human HCCs grouped into tumors positive or negative regarding vascular invasion (91 samples; GSE20238)

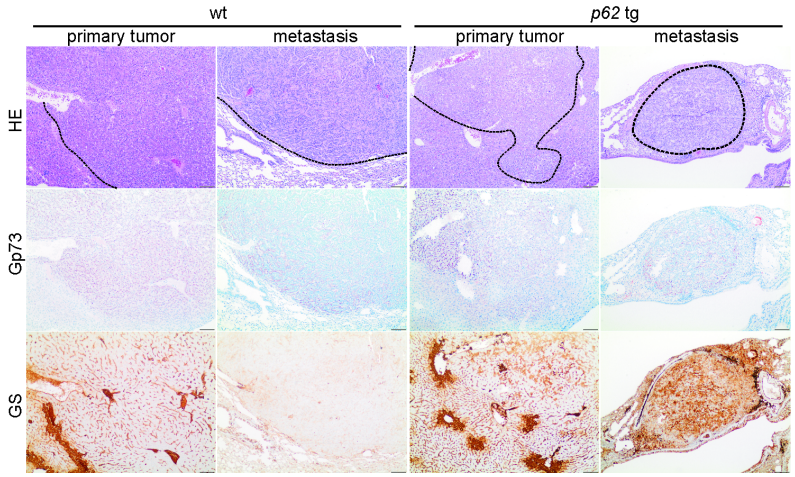

Figure 5.20: Representative HE and Gp73 and GS IHC of primary liver tumor and lung metastasis in wt (n=21) and *p62* tg (n=18) mice in the metastatic phase

In the metastatic phase also some wild−type tumors showed positive GS staining (data not shown). aCGH analysis confirmed clonality of primary tumors and both intrahepatic ($p < 10-5$) as well as lung metastases ($p < 10-5$) (Figure 5.21).

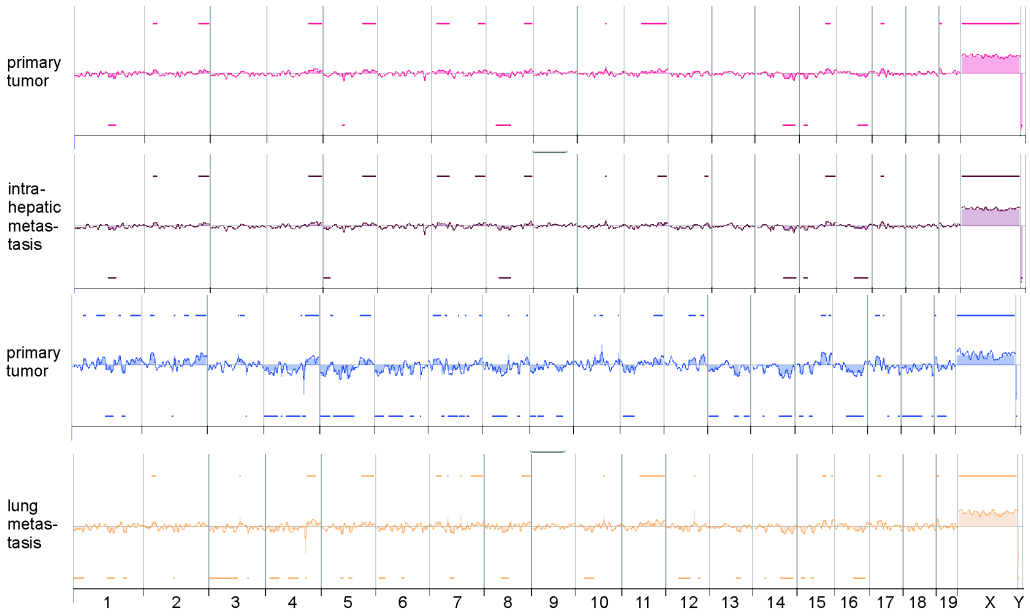

Figure 5.21: Representative aCGH plots of primary HCC and corresponding intrahepatic metastasis (left) and of primary HCC and corresponding lung metastasis (right) of *p62* transgenic mice

***p62* transgenic mice are more susceptible to chromosomal aberrations**
aCGH analysis (Figure 5.22) revealed increased alterations in tumors of transgenic (lower panel) compared to wild-type animals (upper panel). Significant gains were only observed in transgenic tumors and significant losses were stronger in transgenic compared to wild−type tumors (Figure 5.23). Some loci only showed aberrations in *p62* transgenic mice. Gene Ontology analysis revealed that the affected loci harbour genes, which are involved in growth, proliferation, negative apoptosis signalling, and angiogenesis. Interestingly, the distal mouse 15B3.1−C region, amplified only in *p62* transgenics and corresponding to the human distal chromosome 8q23.1−23.3, is the second most frequently amplified region in human HCC: array−CGH results from 848 HCC samples show an amplification in about 45% of cases (www.progenetix.net). This region comprises genes commonly mutated in cancer.

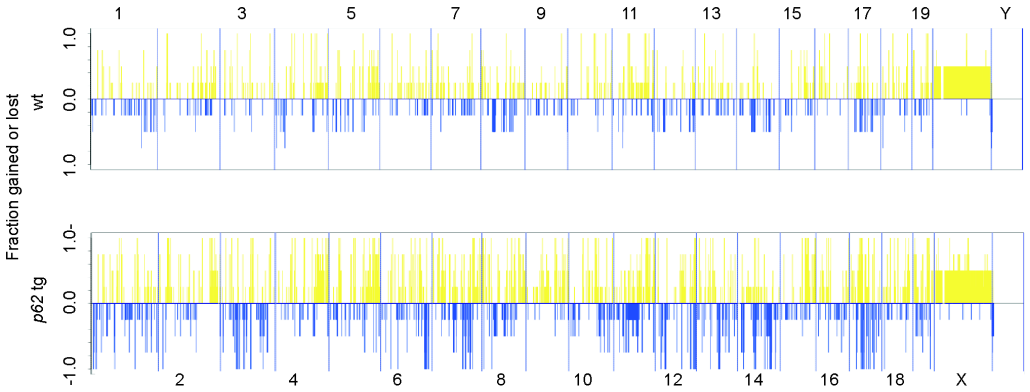

Figure 5.22: Frequency plot of fractions gained or lost along the genome of primary tumors in wt (n=4; upper panel) and *p62* tg (n=4; lower panel) mice in the late tumor stage

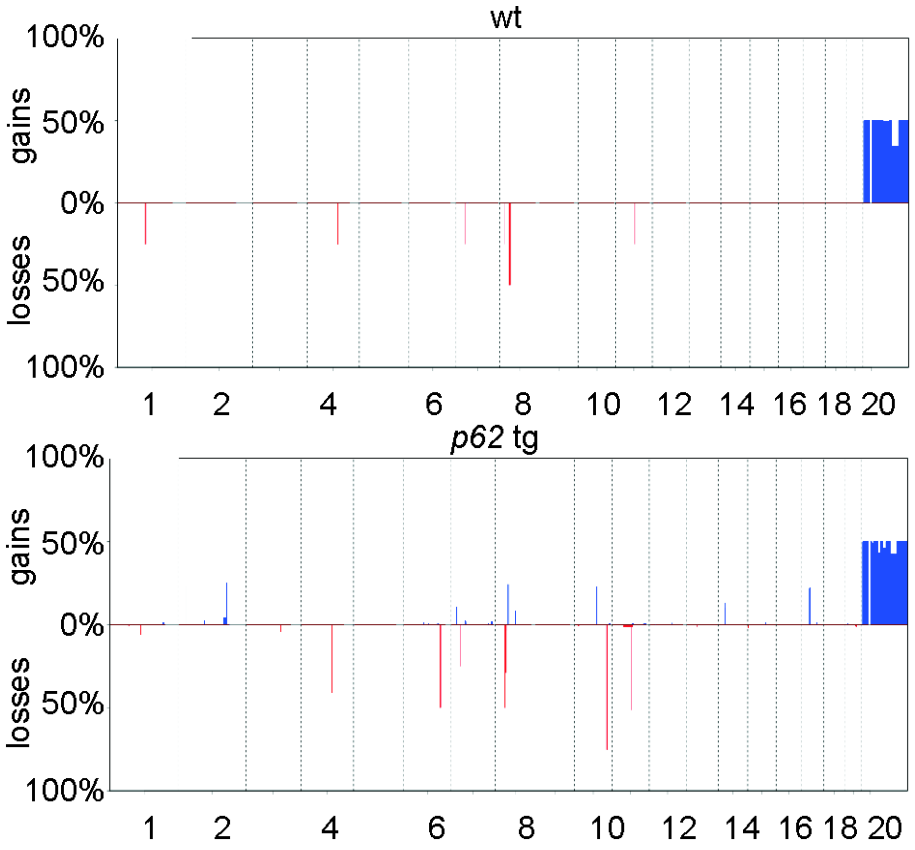

Figure 5.23: Most significant alterations in primary tumors of wt (top) and p62 tg (bottom) animals during the late tumor stage. Shown are percentages of gains and losses for individual altered segments obtained with the CGHcall package

Tumor-promoting DLK1 drives RAC1−induced ROS formation

We sought to identify the reason for p62−induced increased genomic instability and found significantly elevated levels of TBARS as indicators of oxidant stress in *p62* transgenic animals after short−term treatment with DEN (Figure 5.24).

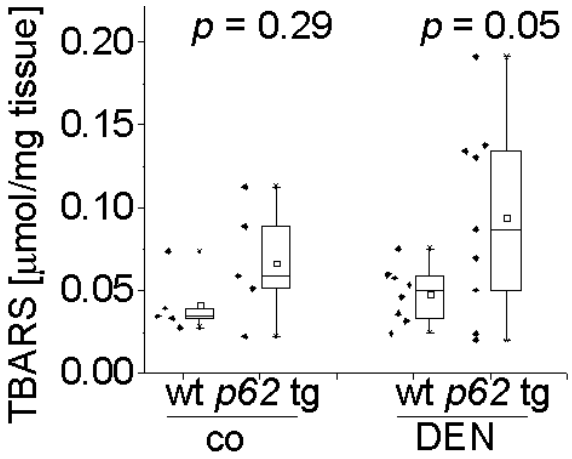

Figure 5.24: Hepatic TBARS levels in wt (n=5) and p62 tg (n=5) livers of untreated (co) and 48 h DEN-treated (DEN) animals (wt:n=8; tg:n=9). Error bars show the interquartile range

ROS are important inducers of DNA damage and chromosomal instability [104]. NADPH oxidase represents an ROS−generating enzyme complex that contributes to DEN−induced carcinogenesis [131]. NADPH oxidase is activated by the small GTPase RAC1 [120] and DLK1 was previously shown to induce RAC1 [132]. We observed increased levels of both Dlk1 and Rac1 mRNA in *p62* transgenic livers and a strong correlation between each other (Pearson R2=0.56, p=0.015) (Figures 5.21-bottom, 5.22, 5.23-top). The secreted form of DLK1 was elevated in *p62* transgenic mice (Figure 5.26).

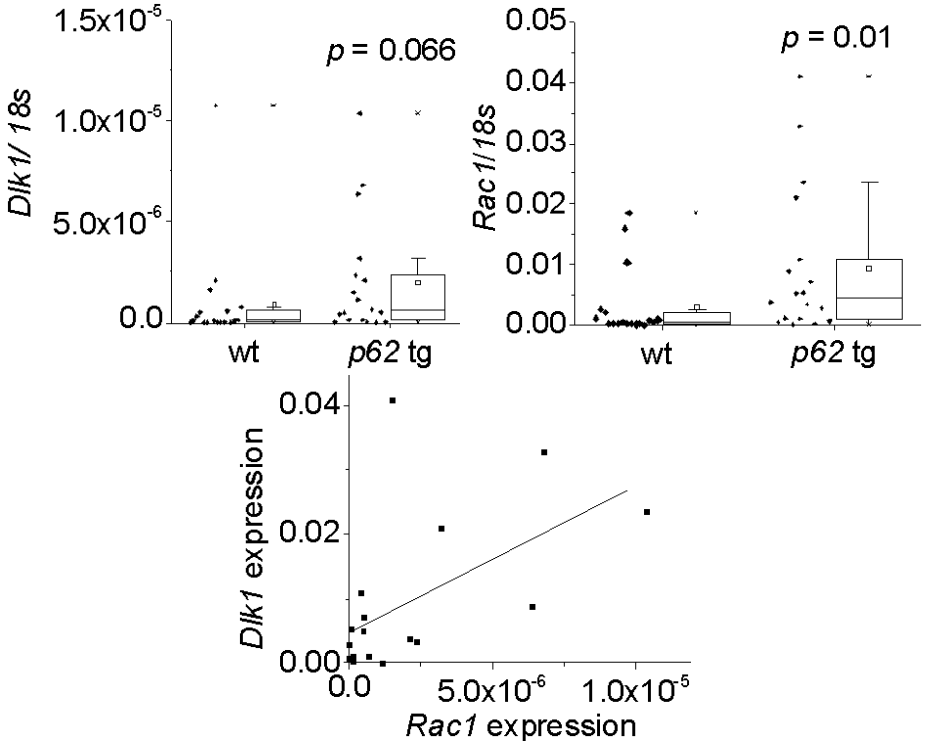

Figure 5.25: *DLK1* (top left) and *RAC1* (top right) mRNA expression as well as correlation of both (bottom) was investigated after 8 months (wt: n=18; tg:n=18). Error bars show the interquartile range

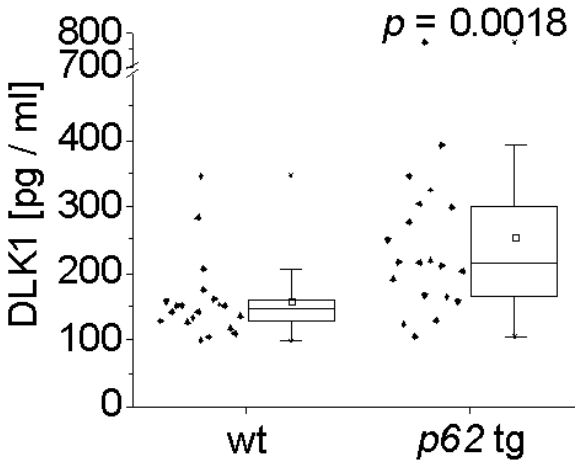

Figure 5.26: Secreted *DLK1* protein serum levels were measured by ELISA. Error bars show the interquartile range

In order to test the causal effect of *p62* and DLK1 on RAC1, in vitro experiments were performed. DLK1 treatment increased RAC1 mRNA levels as well as activated RAC1 protein as detected by pull−down assay in HepG2 cells (Figure 5.27).

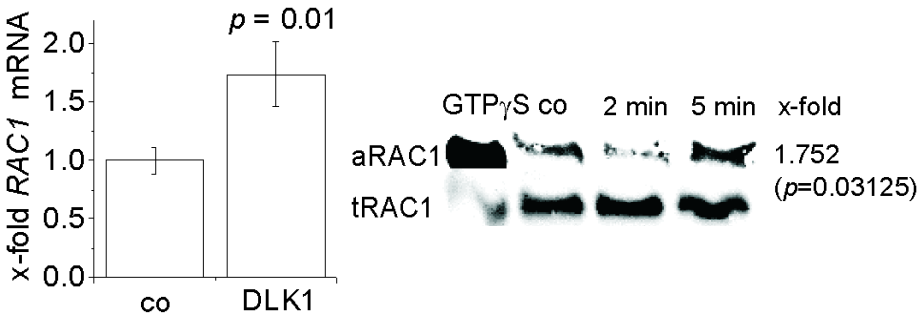

Figure 5.27: Levels of RAC1 mRNA presented as mean +/- sem (top) and activated RAC1 protein levels determined by pulldown assay (bottom) in HepG2 cells after treatment with 1 μg/ml DLK1 protein (n=3 in duplicate). Bottom: Representative pull-down assay with activated RAC1 (aRAC1) and total RAC1 (tRAC1) is shown. X-fold signal intensities of 5 min treatment with DLK1 were normalized to untreated control (co)

Furthermore, DLK1 treatment increased ROS levels, which was completely abrogated by pre−incubation with the RAC1 inhibitor NSC23766 (Figure 5.28). Also *p62* overexpression increased RAC1 expression (Figure 5.29). Vice versa, knockdown of *p62* led to decreased RAC1 mRNA levels (Figure 5.29). ROS levels were elevated after *p62* overexpression by 9.46± 1.24% (p=0.045) 48 h after transfection, which was abrogated by the RAC1 inhibitor (p=0.0046). Finally, the human HCC cohort, which showed differential expression of IMP2 and DLK1 (Figure 5.2), significantly overexpressed RAC1 (Figure 5.30) (p<2.2e−16). SNR analysis revealed RAC1 overexpression in class 1, which was characterized by IGF2BP2, AFP, and DLK1 overexpression (Fig. 5.4-left).

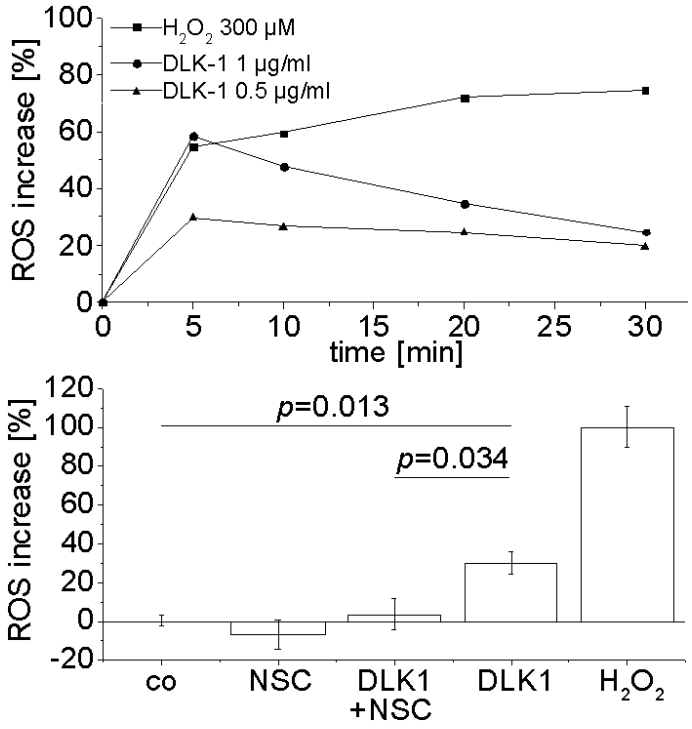

Figure 5.28: ROS levels: representative experiment (quintuplicates) of HepG2 cells treated with 0.5 or 1 μg/ml DLK1 or H2O2 as positive control for 0-30 min (upper part). Data are normalized to untreated HepG2 cells. ROS levels in HepG2 cells treated with either DLK1 or RAC1 inhibitor NSC23766 alone or in combination (lower part). Untreated HepG2 cells served as control. H2O2-induced ROS formation was set to 100% (n=2, quintuplicate). Data are presented as mean +/− sem.

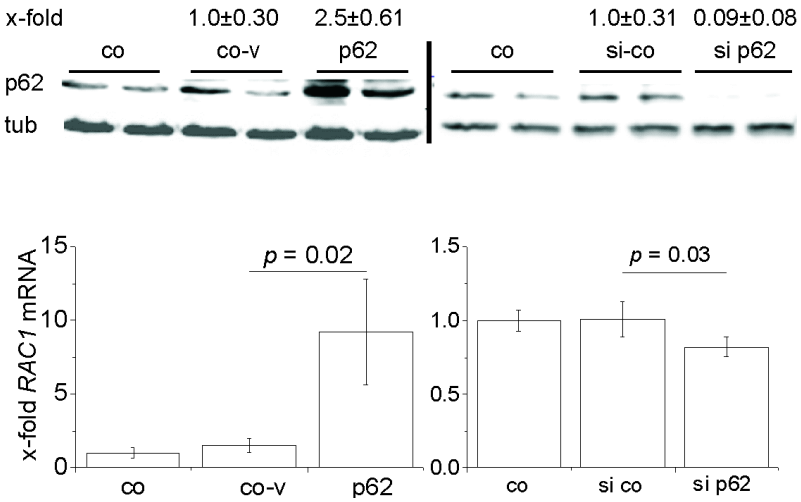

Figure 5.29: RAC1 expression in HepG2 cells overexpressing (top) p62-sense plasmid (p62) compared to antisense-plasmid (co-v), untreated control (co), and siRNA knockdown (bottom) of p62 (si p62) compared to random siRNA (si co) (n=3 triplicate/quadruplicate). Data show mean +/− sem. Western blot knockdown/overexpression control was densitometrically quantified (n=4 triplicate/quadruplicate; upper part)

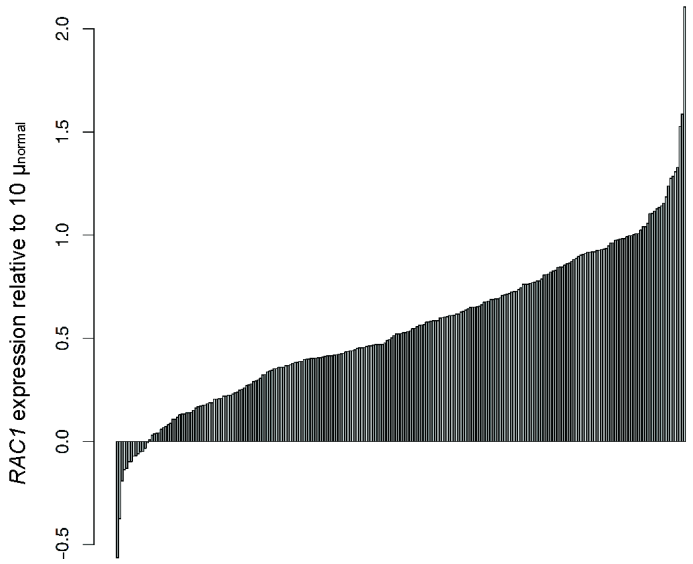

Figure 5.30: *RAC1* expression in human HCC (GSE14520) normalized to the mean of normal samples

5.1.4 Discussion

The IMP *p62* was originally identified as a tumor−associated auto−antigen with auto−antibodies against *p62* detected in HCC patients [106] and in several other types of cancer [133][134]. Interestingly, despite several investigations of *p62* autoantibodies as a potential tumor marker and a recently suggested resistance of *IMP2* knockout mice towards malignancy [135], functional implications of the *p62* protein in carcinogenesis are widely unknown. Our analysis of a large homogenous human HCC cohort with about 250 viral HCC samples showed strongly increased expression of IMP2 in the majority of HCC patients. These data are supported by other reports suggesting elevated levels of *p62* in HCC tissue in rather small patient cohorts[111][136]. According to the classification performed in this study, overexpression of AFP and IGF2, both correlating with IMP2 expression in HCC (present findings and [111]) marks Hoshida's S2 class of aggressive HCC [137]. Positivity of the stem cell surface antigen EpCAM and vascular invasion, which we observed to be linked to IMP2 overexpression, was used as a classification system by others [138][139]. In fact, EpCAM expression is associated with early recurrence and short survival time [140].

Regarding the classification from Boyault et al., IMP2 overexpressing samples probably belong to the G1 subset, which is characterized by an increased expression of AFP and the imprinted gene products IGF2, and H19 [141]. *p62* transgenic mice were shown to overexpress both imprinted genes [112]. Finally, IMP2 overexpressing samples match the molecular pattern of Cairo's aggressive hepatoblastoma, in which AFP, Krt19, and EpCAM are elevated. In the same study the authors provide data from Myc−induced murine tumors highly expressing DLK1, IGF2, and AFP [142].

Interestingly, we observed a correlation of IMP2 expression with the oval/stem cell marker DLK1 [128]. DLK1 was previously shown to correspond with poor survival in HCC [143]. Oval cells share phenotypic markers with embryonic hepatoblasts, in which DLK1 is also highly expressed [144]. The cytoplasmatic appearance of DLK1 in *p62* transgenic mice reveals a fetal phenotype (Figure 4H) as previously reported for HCC and hepatoblastoma tissue [145].

Secreted DLK1, suggested as a serum marker for hepatoblastoma [146], was elevated in sera of *p62* transgenic mice. Secreted DLK1 was suggested to have paracrine functions, i.e. inducing the secretion of inflammatory cytokines, such as TNFα and IL6 in monocytes and adipocytes [147]. Recently, *p62* expression was shown to promote liver disease by amplifying inflammatory processes [111][113][116][119][148]. HCC mostly develops within an inflammatory environment, such as viral hepatitis, ASH, and NASH, and inflammatory mediators promote hepatocarcinogenesis [117]. We here present a transgenic mouse model, which develops HCC out of an inflammatory state involving elevated IL6 and TNFα production. We observed an early onset and an accelerated progression of HCC in *p62* transgenic mice.

There are two different models using the carcinogen DEN to induce liver tumors. DEN is either given as a single dose by itself or in combination with the tumor−promoting agent phenobarbital to induce tumors with β−catenin mutations, which are linked to GS positivity [130]. Interestingly, employing *p62* transgenic mice, we observed GS−positive tumors in the DEN model without using phenobarbital. The expression of DLK1 is closely linked to *WNT10B* , a member of the canonical WNT pathway, leading to β−catenin accumulation in the cytoplasm and the nucleus, which can be altered by DLK1 [149]. Both elevated *WNT10B* and cytoplasmatic/nuclear localization can be found in *p62* transgenic tumors.

In tumors positive for the stem cell marker EpCAM, co−expression of DLK1 and AFP was defined by poor prognosis [150]. Tumors of *p62* transgenic livers were more susceptible to chromosomal aberrations than tumors of wild−type animals and showed more pronounced alterations. Increasing levels of chromosomal instability correlate with progression of HCC, suggesting that marked genomic instability characterizes more advanced stages of the disease. The homologue of human 8q23, amplified specifically in *p62* transgenic animals, is frequently gained in human HCC tissues [151]. Interestingly, amplification of the homologue of human chromosome 3q, which was gained in *p62* transgenic tumors, is correlated with advanced−stage disease in cervical carcinomas [152]. The losses specifically observed in *p62* transgenics on the homologues of human chromosomes 9q33.3−34.3, 11q23.1−24.1, 16q42.13−42.2, and 21q22.11−3 were reported to be deleted in different types of cancer including HCC [153][154].

Genomic instability can be induced through ROS production [104]. A major ROS−generating enzyme complex, the NADPH oxidase, is activated by the small GTPase RAC1 [120]. We found RAC1 to be highly overexpressed in a large proportion of HCC tissues. RAC1 itself has been described to play a role in HCC [155] and might act at least partly via ROS production [156]. Also Ras−induced ROS production and DNA damage has been linked to RAC1 activation [157]. Our data functionally link the aggressive and de−differentiated phenotype of the tumors in *p62* transgenic livers to DLK1−facilitated induction of RAC1. The stem cell marker and paracrine factor DLK1 was previously reported to induce RAC1 activation in 3T3−L1 cells [132]. We here report that DLK1−induced RAC1 activation leads to elevated ROS levels (Figure 5.31).

