Aus dem Lehrstuhl für Klinische Bioinformatik
Theoretische Medizin und Biowissenschaften
der Medizinischen Fakultät
der Universität des Saarlandes, Homburg/Saar

Multi-scale analysis of non-synonymous single nucleotide variant sets and candidate identification in genetic disorders

**Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften
der Medizinischen Fakultät
der UNIVERSITÄT DES SAARLANDES**

2015

vorgelegt von

Sabine Christel Müller

geb. am 10.08.1984 in Saarbrücken

Datum des Kolloquiums: 29. April 2016
Dekan der Medizinischen Fakultät: Prof. Dr. Michael D. Menger

Mitglieder des Prüfungsausschusses:
Vorsitzender: Prof. Dr. Veit Flockerzi
Erster Gutachter: Prof. Dr. Andreas Keller
Zweiter Gutachter: Prof. Dr. Eckart Meese
Wissenschaftlicher Beirat: Prof. Dr. Steffi Urbschat

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertig habe. Die aus den anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Saarbrücken, 19.Oktober 2015

(Sabine C. Müller)

Abstract

The study of genetic factors in human diseases has become a central part of medical research. Non-synonymous single nucleotide variants (nsSNVs) in coding regions of the human genome alter a protein's amino acid sequence and thus, are frequently associated with pathogenic phenotypes. For Mendelian disorders, where a variation in one single gene is causative, various disease-causing genes with corresponding nsSNVs have been identified already. Common or complex diseases such as diabetes or cancer, however, are caused by a varying number of genetic variants modulating a disease's severity and type. Furthermore, a human individual inheres more than one nsSNV. Especially the individual combination of nsSNVs may play a fundamental role in clinical diagnostics to tailor patient-specific treatment. Methods to study these, though, are currently limited. In this thesis, we present bioinformatic approaches to analyze putative synergetic effects of nsSNV sets and identify candidates for disease relevance. First, we performed a preliminary inventory of current analysis methods, their capabilities as well as their contribution to medical research, and in particular, their shortcomings. Based on these findings and additional next-generation sequencing (NGS) studies, we developed approaches covering the two main scenarios: multiple nsSNVs in one gene and nsSNVs in multiple genes, respectively. With the presented software tool BALL-SNP, we combine genetic and structural information in a three-dimensional visualization to enable the assessment of nsSNV sets in a single gene for diagnostic candidate identification. To study the functional impact of nsSNVs in multiple genes, we constructed a multi-scale pipeline comprising 3D content, interaction information and functional cascades of nsSNV-inherited genes and their encoded proteins.
The developed approaches have been used to analyze a high-quality, clinical data set of dilated cardiomyopathy (DCM) patients. In consequence, we have been able to identify nsSNV sets putatively contributing to DCM and consequently demonstrate the ability of the thesis to promote nsSNV analysis towards computational diagnostics.

German Abstract

Das Studium genetischer Faktoren bei humanen Erkrankungen hat sich zu einem zentralen Punkt in der medizinischen Forschung entwickelt. Nicht-synonyme Einzelnukleotid-Varianten (kurz nsSNVs) in kodierenden Bereichen des menschlichen Genoms verändern die Aminosäuresequenz eines Proteins und stehen somit häufig in Zusammenhang mit pathogenen Phänotypen. Bei Mendelschen Erkrankungen, die auf dem Defekt eines einzelnen Gens beruhen, konnten bereits einige kausative Gene identifiziert werden. Komplexe Erkrankungen wie Diabetes oder Krebs allerdings werden von einer Vielzahl genetischer Varianten, die Schweregrad und Art einer Erkrankung beeinflussen, verursacht. Zudem trägt ein Mensch mehrere nsSNVs in seinem Genom. Aus medizinischer Sicht könnte gerade diese individuelle Kombination von nsSNVs eine fundamentelle Rolle hinsichtlich patientenspezifischer Therapien spielen. Methoden, um diese Kombinationen zu analysieren, sind zur Zeit allerdings limitiert.