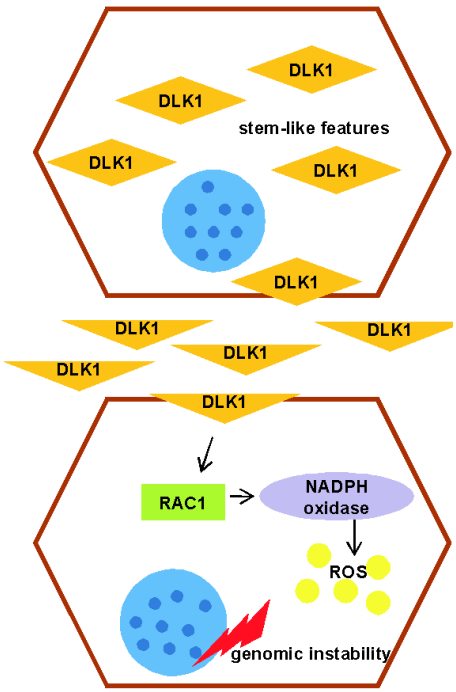

Figure 5.31: Overview of p62-promoted DLK1-RAC1-induced genomic instability. DLK1 overexpressing cells with stem-cell-like features secrete DLK1 protein, which activates RAC1 in a paracrine fashion, in turn leading to ROS generation via NADPH oxidase. Elevated ROS levels finally result in genomic instability

We suggest that the DLK1/RAC1−induced increase in ROS is the cause of chromosomal instability [104], which in turn leads to more undifferentiated tumors [158]. Interestingly, RAC1 activation was shown to drive proliferation of intestinal stem cells [159] and targeting RAC1 suppresses cancer cell viability [160], cancer stem cell activities [161], and metastasis [162]. Wang and colleagues reported that RAC GTPase−activating protein 1 is associated with early recurrence in HCC [163].

Taken together, our in vivo, in vitro, and in silico analyses show that IMP2/p62 plays an important role in HCC initiation and progression and characterizes human HCC prognosis.

5.2 Overexpression of IGF2 mRNA-Binding Protein 2 (IMP2/p62) as a Feature of Basal-like Breast Cancer Correlates with Short Survival

Recent evidence suggested that autoantibodies against IMP2/p62 may be useful serum biomarkers for early-stage breast cancer screening and diagnosis [164]. The study by Liu et al. [164] elegantly demonstrated that the frequency of autoantibodies against IMP2 and IMP2 expression itself is significantly increased in breast tumour tissues compared to normal tissues. An autoimmune response to IMP2/p62 is also known for other tumours, for example colon carcinoma and hepatocellular carcinoma (HCC) [133][165]. To test the relevance of IMP2 expression for prognosis in breast cancer, we analysed a large human Gene Omnibus (GEO) data set (GSE42568 [166]). Interestingly, high IMP2 expression correlated with short survival (Fig. 5.32). Therefore, IMP2 expression could serve not only as a diagnostic but also as a prognostic biomarker. It is well known that there are different

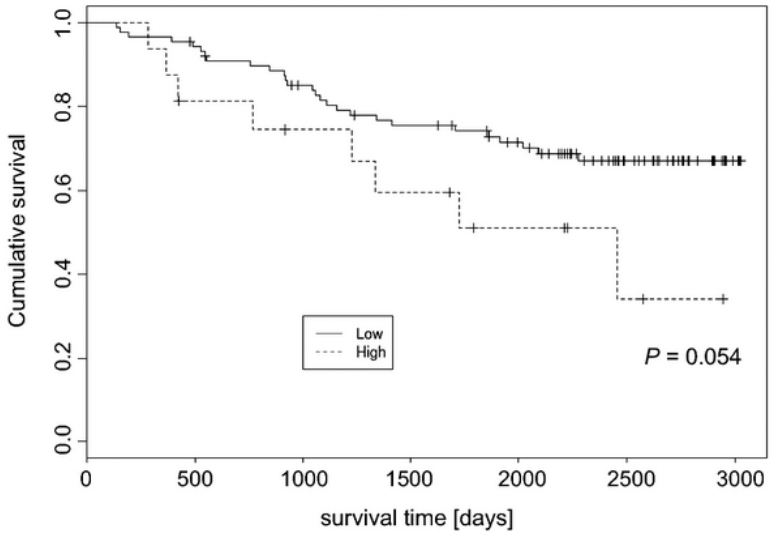

Figure 5.32: Kaplan–Meier survival plot referring to "low" and "high" IMP2 expression levels in data set GSE42568 (n = 104). High expression are those samples with IMP2 expression higher than 5, and low expression smaller than 5, respectively

classes of human breast tumours, which are characterized by different molecular patterns [167]. Luminal cancers are the most common subtype. The basal-like subtype, which mostly corresponds to the triple-negative subtype, stands for about 20% of breast cancer cases with a shorter survival than the luminal subtype [168]. To test whether IMP2 expression might be a feature of a specific breast cancer subtype, we analysed an additional human data set, which provided subtype-classified samples (GDS1329 [169]). IMP2 was especially elevated in tissues of basal-like cancer compared to the luminal or apocrine subtype (5.33). The overexpression of IMP2 was confirmed in another set of basal-like breast cancer tissues (GDS2250 [170]) compared to non-basal-like samples and normal tissues (Fig. 5.34). All data sets analyzed in this work are RMA-normalized and downloaded from Gene Omnibus (GEO). Statistical significance was determined by Kolmogorov–Smirnov test. In conclusion, detection of IMP2/p62 expression in breast cancer presented by Liu and colleagues [164] might even be of prognostic relevance as already reported for HCC [6]. Furthermore, the high expression of IMP2 in the basal-like breast cancer subtype might lead to new individualized therapeutics.

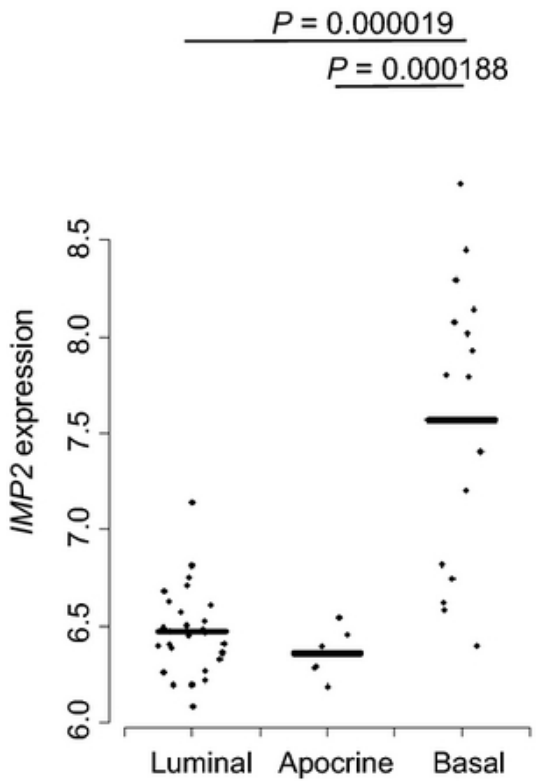

Figure 5.33: IMP2 expression in basal-like breast cancer tissues compared to luminal and apocrine in data set GDS1329

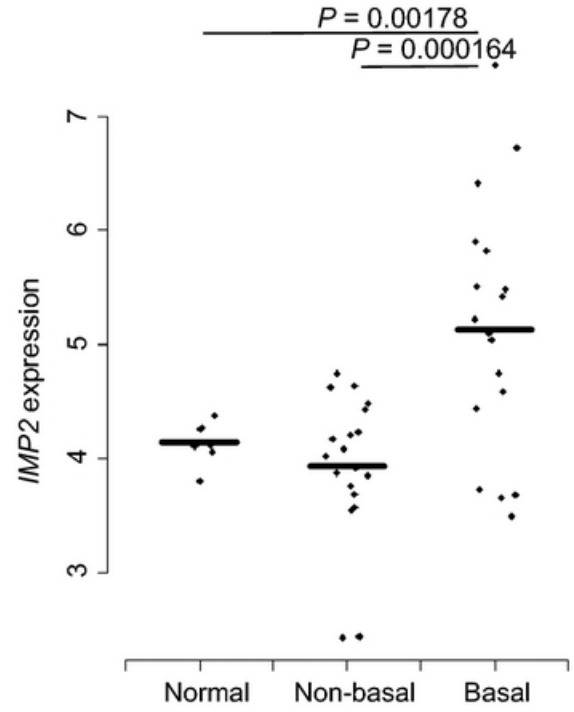

Figure 5.34: IMP2 expression in basal-like breast cancer tissues compared to non-basal-like and normal tissue in data set GDS2250

5.3 Hepatic Hepcidin Expression is Decreased in Cirrhosis and HCC

Recent evidence showed the protective role of the iron homeostasis regulator hepcidin (Hamp) in iron overload-related liver diseases [171]. The study by Lunova et al. elegantly demonstrated that the knockdown of hepcidin promotes hepatic inflammation and fibrogenesis after feeding mice an iron-rich diet [171]. It is well known that perturbations of the iron metabolism, as it is the case in hemochromatosis, can lead to hepatocellular carcinoma (HCC). HCC represents the second most common cancer related death worldwide and displays also the end-stage of liver diseases related to chronic viral or non-viral hepatitis. As hepcidin deficient mice were more prone to develop fibrosis [171], which is itself a risk factor for HCC, deregulation of Hamp might also play a role in the progression of chronic liver disease to HCC development. Also alcohol intake, another risk factor for HCC development, lowers hepatic Hamp expression in a murine model of alcoholic steatohepatitis [172]. Regarding HCC, low Hamp levels have been reported in late stage murine and rat tumors [173][174]. As this downregulation might display a late, secondary, rather than an initial effect of carcinogenesis, we aimed at deciphering whether Hamp expression is already decreased in early hepatocarcinogenesis. We observed that mice treated with the carcinogen diethylnitrosamine (DEN), to induce hepatocarcinogenesis, showed decreased hepatic Hamp expression already in an early stage of tumor development (Fig. 5.35).

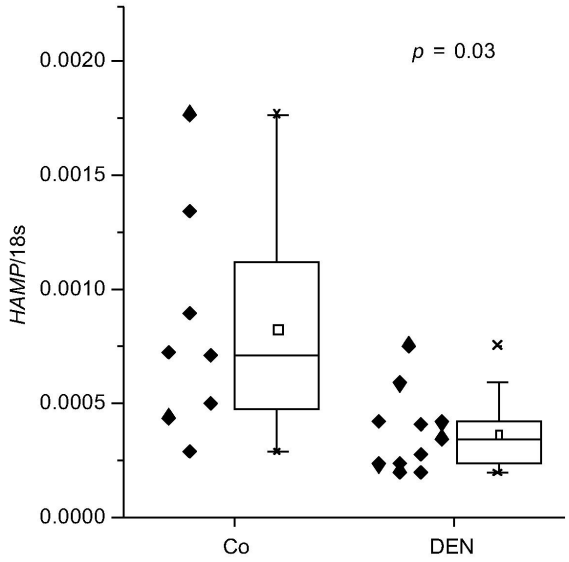

Figure 5.35: Hepatic Hamp expression in non-tumorous murine liver tissue, 6 months after intraperitoneal injection of 5 mg/kg BW diethylnitrosamine (DEN) at the age of 2 weeks, compared to untreated control (co). Data are presented as individual values and box plots with median (-) and mean (Small box) of untreated control (co, n = 8) and DEN-treated (DEN, n = 11) animals. p-values from MannWhitney U test

Hamp expression was also reduced in tumor tissues, compared to matched adjacent normal liver tissues, in a later stage of murine tumorigenesis (Fig. 5.36). Hamp expression was normalised to 18s expression in figures 5.35 and 5.36. To test the relevance of the observed decreased hepcidin in rodent HCC for human disease, we analyzed a large human Gene Omnibus (GEO) dataset (GSE14520 [125]), mostly consisting of hepatitis B virus (HBV)-related HCC samples.

Several datasets were downloaded from Gene Omnibus (GEO) to analysze the Hamp expression. The statistical significance was determined by MannWhitney U test (Figure 5.35), paired sample t test (Figure 5.36), subsequent to confirmation of normal distribution, or Kolmogorov-Smirnov test (Figures 5.37, 5.38, 5.39). Hamp expression was strongly

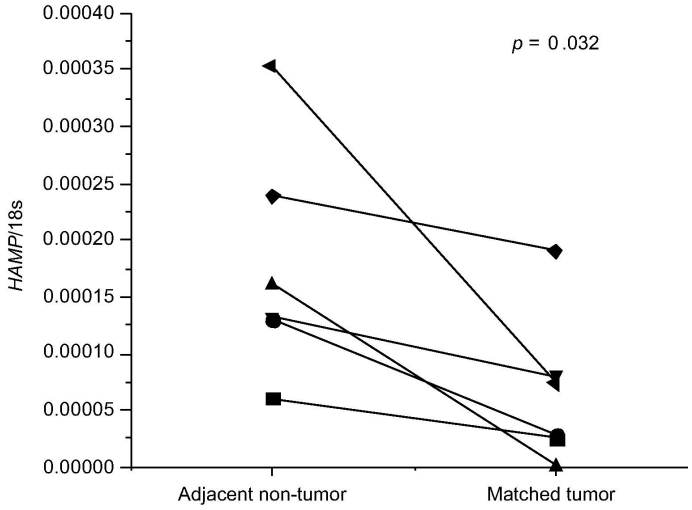

Figure 5.36: Hamp expression in adjacent non-tumorous murine liver tissues and matched tumor tissues (n = 6), 8 months after DEN injection as described in 5.35

decreased in the majority of tumors compared to normal liver samples (Fig. 5.37). This is in line with results from a small HCC cohort with mixed etiology [175].

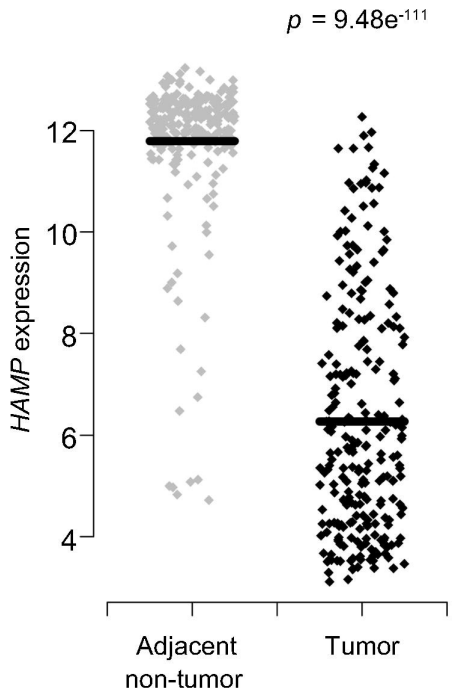

Figure 5.37: Gene expression of Hamp in human dataset GSE14520 (adjacent non-tumor samples n = 247, tumor samples n = 239)

Interestingly, serum hepcidin levels were shown to be decreased in patients with chronic hepatitis C [176]. To test for hepatic hepcidin expression in cirrhosis, we analysed two additional datasets containing cirrhotic liver samples. Cirrhotic tissues showed lower Hamp expression compared to healthy liver samples in an HBV-related cohort (Fig. 5.38) as well as in HCV-infected patients (Fig. 5.39)

Furthermore, Hamp mRNA levels were even lower in tumor tissue (Fig. 5.38 and 5.39). Interestingly, hepatitis C virus (HCV) has been described to suppress hepcidin expression via generation of reactive oxygen species [177].

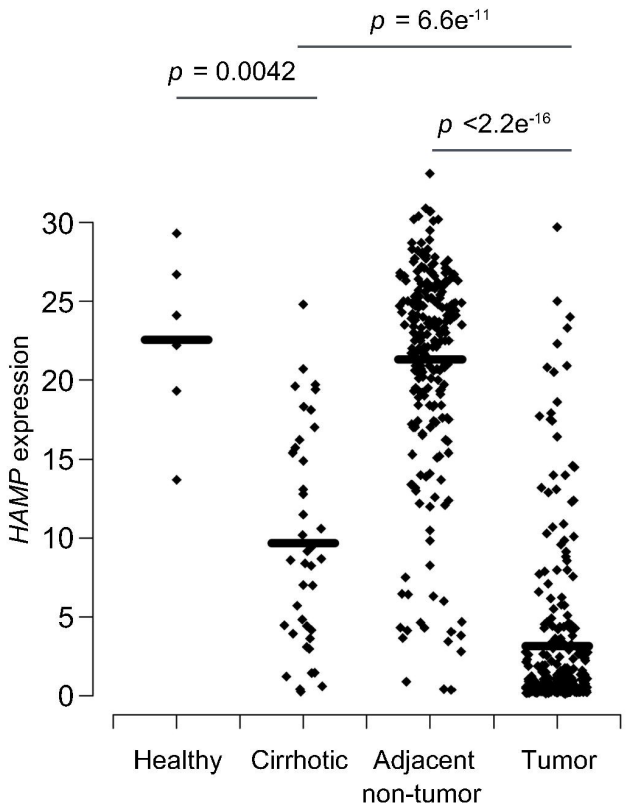

Figure 5.38: Gene expression of Hamp in human dataset GSE25097 (healthy samples n = 6, cirrhotic samples n = 40, adjacent non-tumor samples 243, tumor samples n = 268)

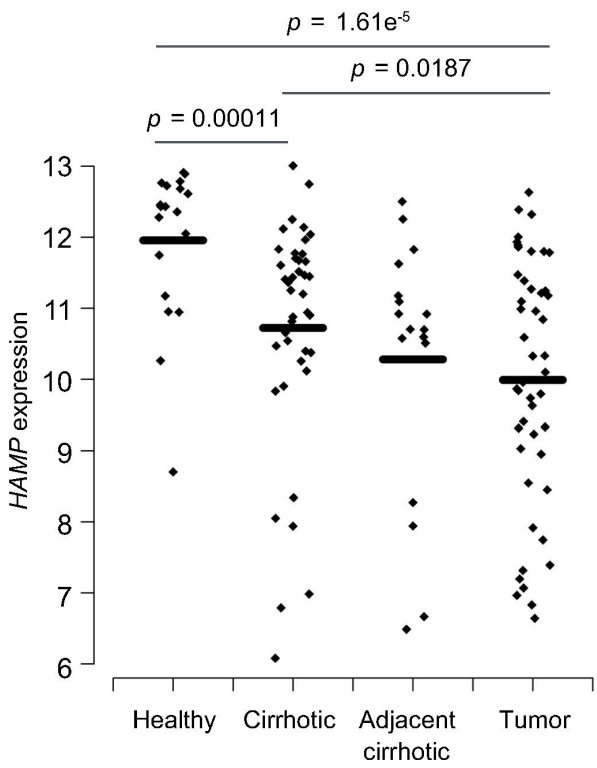

Figure 5.39: Gene expression of Hamp in human dataset GSE14323 (healthy samples n = 19, cirrhotic samples n = 41, adjacent cirrhotic non-tumor samples n = 17, tumor samples n = 47)

With HBV also inducing oxidant stress, this might also be true for HBV. Furthermore, Hamp expression can be transcriptionally activated by the tumor suppressor p53 [178]. As

p53 is frequently suppressed in HCC [179], downregulation of hepcidin might be linked to p53 suppression. In conclusion, these findings in the DEN mouse model and three human HCC cohorts strongly support a role of hepcidin deficiency not only as a model for iron-related liver disease, but also for other liver diseases leading to HCC. Therefore, hepcidin knockout mice presented by Lunova and colleagues [171] might be an interesting model to study progression of various liver diseases towards HCC.

5.4 Lipid Metabolism Signatures in NASH-Associated HCC

An article published recently in Cancer Research elegantly performed lipidomic and gene expression analyses in a murine model of nonalcoholic steatohepatitis (NASH)–associated hepatocellular carcinoma (HCC) and compared the findings with serum samples from patients with fibrosis and HCC [180].

The study reports that the expression of the C18 fatty acid producing elongase (*ELOVL6*) is elevated in a mouse NASH model. The animals also exhibited elevated oleic acid (18:1n9) and vaccenic acid (18:1n7) abundance in livers and serum. Thereby, the study supports findings about increased hepatic *ELOVL6* expression in other models of NASH, such as a fructose feeding model [181] and low-density lipoprotein receptor (LDLR) knockout animals fed on a Western-type diet [182]. In line with these findings, a causal role for *ELOVL6* in the development of NASH was published recently in a comprehensive work using overexpression and knockdown strategies [183].

HCC represents a rare but important complication of NASH [184]. The study by Muir and colleagues reports an increased expression of *ELOVL6* not only in murine NASH but also in murine NASH–associated HCC. Because lipidomic analyses of sera of 15 patients with HCC showed a higher prevalence of the C18 vaccenic acid (18:1n7) than serum of patients with cirrhosis, the authors suggested elevated *ELOVL6* expression in human HCC. Although they observed lower levels of the more abundant linoleic acid (18:2n6) and they do not show any data on *ELOVL6* expression in patients with HCC, they propose *ELOVL6* as a pharmacologic target for patients predisposed to HCC.

We investigated differential *ELOVL6* gene expression between HCC (n = 247) and nontumor (n = 239) samples of a Gene Expression Omnibus dataset (GSE14520; see Fig. 5.40). Interestingly, in contrast to Muir and colleagues, our results from this large dataset revealed significantly decreased levels of *ELOVL6* gene expression in the majority of human liver tumors compared with nontumorous tissue. We also observed a decreased expression of *ELOVL6* in the widely accepted murine diethylnitrosamine (DEN) HCC model (see Fig. 5.41; ref. [184]).

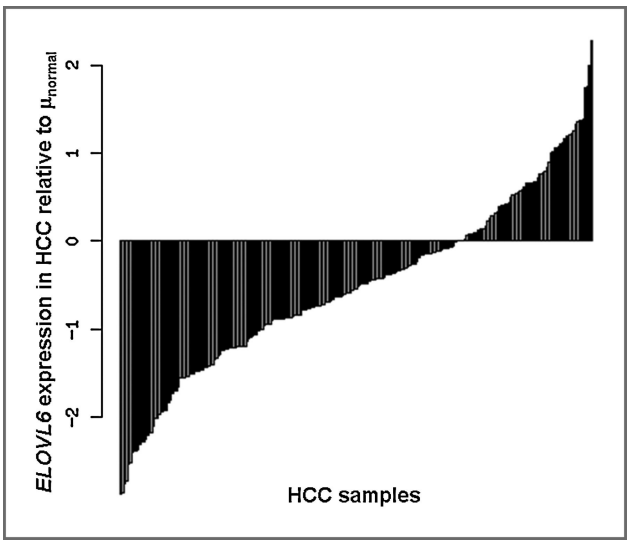

Figure 5.40: mRNA levels of *ELOVL6* in 247 human HCC samples relative to the mean of 239 nontumor liver tissue (μ_normal). Samples of dataset GSE14520 [$\log_2$ (expression) values from GEO after Robust Multi-array Average normalization] were mapped to hgu133a.db using bioconductor. Significance values: P = 3.8E$^-$11, Kolmogorov–Smirnov test; P = 6.7E$^-$11, t test; 5.1E$^-$11, Mann–Whitney U test

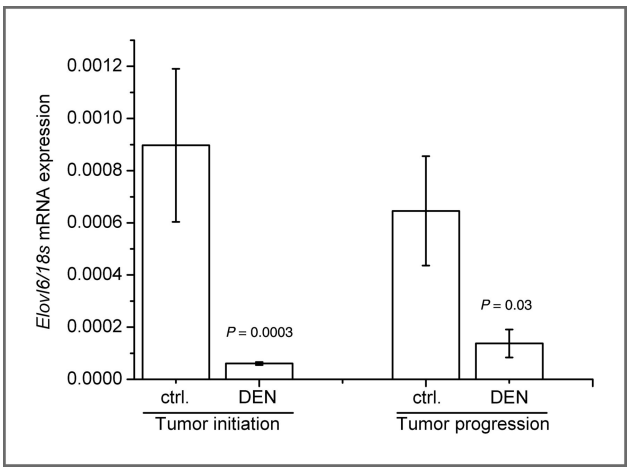

Figure 5.41: Wild-type mice were treated with the carcinogen DEN at the age of 2 weeks. Livers were analyzed after 24 weeks to assess the tumor initiation state. Analyses in the tumor progression stadium were done after 36 weeks. *Elovl6* mRNA expression as determined by real-time reverse transcriptase PCR with $n = 8$–18 per group. Data were normalized to *18S*. Statistical differences compared with untreated animals of the same age (ctrl.) were calculated by Mann–Whitney U test

Taken together, different recent reports from the literature suggest a pathophysiologic role for *ELOVL6* in steatohepatitis. Still, a role for *ELOVL6* in HCC is as yet elusive and our data show *ELOVL6* expression to be reduced in a common murine non-NASH–associated HCC model as well as in a large proportion of patients with HCC. In our opinion, the data available on *ELOVL6* in HCC do not justify proposing *ELOVL6* as a therapeutic target in either prevention or treatment of HCC.

5.5 Fatty Acid Elongation in Non-Alcoholic Steatohepatitis and Hepatocellular Carcinoma

5.5.1 Abstract

Non-alcoholic steatohepatitis (NASH) represents a risk factor for the development of hepatocellular carcinoma (HCC) and is characterized by quantitative and qualitative changes in hepatic lipids. Since elongation of fatty acids from C16 to C18 has recently been reported to promote both hepatic lipid accumulation and inflammation we aimed to investigate whether a frequently used mouse NASH model reflects this clinically relevant feature and whether C16 to C18 elongation can be observed in HCC development. Feeding mice a methionine and choline deficient diet to model NASH not only increased total hepatic fatty acids and cholesterol, but also distinctly elevated the C18/C16 ratio, which was not changed in a model of simple steatosis (*ob/ob* mice). Depletion of Kupffer cells abrogated both quantitative and qualitative methionine-and-choline deficient (MCD)-induced alterations in hepatic lipids. Interestingly, mimicking inflammatory events in early hepatocarcinogenesis by diethylnitrosamine-induced carcinogenesis (48 h) increased hepatic lipids and the C18/C16 ratio. Analyses of human liver samples from patients with NASH or NASH-related HCC showed an elevated expression of the elongase *ELOVL6*, which is responsible for the elongation of C16 fatty acids. Taken together, our findings suggest a detrimental role of an altered fatty acid pattern in the progression of NASH-related liver disease.