In der vorliegenden Arbeit stellen wir Ansätze der Bioinformatik vor, um nsSNV Kombinationen zu untersuchen, synergetische Effekte zu detektieren und Kandidaten zu identifizieren, die möglicherweise krankheitsrelevant sind. In einem ersten Schritt wurden bestehende Methoden zur Pathogenitätsvorhersage einzelner nsSNVs bezüglich ihrer Stärken und Schwächen sowie ihrer medizinischen Relevanz untersucht. Basierend auf diesen Ergebnissen und weiteren Next-Generation Sequencing (NGS) Studien wurden Methoden entwickelt, deren Fokus auf der Analyse der folgenden beiden Szenarien liegt: Mehrere nsSNVs in einem einzelnen Gen sowie nsSNVs in mehreren Genen. Die entwickelte Software BALL-SNP kombiniert genetische und strukturelle Informationen in einer drei-dimensionalen (3D) Darstellung und ermöglicht somit die Bewertung von Kombinationen aus nsSNVs in einem einzelnen Gen bzw. dem kodierten Protein hinsichtlich ihrer krankheitsassoziierten Relevanz. Um auch den pathogenen Einfluss von nsSNVs in mehreren Genen untersuchen zu können, wurde ein Mehrskalenansatz entwickelt, der 3D Inhalte, Interaktionsinformationen und funktionelle Kaskaden von Genen mit nsSNVs sowie den kodierten Proteinen umfasst.

Die in dieser Arbeit vorgestellten Methoden wurden verwendet, um einen hochqualitativen klinischen Datensatz an Dilatativer Kardiomyopathie (DCM) erkrankter Patienten zu analysieren. Dadurch war es möglich nsSNV Kombinationen zu identifizieren, die möglicherweise kausativ zur Erkrankung an DCM beitragen. Insbesondere konnte damit gezeigt werden, dass die vorliegende Arbeit einen wichtigen Teil zur Analyse von nsSNVs bezüglich computergestützter Diagnostikverfahren beiträgt.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Introduction

In the last decade, rapid development and advances in experimental high-throughput techniques like next-generation sequencing (NGS) have enabled the reliable detection of individual sequence variants in the human genome [1]. Along with computational methods, NGS allows the discovery and genotyping of hundreds to thousands of genetic variants in different species, while becoming less expensive. Figure 1.1 illustrates the course of this trend.

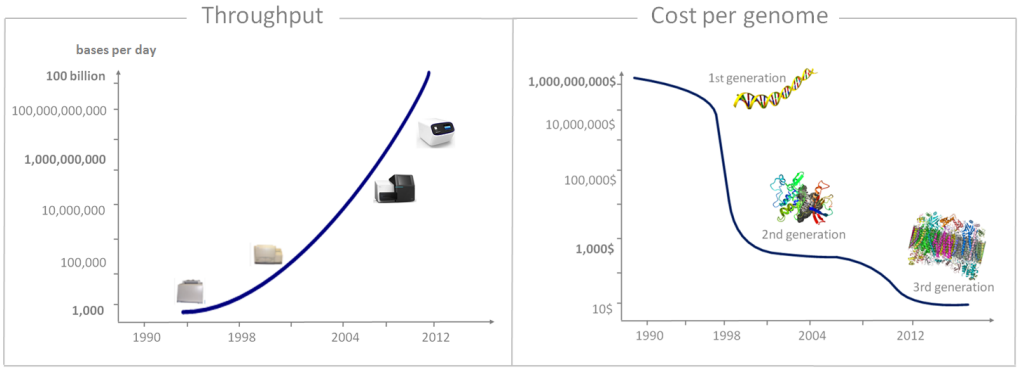


Abbildung 1.1: In the last decade, genome-wide sequencing techniques achieved both, higher throughput and significant cost reduction [based on the National Institutes of Health (NIH)].

Since the sequencing of an individual's whole genome becomes affordable and less time-consuming, clinicians can get more and more access to genetic data to customize medical treatment. In particular, understanding the genotype-phenotype relationship is expected to promote diagnosis, treatment, and prevention in health care [2]. Knowing the genetic basis of a disease may permit the early detection of patients with a high risk of developing this particular disease.
An extensive number of genetic variations can be identified in whole genome sequencing studies. Genetic alterations such as non-synonymous variants, for example, are known to play a critical role in human diseases [3]. Although only a small fraction of genetic variations are *non-synonymous single nucleotide variants* (nsSNVs), over 85% of such mutations are associated with a specific disease [4]. NsSNVs refer to single base changes in DNA coding regions altering a protein's amino acid sequence. A pathogenic