5.5.2 Introduction

Non-alcoholic fatty liver disease (NAFLD) is regarded as the most common liver disorder [185]. Although NAFLD is often an asymptomatic disease and therefore difficult to detect, the prevalence appears to be around 20%–35% of the adult population in Western countries [186][187][188]. NAFLD/NASH (non-alcoholic steatohepatitis) strongly correlates with characteristics of the metabolic syndrome, such as obesity and diabetes mellitus, and NAFLD/NASH [189][190][191]. Liver pathogenesis of NAFLD is widely believed to start with simple steatosis, which is characterized by excessive lipid accumulation [186][192]. The progression from simple steatosis to NASH is mediated by the release of inflammatory cytokines [193] and can result in hepatic cirrhosis and finally in hepatocellular carcinoma (HCC) [194]. Due to this inflammatory environment 4% to 27% of individuals with NASH and cirrhosis [195] develop HCC. There is increasing evidence that in steatosis besides the total amount of accumulated lipids the composition of lipids has an impact on pathophysiology [196][197]. In fact, human NAFLD is characterized by numerous changes in hepatic lipid composition and free fatty acid ratios [198][199]. Viral hepatitis has also been described to lead to strongly altered hepatic lipid content and composition [200][201][202][203]. In HCC a decreased stearic acid (C18:0) to oleic acid (C18:1) ratio compared to normal tissue has been reported, suggesting the importance of desaturation in HCC development [204]. Regarding changes in hepatic fatty acid pattern it is important to note that the ELOVL fatty acid elongase 6 (ELOVL6), which catalyzes the elongation of C16 to C18 fatty acids [205], has been shown to promote NASH [183][180]. A role of*ELOVL6*in murine NASH-related HCC has recently been suggested, but still remains unclear in human NASH-associated HCC [180]. The aims of our study were to investigate the occurrence of fatty acid elongation in lipid metabolism in different NAFLD mouse models and to elucidate its relevance in human NASH and NASH-associated HCC

5.5.3 Experimental Section

Animals

All animal procedures were performed in accordance with the local animal welfare committee (#13/2009, 09/06/2009; #34/2010, 15/11/2010; Landesamt für Soziales, Gesundheit und Verbraucherschutz Saarland). Mice were kept under stable conditions regarding temperature, humidity, food delivery, and 12 h day/night rhythm. At the age of 3 weeks mice (DBA2/Bl6/J background) were fed either a methionine-choline deficient (MCD) or a methionine-choline supplemented control (crtl) diet for 3 weeks. Intraperitoneal clodronate or empty liposome injections [206] were started two days prior to MCD or control diet and repeated every five days to ensure Kupffer cell depletion. Leptin deficient mice *ob/ob* (Bl6:Cg-Lep^{ob}/J) and lean control mice (*ob/+*) were obtained from Charles River and sacrificed at an age of 10 weeks. DEN treatment of mice (DBA2/Bl6/J background) on regular chow was performed by a single intraperitoneal injection of 100 mg/kg body weight at the age of 2.5 weeks. Mice were sacrificed 48 h after DEN injection. Animals of all experimental groups were sacrificed in a non-fasted state.

Human Liver Tissue

Paraffin-embedded liver samples from randomly selected pseudonymized HCC patients who underwent liver resection at the Saarland University Medical Center between 2005 and 2010 were obtained as described previously [6]. The study protocol was approved by the local Ethics Committee (#47/07). Samples had a mixed etiology including NASH, alcoholic liver disease, viral hepatitis, hemochromatosis, porphyria, and cryptogenic [6].

Fatty Acid Measurement by Gas Chromatography-Mass Spectrometry (GC-MS)

Murine liver samples were lyophilized and analyzed according to Bode et al. [207]. In short, lyophilized samples were dissolved in a mixture of 500 μL methanol/toluene/-sulfuric acid (50:50:2, *v/v/v*) and incubated at 55 °C overnight. Subsequently, 400 μL of a 0.5 M NH_4CO_3, 2 M KCl solution were added and samples were centrifuged. The organic phase was transferred into a new glass vial, derivatized with 25 μL N-methyl-N-(trimethylsilyl)trifluoroacetamide at 37 °C for 1 h. Fatty acid separation was performed on an Agilent 6890N gas chromatograph coupled to an Agilent 5973N mass selective detector and equipped with a non-polar J&WDB-5HT capillary column (Agilent Technologies, Böblingen, Germany). The column temperature was kept at 130 °C for 2.5 min, increased to 240 °C at a rate of 5 °C/min, and then ramped to 300 °C at 30 °C/min, and held at 300 °C for 5 min. Helium was used as the carrier gas at a flow rate of 1 mL/min. The mass selective detector was operated in scan mode, average spectra were acquired in the *m/z* range of 40–700 *m/z* and were recorded at a scan speed of 2.24 scans/s. Scan control, data acquisition, and processing were performed by MSD ChemStation (Agilent Technologies, Böblingen, Germany) and AMDIS software based on the fragmentation patterns and retention time, in comparison with the reference standards Supelco 37 Component FAME Mix (Sigma-Aldrich, Taufkirchen, Germany), and NIST 08 library. Methyl-nonadecanoate (74208, Sigma-Aldrich, Taufkirchen, Germany) was used as an internal standard. The method detects both free and bound fatty acids.

Histochemistry and Immunohistochemistry

Hematoxylin-eosin staining of paraffin-embedded tissues was performed as previously reported [113][112]. Immunohistochemical F4/80 detection was achieved using the Vectas-

tain Peroxidase Elite ABC kit/DAB with anti-F4/80 antibody (AbD Serotec, Puchheim, Germany) 1:1000 overnight at 4 °C. Epitopes were demasked with citrate buffer pH 6.0 for 10 min in a waterbath at 95 °C.

Analysis of the Public Gene Omnibus (GEO) Datasets

Datasets GSE48452 and GSE37031 [208][209] normalized using log2-RMA and log2-GCRMA respectively, were downloaded from Gene omnibus (GEO) [210]. Dataset GSE48452 with samples from different stages of NAFLD contained 18 NASH and 14 steatosis samples while dataset GSE37031 included 8 NASH and 7 control samples. The statistical significance was determined by Kolmogorov-Smirnov test.

Statistical Analysis

Results are expressed as means ± SEM. The statistical significance was determined by independent two-sample t-test. Expression data of human tissues were analyzed using Mann-Whitney U tests. The results were considered as statistically significant when p value was less than 0.05.

5.5.4 Results and Discussion

Fatty Acid Elongation in Murine Non-Alcoholic Steatohepatitis (NASH) Is Kupffer Cell Dependent

Fatty acid elongation plays a role in murine and human NASH development [183][180]. Feeding mice a methionine-choline deficient diet (MCD) led to increased total fatty acid and cholesterol levels and profound inflammation (Figure 5.42A,B). Having a closer look at the composition of the fatty acids, we observed that the chain length of fatty acids was different in MCD-fed livers compared to tissues from a control diet: NASH livers exhibited an increased ratio of C18 to C16 fatty acids (Figure 5.42B). Increased *ELOVL6* mRNA expression has been observed in NASH animal models, such as low-density lipoprotein receptor knockout animals fed a western type diet [180] or a fructose diet [211]. In order to study whether this increased fatty acid elongation from C16 to C18 could also be observed in simple steatosis we investigated the well-established leptin-deficiency (*ob/ob*) mouse model [211][212]. As expected, livers of *ob/ob* mice showed excessive hepatic lipid accumulation compared to lean controls (Figure 5.42C,D). However, no distinct inflammation was observed and the ratio of C18 to C16 fatty acids was not changed compared to wild-type animals (Figure 5.42D). Adipose tissue of *ob/ob* mice is known to show inflammation [213]. Concordantly, adipose tissue macrophages of ob/ob mice exhibit increased *ELOVL6* mRNA expression [214]. Another study reported increased hepatic *ELOVL6* expression in older ob/ob mice, whereby some of the animals were in a fasted condition [215]. The study does not contain any information on fatty acid composition.

Leroux et al. reported that lipid storage by Kupffer cells correlates with a pro-inflammatory phenotype in NASH [216]. In order to clarify whether the altered fatty acid elongation in murine NASH is due to co-existing inflammation, we depleted Kupffer cells by clodronate liposomes. This intervention is known to attenuate inflammatory and metabolic events in the MCD model [217]. In line with these findings we observed a strong decrease of hepatic lipid accumulation after Kupffer cell depletion (Figure 5.43A,B). Besides hepatocytes also macrophages are known to express *ELOVL6* [214], which was shown to be relevant for lipid storage [218]. After Kupffer cell depletion the MCD-induced changes in C18 to C16 fatty acids and cholesterol were completely abrogated (Figure 5.43A,B). Kupffer cell depletion was confirmed by immunohistochemical F4/80 staining (Figure 5.43A).

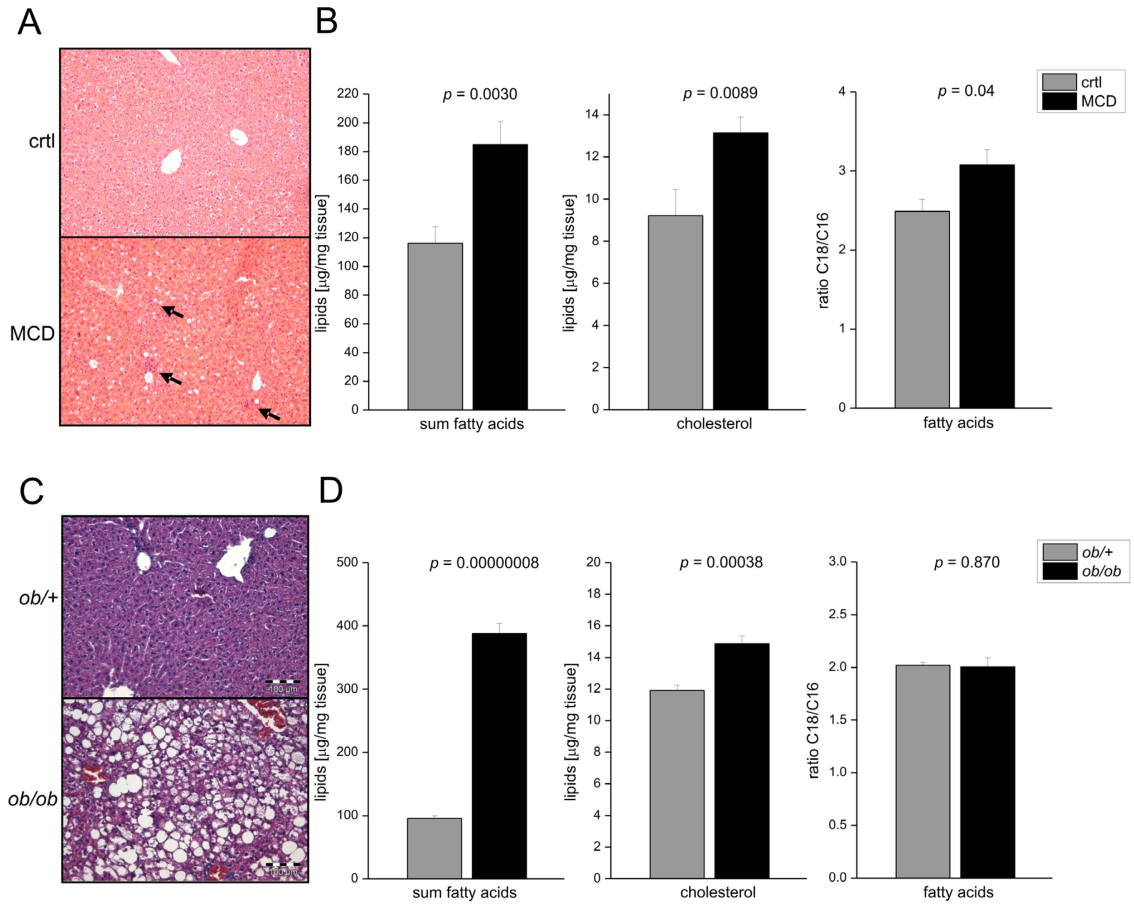

Figure 5.42: Non-alcoholic steatohepatitis (NASH), but not non-alcoholic fatty liver disease (NAFLD), is accompanied by elevation of C18 over C16. (A) Representative liver sections stained with hematoxylin-eosin (HE) from animals fed with either a methionine-choline deficient (MCD) or a control (crtl) diet for 3 weeks (original magnification 200×). Arrows denote inflammatory foci; (B) Sum of all hepatic fatty acids, hepatic cholesterol, and ratio of hepatic C18/C16 fatty acids of MCD fed animals compared to ctrl were analyzed by GC-MS (gas chromatography-mass spectrometry) (n = 9–10); and (C,D) Representative HE-stained liver sections (C), hepatic fatty acids as well as hepatic cholesterol, and ratio of hepatic C18/C16 fatty acids (D) of *ob*/+ and *ob*/*ob* mice (n = 8)

Role of Fatty Acid Elongation in NASH-Related Hepatocellular Carcinoma (HCC) and Human NASH

To further investigate the role of fatty acid elongation in hepatocarcinogenesis, we used short-term (48 h) treatment with the carcinogen diethylnitrosamine (DEN) to model early inflammatory events associated with hepatocarcinogenesis [219][117]. In fact, we histologically observed inflammatory foci in the livers exposed to DEN (Figure 5.44A). Little is known about increased lipid accumulation after DEN treatment: Histologically detected hepatic lipid deposition by DEN was reported in fish (*Oryzias lapites*) [220][221]. Changes in the lipid composition have been reported for cancerous tissues compared to normal tissue in DEN-induced hepatocarcinogenesis [222][223][224], but not in the precancerous short-term protocol. We observed that inflammatory events were paralleled by distinct metabolic alterations similar to the murine NASH model: fatty acids, cholesterol levels, and the C18/C16 ratio were elevated upon 48 h DEN treatment (Figure 5.44B). This short-term model might therefore resemble NASH-related hepatocarcinogenesis. In later stages of DEN-induced carcinogenesis and during tumor progression we observed that fatty acid elongation was rather repressed [118]. This can be explained by the fact that the long-term DEN model predominantly acts via genotoxic effects of the carcinogen and therefore no

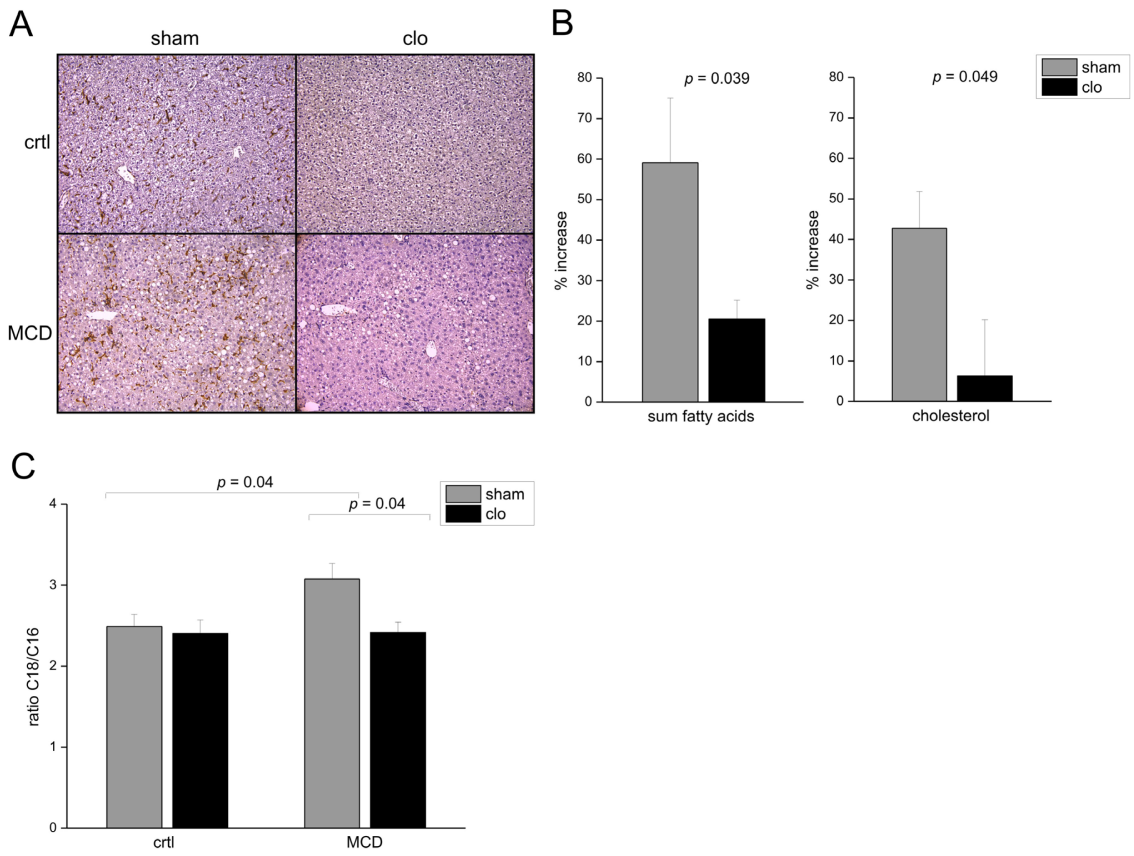

Figure 5.43: Kupffer cell depletion abrogated elevation of C18 over C16. (A) Representative liver sections immunohistologically stained against F4/80 as Kupffer cell marker from animals fed with the respective diet for 3 weeks with simultaneous administration of clodronate (clo) or empty (sham) liposomes (original magnification 200×); and (B,C) Increase of the sum of all hepatic fatty acids, hepatic cholesterol (B), and ratio of hepatic C18/C16 fatty acids (C) of MCD fed animals treated with clodronate (clo) or empty (sham) liposomes compared to ctrl analyzed by GC-MS (n = 9–10)

NASH-related HCCs are induced.

In order to study the relevance of increased fatty acid elongation in human NASH and human NASH-related HCC, we analyzed the hepatic mRNA expression of the enzyme ELOVL6, which is responsible for the elongation of C16 fatty acids. We observed increased levels of *ELOVL6* in NASH *versus* steatosis samples (GSE48452 [208]) (Figure 5.44C) as well as in NASH compared to healthy control tissues (GSE37031 [209]) (Figure 5.44D). Interestingly, expression of *ELOVL6* was also altered in NASH-associated HCCs compared to HCC tissues of mixed etiology (Figure 5.44E): human NASH-related HCCs express increased levels of *ELOVL6*, indicating fatty acid elongation to play a critical role in this particular HCC subtype.

5.5.5 Conclusions

In the present study, we identified that NASH-induced fatty acid elongation is an inflammation-associated pathophysiological step in liver disease. Furthermore, the fatty acid elongase *ELOVL6* is elevated in human NASH and NASH-related HCC.

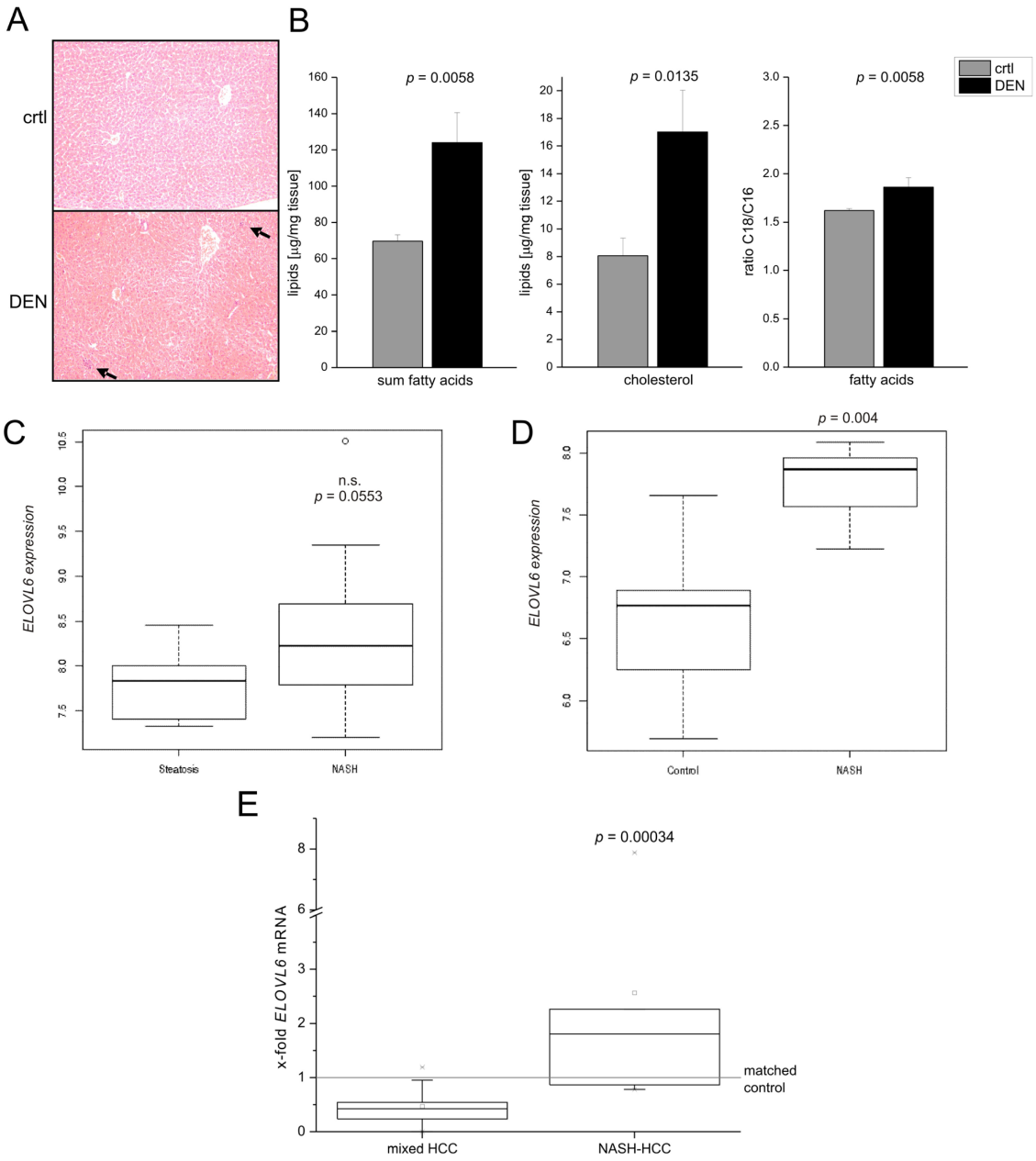

Figure 5.44: (A) Representative liver sections stained with HE from animals treated with DEN (DEN) compared to untreated control (crtl) (original magnification 200×). Arrows denote inflammatory foci; (B) Sum of all fatty acids, hepatic cholesterol, and ratio of C18/C16 fatty acids of DEN treated animals compared to untreated control (ctrl) are displayed (n = 6–15); (C,D) Expression of *ELOVL6* in human NASH (n = 18) compared to steatosis (n = 14) (GSE48452) (n.s. = not statistically significant) (C) as well as healthy control samples (n = 8 for NASH; n = 7 for control; GSE37031) (D); and (E) *ELOVL6* mRNA expression in human NASH-related HCC samples (NASH-HCC) (n = 6) compared to HCC with mixed etiology (mixed HCC) (n = 26) [6]. Expression of tumor tissues was normalized to matched normal liver tissue (matched control)

Chapter 6

Cross-talk Between Intragenic Epigenetic Modifications and Exon Usage Across Developmental Stages of Human Cells

The results of this chapter were jointly obtained by Dr. Siba Shanak and the author. My contribution was late normalization and preparing data in tables, calculation of Correlation and production of most figures. The main contribution of Dr. Shanak was data retrieval, establishing early calculations and data preprocessing, calculation of read count on the exon level, annotating genes according to exon count and early normalization stages. Results were jointly analyzed with Prof. Volkhard Helms.

6.1 Abstract

Differential exon usage has been reported to affect the large majority of genes in mammalian genomes. It has been shown that different splice forms sometimes have distinctly different protein function. Epigenetics is well associated with alternative splicing in the gene body, but the connection between differential exon usage and the distinct developmental stages has not been addressed so far. Here, we present an analysis of the Human Epigenome Atlas (version 8) to connect the differential usage of exons in various developmental stages of human cells/tissues to differential epigenetic modifications at the exon level. We found that the differential incidence of protein isoforms across developmental stages is often associated with changes in histone marks as well as changes in DNA methylation in the gene body or the promoter region. Many of the genes that are differentially regulated at the exon level were found to be associated with development and metabolism.

6.2 Introduction

Differential exon usage is reported to occur in 90-95% of all human multi-exon genes [225][226]. Different splice variants of a gene may lead to different protein products that exert different functions. As a result, differential exon usage leads to a strong expansion of the eukaryotic proteome [227]. An example for this is the well-known Nanog gene; where alternative splicing results in two variants of the Nanog protein with different capabilities for self-renewal and pluripotency in embryonic stem (ES) cells [228]. An alternative scenario takes place when genes coding for different proteins occupy the same position on a chromosome. In such cases, differential exon usage even controls the expression of different proteins. A well-characterized example for this case are the overlapping imprinted genes PEG3/ZIM2 that are exclusively expressed from the paternal allele [229][230]. Gamazon et al showed that 90% of human genes are so far known to undergo alternative splicing [231]. However, the notion of alternative splicing across tissues should not be considered as an exclusive either/or mechanism. Thanks to recent advances in RNA-Seq technology [232], it is now possible to study the expression of genes at the level of single exons. The granulity of exon usage can thus be increased from the basic classification of a one-or-none

expression per gene (alternative splicing) to fine-tuned quantitative read counts that can be accounted for per individual exon.

A recent study reported that differential exon usage in primates shows more profound differences across species than on the intra-species level. It was targeted at adult tissues (brain, cerebellum, heart, kidney, and liver), and did not analyze the effect of differential usage of exons in terms of organismal development [233]. The state of the art next-generation sequencing (NGS) studies emphasize to examine the link of alternative splicing, through differential exon usage, to development and to epigenetic modifications. Furthermore, there is also a strong connection established in literature between development and epigenetic modifications (see figure 6.1).