phenotype may arise when an amino acid substitution affects structurally important residues and sites relevant for function, such as residues in catalytic sites of enzymes. Experimentally gained knowledge about genetic variations is deposited and curated in different databases, where the Single Nucleotide Polymorphism Database (dbSNP) currently refers to the largest database concerning SNV annotation with 29,901,117 deposited SNVs (dbSNP build 138) [5]. The most popular database for known pathogenic mutations denotes the Human Gene Mutation Database (HGMD), a comprehensive repository of mutations associated with human inherited diseases [6]. At present, 94,860 disease-associated nsSNVs are published in the HGMD [7]. Due to the advent of high-throughput variant detection, though, the amount of identified nsSNVs is growing rapidly.

To gain knowledge concerning the pathogenicity of nsSNVs via experimental analysis is laborious and time-consuming, and often even not possible. Thus, computational approaches have been developed to study the functional impact of nsSNVs *in silico*. To date, *genome-wide association studies* (GWAS) are commonly performed to assess the statistical association of genetic alterations with common diseases [8]. GWAS compare regions of the genome between cohorts to identify common genetic variants, statistically associated with a trait. Although hundreds of GWAS on particular genetic variations and various diseases have been conducted today, GWAS is susceptible to many limitations. The application of GWAS, for example, requires large sample sizes, which are often not available. Multi-marker approaches considering nsSNV combinations in GWAS are of high complexity increasing the curse of dimensionality. Furthermore, GWAS ignore known disease pathobiology and prior biological knowledge, however, the inclusion of this information into data analysis may improve the results [9]. Most nsSNVs discovered in GWAS have been recognized not to reveal the genetic basis of disease susceptibility and etiology. In consequence, the majority of these nsSNVs are not suitable to improve medical health care by genetic testing.

Due to these limitations, prior knowledge about the investigated disease is used to concentrate on target genomic regions including genes previously associated with the disease to identify causative variants. Methods to predict the pathogenicity of inherited nsSNVs cover a wide range of ideas from evolutionary and sequence-based predictions to detailed atomic energy-based methods [10]. In the following, we summarize the basic ideas of available computational approaches using different algorithms and features to predict whether an nsSNV is disease-associated or not.

The majority of the existing computational approaches predict the pathogenic effect using statistical methods, machine learning techniques or protein evolution models, based on derived features [11]. These features correspond in general to sequence homology, biochemical properties and structural information (hydrogen-bond network, solvent accessibility, etc.). Besides, there are computational methods based on potential energy functions, force fields and molecular dynamics, which analyze the change in a protein's stability, dynamics and interactions to consequently derive the impact of an

amino acid substitution [12][13]. These methods, however, can also be time-consuming and are generally used for small-scale investigations [14].
Among others, the main problem of all these approaches refers to the 'one-SNV, one-phenotype'-paradigm. For Mendelian disorders, where variation in a single gene is responsible for the phenotypic consequence, thousand such causative genes have been identified already [15]. Detecting the causative genes in common diseases such as hypertension, diabetes or cancer, however, still remains a challenge. In general, these diseases are caused by a varying number of genetic alterations and are influenced by environmental factors that modulate the severity and type of disease-related phenotypes [16]. Several nsSNVs in a single gene may exert a synergetic effect on a protein's function, whereas a phenotype can also result from combined action of nsSNVs in many genes [17]. In consequence, the assumption of nsSNVs as single entities, independent from each other, may not substantially contribute to the improvement of computational diagnosis, especially for common diseases. Furthermore, from a medical point of view, especially the individual combination of nsSNVs may play a crucial role in clinical diagnostics regarding personalized medicine, since genetic variations have been identified to influence selection, dosing and adverse events of medical drugs [18].

In this work, we focused on the development of approaches to assess the phenotypic impact of multiple nsSNVs, so-called nsSNV sets, on disease phenotypes and propose strategies to analyze putative synergetic effects. To the best of our knowledge, computational methods to predict disease association for nsSNV sets are currently limited. Figure 1.2 gives an overview of the main analysis strategies developed in this thesis.