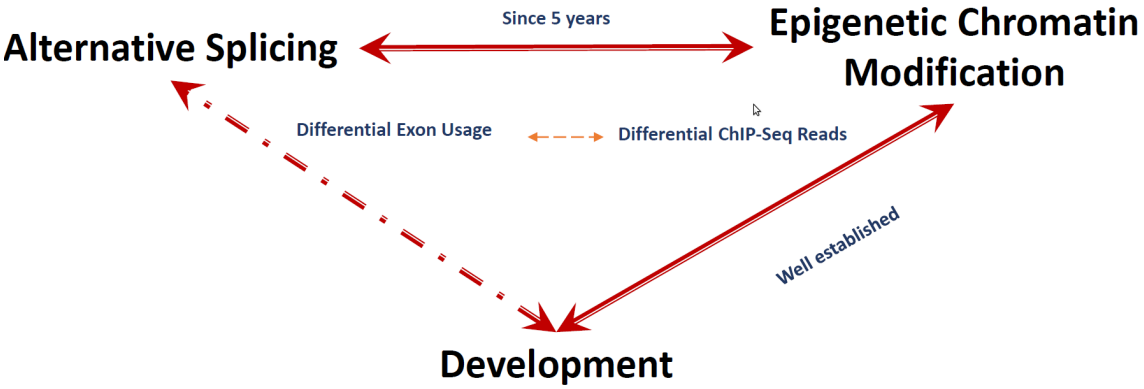

Figure 6.1: The triad of 'Alternative Splicing- Epigenetic Chromatin Modifications- Development'. We reviewed the connections established per each two of the three and found that only few studies have addressed this triangle

Over the past five years, relating alternative splicing events with epigenetic modifications has become a very active research field. Zhou et al. studied the relationship between alternative splicing and histone marks [234]. Another study by Schwartz et al. addressed the interplay between chromatin structure and the exon-intron architecture. They showed that histone modifications within the gene body are more pronounced in exon regions than in intron regions, and thus may serve to define the exon-intron boundaries [235]. Alternative splicing plays a fundamental role in development; where it affects organ morphogenesis, stem cell differentiation as well as neuronal development [236]. The relationship between alternative splicing and development has been an active research field since the early 1980s [237][238] [239][240] [241][242]

For more than a decade, it has been understandable that development is under the control of epigenetic modifications. One form of epigenetic modifications, DNA methylation, affects gene expression via inhibition of transcription factor binding or by recruiting histone modifiers that induce DNA supercoiling. Histone modifications, through methyl or acetyl marks, is another epigenetic modifications known to impact development [243]. These modifications play crucial role in the mitotic memory of human cells during development. Recent studies highlighted chromatin-based and DNA-based changes accompanying the erasure, reprogramming and the reacquisition of pluripotency in model mammalian organisms. Paternal genome in the zygote is modelled differently to the maternal genome, with hyperacetylation and hypomethylation that makes the paternal genome prone to excessive remodeling. This asymmetry is perceived up to the 4-cell stage. In this respect, DNA demethylation occurs until the blastocyst stage, while chromatin protects some maternal genes against this event until the implantation stage [244]

The notion of splicing was thought to work by transcribing either the full (constitutive)

genes or the alternatively spliced forms. Until recently, the relation between exon usage and transcript abundance has been scarcely analyzed. To our knowledge, there has been so far no attempt to study the relationship between differential usage of exons and various types of epigenetic modifications at the exon level and to connect this with different developmental stages of human. This is precisely the aim of this study. Based on data for human development across different stages from the Human Epigenome Atlas [245] [246] [247] [248], we show a correlation between differential exon usage and several epigenetic modifications at the exon/intron/promoter level, namely DNA methylation and several histone marks. The correlation is significant for both the constitutive genes and for gene clusters. Furthermore, we could associate the occurrence of differential exon usage with functional annotations that, indeed, often relate to regulation of signaling and developmental processes.

6.3 Methods

6.3.1 Datasets Used

Data for this study was retrieved from the Human Epigenome Atlas (up to release 8) that is part of the Roadmap Epigenomics project [245] [246] [247] [248]. Table 6.1 introduces the assays and the epigenetic modifications analyzed in our study. The aim of this study was to find the link between the differential usage of exons at the expression level and that for specific epigenetic marks.

Table 6.1: assays used in this study to evaluate the levels of expression, chromatin organization and DNA methylation in the human genome during different developmental stages.

Expression	Chromatin organization	DNA methylation
mRNA-Seq siRNA-Seq	ChIP-Seq Input DNAse hypersensitivity H3K27ac/H3K27me3 H3K36me3 K3K4me1/H3K4me3 H3K9ac/H3K9me3	Bisulfite-Seq/RRBS MeDIPS-Seq

We further aimed to know how this correlates to different stages of human development. Thus, we only studied sample types for which release 8 provided complete data sets according to table 6.1. These stages included stem cells, early developmental stages, induced differentiated cells, fetus, and adult tissues. Figure 6.2 lists the studied tissues.

We downloaded the human UCSC hg19 reference genome, retrieved the exons of each gene, and prepared them in a temporary annotation file. To account for possible ambiguity, each gene should only be mapped to one genomic region. As a result, we dropped a small set of 100 genes spanning more than one genomic region from our analysis. Furthermore, we clustered genes that mapped to the same genomic region into one gene cluster to prevent redundancy in mapping, see figure 6.3. Following the strategy of Anders et al, we sorted the group of exons belonging to the genes of one gene cluster, and extracted the unique exons [249]. If any two exons from different genes mapped to the same genomic region, we rearranged them and assigned them to a new non-overlapping classification of exons that mapped to the same region, see figure 6.3 for illustration. After that, we mapped introns and promoters accordingly. We defined the promoter region as the region between -2000 bp upstream of the transcriptional start site and 0 bp of the gene/gene cluster region.

For data processing of the ChIP-Seq assays, we called the peaks associated with the reads in the retrieved bed files. To this aim, we used ChIP-Seq Input assay to check for the background effect. Peak calling of the different histone marks was performed using MACS

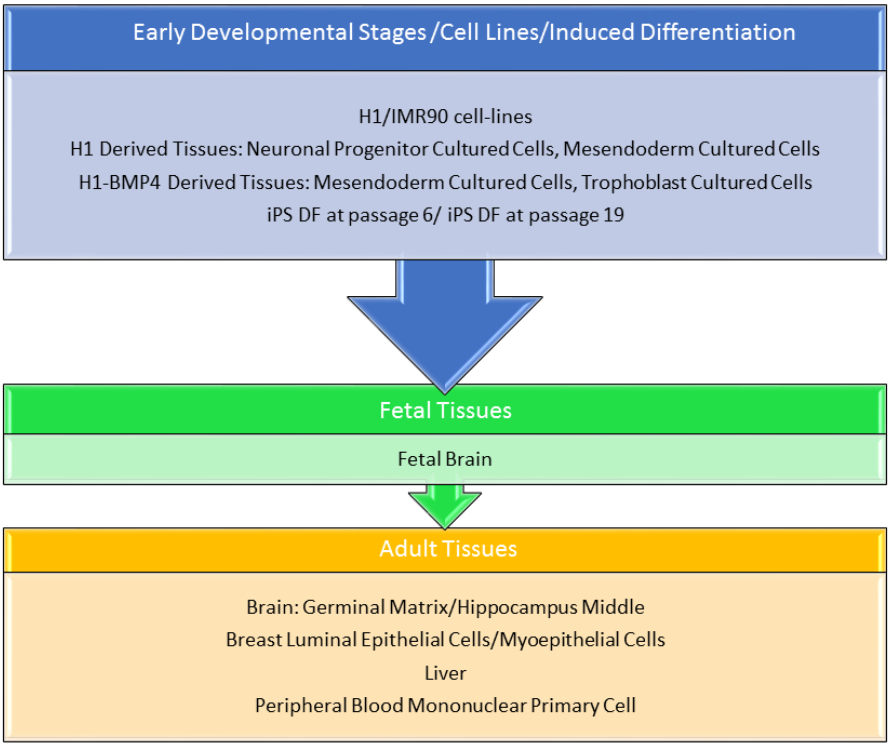

Figure 6.2: The 14 different tissues that were investigated in this study belong to the three different main developmental stages

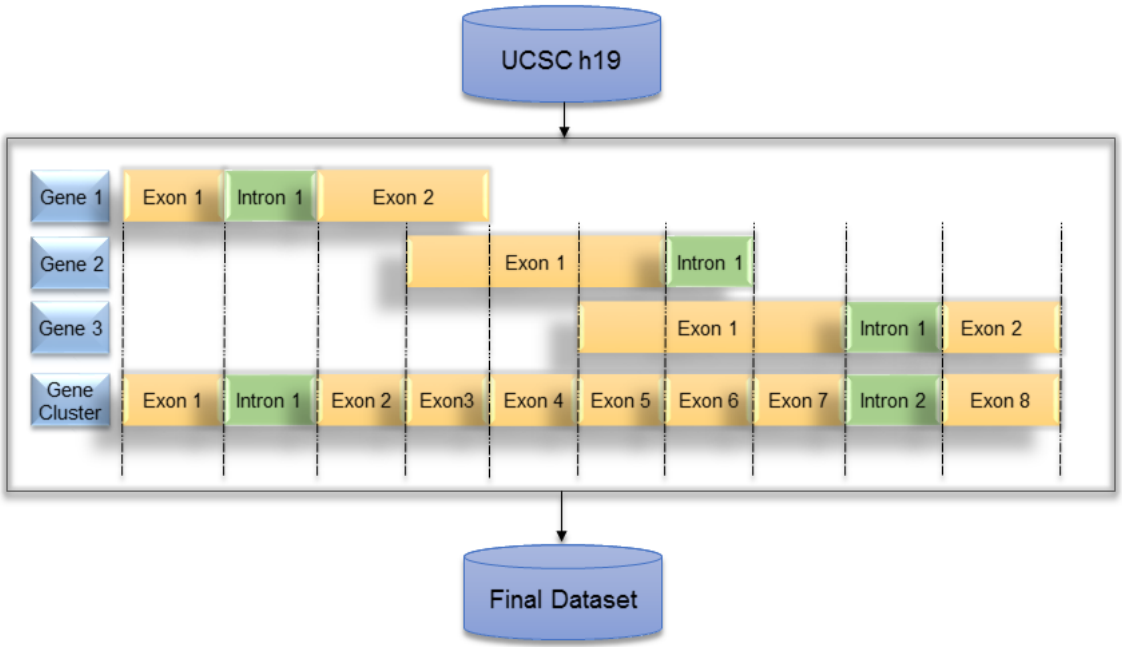

Figure 6.3: A schematic representation of the exon architecture of three exemplary genes that show partial overlap. The virtual gene cluster shown in the bottom row consists of shorter exons 2-7 in order to resolve the overlapping issue. Also shown is how Exon 6 is assigned to resolve a conflict of the overlapping Gene 2:Intron 1 and Gene 3:Exon 1 case. See the main text for further explanation

zhang2008model.

6.3.2 Data Normalization

In order to account for putative technical noise in the data and to check for differential read usage, we performed pair-wise comparisons of the reads in different tissues. To this aim, we modeled read counts using regression analysis to detect noise in tissues in a pairwise manner.

The peak calls in ChIP-Seq data were normalized using MAnorm, where linear regression analysis is performed [250]. Normalization of MeDIP-Seq data was done with the MeDIPS Bioconductor Package that uses negative binomial regression [251]. The methylation datasets from the bisulfite-seq and RRBS datasets were normalized using the Biocondutor package methylKit [252]. For each basepair position, logistic regression was applied to check for differential methylation per base. These results were processed to obtain the mean methylation ratio per exon. To normalize the mRNA and smRNA data, we first obtained the transcript and exon abundance. We generated SAM files from the supplier's BED files via BedTools and SamTools [253][254] and sorted the SAM files lexicographically. Read counts of genes and exons were prepared from the SAM files using the HTSeq package [255] and used as an input for the Bioconductor DEXSeq package [249] to reduce noise in the data.

Data annotation for the normalized ChIP-Seq and methylation data was performed using BedTools [253]. Expression data were already annotated by the HTSeq package [255]. After that, we mapped the whole set of normalized reads, including the read numbers for expression, the different histone marks, and the methylation status for each exon in a gene/gene cluster per tissue into a final table per read type. The table consisted of one read value per tissue per exon. If for a read type different read numbers mapped to the same exon, we averaged them. After that, we normalized all read numbers for a single gene to a final range of log values between -1 to +1.

6.3.3 Differential Usage of Exons

Differential usage of exons was analyzed using the strategy described in figure 6.4. We aimed at identifying genes for which differential usage of their exons across developmental stages in terms of exon expression is associated with clear differences in epigentic marks. To achieve this, we followed two different strategies to examine correlations between different epigenetic marks (at the exon level) and the expression levels of exons. Both marks needed to map to the same exon, to a directly adjacent intron, or to the promoter region for the genes/gene clusters that we defined.

The first strategy checks for anticorrelations in read counts on the gene level. We calculated these anticorrelations for exons that belong to a single gene/gene cluster in a pair-wise manner between tissues. With this, we wished to identify differential changes of exon usage (for both expression and epigenetic modifications) at the tissue level. To this aim, we explored all genes with ≥ 4 exons in all possible pairwise combinations among the 14 tissues studied here. We set the threshold of the Pearson correlation coefficient (PCC) to a tight bound of $\leq$ -0.7. We followed this strategy for read counts of mRNA, different histone marks, DNAse hypersensitivity, siRNA, or methylation levels. Additionally, we applied the same strategy to examine anticorrelations on the intron level. This test yielded lists of genes with anticorrelated levels of exon expression and one or more of the associated epigenetic marks, or epi-spliced genes (see supplementary material and figure 6.5 for details). Furthermore, if the anticorrelation of expression coincided with both anticorrelations in histone marks and methylation, this was documented as well.

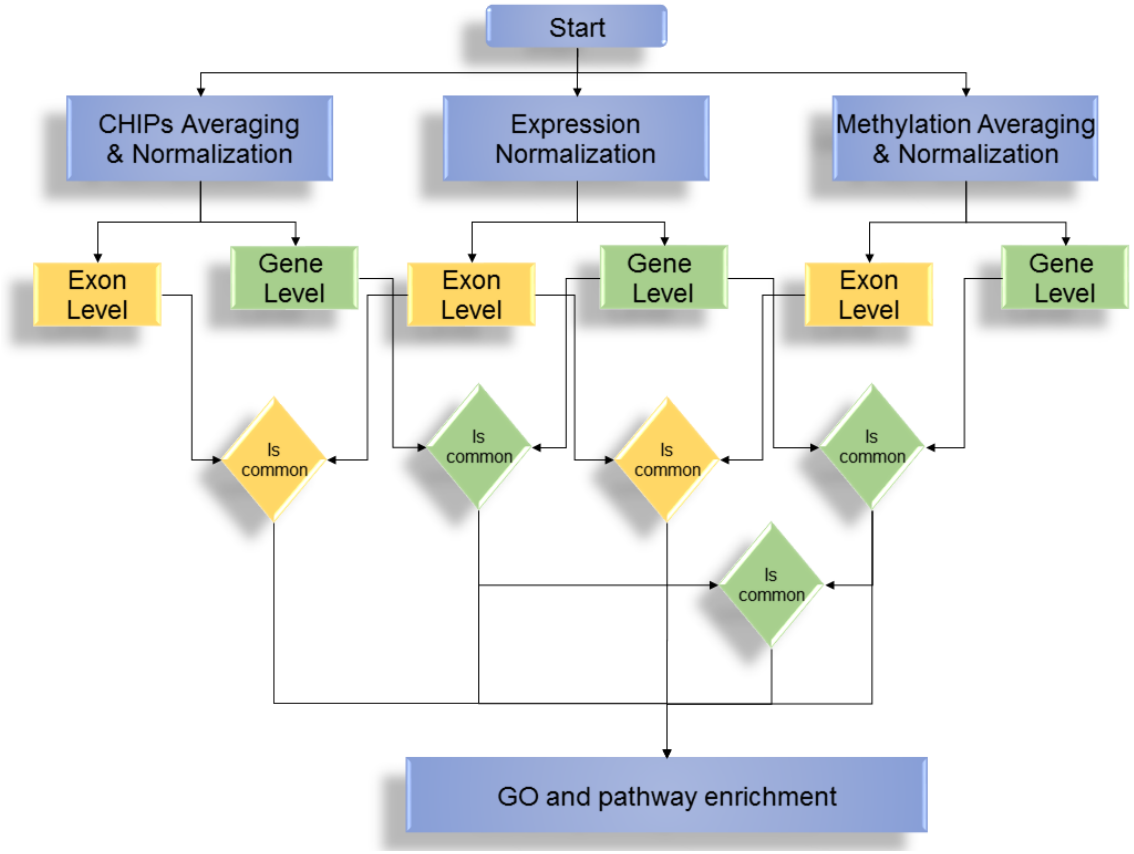

Figure 6.4: A schema for the pipeline of studying the gene- and exon- levels of differential exon usage across developmental stages and correlating this to the differential epigenetic marks. See main text for further explanation

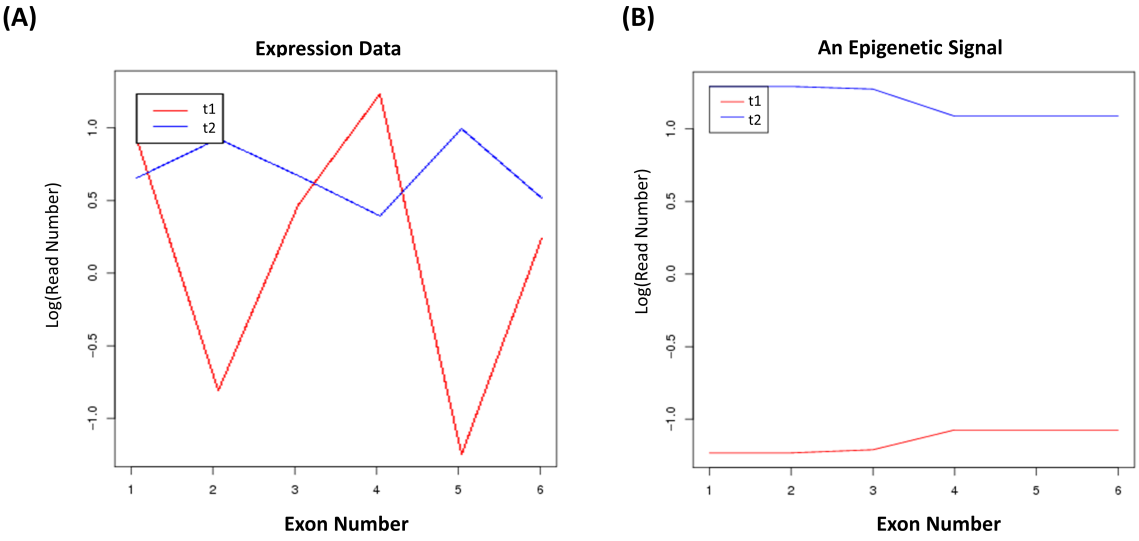

Figure 6.5: An example of anticorrelations

The second strategy identified changes of the read number on the exon level in all possible genes across developmental stages. The results were then correlated with the changes in read counts for the different epigenetic marks described above. With this strategy, we aimed to understand the possible functional association with a single epigenetic mark and expression and at the exon level. We set the Pearson correlation coefficient to a tight bound with an absolute value of at least 0.7.

After that, we checked for the enrichment in GO terms using the GOSim package [123].

We examined functional similarity in two sets of genes. The first set included genes that were identified in the same pairwise tissue comparison. For this analysis, we only considered tissue pairs that have at least 10 genes that are both differentially expressed and show differential epigenetic marks between those two tissues. In the second set, we grouped the genes that showed correlated changes of expression for a single exon and one epigenetic modification across tissues.

For completeness, we also analyzed positive correlations in read counts on the gene level. To identify cases of constitutive gene expression, we calculated these correlations for exons in a single gene/gene cluster in a pair-wise manner between tissues. As before, we explored all genes in all possible combinations of the 14 tissues we studied. Again, we set the threshold of Pearson correlation coefficient (PCC) to a tight bound of ≥ 0.7.

6.4 Results and Discussion

We wanted to check for the differential usage of exons across developmental stages. We also correlated this to the differential changes in read count at the exon level for the different epigenetic marks and through development. This latter analysis was performed for several histone marks as well as DNA methylation levels. To this aim, each type of differential changes (i.e., expression, DNA methylation, histone binding) was estimated alone then the several marks linked to expression. Our analysis considered 7960 constitutive genes and 14668 gene clusters. Figure 6.6 shows results on the gene level for the number of genes/gene clusters with differential exon usage that was negatively correlated with epigenetic marks between each pair of tissues (PCC $<$ -0.7, see methods for definition of differential exon usage and for the definition of gene clusters). Figure 6.6-a shows the dissimilarity of exon expression. The largest differences were found between trophoblast cultured cells and mesendoderm cultured cells, as well as iPS passage 19 (dark blue). Differences between later developmental stages were rather small in comparison. Figure 6.6-b shows the dissimilarity of DNA methylation for all genes. Notably, trophoblast cultured cells were the least similar to all other tissues. A seemingly peculiar similarity was found between fetal brain and all other tissues as well as for the two breast tissues (straight light grey bars). This could be traced back to the fact that very few genes showed differential methylation of their exons for these three tissues.

Figure 6.6-c shows the dissimilarity of H3K36me3 as an example of the respective histone analysis. H3K36me3 was selected for this because it showed the largest number of histone marks in the gene body, as has been reported before [256]. In contrast to Figures Figure 6.6-(a,b), rather balanced differences were found between all tissues.

Figure 6.6−d shows the results from an integrated analysis, where the set of genes showing differential exon usage (measured by expression, see Fig. Figure 6.6-a) was intersected with the set of genes showing either anticorrelation (PCC $\leq$ -0.7) in DNA methylation (Fig. Figure 6.6-b) or in histone marks (Figure 6.6-c and other histone marks that are not shown). We will refer to such genes as "epi-spliced" genes. Dissimilarity was measured as the ratio of genes in the intersection set over the max number of intersections in all studied tissues. Clearly, trophoblast cells were the most different from all other tissues. Two different passages of iPS cells showed very similar combined expression/epigenetic marks, whereas ESCs are more distant to the iPS cells. Fetal brain and adult brain also showed similarity. As expected, we found that mesendoderm cultured cells and trophoblast cultured cells exhibited large differences in their epi-spliced genes. Based on the data from figure 6.6−d, we generated a cluster dendogram by average-linkage hierarchical clustering. Trophoblast cells showed by far the largest dissimilarity to all other tissues (note that figure 6.7 shows the logarithm of the distance). As expected, breast tissues showed high similarity

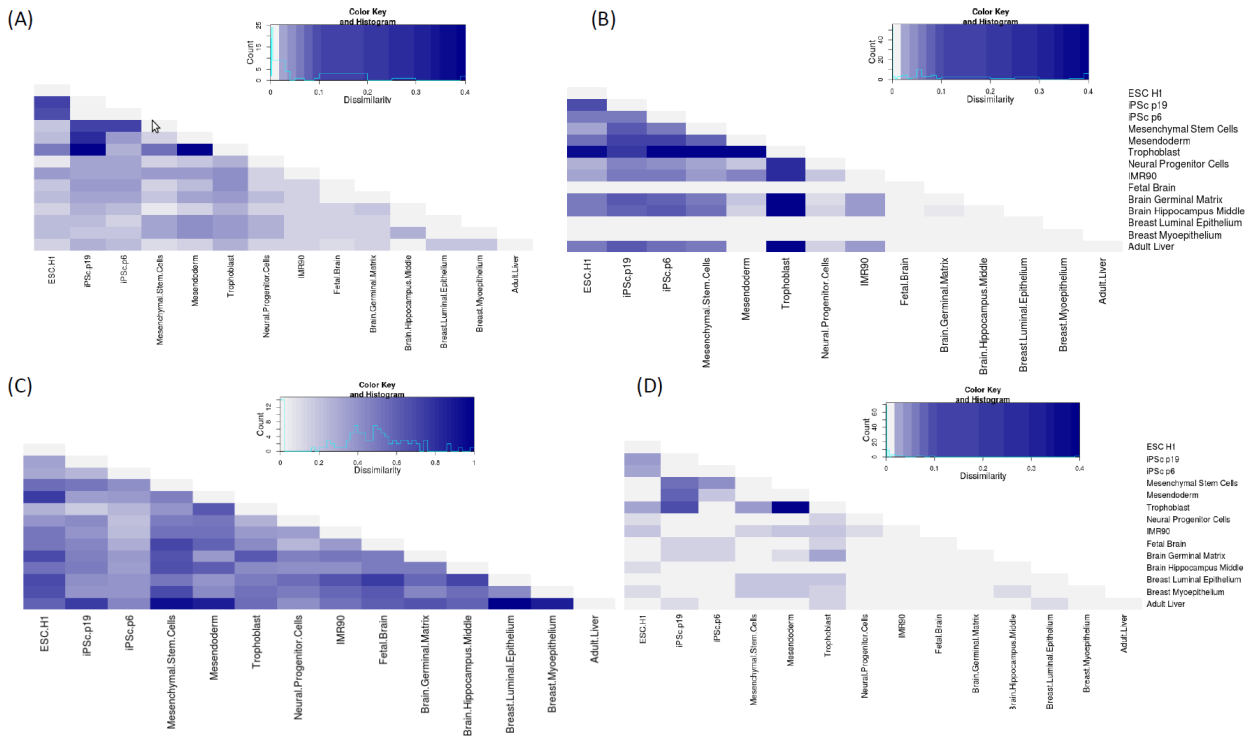

Figure 6.6: Heatmaps of the number of the resulting pairwise negative correlations for (a) expression data, (b) methylation data, (c) histone modifications, here H3K36me3, (d) the above mentioned union

in terms of associations with epigenetic marks. This was also the case for some of the brain tissues, various stem cell−like stages, and for the two passages of the iPS cells.

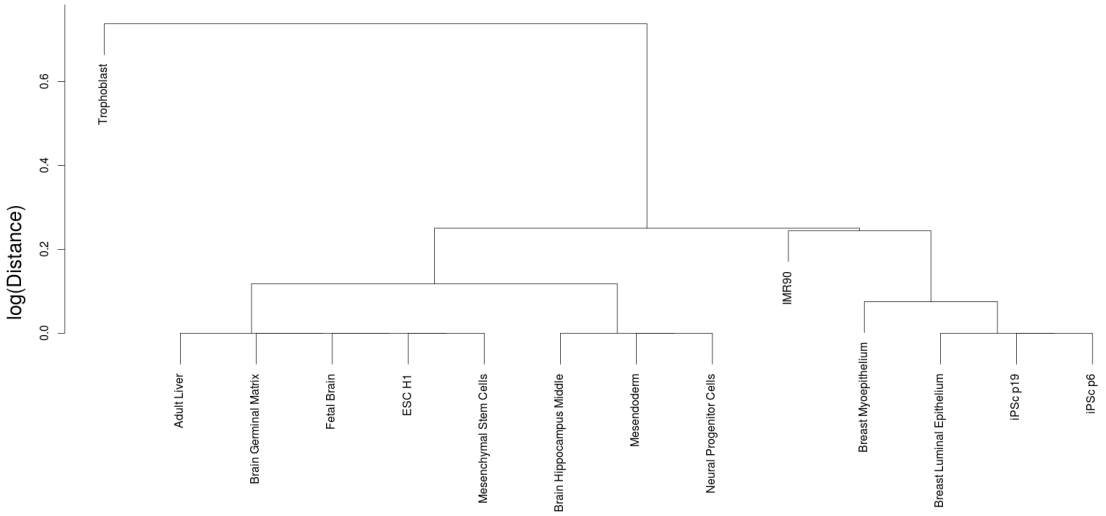

Figure 6.7: Hierarchical clustering for the set of genes that were analyzed in figure 6.7-d

For a better representation, we selected mutual negative correlations in the pairwise tissue comparisons with at least 10 epi-spliced genes (see Chapter 6.4.3). Within the constitutive genes, we found a total of 1529 epi-spliced genes. From this list, only 81 genes/gene clusters showed common modulation at the level of histone modification and methylation at the same time.

Next, we wished to understand the functional effect a single epigenetic mark can promote on the expression of a single exon, i.e. on the exon level. To this aim, we further investigated

the list of exons where changes in expression were associated with changes in any of the epigenetic marks across the studied developmental stages. We only considered epigenetic changes in the gene body of the same exon or in the promoter. Such changes can help assign an effect of a single histone mark, methylation state or siRNA regulation to human development by investigating crucial genes/gene clusters which are convolute with a single epigenetic modification. Table 6.2 lists the number of exons showing high correlations ($r \geq 0.7$) between an epigenetic modification and expression as well as the number of genes containing these exons. We did not account for putative exons that can be both positively and negatively associated with the same epigenetic mark, but only identified those showing either one of the two trends. However, the same gene can contain exons that are either positively or negatively associated with the same epigenetic mark.

Table 6.2: Number of exons/genes with significant correlation of exon-level expression and an epigenetic mark.

Epigenetic modification	Chromatin accessibility	H3K27ac	H3K27me3	H3K36me3	H3K4me1
Exon	2673	9081	568	3990	2267
Gene	725	452	187	1416	886
Epigenetic modification	H3K4me3	H3K9ac	H3K9me3	Methylation	siRNA
Exon	3519	319	121	318	0
Gene	942	145	44	122	0

6.4.1 Functional Classification of Epi-spliced Genes

We wished to classify our 'epi-spliced' genes into groups based on their functional similarity. Based on the result of the correlation analysis, we identified enriched GO terms for the resulting gene sets, both on the gene and exon levels. In doing so, we ignored the fact that different splice variants of a gene may sometimes promote very different functions [257].

To reach our aim of the functional classification of the 'epi-spliced' genes, we first analyzed the results for the negative associations on the gene level in a pair-wise manner, and considered enriched gene groups in terms of pair-wise tissue allocation. We aimed to identify the possible interplay between DNA methylation and the several histone marks in terms of regulation of 'differential exon usage'. For this, we identified genes where changes of both a histone mark and DNA methylation state coincide significantly with differential exon usage for the same gene. Such cases were exclusively found for combinations between the trophoblast cultured cells, mesendoderm cultured cells and induced pluripotent stem cells. GO terms associated with epi-spliced genes in those stages were associated with chromatin organization, (e.g.; the introduction of the heterochromatin and telomere structuring; growth of the ovarian follicle, oocyte, etc; transport processes of organic and inorganic molecules;) with metabolism; with transcriptional/translational and post translational regulation (e.g., K48- or K63-linked deubiquitination) and with homeostasis by regulation of embryonic hormones, interferons and Rac GTPase gene. The Rac protein has a role in growth and epithelial tissue differentiation and also a well established role in cancer. One further enriched GO term was H3K4 methylation.

We wanted to check whether differential changes in expression levels (namely differential exon usage) correlate with a single epigenetic mark. We thus analyzed GO terms for groups of genes with differential exon usage showing significant common changes of either histone marks or the DNA methylation state. We wished to categorize the GO terms at the tissue-level. We therefore grouped genes according to the differential pairwise tissue

allocation. The majority of significant intersections in histone modifications show early in development. Apart from the trophoblast cultured cells, the mesendoderm cells and the iPS cells, we also found significant changes between the trophoblast cultured cells and any of mesenchymal stem cells, the H1 embryonic stem cells, and the brain germinal matrix. iPS cells also display significant differences from the H1-derived mesenchymal stem cells. Apparently, using different passages (passage 6 and passage 19) of iPS- cells results in significant differences.

To understand the impact of 'episplicing' at the level of tissues, we grouped the GO terms into seven broad functional categories, see figure 6.8, namely development, DNA and chromatin organization, regulation of transcription and translation, signaling pathways, metabolism, regulation, and others.

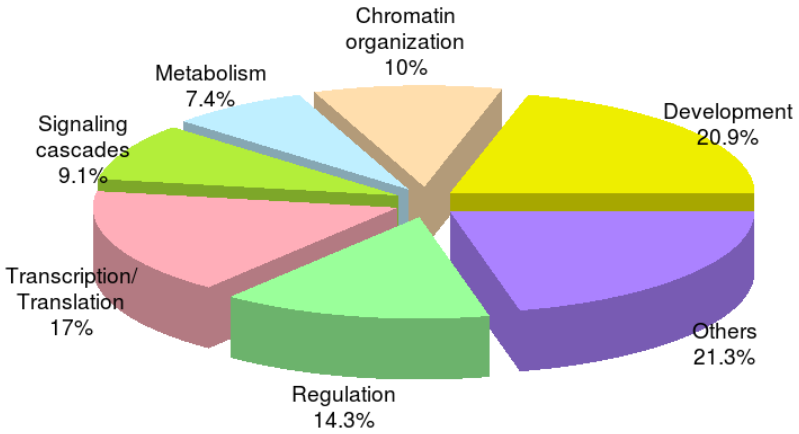

Figure 6.8: Frequency of Gene Ontology terms belonging to epi-spliced genes to seven according to seven manually defined biological categories. Epi-spliced genes are those showing a common negative correlation on the expression/epigenetic modification level across pair-wise tissue comparisons

Epi-spliced genes were overrepresented in developmental processes associated with the following tissues: blood vessels, chondrocytes, cytotoxic T cell, keratinocyte, oogenesis, organelle assembly, and several others. The biological processes related to chromatin organization that involve epi-spliced genes include processes associated with M-phase of the cell cycle, and several preparatory processes of the G1/S/G2 phases of the cell cycle. The category of transcription and translation involved many regulatory processes at the level of transcription, translation and post-translational modifications. The identified metabolic processes were associated with sugar metabolism, e.g., fructose 6-phosphate and fructose 1,6-phosphate metabolism, with phosphate metabolism, fatty acid metabolism, growth factors production, etc. The category of regulation included Ras GTPase activity, neuron migration, keratinocytes migration, and several others. As for signaling cascades, this category included for instance the regulation of the MAPK cascade, bone morphogenic protein (BMP) signaling, signal transduction, involving Rac and Rho proteins and nerve growth factor receptor signaling pathways, as well as SMAD proteins. Rac and Rho proteins belong to the Ras family and regulate important cellular processes as cytoskeleton remodelling, gene expression, cell proliferation and organelle development [258],[259]. SMADs are involved in TGF-β signalling from the cell membrane to the nucleus [260].

6.4.2 Linking Epi-spliced Genes to Particular Epigenetic Modifications

We then aimed at understanding the functional association at the exon level of a single epigenetic mark with exon expression. This helps identify the biological significance of a single epigenetic mark at the exon level. We accordingly identified GO terms of epi-spliced genes that were significantly linked to individual epigenetic modifications. This grouping

was based on significant correlations at the single exon level across developmental stages. We used the same functional cataloguing of GO terms used at the tissue-level. Figure 6.9 illustrates the set of biological processes and their modulation via epigenetic signals. Overall, more GO terms were associated with differential histone marks than with differential DNA methylation. Additionally, H3K36me3 showed the strongest association with regulation processes related to transcription/translation/post-translational modification, chromatin modeling, and development. Whereas several histone modifications showed strong effects on genes of the given biological categories, others exhibited weak correlations in the same context, namely H3K9me3 and H3K27me3. Table S2 lists the set of GO terms that are enriched for each studied epigenetic modification.

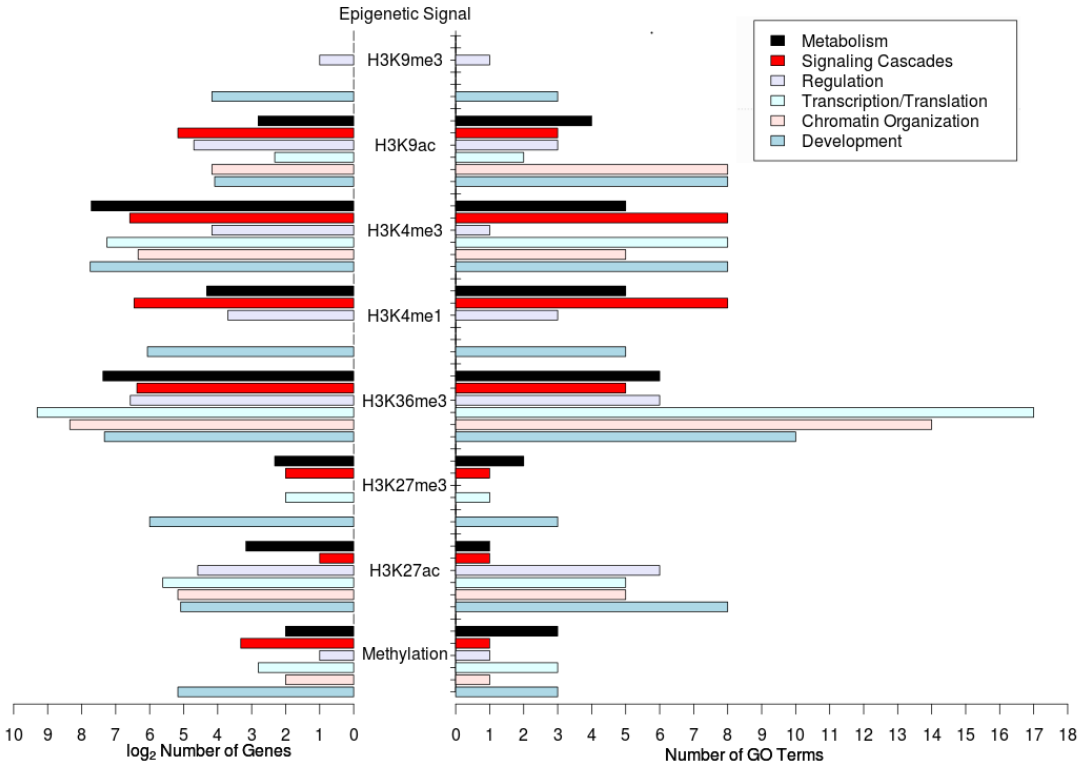

Figure 6.9: The numbers of selected GO terms belonging to genes showing differential regulation of the exon level for different types of epigenetic modifications

Association of epi-splicing with developmental stages In the epi-spliced genes under microscope, we wished to investigate intensely the functional GO terms correlated with development. GO terms associated with development and growth were enriched in genes that showed correlation between exon usage and DNA methylation, see Figure 6.9. The most pronounced developmental effects related to DNA methylation were associated with nervous system development. Interestingly, we observed that genes for which high methylation levels of their exons were correlated with their expression had an important effect on DNA conformation, what is a well-known effect documented from experiments [261]. We also noticed that differential methylation was associated with crucial regulatory processes, including the regulation of protein phosphatase 2B, GTP catabolism, and Rho protein signal transduction. Next, we examined the biological processes enriched in epi-spliced genes that are associated with DNAse hypersensitivity. In this context, we found that this assay targets genes enriched with GO terms of open/closed chromatin organization. A striking example for this effect is the H3K4 histone methylation level.

Next, we studied epi-spliced genes showing histone modifications at the exon level across several developmental stages and analyzed GO terms enriched in these genes. The inhibitory/activating marks of H3K4 methylation/tri-methylation are associated on the

developmental level with notochord regression, neuron projection regeneration and several morphogenetic processes. We also found that these histone modifications are associated with female pregnancy and hippo signaling pathways that are prominent in the regulation of cell proliferation and apoptosis [262]. Hippo signaling also serves the organisms to stop growth at a specific point, thus aiding in size control [263]. H3K4 methylation/tri-methylation have a relevant effect on the levels of two proteins that act as a heterodimer, namely TLR1 and TLR2 that have roles in immune response [264]. Differential H3K4 methylation correlated to exon usage is associated with the signaling cascade of the oncogene smoothened protein, with the bone morphogenic protein signaling cascade, with cascades including the SMAD protein and the transforming growth factor proteins, and with the nerve growth factor signaling cascade, including the well known BDNF protein that is also controlled through DNA methylation [265],[266]. In terms of post-translational modifications, we found that these histone modifications also control the phosphoprotein phosphatase activity.

With respect to modifications of H3K9 associated with differential exon usage at the developmental level, we found that H3K9 acetylation is strongly connected to DNA and chromatin organization, cell cycle events, bone morphogenesis and differentiation, and with post-translational modifications, including for example Hsp90 chaperon acetylation. H3K9 tri-methylation, on the other hand, is mainly associated with nervous system development. Acetylation of H3K27 is directly associated with GO terms related to histone acetylation, suggesting a possible negative/positive feedback effect. It is also associated with the developmental control on the level of embryonic heart muscles and with processes related to chromatin organization. On the other hand, H3K27 trimethylation is associated with nervous system development and platelet-derived growth factor (PDGF) receptor signaling pathways. Lastly, we found that H3K36 trimethylation is associated with DNA replication and repair, chromatin organization, cell cycle events (e.g. G2/M phase checkpoints and mitotic cell division), regulation of transcription, and several developmental stages. We found that H3K36 trimethylation also modulates signaling cascades together with H3K4 methylation.

The biological processes just discussed only involved associations of epi-spliced genes and individual or several synchronously changing epigenetic marks. Next, we selected several groups of genes known previously to have direct or indirect associations with one or several epigenetic modifications. We wished to know whether these genes are epi-spliced. The identified categories included several imprinted genes, chromatin remodelers, protein kinases, and transcription factors and cofactors. We performed this analysis for epi-genes both on the gene level in pairwise tissue comparisons (Table S3) and on the exon level across developmental stages (Table S4). As an example of the imprinted genes, we found that four different paternally expressed genes vary their exon usage in a common manner due to a synchronous change of histone modifications and DNA methylation between mesendoderm cultured cells and trophoblast cultured cells. These genes are PEG3/ZIM2 and SNRPN/SNURF that are all known to undergo alternative splicing [229][230][267]. Another exciting example of a chromatin remodeler gene that varies the expression of its isomers in the same manner and in the same tissues is DNMT3L methyltransferase that is well known to recruit chromatin remodelers, especially histone deacetylases [268][269]. This enzyme is also known to have crucial roles in early developmental stages, especially in the establishment of imprints together with de novo methyltransferases [270]. A splice variant has been introduced for this gene in Ref-Seq genes, 2012.

Furthermore, and for the heretofore-mentioned gene categories, e.g., imprinted genes and chromatin remodelers, we explored the list of genes for which exon expression is associated with a specific epigenetic mark, i.e. at the exon level. Interestingly, exon expression of a few imprinted genes changes across developmental stages, and this expression was modulated by

several epigenetic marks. For example, the maternally expressed gene SLC22A18 changed its expression according to the padding at the chromatin structure and is modulated by H3K4 mono-methylation. This gene has been linked to alternative splicing events before [271]. Transcription factors are another example for genes with documented modulation at the exon level. Here, we found two well-known transcription factors, ZFP42 and NANOG, that regulate pluripotency and differentiation in the embryonic stem cells [272]. For example, we found that ZFP42, on the exon level, changes its chromatin organization (DNAse hypersensitivity) and is modulated by H3K4 trimethylation. This gene was shown to undergo changes on the exon level in early development [273][274]. Moreover, we found that NANOG, a regulating transcription factor of the ZFP42 gene [275] which is also known to undergo alternative splicing [228], is also modulated by the same epigenetic mark, H3K4me3, on the exon level.

6.4.3 Positive Correlations

To complete our analysis, we finally searched for common positive correlations in the expression level of exons across tissues with PCC $\geq$0.7. We aimed to find possible impact of constitutive genes on development. Interestingly, we found that such constitutively expressed genes were not usually ubiquitously expressed across tissues. Additionally and as expected, coexpression was predominately found for highly similar tissues (lower left half of figure 6.10-b), thus arguing against an important role of constitutive genes in development. The two genes that showed the largest number of abundant constitutive expression, CA2 and FOXO4, also showed the highest abundance in alternative splicing. Figure 6.10-(a,b) shows a comparison of the numbers and allocation of positive and negative correlations of gene expression. The range of the number of genes in both matrices is similar on average. However, the anticorrelations involved mostly genes/gene clusters from early developmental stages. Figures 6.10-(c,d) show the normalized expression of the exons contained in the set of genes that show anticorrelations and correlations in at least 26 combinations of tissues, respectively. In general, where changes do occur for exons in the anticorrelated genes, they do not occur at the level of the full genes. Rather specific exons are responsible for the variation, and other exons of these genes are more constitutively expressed.

6.4.4 Conclusion

Exon-intron boundaries set by histones/epigenetic marks are not only used to define the ends of the elements for the mRNA transcript to be expressed. Rather, they can also be considered as a part of a machinery for regulating and controlling the relative abundance of the several transcripts or protein isoforms that map to the same chromosomal region across tissues. This relationship seems to be most prominent in early developmental stages, and this suggests differential regulation across developmental stages, brought about by the distinct epi-genes. Additionally, exon-body epigenetic effect is more pronounced than that of intronic or promoter effects.

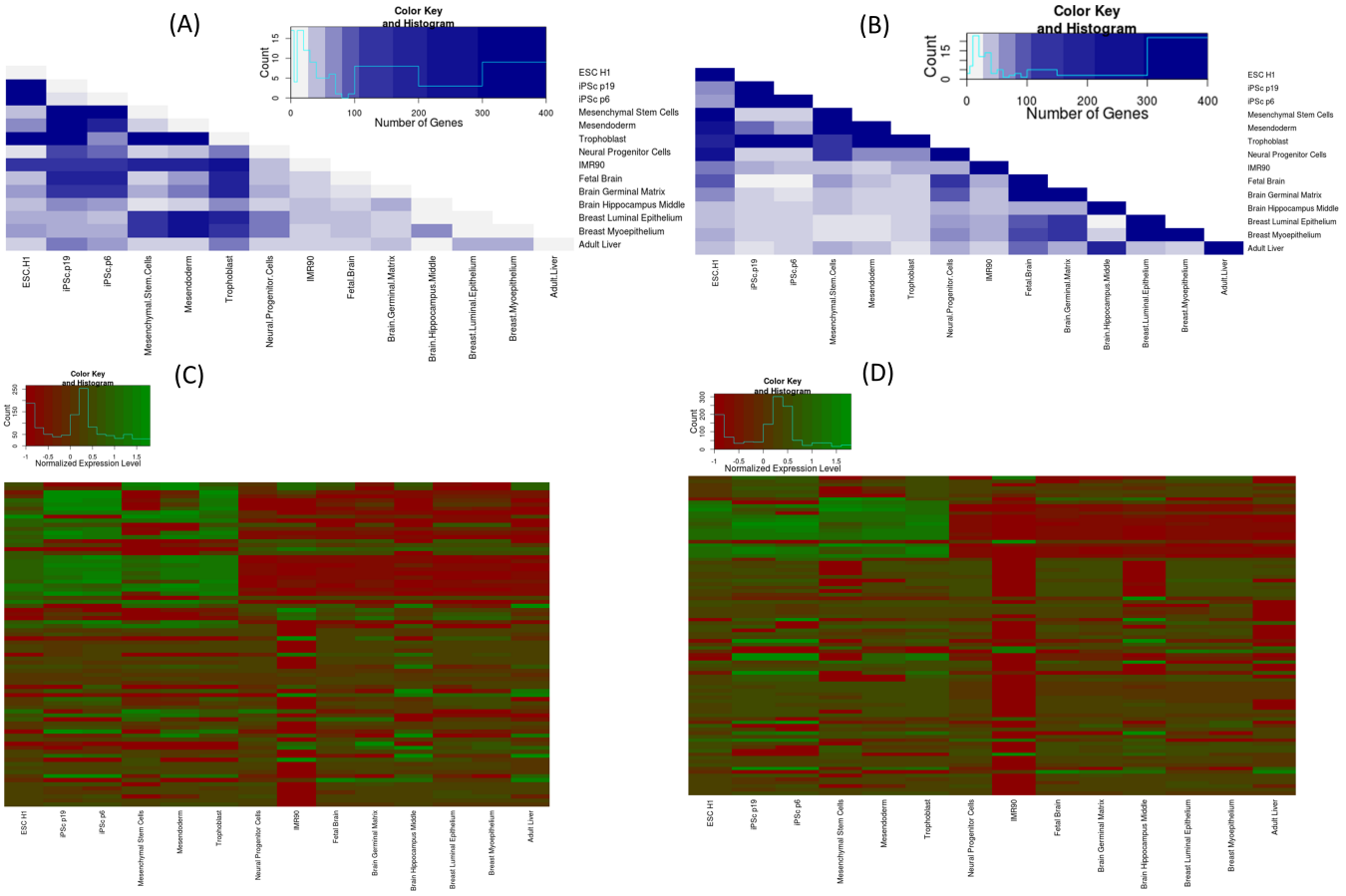

Figure 6.10: Heatmaps for the expression levels and gene numbers in pair-wise tissue correlations for (a) alternatively spliced genes, (b) constitutively expressed genes, (c) expression levels of individual exons belonging to 11 selected genes that show strong anticorrelation, and (d) expression levels of individual exons belonging to 10 selected genes that show strong positive correlation

Chapter 7

Summary and Outlook

In this work we addressed from various angles how one may suggest genes of certain functions as tumor markers and then study them at the basic exon level. In the first case, we concentrated on the function of gene products that have a chance of being tumor markers. We started from the fact that it is costly and time consuming to study the function of individual proteins. Therefore, once a protein function is identified, it is desirable to transfer it to other proteins sharing certain sequence or structural similarity. In chapter 3, we presented a combined approach for transferring functional annotations of certain proteins (Transporters in this work) between transporters sharing certain sequence similarity from different species. The accumulative decision was based on decisions from BLAST, HMMER and MEME. All tools search for sequence similarity but MEME performs motif searches. We found that up to certain thresholds of E-values, the transporter function can be transferred to putative transporters from other organisms. Among the three combined methods, the functional annotations based on MEME results were less reliable than those based on BLAST and HMMER results.

This approach worked nicely for transporter families created based on phylogeny (TC families). However, it did not achieve comparable results when applied to families we created based on substrates being transferred. This gives a hint that phylogenetic inference is a more sensitive indicator of homology compared to sequence similarity. Apart from TC families, we suggest that using other sequence analysis methods like Amino acid Composition (AAC) or even incorporating structure based methods side by side with sequence based methods might aid in the substrate prediction.

The next component was a preparatory phase before analysing possible marker genes based on their expression or methylation profiles in normal and tumor samples. As these datasets frequently suffer from outlier values leading to misleading results, this component concentrated on detecting such outliers and removing the clear cases. Outliers in such datasets can be samples or genes with some outlier expression/ methylation values. To detect outlier samples we used hierarchical clustering. At the gene level, we checked first whether the expression/methylation of the gene follows a normal distribution. If this is the case, we applied the GSED algorithm. For other genes we applied boxplot and MAD algorithms. Moreover, it is reported that some algorithms might label one outlier even if none exists. Here, aroused the need for establishing an outlier margin where a gene is labeled as outlier only if it has expression/methylation outlier values exceeding certain threshold. We found that 2 outlier observations might ruin a perfect co-expression and thus used 2 as the outlier margin.

Some outlier genes appeared to carry interesting details behind this outlier behaviour. Therefore, we tested for semantic similarity between outliers and kept groups of functionally similar outliers for further analysis. Only pure non similar outliers where labelled for removal. Although the whole analysis was completed in R-cran, Taner Arslan developed a GUI python stand alone tool for outlier detection under the supurvision of the thesis author.

The next step in this field would be to test this outlier detection approach on other

datasets for example MicroRNA expression datasets. We also suggest that testing for the commonly found distributions in different types of datasets would be beneficial for detecting outliers. For instance, normal distribution was common in expression datasets but not that common in methylation datasets that we analyzed.

The third component was the core one for suggesting and analyzing new tumor marker genes. Our research was mainly focused on hepatocellular carcinoma (HCC) as it is the second most common cancer related death worldwide. Additionally, it might be the end-stage of untreated liver diseases like hepatitis. Generally, we incorporated expression and methylation analysis with the wet-lab testing by our collaborators to analyze the behaviour of the tumor marker.

For example, our analysis suggests that IMP2 plays an important role in initiating HCC and also in its progression. One hint was that it had increased expression in the majority of HCC patients. IMP2 had a similar behaviour in one of the breast cancer subtypes. IMP2 expression was elevated in tissues of basal-like cancer compared to the luminal or apocrine subtypes.

On the other hand, the gene Hamp had a different behaviour because its expression was reduced in tumor tissues compared to adjacent normal liver tissues. We found that Hamp expression is low in liver disease samples (chronic hepatitis C and cirrhosis) compared to healthy liver samples. Along the same side, our analysis showed that the expression levels of *ELOVL6* are significantly decreased in the majority of human liver tumors compared to nontumorous tissues. However, *ELOVL6* expression is elevated in human NASH and NASH-related HCC samples. The next step here would be to map the suggested marker genes to other cancer types to get a broader overview.

In the last component, we presented an approach to study certain genes at the basic level of their exons. Although we presented a genome wide study, this approach of course can be used for a set of genes of interest (the tumor marker genes for example). Differential exon usage helps to express several proteins from the same genomic location via the mechanism of alternative splicing. In this work we showed that epigenetic modifications are strongly associated with alternative splicing especially for genes that are essential for development. The next step here would be to establish exon usage relations based on exon expression and epigenetic modifications.

In summary, the work presented in this thesis led to transferring functional annotations of specific proteins across species. Such proteins might be options for advanced tumor markers. As tumor markers are often identified according to their expression/ methylation profiles, this work insisted on cleaning the needed datasets from outliers before analysis. Once the markers are identified and validated in the lab, the rest of this work presents a method for intensive analysis of the differential exon usage of the genes of interest.

Chapter 8

Appendix

A.1 Supplementary Data for Chapter 3

		BLAST					HMMER					MEME				
		1e-20	1e-16	1e-12	1e-8	1e-4	1e-20	1e-16	1e-12	1e-8	1e-4	1e-20	1e-16	1e-12	1e-8	1e-4
Ec-Hs	Precision	79.2	73.2	75.4	75.7	35.1	65.3	70.8	75.4	75.4	68.2	79.2	79.2	65.3	42.0	25.0
	Recall	79.2	79.2	79.2	75.7	58.8	79.2	79.2	79.2	79.2	79.2	79.2	79.2	79.2	63.4	39.8
	F-measure	79.2	76.0	77.2	75.7	43.5	70.8	74.5	77.2	77.2	73.2	79.2	79.2	70.8	45.0	29.1
	Unclassified	87.5	86.1	80.6	77.8	30.6	91.7	88.9	80.6	80.6	76.4	94.4	94.4	91.7	44.4	0.0
Hs-At	Precision	66.7	63.8	57.6	56.7	13.8	66.7	66.7	66.7	63.8	57.6	66.7	66.7	57.6	26.1	25.0
	Recall	66.7	66.7	66.7	66.7	100.0	66.7	66.7	66.7	66.7	66.7	66.7	66.7	43.3	25.5	33.5
	F-measure	66.7	65.1	60.8	60.0	24.2	66.7	66.7	66.7	65.1	60.8	66.7	66.7	45.2	25.3	27.3
	Unclassified	66.7	63.3	60.0	60.0	0.0	66.7	66.7	63.3	60.0	60.0	70.0	70.0	63.3	0.0	0.0

Table A1: Complete annotation results of the pairs (*Hs*, *At*) and (*Ec*, *Hs*).