In order to analyze the current potential of single nsSNV pathogenicity prediction, we first performed comprehensive evaluation studies on both, prediction concordance and prediction quality of existing tools on a high quality clinical data set of cardiomyopathy samples.
The choice and quality of test data plays a crucial role and since the coverage rate of the cardiomyopathy data exceeds classical exome capture studies by several orders of magnitude, these data were used for several analyses within this thesis. In consequence, we also studied genetic factors putatively contributing to cardiovascular diseases.
The evaluation study of current single nsSNV pathogenicity prediction strategies revealed several drawbacks with respect to performance, congruency, applicability and clinical relevance. However, we have been able to demonstrate the importance of structural information when analyzing the functional impact of nsSNVs [11].
A protein's structure, dynamics and interactions are interrelated. In addition, the effects of genetic differences on protein function are various [19]. Hence, nsSNVs may change several properties of a protein, simultaneously. The visual inspection of the three dimensional (3D) structure of proteins affected by nsSNV-introduced amino acid substitutions, thus, may reveal crucial insights in effects altering protein function.
Based on the gained information, we developed BALL-SNP - a tool to identify candidate

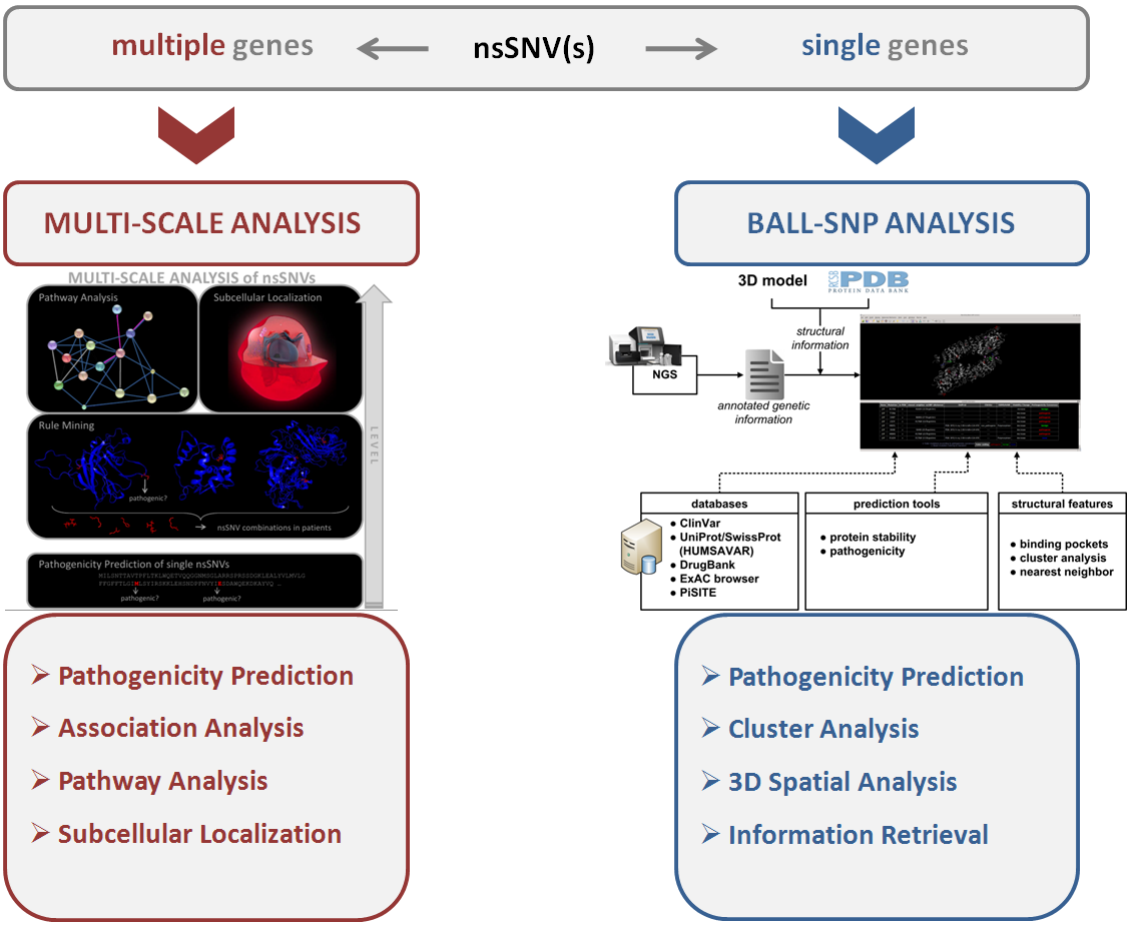


Abbildung 1.2: In this thesis, we developed approaches to assess the functional impact of nsSNV sets. The analysis of nsSNV sets comprises two parts: the analysis of nsSNVs accumulating in one gene and the multi-scale analysis of nsSNVs inhered in multiple genes.

nsSNVs for computational diagnostics. BALL-SNP combines genetic and structural information, respectively, and in particular, offers a visualization of the generated information as well as an intuitive user interface. Furthermore, we systematically incorporated information retrieval and sophisticated analysis methods, such as e.g. cluster analysis, to collect and generate crucial information in one tool. BALL-SNP relies on simple input formats: the output of standard SNP annotation software such as ANNOVAR or a simple tab-separated input file, which prevents users from substantial re-formatting.
Analyses of nsSNVs in the cardiomyopathy samples impressively demonstrated the ability of BALL-SNP to uncover spatial relations and putative synergetic effects introduced by genetic alterations.
Besides the analysis of nsSNV sets in one gene, the study of nsSNVs in multiple genes may play an important role in computational diagnostics. The genetic basis of most common diseases refers to multiple genetic factors such as gene-gene and gene-environment interactions [20]. To study nsSNVs in multiple genes, we furthermore developed a multi-scale approach integrating 3D context, interaction information and functional cascades from gene sets with nsSNVs. Since we used data from cardiomyopathy patients, we were also able to study synergetic effects of nsSNVs related to cardiomyopathy. Our multi-scale analysis of this data set revealed associated nsSNV combinations in seven genes, putatively related to cardiomyopathy.

In summary, the impact of nsSNVs in coding genes on the cause and the severity of a disease has become a key task in human health care. A pathogenic phenotype may result from nsSNV sets clustering in one single gene as well as from the combination of nsSNVs in many genes. In this thesis, we developed approaches to address both, the study of nsSNVs sets in one gene and in multiple genes. BALL-SNP, a freely available software tool, enables the assessment of the impact of nsSNV clusters on a protein's function and stability, and consequently assists the selection of candidate nsSNVs for experimental validation. Though further improvement is needed to meet requirements of the clinical application, BALL-SNP already decisively contributes to existing instruments of candidate nsSNV analysis. The constructed multi-scale analysis pipeline for nsSNV sets in multiple genes supports the computational study of cumulative effects and their impact on pathogenicity. We have been able to demonstrate this on the example of a cardiac phenotype, however, the analysis can be likewise applied to other diseases such as cancer.

The outline of this thesis comprises: in the next chapter, a summary of important concepts and methods used for this thesis. Next, we present in chapter 3-5 our findings concerning the pathogenicity prediction of single nsSNVs, sets of nsSNVs within one gene and nsSNV combinations in multiple genes. Finally, we discuss our contribution to the analysis of nsSNVs towards the applicability of NGS data in clinical routine.

2 Material and Methods

In this chapter, we briefly summarize important concepts as well as already available information and software tools, applied or included within this work. Since this thesis aims to represent a bridge between bioinformatic strategies and clinical application scenarios, we focused on comprehensibility for both, bioinformaticians and clinicians.

2.1 Single Nucleotide Variants

The genome comprises an organism's complete set of deoxyribonucleic acid (DNA) [21]. The human DNA resides in the cell nucleus and is organized into chromosomes. It consists of two strands made up by four types of bases: adenine (A), thymine (T), guanine (G) and cytosine (C). These bases link into pairs, forming the double-helical DNA structure. The genetic information is coded in the order of the bases in the strands. The human genome is composed of about 3 billion base pairs including circa 23.000 genes. Interestingly, only approximately 2% of the genome code for proteins. Three bases, so-called triplets or codons, code for one amino acid constituting a protein's primary sequence. Figure 2.1 schematically illustrates the architecture of the human genome.
The genome sequences of two individuals are 99.9% identical, the remaining 0.1% DNA accounts for natural genetic variation between and within populations [22], which results in different traits or phenotypes. The most common type of genetic variation is defined as a single nucleotide polymorphism (SNP), a position where two alternative bases occur with $> 1\%$ in the human population [1]. More generally, we speak of single nucleotide variants (SNVs), if no information about the 1% criterion fit is available. Figure 2.2 illustrates a SNV.
SNVs may occur in gene coding regions as well as in non-coding or intergenic regions of the DNA. SNVs not located in protein-coding regions may affect gene splicing, transcription factor binding or messenger RNA degradation. In the coding region, we discriminate between synonymous and non-synonymous SNVs (nsSNVs). Synonymous SNVs do not change the amino acid sequence of the corresponding protein due to the degeneracy of the genetic code, where several triplets can code for one amino acid (see Figure 2.3). In contrast, nsSNVs alter the protein sequence. They, for example, can introduce premature stop codons, consequently producing functionally incompetent truncated proteins, and hence, may possibly be lethal. The more interesting variations refer to the viable nsSNVs. They frequently result in a single amino acid substitution within a protein sequence and thus, can alter protein function comprising folding,