	Ec-Hs					*Sc-At*					*Ec-At*				
	1e-16	1e-8	1e-4	1e-3	1e-2	1e-16	1e-8	1e-4	1e-3	1e-2	1e-16	1e-8	1e-4	1e-3	1e-2
Precision[%]	75.4	72.2	67.7	57.9	41.8	69.3	41.1	25.5	22.2	19.8	63.6	64.8	45.7	33.0	31.6
Recall[%]	79.2	79.2	72.8	66.2	60.3	100.0	100.0	100.0	100.0	100.0	86.9	83.7	73.0	58.0	51.6
F-measure[%]	77.2	75.4	70.1	59.8	48.5	79.0	57.0	40.2	35.9	32.7	69.5	69.5	46.0	37.2	33.9
Unclassified[%]	72.6	66.5	53.8	49.9	34.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

	Ec-Sc					*Hs-At*					*Hs-Sc*				
	1e-16	1e-8	1e-4	1e-3	1e-2	1e-16	1e-8	1e-4	1e-3	1e-2	1e-16	1e-8	1e-4	1e-3	1e-2
Precision[%]	63.0	57.5	39.9	36.4	32.0	56.7	33.0	16.6	15.5	15.6	56.7	43.3	29.6	26.2	24.7
Recall[%]	93.1	81.3	59.5	53.1	45.4	66.7	66.7	93.3	93.3	100.0	66.7	63.8	39.4	37.6	33.2
F-measure[%]	73.4	62.1	43.0	37.1	32.6	60.0	43.8	27.8	26.5	26.8	60.0	51.5	31.3	28.6	26.3
Unclassified	0.0	0.0	0.0	0.0	0.0	60.0	53.3	23.3	10.0	3.3	60.0	60.0	23.3	13.3	3.3

Table A2: Results of FASTA global searches.

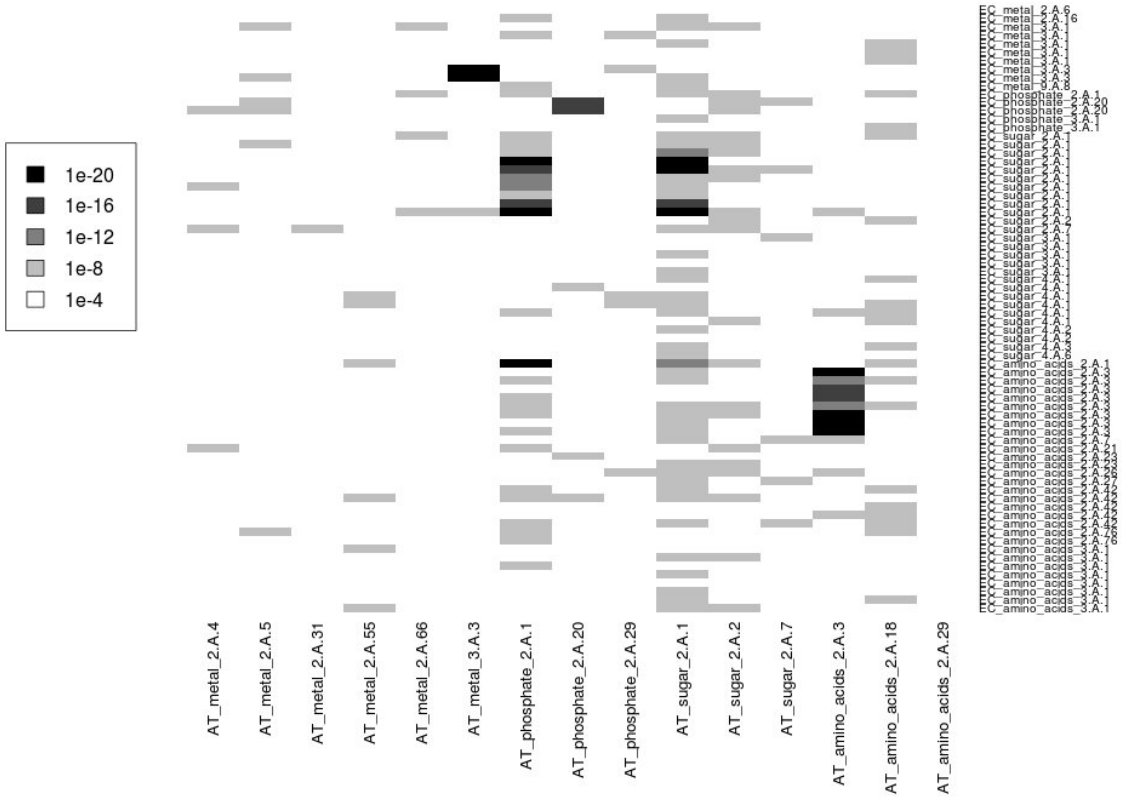

Figure A1: Heatmap of BLASTing *Ec* substrate-TC families against *At* families

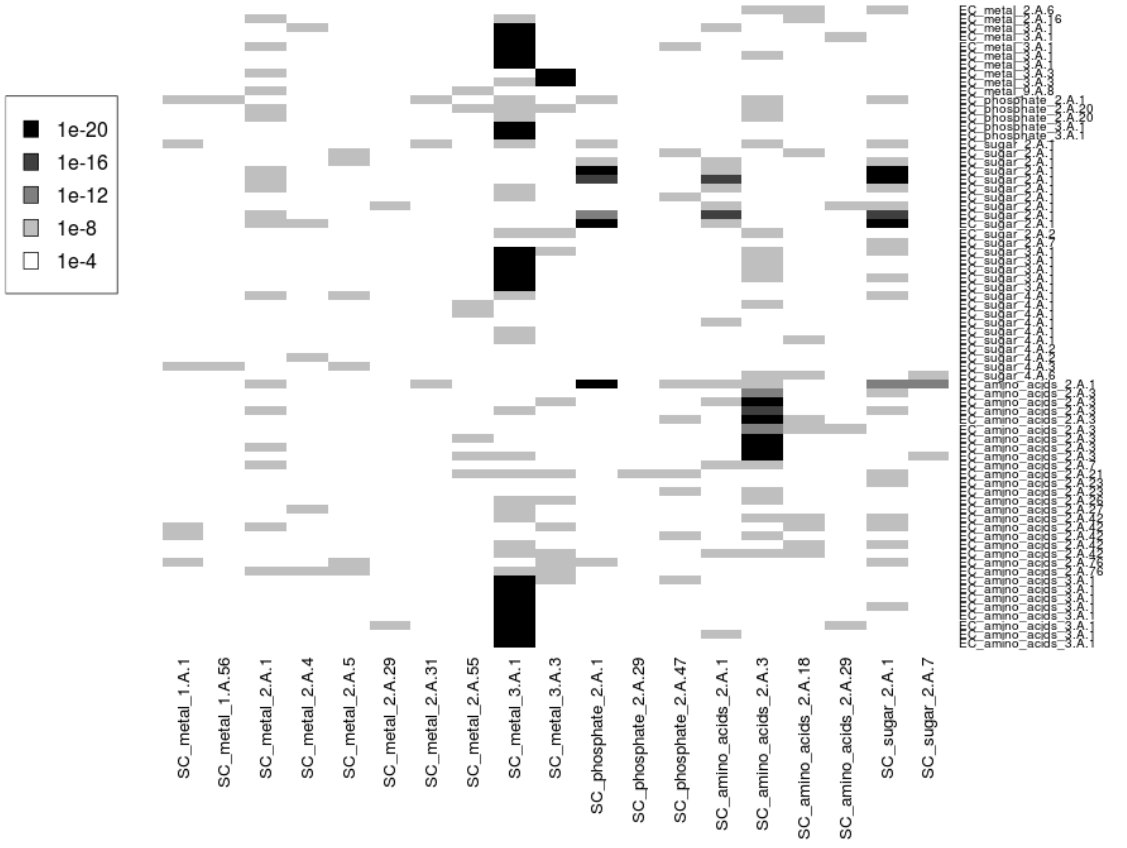

Figure A2: Heatmap of BLASTing *Ec* substrate-TC families against *Sc* families

A.2 Supplementary Data for Chapter 5

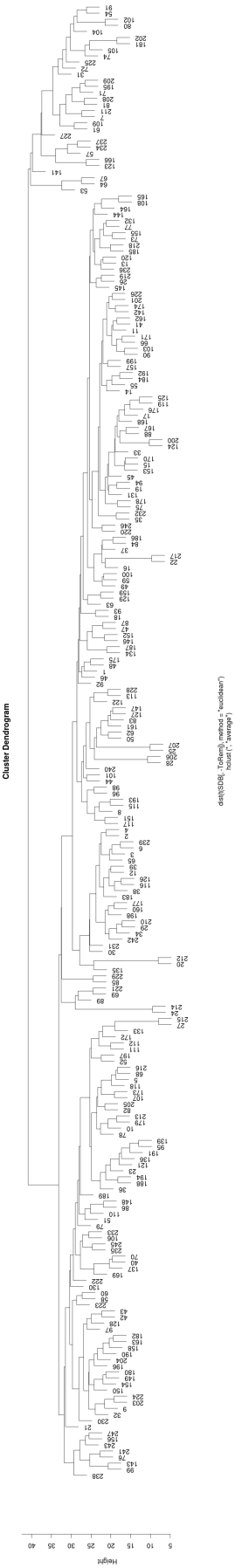

Figure A3: Cluster dendogram of complete hierarchical clustering analysis of dataset GSE14520 using marker genes presented by Hoshida et al. [7]. Two major subclasses were identified

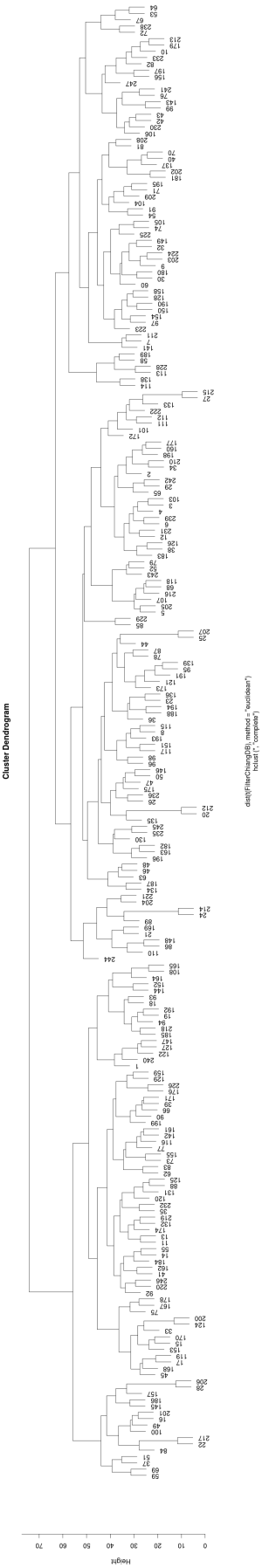

Figure A4: Cluster dendogram of complete hierarchical clustering analysis of dataset GSE14520 using marker genes presented by Chiang et al. [8]. Three major subclasses were identified

References

[1] K. D. Robertson, "DNA methylation and human disease," *Nature Reviews Genetics*, vol. 6, pp. 597–610, 2005.

[2] R. Kumar, A. Sharma, A. K. Pattnaik, and P. K. Varadwaj, "Stem cells: An overview with respect to cardiovascular and renal disease," *Journal of Natural Science, Biology, and Medicine*, vol. 1, pp. 43–52, 2010.

[3] Source: Boundless. "Electrochemical Gradient." Boundless Biology. Boundless, 14 Nov. 2014. Retrieved 07 Apr. 2015 from https://www.boundless.com/biology/textbooks/boundless-biology-textbook/structure-and-function-of-plasma-membranes-5/active-transport-66/electrochemical-gradient-336-11473/.

[4] A. S. Alex Mitchell, "Introduction to protein classification at the ebi." Train Online Course. https://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/what-are-protein-signatures/signature-types/what-ar-1.

[5] J. Fridlyand, A. M. Snijders, D. Pinkel, D. G. Albertson, and A. N. Jain, "Hidden Markov models approach to the analysis of array CGH data," *Journal of Multivariate Analysis*, vol. 90, pp. 132–153, 2004.

[6] S. M. Kessler, J. Pokorny, V. Zimmer, S. Laggai, F. Lammert, R. M. Bohle, and A. K. Kiemer, "IGF2 mRNA binding protein p62/IMP2-2 in hepatocellular carcinoma: antiapoptotic action is independent of IGF2/PI3K signaling," *American Journal of Physiology-Gastrointestinal and Liver Physiology*, vol. 304, pp. G328–G336, 2013.

[7] Y. Hoshida, S. M. Nijman, M. Kobayashi, J. A. Chan, J.-P. Brunet, D. Y. Chiang, A. Villanueva, P. Newell, K. Ikeda, M. Hashimoto, *et al.*, "Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma," *Cancer Research*, vol. 69, pp. 7385–7392, 2009.

[8] D. Y. Chiang, A. Villanueva, Y. Hoshida, J. Peix, P. Newell, B. Minguez, A. C. LeBlanc, D. J. Donovan, S. N. Thung, M. Solé, *et al.*, "Focal gains of VEGFA and molecular classification of hepatocellular carcinoma," *Cancer Research*, vol. 68, pp. 6779–6788, 2008.

[9] E. Suárez, C. A. Sariol, A. Burguete, and G. Mclachlan, "A tutorial in genetic epidemiology and some considerations in statistical modeling," *Puerto Rico Health Sciences Journal*, vol. 26, pp. 1145–1160, 2007.

[10] F. Ducray, J. Honnorat, and J. Lachuer, "DNA microarray technology: principles and applications to the study of neurological disorders," *Revue Neurologique*, vol. 163, pp. 409–420, 2007.

[11] J. K. Cowell and L. Hawthorn, "The application of microarray technology to the analysis of the cancer genome," *Current Molecular Medicine*, vol. 7, pp. 103–120, 2007.

[12] Y. Chu and D. R. Corey, "RNA sequencing: platform selection, experimental design, and data interpretation," *Nucleic Acid Therapeutics*, vol. 22, pp. 271–274, 2012.

[13] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, T. P. Speed, *et al.*, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, pp. 249–264, 2003.

[14] M. E. Ritchie, J. Silver, A. Oshlack, M. Holmes, D. Diyagama, A. Holloway, and G. K. Smyth, "A comparison of background correction methods for two-colour microarrays," *Bioinformatics*, vol. 23, pp. 2700–2707, 2007.

[15] G. K. Smyth, " Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, pp. 1–25, 2004.

[16] F. J. Massey Jr, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, pp. 68–78, 1951.

[17] M. A. P. Mohammadmersad Ghorbani, Simon J. E. Taylor and A. Payne, "Comparative (computational) analysis of the DNA methylation status of trinucleotide repeat expansion diseases," *Journal of Nucleic Acids*, p. 12–19, 2013.

[18] A. Bird, "Perceptions of epigenetics," *Nature*, vol. 447, pp. 396–398, 2007.

[19] A. P. W. Adrian P Bird, "Methylation-induced repression— belts, braces, and chromatin," *Cell*, vol. 99, pp. 451–454, 1999.

[20] J. H. Kim, S. M. Dhanasekaran, J. R. Prensner, X. Cao, D. Robinson, S. Kalyana-Sundaram, C. Huang, S. Shankar, X. Jing, M. Iyer, M. Hu, L. Sam, C. Grasso, C. A. Maher, N. Palanisamy, R. Mehra, H. D. Kominsky, J. Siddiqui, J. Yu, Z. S. Qin, and A. M. Chinnaiyan, "Deep sequencing reveals distinct patterns of dna methylation in prostate cancer," *Genome Research*, vol. 21, pp. 1028–1041, 2011.

[21] P. J. Hurd and C. J. Nelson, "Advantages of next-generation sequencing versus the microarray in epigenetic research," *Briefings in Functional Genomics and Proteomics*, vol. 8, pp. 174–183, 2009.

[22] T. Kouzarides, "Chromatin modifications and their function," *Cell*, vol. 128, pp. 693–705, 2007.

[23] S. L. Berger, "Histone modifications in transcriptional regulation," *Current Opinion in Genetics & Development*, vol. 12, pp. 142–148, 2002.

[24] J.-H. Lee, S. R. Hart, and D. G. Skalnik, "Histone deacetylase activity is required for embryonic stem cell differentiation," *Genesis*, vol. 38, pp. 32–38, 2004.

[25] V. Calvanese, A. F. Fernández, R. G. Urdinguio, B. Suarez-Alvarez, C. Mangas, V. Pérez-García, C. Bueno, R. Montes, V. Ramos-Mejía, P. Martínez-Camblor, *et al.*, "A promoter DNA demethylation landscape of human hematopoietic differentiation," *Nucleic Acids Research*, vol. 40, pp. 116–131, 2012.

[26] K. Furge, M. Tan, K. Dykema, E. Kort, W. Stadler, X. Yao, M. Zhou, and B. Teh, "Identification of deregulated oncogenic pathways in renal cell carcinoma: an integrated oncogenomic approach based on gene expression profiling," *Oncogene*, vol. 26, pp. 1346–1350, 2007.

[27] M. A. Virji, D. W. Mercer, and R. B. Herberman, "Tumor markers in cancer diagnosis and prognosis," *CA: a Cancer Journal for Clinicians*, vol. 38, pp. 104–126, 1988.

[28] K. Adekola, S. T. Rosen, and M. Shanmugam, "Glucose transporters in cancer metabolism," *Current Opinion in Oncology*, vol. 24, pp. 650–654, 2012.

[29] N. Alexander, M. Wankerl, J. Hennig, R. Miller, S. Zänkert, S. Steudte-Schmiedgen, T. Stalder, and C. Kirschbaum, "DNA methylation profiles within the serotonin transporter gene moderate the association of 5-httlpr and cortisol stress reactivity," *Translational Psychiatry*, vol. 4, p. e443, 2014.

[30] W. Busch and M. H. Saier, "The IUBMB-endorsed transporter classification system," *Molecular Biotechnology*, vol. 27, pp. 253–262, 2004.

[31] M. H. Saier, M. R. Yen, K. Noto, D. G. Tamang, and C. Elkan, "The transporter classification database: recent advances," *Nucleic Acids Research*, vol. 37, pp. D274–D278, 2009.

[32] Q. Ren, K. Chen, and I. T. Paulsen, "TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels," *Nucleic Acids Research*, vol. 35, pp. 274–279, 2007.

[33] R. Schwacke, A. Schneider, E. van der Graaff, K. Fischer, E. Catoni, M. Desimone, W. B. Frommer, U.-I. Flügge, and R. Kunze, "ARAMEMNON, a novel database for Arabidopsis integral membrane proteins," *Plant Physiology*, vol. 131, pp. 16–26, 2003.

[34] W. R. Pearson, "An introduction to sequence similarity ("homology") searching," *Current Protocols in Bioinformatics*, pp. 1–3, 2013.

[35] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, pp. 403–410, 1990.

[36] S. R. Eddy *et al.*, "A new generation of homology search tools based on probabilistic inference," in *Genome Inform*, vol. 23, pp. 205–211, World Scientific, 2009.

[37] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," *Proceedings of the National Academy of Sciences*, vol. 85, pp. 2444–2448, 1988.

[38] A. Marsico, K. Scheubert, A. Tuukkanen, A. Henschel, C. Winter, R. Winnenburg, and M. Schroeder, "MeMotif: a database of linear motifs in α-helical transmembrane proteins," *Nucleic Acids Research*, vol. 38, pp. 181–189, 2009.

[39] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in bipolymers," *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28–36, 1994.

[40] D. Wheeler and M. Bhagwat, "BLAST QuickStart," 2007.

[41] W. R. Pearson, "Comparison of methods for searching protein sequence databases.," *Protein Science: a Publication of the Protein Society*, vol. 4, pp. 1145–1160, 1995.

[42] S. R. Eddy, "Hidden markov models," *Current Opinion in Structural Biology*, vol. 6, pp. 361–365, 1996.

[43] S. R. Eddy, "Accelerated profile HMM searches," *PLoS Computational Biology*, vol. 7, p. e1002195, 2011.

[44] L. Cheng and G. Butler, "Accelerating Search of Protein Sequence Databases using CUDA-Enabled GPU," in *Database Systems for Advanced Applications* (M. Renz, C. Shahabi, X. Zhou, and M. A. Cheema, eds.), vol. 9049 of *Lecture Notes in Computer Science*, pp. 279–298, Springer International Publishing, 2015.

[45] T. L. Bailey and M. Gribskov, "Combining evidence using p-values: application to sequence homology searches.," *Bioinformatics*, vol. 14, pp. 48–54, 1998.

[46] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, T. P. Speed, *et al.*, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, pp. 249–264, 2003.

[47] R. P. Grant, *Computational genomics: theory and application.* Horizon Bioscience, 2004.

[48] E. Suárez, A. Burguete, and G. J. Mclachlan, "Microarray data analysis for differential expression: a tutorial," *Puerto Rico Health Sciences Journal*, vol. 28, pp. 89–104, 2009.

[49] R. Shamir, A. Maron-Katz, A. Tanay, C. Linhart, I. Steinfeld, R. Sharan, Y. Shiloh, and R. Elkon, "EXPANDER–an integrative program suite for microarray data analysis," *BMC Bioinformatics*, vol. 6, p. 232, 2005.

[50] S. V. Yi and M. A. Goodisman, "Computational approaches for understanding the evolution of DNA methylation in animals," *Epigenetics*, vol. 4, pp. 551–556, 2009.

[51] F. A. Feltus, E. K. Lee, J. F. Costello, C. Plass, and P. M. Vertino, "DNA motifs associated with aberrant CpG island methylation," *Genomics*, vol. 87, pp. 572–579, 2006.

[52] H. Zheng, H. Wu, J. Li, and S.-W. Jiang, "CpGIMethPred: computational model for predicting methylation status of CpG islands in human genome," *BMC Medical Genomics*, vol. 6, p. S13, 2013.

[53] A. E. Jaffe, P. Murakami, H. Lee, J. T. Leek, M. D. Fallin, A. P. Feinberg, and R. A. Irizarry, "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies," *International Journal of Epidemiology*, vol. 41, pp. 200–209, 2012.

[54] D. Sun, Y. Xi, B. Rodriguez, H. J. Park, P. Tong, M. Meong, M. A. Goodell, and W. Li, "MOABS: model based analysis of bisulfite sequencing data," *Genome Biology*, vol. 15, p. 38, 2014.

[55] Y. Assenov, F. Müller, P. Lutsik, J. Walter, T. Lengauer, and C. Bock, "Comprehensive analysis of DNA methylation data with RnBeads," *Nature Methods*, vol. 11, pp. 1138–1140, 2014.

[56] A. D. Shieh and Y. S. Hung, "Detecting outlier samples in microarray data," *Statistical Applications in Genetics and Molecular Biology*, vol. 8, pp. 1–24, 2009.

[57] A. C. S. Rao, D. Somayajulu, H. Banka, and R. Chaturvedi, "Outlier Detection in Microarray Data Using Hybrid Evolutionary Algorithm," *Procedia Technology*, vol. 6, pp. 291–298, 2012.

[58] X. Lu, Y. Li, and X. Zhang, "A simple strategy for detecting outlier samples in microarray data," in *Control, Automation, Robotics and Vision Conference, 2004. ICARCV 2004 8th*, vol. 2, pp. 1331–1335, IEEE, 2004.

[59] B. Rosner, "Percentage points for a generalized ESD many-outlier procedure," *Technometrics*, vol. 25, pp. 165–172, 1983.

[60] M. A. Van de Wiel, K. I. Kim, S. J. Vosse, W. N. Van Wieringen, S. M. Wilting, and B. Ylstra, "CGHcall: calling aberrations for array CGH tumor profiles," *Bioinformatics*, vol. 23, pp. 892–894, 2007.

[61] J. M. Cherry, E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, E. T. Chan, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel, *et al.*, "Saccharomyces Genome Database: the genomics resource of budding yeast," *Nucleic Acids Research*, vol. 40, pp. D700–D705, 2011.

[62] N. S. Schaadt, J. Christoph, and V. Helms, "Classifying substrate specificities of membrane transporters from Arabidopsis thaliana," *Journal of Chemical Information and Modeling*, vol. 50, pp. 1899–1905, 2010.

[63] N. S. Schaadt and V. Helms, "Functional classification of membrane transporters and channels based on filtered TM/non-TM amino acid composition," *Biopolymers*, vol. 97, pp. 558–567, 2012.

[64] X. D. Haiquan Li and X. Zhao, "A nearest neighbor approach for automated transporter prediction and categorization from protein sequences," *Bioinformatics*, vol. 24, pp. 1129–1136, 2008.

[65] H. Li, V. A. Benedito, M. K. Udvardi, and P. X. Zhao, "TransportTP: a two-phase classification approach for membrane transporter prediction and characterization," *BMC Bioinformatics*, vol. 10, p. 418, 2009.

[66] M. M. Gromiha and Y. Yabuki, "Functional discrimination of membrane proteins using machine learning techniques," *BMC Bioinformatics*, vol. 9, p. 135, 2008.

[67] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn, "The pfam protein families database," *Nucleic Acids Research*, vol. 40, pp. D290–D301, 2012.

[68] A. B. Chang, R. Lin, W. K. Studley, C. V. Tran, and M. H. Saier, Jr, "Phylogeny as a guide to structure and function of membrane transport proteins (Review)," *Molecular Membrane biology*, vol. 21, pp. 171–181, 2004.

[69] F. Chen, A. J. Mackey, J. K. Vermunt, and D. S. Roos, "Assessing performance of orthology detection strategies applied to eukaryotic genomes," *PloS one*, vol. 2, p. e383, 2007.

[70] M. Remm, C. E. Storm, and E. L. Sonnhammer, "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons," *Journal of Molecular Biology*, vol. 314, pp. 1041–1052, 2001.

[71] T. A. G. Iniative, "Analysis of the genome sequence of the fowering plant arabidopsis thaliana," *Nature*, vol. 408, pp. 796–815, 2000.

[72] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, and A. Bateman, "The pfam protein families database," *Nucleic Acids Research*, vol. 38, pp. D211–D222, 2010.

[73] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, *et al.*, "ChEMBL: a large-scale bioactivity database for drug discovery," *Nucleic Acids Research*, vol. 40, pp. D1100–D1107, 2012.

[74] T. U. Consortium, "Update on activities at the universal protein resource (uniprot) in 2013," *Nucleic Acids Research*, vol. 41, pp. D43–D47, 2013.

[75] D. J. Eide, "The molecular biology of metal ion transport in Saccharomyces cerevisiae," *Annual Review of Nutrition*, vol. 18, pp. 441–469, 1998.

[76] C. W. Peterson, S. S. Narula, and I. M. Armitage, "3D solution structure of copper and silver-substituted yeast metallothioneins," *FEBS Letters*, vol. 379, pp. 85–93, 1996.

[77] L. C. Williamson, S. P. Ribrioux, A. H. Fitter, and H. O. Leyser, "Phosphate availability regulates root system architecture in Arabidopsis," *Plant Physiology*, vol. 126, pp. 875–882, 2001.

[78] D. P. Schachtman, R. J. Reid, and S. M. Ayling, "Phosphorus uptake by plants: from soil to cell," *Plant Physiology*, vol. 116, pp. 447–453, 1998.