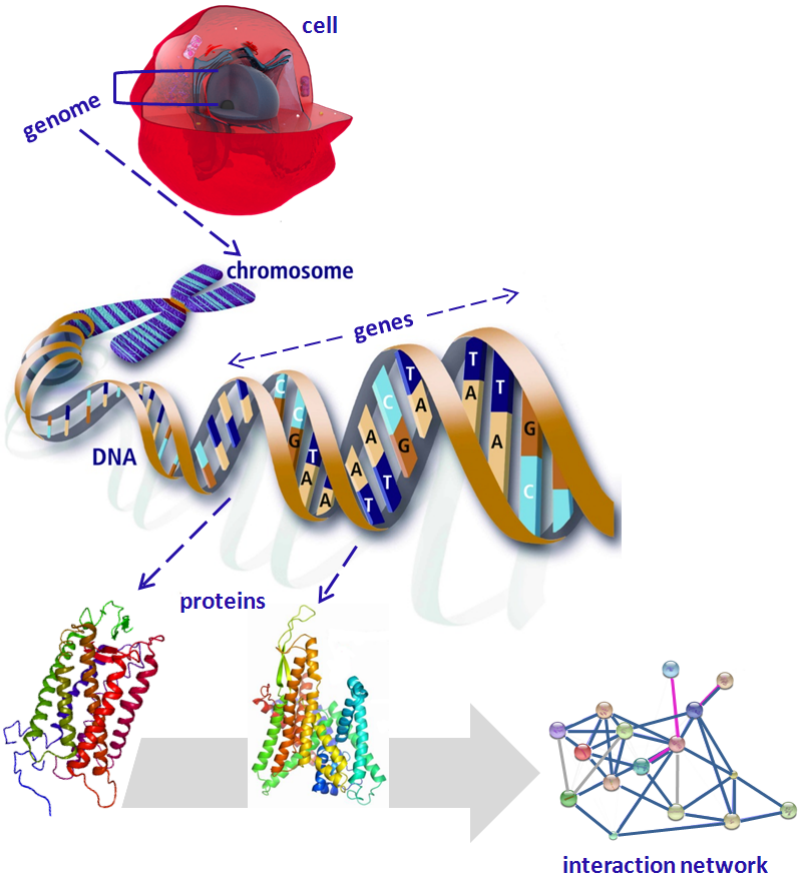


Abbildung 2.1: The human genome comprises all genetic instructions of an organism.

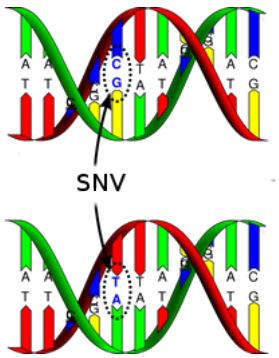


Abbildung 2.2: A single nucleotide variant (SNV) refers to a single base change within the DNA.

stability and binding of other proteins or ligands. From a medical point of view, nsSNVs in coding DNA can be neutral, associated with a disease by exerting a small effect on a specific trait, or they can be the cause for a distinct disease [11] [7].