[79] H. Shin, H.-S. Shin, G. R. Dewbre, and M. J. Harrison, "Phosphate transport in Arabidopsis: Pht1; 1 and Pht1; 4 play a major role in phosphate acquisition from both low-and high-phosphate environments," *The Plant Journal*, vol. 39, pp. 629–642, 2004.

[80] L. E. Williams, R. Lemoine, and N. Sauer, "Sugar transporters in higher plants–a diversity of roles and complex regulation," *Trends in Plant Science*, vol. 5, pp. 283–290, 2000.

[81] M. C. Frith, N. F. Saunders, B. Kobe, and T. L. Bailey, "Discovering sequence motifs with arbitrary insertions and deletions," *PLoS Computational Biology*, vol. 4, p. e1000071, 2008.

[82] K. Khafizov, R. Staritzbichler, M. Stamm, and L. R. Forrest, "A study of the evolution of inverted-topology repeats from LeuT-fold transporters using AlignMe," *Biochemistry*, vol. 49, pp. 10702–10713, 2010.

[83] I. T. Paulsen, M. K. Sliwinski, and M. H. Saier, "Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities," *Journal of Molecular Biology*, vol. 277, pp. 573–592, 1998.

[84] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, 1999.

[85] F. Bertucci, S. Salas, S. Eysteries, V. Nasser, P. Finetti, C. Ginestier, E. Charafe-Jauffret, B. Loriod, L. Bachelart, J. Montfort, *et al.*, "Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters," *Oncogene*, vol. 23, pp. 1377–1391, 2004.

[86] K. Birkenkamp-Demtroder, L. L. Christensen, S. H. Olesen, C. M. Frederiksen, P. Laiho, L. A. Aaltonen, S. Laurberg, F. B. Sørensen, R. Hagemann, and T. F. Ørntoft, "Gene expression in colorectal cancer," *Cancer Research*, vol. 62, pp. 4352–4363, 2002.

[87] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, *et al.*, "NCBI GEO: archive for functional genomics data sets—update," *Nucleic Acids Research*, vol. 41, pp. D991–D995, 2013.

[88] X. Lu, Y. Li, and X. Zhang, "A simple strategy for detecting outlier samples in microarray data," in *Control, Automation, Robotics and Vision Conference, 2004. ICARCV 2004 8th*, vol. 2, pp. 1331–1335, IEEE, 2004.

[89] J. P. Mpindi, H. Sara, S. Haapa-Paananen, S. Kilpinen, T. Pisto, E. Bucher, K. Ojala, K. Iljin, P. Vainio, M. Björkman, *et al.*, "GTI: a novel algorithm for identifying

outlier gene expression profiles from integrated microarray datasets," *PloS One*, vol. 6, p. e17259, 2011.

[90] I. Pawlikowska, G. Wu, M. Edmonson, Z. Liu, T. Gruber, J. Zhang, and S. Pounds, "The most informative spacing test effectively discovers biologically relevant outliers or multiple modes in expression," *Bioinformatics*, vol. 30, pp. 1400–1408, 2014.

[91] F. R. Hampel, "The influence curve and its role in robust estimation," *Journal of the American Statistical Association*, vol. 69, pp. 383–393, 1974.

[92] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, "GOSemSim: an R package for measuring semantic similarity among GO terms and gene products," *Bioinformatics*, vol. 26, pp. 976–978, 2010.

[93] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, vol. 96, pp. 6745–6750, 1999.

[94] B. Shokirov, "Test for Normality of the Gene Expression Data," in *Statistical Methods for Microarray Data Analysis*, pp. 193–208, Springer, 2013.

[95] W. Pan, J. Lin, C. T. Le, *et al.*, "Model-based cluster analysis of microarray gene-expression data," *Genome Biology*, vol. 3, pp. 1–0009, 2002.

[96] M.-L. T. Lee, F. C. Kuo, G. Whitmore, and J. Sklar, "Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations," *Proceedings of the National Academy of Sciences*, vol. 97, pp. 9834–9839, 2000.

[97] J. Berkson, "Minimum chi-square, not maximum likelihood!," *The Annals of Statistics*, vol. 8, pp. 457–487, 1980.

[98] R. Akulenko and V. Helms, "DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples," *Human Molecular Genetics*, pp. 3016–3022, 2013.

[99] B. Iglewicz and D. C. Hoaglin, *How to detect and handle outliers*, vol. 16. Asq Press, 1993.

[100] M. Hubert and E. Vandervieren, "An adjusted boxplot for skewed distributions," *Computational Statistics & Data Analysis*, vol. 52, pp. 5186–5201, 2008.

[101] K. Carling, "Resistant outlier rules and the non-Gaussian case," *Computational Statistics & Data Analysis*, vol. 33, pp. 249–258, 2000.

[102] A. Barghash and V. Helms, "Transferring functional annotations of membrane transporters on the basis of sequence similarity and sequence motifs," *BMC Bioinformatics*, vol. 14, p. 343, 2013.

[103] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA: a Cancer Journal for Clinicians*, vol. 61, pp. 69–90, 2011.

[104] K. Block and Y. Gorin, "Aiding and abetting roles of NOX oxidases in cellular transformation," *Nature Reviews Cancer*, vol. 12, pp. 627–637, 2012.

[105] J. Christiansen, A. M. Kolte, F. C. Nielsen, *et al.*, "IGF2 mRNA-binding protein 2: biological function and putative role in type 2 diabetes," *Journal of Molecular Endocrinology*, vol. 43, pp. 187–195, 2009.

[106] J.-Y. Zhang, E. K. Chan, X.-X. Peng, and E. M. Tan, "A novel cytoplasmic protein with RNA-binding motifs is an autoantigen in human hepatocellular carcinoma," *The Journal of Experimental Medicine*, vol. 189, pp. 1101–1110, 1999.

[107] T. Gutschner, M. Hämmerle, N. Pazaitis, N. Bley, E. Fiskin, H. Uckelmann, A. Heim, M. Groβ, N. Hofmann, R. Geffers, *et al.*, "Insulin-like growth factor 2 mRNA-binding protein 1 (IGF2BP1) is an important protumorigenic factor in hepatocellular carcinoma," *Hepatology*, vol. 59, pp. 1900–1911, 2014.

[108] Y.-M. Jeng, C.-C. Chang, F.-C. Hu, H.-Y. E. Chou, H.-L. Kao, T.-H. Wang, and H.-C. Hsu, "RNA-binding protein insulin-like growth factor II mRNA-binding protein 3 expression promotes tumor invasion and predicts early recurrence and poor prognosis in hepatocellular carcinoma," *Hepatology*, vol. 48, pp. 1118–1127, 2008.

[109] T. R. Vargas, S. Boudoukha, A. Simon, M. Souidi, S. Cuvellier, G. Pinna, and A. Polesskaya, "Post-transcriptional regulation of cyclins D1, D3 and G1 and proliferation of human cancer cells depend on IMP-3 nuclear localization," *Oncogene*, vol. 33, pp. 2866–2875, 2014.

[110] I. Elcheva, R. Tarapore, N. Bhatia, and V. Spiegelman, "Overexpression of mRNA-binding protein CRD-BP in malignant melanomas," *Oncogene*, vol. 27, pp. 5069–5074, 2008.

[111] S. Laggai, S. M. Kessler, S. Boettcher, V. Lebrun, K. Gemperlein, E. Lederer, I. A. Leclercq, R. Mueller, R. W. Hartmann, J. Haybaeck, *et al.*, "The IGF2 mRNA binding protein p62/IGF2BP2-2 induces fatty acid elongation as a critical feature of steatosis," *Journal of Lipid Research*, vol. 55, pp. 1087–1097, 2014.

[112] E. Tybl, F.-D. Shi, S. M. Kessler, S. Tierling, J. Walter, R. M. Bohle, S. Wieland, J. Zhang, E. M. Tan, and A. K. Kiemer, "Overexpression of the IGF2-mRNA binding protein p62 in transgenic mice induces a steatotic phenotype," *Journal of Hepatology*, vol. 54, pp. 994–1001, 2011.

[113] Y. Simon, S. M. Kessler, R. M. Bohle, J. Haybaeck, A. K. Kiemer, E. Dilly, and C. Guth, "The insulin-like growth factor 2 (igf2) mrna-binding protein p62/igf2bp2-2 as a promoter of nafld and HCC?," *Gut*, pp. 861–863, 2013.

[114] A. Gabory, H. Jammes, and L. Dandolo, "The H19 locus: Role of an imprinted non-coding RNA in growth and development," *Bioessays*, vol. 32, pp. 473–480, 2010.

[115] I. O. Fawzy, M. T. Hamza, K. A. Hosny, G. Esmat, H. M. El Tayebi, and A. I. Abdelaziz, "miR-1275: a single microRNA that targets the three IGF2-mRNA-binding proteins hindering tumor growth in hepatocellular carcinoma," *FEBS Letters*, vol. 389, pp. 2257–2265, 2015.

[116] S. M. Kessler, Y. Simon, K. Gemperlein, K. Gianmoena, C. Cadenas, V. Zimmer, J. Pokorny, A. Barghash, V. Helms, N. van Rooijen, *et al.*, "Fatty acid elongation in non-alcoholic steatohepatitis and hepatocellular carcinoma," *International Journal of Molecular Sciences*, vol. 15, pp. 5762–5773, 2014.

[117] E. J. Park, J. H. Lee, G.-Y. Yu, G. He, S. R. Ali, R. G. Holzer, C. H. Österreicher, H. Takahashi, and M. Karin, "Dietary and genetic obesity promote liver inflammation and tumorigenesis by enhancing IL-6 and TNF expression," *Cell*, vol. 140, pp. 197–208, 2010.

[118] S. M. Kessler, S. Laggai, A. Barghash, V. Helms, and A. K. Kiemer, "Lipid metabolism signatures in nash-associated HCC—letter," *Cancer Research*, vol. 74, pp. 2903–2904, 2014.

[119] Y. Simon, S. M. Kessler, K. Gemperlein, R. M. Bohle, R. Müller, J. Haybaeck, and A. K. Kiemer, "Elevated free cholesterol in a p62 overexpression model of non-alcoholic steatohepatitis," *World Journal of Gastroenterology: WJG*, vol. 20, p. 17839, 2014.

[120] R. Fürst, C. Brueckl, W. M. Kuebler, S. Zahler, F. Krötz, A. Görlach, A. M. Vollmar, and A. K. Kiemer, "Atrial natriuretic peptide induces mitogen-activated protein kinase phosphatase-1 in human endothelial cells via Rac1 and NAD (P) H oxidase/Nox2-activation," *Circulation Research*, vol. 96, pp. 43–53, 2005.

[121] J. Fridlyand and P. Dimitrov, "aCGH: classes and functions for array comparative genomic hybridization data," *R package version*, vol. 1, 2010.

[122] J. Haybaeck, N. Zeller, M. J. Wolf, A. Weber, U. Wagner, M. O. Kurrer, J. Bremer, G. Iezzi, R. Graf, P.-A. Clavien, *et al.*, "A lymphotoxin-driven pathway to hepatocellular carcinoma," *Cancer Cell*, vol. 16, pp. 295–308, 2009.

[123] H. Fröhlich, N. Speer, A. Poustka, and T. Beißbarth, "GOSim–an R-package for computation of information theoretic GO similarities between terms and gene products," *BMC Bioinformatics*, vol. 8, p. 166, 2007.

[124] V. E. Seshan and A. Olshen. R package version 1.32.0.

[125] S. Roessler, H.-L. Jia, A. Budhu, M. Forgues, Q.-H. Ye, J.-S. Lee, S. S. Thorgeirsson, Z. Sun, Z.-Y. Tang, L.-X. Qin, *et al.*, "A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients," *Cancer Research*, vol. 70, pp. 10202–10212, 2010.

[126] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, 1999.

[127] R. Dreos, G. Ambrosini, R. C. Périer, and P. Bucher, "The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools," *Nucleic Acids Research*, vol. 43, pp. D92–D96, 2015.

[128] S. Schievenbusch, E. Sauer, H.-M. Curth, S. Schulte, M. Demir, U. Toex, T. Goeser, and D. Nierhoff, "Neighbor of Punc E 11: expression pattern of the new hepatic stem/progenitor cell marker during murine liver development," *Stem Cells and Development*, vol. 21, pp. 2656–2666, 2012.

[129] J. Lee and S. Thorgeirsson, "Comparative and integrative functional genomics of HCC," *Oncogene*, vol. 25, pp. 3801–3809, 2006.

[130] S. Loeppen, D. Schneider, F. Gaunitz, R. Gebhardt, R. Kurek, A. Buchmann, and M. Schwarz, "Overexpression of glutamine synthetase is associated with β-catenin-mutations in mouse liver tumors during promotion of hepatocarcinogenesis by phenobarbital," *Cancer Research*, vol. 62, pp. 5685–5688, 2002.

[131] O. Teufelhofer, W. Parzefall, E. Kainzbauer, F. Ferk, C. Freiler, S. Knasmüller, L. Elbling, R. Thurman, and R. Schulte-Hermann, "Superoxide generation from Kupffer cells contributes to hepatocarcinogenesis: studies on NADPH oxidase knockout mice," *Carcinogenesis*, vol. 26, pp. 319–329, 2005.

[132] Y. Wang, L. Zhao, C. Smas, and H. S. Sul, "Pref-1 interacts with fibronectin to inhibit adipocyte differentiation," *Molecular and Cellular Biology*, vol. 30, pp. 3480–3492, 2010.

[133] W. Liu, Z. Li, W. Xu, Q. Wang, and S. Yang, "Humoral Autoimmune Response to IGF2 mRNA-Binding Protein (IMP2/p62) and Its Tissue-Specific Expression in Colon Cancer," *Scandinavian Journal of Immunology*, vol. 77, pp. 255–260, 2013.

[134] W. N. Rom, J. D. Goldberg, D. Addrizzo-Harris, H. N. Watson, M. Khilkin, A. K. Greenberg, D. P. Naidich, B. Crawford, E. Eylers, D. Liu, *et al.*, "Identification of an autoantibody panel to separate lung cancer from smokers and nonsmokers," *BMC Cancer*, vol. 10, p. 234, 2010.

[135] N. Dai, L. Zhao, D. Wrighting, D. Krämer, A. Majithia, Y. Wang, V. Cracan, D. Borges-Rivera, V. K. Mootha, M. Nahrendorf, *et al.*, "IGF2BP2/IMP2-Deficient Mice Resist Obesity through Enhanced Translation of Ucp1 mRNA and Other mRNAs Encoding Mitochondrial Proteins," *Cell Metabolism*, vol. 21, pp. 609–621, 2015.

[136] M. Lu, R. M. Nakamura, E. D. Dent, J.-Y. Zhang, F. C. Nielsen, J. Christiansen, E. K. Chan, and E. M. Tan, "Aberrant expression of fetal RNA-binding protein p62 in liver cancer and liver cirrhosis," *The American Journal of Pathology*, vol. 159, pp. 945–953, 2001.

[137] Y. Hoshida, S. Toffanin, A. Lachenmayer, A. Villanueva, B. Minguez, and J. M. Llovet, "Molecular classification and novel targets in hepatocellular carcinoma: recent advancements," in *Seminars in Liver Disease*, vol. 30, p. 35, NIH Public Access, 2010.

[138] B. Mínguez, Y. Hoshida, A. Villanueva, S. Toffanin, L. Cabellos, S. Thung, J. Mandeli, D. Sia, C. April, J.-B. Fan, *et al.*, "Gene-expression signature of vascular invasion in hepatocellular carcinoma," *Journal of Hepatology*, vol. 55, pp. 1325–1331, 2011.

[139] T. Yamashita, J. Ji, A. Budhu, M. Forgues, W. Yang, H.-Y. Wang, H. Jia, Q. Ye, L.-X. Qin, E. Wauthier, *et al.*, "EpCAM-positive hepatocellular carcinoma cells are tumor-initiating cells with stem/progenitor cell features," *Gastroenterology*, vol. 136, pp. 1012–1024, 2009.

[140] Z. Guo, L.-Q. Li, J.-H. Jiang, C. Ou, L.-X. Zeng, and B.-D. Xiang, "Cancer stem cell markers correlate with early recurrence and survival in hepatocellular carcinoma," *World Journal of Gastroenterology: WJG*, vol. 20, p. 2098, 2014.

[141] S. Boyault, D. S. Rickman, A. De Reynies, C. Balabaud, S. Rebouissou, E. Jeannot, A. Hérault, J. Saric, J. Belghiti, D. Franco, *et al.*, "Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets," *Hepatology*, vol. 45, pp. 42–52, 2007.

[142] S. Cairo, C. Armengol, A. De Reyniès, Y. Wei, E. Thomas, C.-A. Renard, A. Goga, A. Balakrishnan, M. Semeraro, L. Gresh, *et al.*, "Hepatic stem-like phenotype and interplay of Wnt/β-catenin and Myc signaling in aggressive childhood liver cancer," *Cancer Cell*, vol. 14, pp. 471–484, 2008.

[143] Z.-h. Jin, R.-j. Yang, B. Dong, and B.-c. Xing, "Progenitor gene DLK1 might be an independent prognostic factor of liver cancer," *Expert Opinion on Biological Therapy*, vol. 8, pp. 371–377, 2008.

[144] C. Floridon, C. H. Jensen, P. Thorsen, O. Nielsen, L. Sunde, J. G. Westergaard, S. G. Thomsen, and B. Teisner, "Does fetal antigen 1 (FA1) identify cells with regenerative, endocrine and neuroendocrine potentials? A study of FA1 in embryonic, fetal, and placental tissue and in maternal circulation," *Differentiation*, vol. 66, pp. 49–59, 2000.

[145] H. Yanai, K. Nakamura, S. Hijioka, A. Kamei, T. Ikari, Y. Ishikawa, E. Shinozaki, N. Mizunuma, K. Hatake, and A. Miyajima, "Dlk-1, a cell surface antigen on foetal hepatic stem/progenitor cells, is expressed in hepatocellular, colon, pancreas and

breast carcinomas at a high frequency," *Journal of biochemistry*, vol. 148, pp. 85–92, 2010.

[146] F. A. Falix, D. C. Aronson, W. H. Lamers, J. K. Hiralall, and J. Seppen, "DLK1, a serum marker for hepatoblastoma in young infants," *Pediatric Blood & Cancer*, vol. 59, pp. 743–745, 2012.

[147] M. Chacon, M. Miranda, C. Jensen, J. Fernandez-Real, N. Vilarrasa, C. Gutierrez, S. Näf, J. Gomez, and J. Vendrell, "Human serum levels of fetal antigen 1 (FA1/Dlk1) increase with obesity, are negatively associated with insulin sensitivity and modulate inflammation in vitro," *International Journal of Obesity*, vol. 32, pp. 1122–1129, 2008.

[148] S. Laggai, Y. Simon, T. Ranssweiler, A. K. Kiemer, and S. M. Kessler, "Rapid chromatographic method to decipher distinct alterations in lipid classes in NAFLD/NASH," *World Journal of Hepatology*, vol. 5, p. 558, 2013.

[149] N.-L. Zhu, K. Asahina, J. Wang, A. Ueno, R. Lazaro, Y. Miyaoka, A. Miyajima, and H. Tsukamoto, "Hepatic stellate cell-derived delta-like homolog 1 (DLK1) protein in liver regeneration," *Journal of Biological Chemistry*, vol. 287, pp. 10355–10367, 2012.

[150] T. Yamashita, M. Forgues, W. Wang, J. W. Kim, Q. Ye, H. Jia, A. Budhu, K. A. Zanetti, Y. Chen, L.-X. Qin, *et al.*, "EpCAM and α-fetoprotein expression defines novel prognostic subtypes of hepatocellular carcinoma," *Cancer Research*, vol. 68, pp. 1451–1461, 2008.

[151] H. Okamoto, K. Yasui, C. Zhao, S. Arii, and J. Inazawa, "PTK2 and EIF3S3 genes may be amplification targets at 8q23–q24 and are associated with large hepatocellular carcinomas," *Hepatology*, vol. 38, pp. 1242–1249, 2003.

[152] K. Heselmeyer, M. Macville, E. Schröck, H. Blegen, A.-C. Hellström, K. Shah, G. Auer, and T. Ried, "Advanced-stage cervical carcinomas are defined by a recurrent pattern of chromosomal aberrations revealing high genetic instability and a consistent gain of chromosome arm 3q," *Genes Chromosomes and Cancer*, vol. 19, pp. 233–240, 1997.

[153] S.-M. Ahn, S. J. Jang, J. H. Shim, D. Kim, S.-M. Hong, C. O. Sung, D. Baek, F. Haq, A. A. Ansari, S. Y. Lee, *et al.*, "Genomic portrait of resectable hepatocellular carcinomas: implications of RB1 and FGF19 aberrations for patient stratification," *Hepatology*, vol. 60, pp. 1972–1982, 2014.

[154] S. Sinha, R. K. Singh, N. Bhattacharya, N. Mukherjee, S. Ghosh, N. Alam, A. Roy, S. Roychoudhury, and C. K. Panda, "Frequent alterations of LOH11CR2A, PIG8 and CHEK1 genes at chromosomal 11q24. 1-24.2 region in breast carcinoma: clinical and prognostic implications," *Molecular Oncology*, vol. 5, pp. 454–464, 2011.

[155] T. K. Lee, R. T. Poon, A. P. Yuen, K. Man, Z. F. Yang, X. Y. Guan, and S. T. Fan, "Rac activation is associated with hepatocellular carcinoma metastasis by up-regulation of vascular endothelial growth factor expression," *Clinical Cancer Research*, vol. 12, pp. 5082–5089, 2006.

[156] F. V. Rassool, T. J. Gaymes, N. Omidvar, N. Brady, S. Beurlet, M. Pla, M. Reboul, N. Lea, C. Chomienne, N. S. Thomas, *et al.*, "Reactive oxygen species, DNA damage, and error-prone repair: a model for genomic instability with progression in myeloid leukemia?," *Cancer Research*, vol. 67, pp. 8762–8771, 2007.

[157] M. Ogrunc, R. Di Micco, M. Liontos, L. Bombardelli, M. Mione, M. Fumagalli, V. Gorgoulis, and F. d. di Fagagna, "Oncogene-induced reactive oxygen species fuel hyperproliferation and DNA damage response activation," *Cell Death & Differentiation*, vol. 21, pp. 998–1012, 2014.

[158] C. E. Pasi, A. Dereli-Öz, S. Negrini, M. Friedli, G. Fragola, A. Lombardo, G. Van Houwe, L. Naldini, S. Casola, G. Testa, *et al.*, "Genomic instability in induced stem cells," *Cell Death & Differentiation*, vol. 18, pp. 745–753, 2011.

[159] K. Myant, A. Scopelliti, S. Haque, M. Vidal, O. Sansom, and J. Cordero, "Rac1 drives intestinal stem cell proliferation and regeneration," *Cell Cycle*, vol. 12, pp. 2973–2977, 2013.

[160] L. Gonzalez-Santiago, Y. Suárez, N. Zarich, M. Munoz-Alonso, A. Cuadrado, T. Martinez, L. Goya, A. Iradi, G. Saez-Tormo, J. Maier, *et al.*, "Aplidin® induces jnk-dependent apoptosis in human breast cancer cells via alteration of glutathione homeostasis, rac1 gtpase activation, and mkp-1 phosphatase downregulation," *Cell Death & Differentiation*, vol. 13, pp. 1968–1981, 2006.

[161] S. Akunuru, J. Palumbo, Q. J. Zhai, and Y. Zheng, "Rac1 targeting suppresses human non-small cell lung adenocarcinoma cancer stem cell activity," *PloS One*, vol. 6, p. e16951, 2011.

[162] P. Wang, L. Chen, J. Zhang, H. Chen, J. Fan, K. Wang, J. Luo, Z. Chen, Z. Meng, and L. Liu, "Methylation-mediated silencing of the miR-124 genes facilitates pancreatic cancer progression and metastasis by targeting Rac1," *Oncogene*, vol. 33, pp. 514–524, 2014.

[163] S. M. Wang, L. L. P. Ooi, and K. M. Hui, "Upregulation of Rac GTPase-activating protein 1 is significantly associated with the early recurrence of human hepatocellular carcinoma," *Clinical Cancer Research*, vol. 17, pp. 6040–6051, 2011.

[164] W. Liu, Y. Li, B. Wang, L. Dai, W. Qian, and J.-Y. Zhang, "Autoimmune Response to IGF2 mRNA-Binding Protein 2 (IMP2/p62) in Breast Cancer," *Scandinavian Journal of Immunology*, vol. 81, pp. 502–507, 2015.

[165] J. Zhang, W. Zhu, H. Imai, K. Kiyosawa, E. Chan, and E. Tan, "De-novo humoral immune responses to cancer-associated autoantigens during transition from chronic liver disease to hepatocellular carcinoma," *Clinical & Experimental Immunology*, vol. 125, pp. 3–9, 2001.

[166] C. Clarke, S. F. Madden, P. Doolan, S. T. Aherne, H. Joyce, L. O'Driscoll, W. M. Gallagher, B. T. Hennessy, M. Moriarty, J. Crown, *et al.*, "Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis," *Carcinogenesis*, vol. 34, pp. 2300–2308, 2013.

[167] C. G. A. Network *et al.*, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, pp. 61–70, 2012.

[168] L. A. Carey, C. M. Perou, C. A. Livasy, L. G. Dressler, D. Cowan, K. Conway, G. Karaca, M. A. Troester, C. K. Tse, S. Edmiston, *et al.*, "Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study," *Jama*, vol. 295, pp. 2492–2502, 2006.

[169] P. Farmer, H. Bonnefoi, V. Becette, M. Tubiana-Hulin, P. Fumoleau, D. Larsimont, G. MacGrogan, J. Bergh, D. Cameron, D. Goldstein, *et al.*, "Identification of molecular apocrine breast tumours by microarray analysis," *Breast Cancer Research*, vol. 7, pp. P2–11, 2005.

[170] A. L. Richardson, Z. C. Wang, A. De Nicolo, X. Lu, M. Brown, A. Miron, X. Liao, J. D. Iglehart, D. M. Livingston, and S. Ganesan, "X chromosomal abnormalities in basal-like human breast cancer," *Cancer Cell*, vol. 9, pp. 121–132, 2006.

[171] M. Lunova, C. Goehring, D. Kuscuoglu, K. Mueller, Y. Chen, P. Walther, J.-C. Deschemin, S. Vaulont, J. Haybaeck, C. Lackner, *et al.*, "Hepcidin knockout mice fed with iron-rich diet develop chronic liver injury and liver fibrosis due to lysosomal iron overload," *Journal of Hepatology*, vol. 61, pp. 633–641, 2014.

[172] D. D. Harrison-Findik, D. Schafer, E. Klein, N. A. Timchenko, H. Kulaksiz, D. Clemens, E. Fein, B. Andriopoulos, K. Pantopoulos, and J. Gollan, "Alcohol metabolism-mediated oxidative stress down-regulates hepcidin transcription and leads to increased duodenal iron transporter expression," *Journal of Biological Chemistry*, vol. 281, pp. 22974–22982, 2006.

[173] P. Holmström, M. Gåfvels, L. C. Eriksson, V. Dzikaite, R. Hultcrantz, G. Eggertsen, and P. Stål, "Expression of iron regulatory genes in a rat model of hepatocellular carcinoma," *Liver International*, vol. 26, pp. 976–985, 2006.

[174] P. Youn, S. Kim, J. H. Ahn, Y. Kim, J.-D. Park, and D.-Y. Ryu, "Regulation of iron metabolism-related genes in diethylnitrosamine-induced mouse liver tumors," *Toxicology letters*, vol. 184, pp. 151–158, 2009.

[175] H. Kijima, T. Sawada, N. Tomosugi, and K. Kubota, "Expression of hepcidin mRNA is uniformly suppressed in hepatocellular carcinoma," *BMC Cancer*, vol. 8, p. 167, 2008.

[176] D. Girelli, M. Pasino, J. B. Goodnough, E. Nemeth, M. Guido, A. Castagna, F. Busti, N. Campostrini, N. Martinelli, I. Vantini, *et al.*, "Reduced serum hepcidin levels in patients with chronic hepatitis C," *Journal of Hepatology*, vol. 51, pp. 845–852, 2009.

[177] K. Miura, K. Taura, Y. Kodama, B. Schnabl, and D. A. Brenner, "Hepatitis C virus–induced oxidative stress suppresses hepcidin expression through increased histone deacetylase activity," *Hepatology*, vol. 48, pp. 1420–1429, 2008.

[178] O. Weizer-Stern, K. Adamsky, O. Margalit, O. Ashur-Fabian, D. Givol, N. Amariglio, and G. Rechavi, "Hepcidin, a key regulator of iron metabolism, is transcriptionally activated by p53," *British Journal of Haematology*, vol. 138, pp. 253–262, 2007.

[179] J.-C. Nault and J. Zucman-Rossi, "Genetics of hepatocellular carcinoma: the next generation," *Journal of Hepatology*, vol. 60, pp. 224–226, 2014.

[180] K. Muir, A. Hazim, Y. He, M. Peyressatre, D.-Y. Kim, X. Song, and L. Beretta, "Proteomic and lipidomic signatures of lipid metabolism in NASH-associated hepatocellular carcinoma," *Cancer Research*, vol. 73, pp. 4722–4731, 2013.

[181] T. Mori, H. Kondo, T. Hase, and T. Murase, "Dietary phospholipids ameliorate fructose-induced hepatic lipid and metabolic abnormalities in rats," *The Journal of Nutrition*, vol. 141, pp. 2003–2009, 2011.

[182] M. Hoekstra, R. J. van der Sluis, J. Kuiper, and T. J. Van Berkel, "Nonalcoholic fatty liver disease is associated with an altered hepatocyte microRNA profile in LDL receptor knockout mice," *The Journal of Nutritional Biochemistry*, vol. 23, pp. 622–628, 2012.

[183] T. Matsuzaka, A. Atsumi, R. Matsumori, T. Nie, H. Shinozaki, N. Suzuki-Kemuriyama, M. Kuba, Y. Nakagawa, K. Ishii, M. Shimada, *et al.*, "Elovl6 promotes nonalcoholic steatohepatitis," *Hepatology*, vol. 56, pp. 2199–2208, 2012.

[184] Y. Li, Z.-Y. Tang, and J.-X. Hou, "Hepatocellular carcinoma: insight from animal models," *Nature Reviews Gastroenterology and Hepatology*, vol. 9, pp. 32–43, 2012.

[185] G. Kanuri and I. Bergheim, "In vitro and in vivo models of non-alcoholic fatty liver disease (NAFLD)," *International Journal of Molecular Sciences*, vol. 14, pp. 11963–11980, 2013.

[186] L. A. Adams, P. Angulo, and K. D. Lindor, "Nonalcoholic fatty liver disease," *Canadian Medical Association Journal*, vol. 172, pp. 899–905, 2005.

[187] J. D. Browning, L. S. Szczepaniak, R. Dobbins, J. D. Horton, J. C. Cohen, S. M. Grundy, and H. H. Hobbs, "Prevalence of hepatic steatosis in an urban population in the United States: impact of ethnicity," *Hepatology*, vol. 40, pp. 1387–1395, 2004.

[188] N. M. W. de Alwis and C. P. Day, "Non-alcoholic fatty liver disease: the mist gradually clears," *Journal of Hepatology*, vol. 48, pp. S104–S112, 2008.

[189] L. A. Adams, O. R. Waters, M. W. Knuiman, R. R. Elliott, and J. K. Olynyk, "NAFLD as a risk factor for the development of diabetes and the metabolic syndrome: an eleven-year follow-up study," *The American Journal of Gastroenterology*, vol. 104, pp. 861–867, 2009.

[190] M. S. Ascha, I. A. Hanouneh, R. Lopez, T. A.-R. Tamimi, A. F. Feldstein, and N. N. Zein, "The incidence and risk factors of hepatocellular carcinoma in patients with nonalcoholic steatohepatitis," *Hepatology*, vol. 51, pp. 1972–1978, 2010.

[191] E. Fabbrini, S. Sullivan, and S. Klein, "Obesity and nonalcoholic fatty liver disease: biochemical, metabolic, and clinical implications," *Hepatology*, vol. 51, pp. 679–689, 2010.

[192] P. Angulo, "Nonalcoholic fatty liver disease," *New England Journal of Medicine*, vol. 346, pp. 1221–1231, 2002.

[193] C. P. Day, "Genetic and environmental susceptibility to non-alcoholic fatty liver disease," *Digestive Diseases*, vol. 28, pp. 255–260, 2010.

[194] H. Malhi and G. J. Gores, "Molecular mechanisms of lipotoxicity in nonalcoholic fatty liver disease," in *Seminars in Liver Disease*, vol. 28, p. 360, NIH Public Access, 2008.

[195] J. C. Cohen, J. D. Horton, and H. H. Hobbs, "Human fatty liver disease: old questions and new insights," *Science*, vol. 332, pp. 1519–1523, 2011.

[196] N. Alkhouri, L. J. Dixon, and A. E. Feldstein, "Lipotoxicity in nonalcoholic fatty liver disease: not all lipids are created equal," *Expert Review of Gastroenterology & Hepatology*, vol. 3, pp. 445–451, 2009.

[197] A. Takaki, D. Kawai, and K. Yamamoto, "Multiple hits, including oxidative stress, as pathogenesis and treatment target in non-alcoholic steatohepatitis (NASH)," *International Journal of Molecular Sciences*, vol. 14, pp. 20704–20728, 2013.

[198] P. Puri, R. A. Baillie, M. M. Wiest, F. Mirshahi, J. Choudhury, O. Cheung, C. Sargeant, M. J. Contos, and A. J. Sanyal, "A lipidomic analysis of nonalcoholic fatty liver disease," *Hepatology*, vol. 46, pp. 1081–1090, 2007.

[199] P. Puri, M. M. Wiest, O. Cheung, F. Mirshahi, C. Sargeant, H.-K. Min, M. J. Contos, R. K. Sterling, M. Fuchs, H. Zhou, *et al.*, "The plasma lipidomic signature of nonalcoholic steatohepatitis," *Hepatology*, vol. 50, pp. 1827–1838, 2009.

[200] K. H. Kim, H.-J. Shin, K. Kim, H. M. Choi, S. H. Rhee, H.-B. Moon, H. H. Kim, U. S. Yang, D.-Y. Yu, and J. Cheong, "Hepatitis B virus X protein induces hepatic steatosis via transcriptional activation of SREBP1 and PPARγ," *Gastroenterology*, vol. 132, pp. 1955–1967, 2007.

[201] H. Miyoshi, K. Moriya, T. Tsutsumi, S. Shinzawa, H. Fujie, Y. Shintani, H. Fujinaga, K. Goto, T. Todoroki, T. Suzuki, *et al.*, "Pathogenesis of lipid metabolism disorder in hepatitis C: polyunsaturated fatty acids counteract lipid alterations induced by the core protein," *Journal of Hepatology*, vol. 54, pp. 432–438, 2011.

[202] K. Moriya, T. Todoroki, T. Tsutsumi, H. Fujie, Y. Shintani, H. Miyoshi, K. Ishibashi, T. Takayama, M. Makuuchi, K. Watanabe, *et al.*, "Increase in the concentration of carbon 18 monounsaturated fatty acids in the liver with hepatitis C: analysis in transgenic mice and humans," *Biochemical and Biophysical Research Communications*, vol. 281, pp. 1207–1212, 2001.

[203] A. Yamaguchi, S. Tazuma, T. Nishioka, W. Ohishi, H. Hyogo, S. Nomura, and K. Chayama, "Hepatitis C virus core protein modulates fatty acid metabolism and thereby causes lipid accumulation in the liver," *Digestive Diseases and Sciences*, vol. 50, pp. 1361–1371, 2005.

[204] C. Wood, N. Habib, K. Apostolov, A. Thompson, W. Barker, M. Hershman, and L. Blumgart, "Reduction in the stearic to oleic acid ratio in human malignant liver neoplasms.," *European Journal of Surgical Oncology: the Journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology*, vol. 11, pp. 347–348, 1985.

[205] T. Matsuzaka, H. Shimano, N. Yahagi, T. Kato, A. Atsumi, T. Yamamoto, N. Inoue, M. Ishikawa, S. Okada, N. Ishigaki, *et al.*, "Crucial role of a long-chain fatty acid elongase, Elovl6, in obesity-induced insulin resistance," *Nature Medicine*, vol. 13, pp. 1193–1202, 2007.

[206] M. Keller, A. L. Gerbes, S. Kulhanek-Heinze, T. Gerwig, U. Grutzner, N. van Rooijen, A. M. Vollmar, and A. Kiemen, "Hepatocyte cytoskeleton during ischemia and reperfusion-influence of ANP-mediated p38 MAPK activation," *World Journal of Gastroenterology*, vol. 11, p. 7418, 2005.

[207] H. B. Bode, M. W. Ring, G. Schwär, R. M. Kroppenstedt, D. Kaiser, and R. Müller, "3-Hydroxy-3-methylglutaryl-coenzyme A (CoA) synthase is involved in biosynthesis of isovaleryl-CoA in the myxobacterium Myxococcus xanthus during fruiting body formation," *Journal of Bacteriology*, vol. 188, pp. 6524–6528, 2006.

[208] M. Ahrens, O. Ammerpohl, W. von Schönfels, J. Kolarova, S. Bens, T. Itzel, A. Teufel, A. Herrmann, M. Brosch, H. Hinrichsen, *et al.*, "DNA methylation analysis in nonalcoholic fatty liver disease suggests distinct disease-specific and remodeling signatures after bariatric surgery," *Cell Metabolism*, vol. 18, pp. 296–302, 2013.

[209] C. López-Vicario, A. González-Périz, B. Rius, E. Morán-Salvador, V. García-Alonso, J. J. Lozano, R. Bataller, M. Cofán, J. X. Kang, V. Arroyo, *et al.*, "Molecular interplay between $\Delta 5/\Delta 6$ desaturases and long-chain fatty acids in the pathogenesis of non-alcoholic steatohepatitis," *Gut*, vol. 63, pp. 344–355, 2014.

[210] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, pp. 207–210, 2002.

[211] K. Imajo, M. Yoneda, T. Kessoku, Y. Ogawa, S. Maeda, Y. Sumida, H. Hyogo, Y. Eguchi, K. Wada, and A. Nakajima, "Rodent models of nonalcoholic fatty liver disease/nonalcoholic steatohepatitis," *International Journal of Molecular Sciences*, vol. 14, pp. 21833–21857, 2013.

[212] V. Trak-Smayra, V. Paradis, J. Massart, S. Nasser, V. Jebara, and B. Fromenty, "Pathology of the liver in obese and diabetic ob/ob and db/db mice fed a standard or high-calorie diet," *International Journal of Experimental Pathology*, vol. 92, pp. 413–421, 2011.

[213] H. Xu, G. T. Barnes, Q. Yang, G. Tan, D. Yang, C. J. Chou, J. Sole, A. Nichols, J. S. Ross, L. A. Tartaglia, *et al.*, "Chronic inflammation in fat plays a crucial role in the development of obesity-related insulin resistance," *Journal of Clinical Investigation*, vol. 112, pp. 1821–1830, 2003.

[214] X. Prieur, C. Y. Mok, V. R. Velagapudi, V. Núñez, L. Fuentes, D. Montaner, K. Ishikawa, A. Camacho, N. Barbarroja, S. O'Rahilly, *et al.*, "Differential lipid partitioning between adipocytes and tissue macrophages modulates macrophage lipotoxicity and M2/M1 polarization in obese mice," *Diabetes*, vol. 60, pp. 797–809, 2011.

[215] T. Matsuzaka, H. Shimano, N. Yahagi, T. Yoshikawa, M. Amemiya-Kudo, A. H. Hasty, H. Okazaki, Y. Tamura, Y. Iizuka, K. Ohashi, *et al.*, "Cloning and characterization of a mammalian fatty acyl-CoA elongase as a lipogenic enzyme regulated by SREBPs," *Journal of Lipid Research*, vol. 43, pp. 911–920, 2002.

[216] A. Leroux, G. Ferrere, V. Godie, F. Cailleux, M.-L. Renoud, F. Gaudin, S. Naveau, S. Prévot, S. Makhzami, G. Perlemuter, *et al.*, "Toxic lipids stored by Kupffer cells correlates with their pro-inflammatory phenotype at an early stage of steatohepatitis," *Journal of Hepatology*, vol. 57, pp. 141–149, 2012.

[217] L. Vonghia, P. Michielsen, and S. Francque, "Immunological mechanisms in the pathophysiology of non-alcoholic steatohepatitis," *International Journal of Molecular Sciences*, vol. 14, pp. 19867–19890, 2013.

[218] R. Saito, T. Matsuzaka, T. Karasawa, M. Sekiya, N. Okada, M. Igarashi, R. Matsumori, K. Ishii, Y. Nakagawa, H. Iwasaki, *et al.*, "Macrophage Elovl6 deficiency ameliorates foam cell formation and reduces atherosclerosis in low-density lipoprotein receptor-deficient mice," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 31, pp. 1973–1979, 2011.

[219] W. E. Naugler, T. Sakurai, S. Kim, S. Maeda, K. Kim, A. M. Elsharkawy, and M. Karin, "Gender disparity in liver cancer due to sex differences in MyD88-dependent IL-6 production," *Science*, vol. 317, pp. 121–124, 2007.

[220] T. A. Braunbeck, S. J. Teh, S. M. Lester, and D. E. Hinton, "Ultrastructural alterations in liver of medaka (Oryzias latipes) exposed to diethylnitrosamine," *Toxicologic Pathology*, vol. 20, pp. 179–196, 1992.

[221] D. J. Laurén, S. J. Teh, and D. E. Hinton, "Cytotoxicity phase of diethylnitrosamine-induced hepatic neoplasia in medaka," *Cancer Research*, vol. 50, pp. 5504–5514, 1990.

[222] S. Abel, C. Smuts, C. De Villiers, and W. Gelderblom, "Changes in essential fatty acid patterns associated with normal liver regeneration and the progression of hepatocyte nodules in rat hepatocarcinogenesis," *Carcinogenesis*, vol. 22, pp. 795–804, 2001.

[223] R. A. Canuto, M. E. Biocca, G. Muzio, and M. U. Dianzani, "Fatty acid composition of phospholipids in mitochondria and microsomes during diethylnitrosamine carcinogenesis in rat liver," *Cell Biochemistry and Function*, vol. 7, pp. 11–19, 1989.

[224] K. Yoshimura, M. K. Mandal, M. Hara, H. Fujii, L. C. Chen, K. Tanabe, K. Hiraoka, and S. Takeda, "Real-time diagnosis of chemically induced hepatocellular carcinoma

using a novel mass spectrometry-based technique," *Analytical Biochemistry*, vol. 441, pp. 32–37, 2013.

[225] A. A. Mironov, J. W. Fickett, and M. S. Gelfand, "Frequent alternative splicing of human genes," *Genome Research*, vol. 9, pp. 1288–1293, 1999.

[226] G. Koscielny, V. Le Texier, C. Gopalakrishnan, V. Kumanduri, J.-J. Riethoven, F. Nardone, E. Stanley, C. Fallsehr, O. Hofmann, M. Kull, *et al.*, "ASTD: the alternative splicing and transcript diversity database," *Genomics*, vol. 93, pp. 213–220, 2009.

[227] T. W. Nilsen and B. R. Graveley, "Expansion of the eukaryotic proteome by alternative splicing," *Nature*, vol. 463, pp. 457–463, 2010.

[228] S. Das, S. Jena, and D. N. Levasseur, "Alternative splicing produces Nanog protein variants with different capacities for self-renewal and pluripotency in embryonic stem cells," *Journal of Biological Chemistry*, vol. 286, pp. 42690–42703, 2011.

[229] J. Kim, V. N. Noskov, X. Lu, A. Bergmann, X. Ren, T. Warth, P. Richardson, N. Kouprina, and L. Stubbs, "Discovery of a novel, paternally expressed ubiquitin-specific processing protease gene through comparative analysis of an imprinted region of mouse chromosome 7 and human chromosome 19q13. 4," *Genome Research*, vol. 10, pp. 1138–1147, 2000.

[230] J. Kim, A. Bergmann, S. Lucas, R. Stone, and L. Stubbs, "Lineage-specific imprinting and evolution of the zinc-finger gene ZIM2," *Genomics*, vol. 84, pp. 47–58, 2004.

[231] E. R. Gamazon and B. E. Stranger, "Genomics of alternative splicing: evolution, development and pathophysiology," *Human genetics*, vol. 133, pp. 679–687, 2014.

[232] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, pp. 57–63, 2009.

[233] A. Reyes, S. Anders, R. J. Weatheritt, T. J. Gibson, L. M. Steinmetz, and W. Huber, "Drift and conservation of differential exon usage across tissues in primate species," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 15377–15382, 2013.

[234] H.-L. Zhou, G. Luo, J. A. Wise, and H. Lou, "Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms," *Nucleic Acids Research*, vol. 42, pp. 701–713, 2014.

[235] S. Schwartz, E. Meshorer, and G. Ast, "Chromatin organization marks exon-intron structure," *Nature Structural & Molecular Biology*, vol. 16, pp. 990–995, 2009.

[236] Y. Wang, J. Liu, B. Huang, Y.-M. Xu, J. Li, L.-F. Huang, J. Lin, J. Zhang, Q.-H. Min, W.-M. Yang, *et al.*, "Mechanism of alternative splicing and its regulation (Review)," *Biomedical Reports*, vol. 3, pp. 152–158, 2015.

[237] P. Leder, "Moving genes.," *Progress in Clinical and Biological Research*, vol. 85, pp. 41–50, 1981.

[238] C. F. Austerberry, C. D. Allis, and M.-C. Yao, "Specific DNA rearrangements in synchronously developing nuclei of tetrahymena," *Proceedings of the National Academy of Sciences*, vol. 81, pp. 7383–7387, 1984.

[239] D. E. Kelley and R. P. Perry, "Transcriptional and posttranscriptional control of immunoglobulin mRNA production during Blymphocyte development," *Nucleic Acids Research*, vol. 14, pp. 5431–5447, 1986.

[240] B. A. Murray, G. C. Owens, E. A. Prediger, K. L. Crossin, B. A. Cunningham, and G. M. Edelman, "Cell surface modulation of the neural cell adhesion molecule resulting from alternative mRNA splicing in a tissue-specific developmental sequence.," *The Journal of Cell Biology*, vol. 103, pp. 1431–1439, 1986.

[241] P. N. Schofield and V. E. Tate, "Regulation of human IGF-II transcription in fetal and adult tissues," *Development*, vol. 101, pp. 793–803, 1987.

[242] A. Campagnoni, B. Sorg, H. Roth, K. Kronquist, S. Newman, K. Kitamura, C. Campagnoni, and B. Crandall, "Expression of myelin protein genes in the developing brain.," *Journal de Physiologie*, vol. 82, pp. 229–238, 1986.

[243] J. C. Kiefer, "Epigenetics in development," *Developmental Dynamics*, vol. 236, pp. 1144–1156, 2007.

[244] I. Cantone and A. G. Fisher, "Epigenetic programming and reprogramming during development," *Nature Structural & Molecular Biology*, vol. 20, pp. 282–289, 2013.

[245] B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, *et al.*, "The NIH roadmap epigenomics mapping consortium," *Nature Biotechnology*, vol. 28, pp. 1045–1048, 2010.

[246] R. A. Harris, T. Wang, C. Coarfa, R. P. Nagarajan, C. Hong, S. L. Downey, B. E. Johnson, S. D. Fouse, A. Delaney, Y. Zhao, *et al.*, "Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications," *Nature Biotechnology*, vol. 28, pp. 1097–1105, 2010.

[247] A. Milosavljevic, "Putting epigenome comparison into practice," *Nature Biotechnology*, vol. 28, pp. 1053–1056, 2010.

[248] A. Milosavljevic, "Emerging patterns of epigenomic variation," *Trends in Genetics*, vol. 27, pp. 242–250, 2011.

[249] S. Anders, A. Reyes, and W. Huber, "Detecting differential usage of exons from RNA-seq data," *Genome Research*, vol. 22, pp. 2008–2017, 2012.

[250] Z. Shao, Y. Zhang, G.-C. Yuan, S. H. Orkin, and D. J. Waxman, "MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets," *Genome Biology*, vol. 13, p. R16, 2012.

[251] M. Lienhard, C. Grimm, M. Morkel, R. Herwig, and L. Chavez, "MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments," *Bioinformatics*, vol. 30, pp. 284–286, 2014.

[252] A. Akalin, M. Kormaksson, S. Li, F. E. Garrett-Bakelman, M. E. Figueroa, A. Melnick, C. E. Mason, *et al.*, "methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles," *Genome Biology*, vol. 13, p. R87, 2012.

[253] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, pp. 841–842, 2010.

[254] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, *et al.*, "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, pp. 2078–2079, 2009.

[255] R. Chandramohan, P.-Y. Wu, J. H. Phan, and M. D. Wang, "Benchmarking RNA-Seq quantification tools," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pp. 647–650, IEEE, 2013.

[256] B. D. Strahl, P. A. Grant, S. D. Briggs, Z.-W. Sun, J. R. Bone, J. A. Caldwell, S. Mollah, R. G. Cook, J. Shabanowitz, D. F. Hunt, *et al.*, "Set2 is a nucleosomal histone H3-selective methyltransferase that mediates transcriptional repression," *Molecular and Cellular Biology*, vol. 22, pp. 1298–1306, 2002.

[257] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge, "Alternative isoform regulation in human tissue transcriptomes," *Nature*, vol. 456, pp. 470–476, 2008.

[258] A. J. Ridley, "Rho GTPases and actin dynamics in membrane protrusions and vesicle trafficking," *Trends in Cell Biology*, vol. 16, pp. 522–529, 2006.

[259] S. I. Ellenbroek and J. G. Collard, "Rho GTPases: functions and association with cancer," *Clinical & Experimental Metastasis*, vol. 24, pp. 657–672, 2007.

[260] C.-H. Heldin, K. Miyazono, and P. Ten Dijke, "TGF-β signalling from cell membrane to nucleus through SMAD proteins," *Nature*, vol. 390, pp. 465–471, 1997.

[261] G. Gupta, M. Bansal, and V. Sasisekharan, "Conformational flexibility of DNA: polymorphism and handedness," *Proceedings of the National Academy of Sciences*, vol. 77, pp. 6486–6490, 1980.

[262] J. M. Luk and K.-L. Guan, "An alternative DNA damage pathway to apoptosis in hematological cancers," *Nature Medicine*, vol. 20, pp. 587–588, 2014.

[263] D. Pan, "Hippo signaling in organ size control," *Genes & Development*, vol. 21, pp. 886–897, 2007.

[264] O. Takeuchi, S. Sato, T. Horiuchi, K. Hoshino, K. Takeda, Z. Dong, R. L. Modlin, and S. Akira, "Cutting edge: role of Toll-like receptor 1 in mediating immune response to microbial lipoproteins," *The Journal of Immunology*, vol. 169, pp. 10–14, 2002.

[265] N. N. Karpova, "Role of BDNF epigenetics in activity-dependent neuronal plasticity," *Neuropharmacology*, vol. 76, pp. 709–718, 2014.

[266] T. Ikegame, M. Bundo, F. Sunaga, T. Asai, F. Nishimura, A. Yoshikawa, Y. Kawamura, H. Hibino, M. Tochigi, C. Kakiuchi, *et al.*, "DNA methylation analysis of bdnf gene promoters in peripheral blood cells of schizophrenia patients," *Neuroscience Research*, vol. 77, pp. 208–214, 2013.

[267] T. A. Gray, S. Saitoh, and R. D. Nicholls, "An imprinted, mammalian bicistronic transcript encodes two independent proteins," *Proceedings of the National Academy of Sciences*, vol. 96, pp. 5616–5621, 1999.

[268] W. A. Burgers, F. Fuks, and T. Kouzarides, "DNA methyltransferases get connected to chromatin," *Trends in Genetics*, vol. 18, pp. 275–277, 2002.

[269] U. Aapola, I. Liiv, and P. Peterson, "Imprinting regulator DNMT3L is a transcriptional repressor associated with histone deacetylase activity," *Nucleic Acids Research*, vol. 30, pp. 3602–3608, 2002.

[270] K. Hata, M. Okano, H. Lei, and E. Li, "Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice," *Development*, vol. 129, pp. 1983–1993, 2002.

[271] M. Simonis, S. S. Atanur, S. Linsen, V. Guryev, F.-P. Ruzius, L. Game, N. Lansu, E. de Bruijn, S. van Heesch, S. Jones, *et al.*, "Genetic basis of transcriptome differences between the founder strains of the rat HXB/BXH recombinant inbred panel," *Genome Biology*, vol. 13, p. r31, 2012.

[272] V. Kashyap, N. C. Rezende, K. B. Scotland, S. M. Shaffer, J. L. Persson, L. J. Gudas, and N. P. Mongan, "Regulation of stem cell pluripotency and differentiation involves a mutual regulatory circuit of the NANOG, OCT4, and SOX2 pluripotency transcription factors with polycomb repressive complexes and stem cell microRNAs," *Stem Cells and Development*, vol. 18, pp. 1093–1108, 2009.

[273] N. Cloonan, A. R. Forrest, G. Kolle, B. B. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, *et al.*, "Stem cell transcriptome profiling via massive-scale mRNA sequencing," *Nature Methods*, vol. 5, pp. 613–619, 2008.

[274] G. W. Yeo, X. Xu, T. Y. Liang, A. R. Muotri, C. T. Carson, N. G. Coufal, and F. H. Gage, "Alternative splicing events identified in human embryonic stem cells and neural progenitors," *PLoS Computational Biology*, vol. 3, p. e196, 2007.

[275] W. Shi, H. Wang, G. Pan, Y. Geng, Y. Guo, and D. Pei, "Regulation of the pluripotency marker Rex-1 by Nanog and Sox2," *Journal of Biological Chemistry*, vol. 281, pp. 23319–23325, 2006.