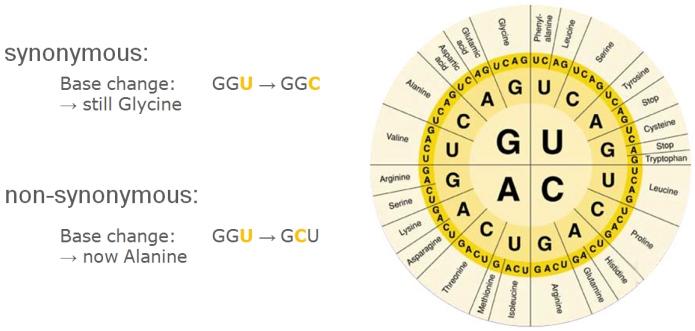


Abbildung 2.3: Due to the degeneracy of the genetic code, particular base changes may introduce different amino acids, whereas others do not alter the protein sequence.

2.2 Databases containing nsSNV information

Experimentally gained knowledge about nsSNVs is deposited and curated in different databases. Table 2.1 lists the main available SNV databases. Due to the raising medical importance, several databases already aim at annotating genetic variants with phenotype association as well as structural and functional information on proteins. In the following paragraphs, we shortly summarize the most important databases, which are also included in the developed approaches presented in this work (see Chapter 4 and 5).

2.2.1 The Single Nucleotide Polymorphism Database (dbSNP)

The Single Nucleotide Polymorphism Database (dbSNP) is currently the largest database concerning SNV annotations with 29,901,117 deposited SNVs (dbSNP build 138) [5]. It refers to a general catalog of genetic variations including sequence information around the variant, descriptions of the population inhering the variant and frequency information by population or individual genotype. Each record is assigned a reference identifier, the so-called *rs ID*, allowing to map variants to external resources or databases. DbSNP is available at `http://www.ncbi.nlm.nih.gov/SNP`.

ClinVar Tightly coupled with dbSNP, ClinVar accessions report human variations and interpretations of the relationship of these variations to human health [23]. It is a

Tabelle 2.1: The table lists currently the most important SNV databases.

Database	Description	Current status
dbSNP	Database of Single Nucleotide Polymorphisms	29,901,117 SNVs
ClinVar	Public archive of relationships among sequence variation and human phenotype	156,921 genetic variations
UniProtKB/SwissProt	Protein sequence database including experimental results, computed features and scientific conclusions	548,872 proteins, including 90,378 variants
HGMD	Human Gene Mutation Database of mutations causing inherited disease	public: 67,439 nsSNVs, commercial: 94,860 nsSNVs

freely accessible, public archive comprising genetic variants identified through clinical testing, research and literature. Entries are labeled according to the variation and its clinical significance. If information is available, ClinVar classifies genetic variants into the following categories:

- unkown
- untested
- non-pathogenic
- probable-non-pathogenic
- probable-pathogenic
- pathogenic
- drug-response
- histocompatibility
- other.

ClinVar is available via dbSNP or at `http://www.ncbi.nlm.nih.gov/clinvar/`.

2.2.2 The UniProt Knowledgebase (UniProtKB)

The UniProt Knowledgebase (UniProtKB) is a very comprehensive data collection of proteins and their functional information, available at `http://www.uniprot.org` [24]. It distincts between manually annotated, reviewed data (Swiss-Prot [25]) and

computationally analyzed, unreviewed data (TrEMBL). Each record lists core information such as amino acid sequence, protein name, etc as well as additional information about biological ontologies, classifications and cross-references to other data sources. In addition, the UniProtKB includes a collection of human polymorphisms and disease mutations assigned according to literature reports on probable disease association, called HUMSAVAR [26].

2.2.3 The Human Gene Mutation Database (HGMD)

The most popular database for known pathogenic mutations is the Human Gene Mutation Database (HGMD) [6]. The HGMD refers to a comprehensive repository of mutations associated with human inherited disease. Since its initiation, the HGMD has become the central disease-associated mutation resource for the scientific community. In contrast to other developed data sources collating variant-disease associations, the HGMD is manually curated to avoid inconsistency as well as biased entries. The HGMD is available at `http://www.hgmd.cf.ac.uk/` in a public version for nonprofit and academic users, and in a professional, commercial version (HGMD and BIOBASE GmbH), respectively.

2.3 The Protein Data Bank (PDB)

The developed software BALL-SNP (Chapter 4) critically depends on 3D structures and is consequently connected to the Protein Data Bank (PDB), the most comprehensive archive for 3D structural data of biological macromolecules world-wide [27]. The publicly available archive contains experimentally resolved structures ranging from small peptides and nucleic acids to large complexes. It is available at `http://www.rcsb.org/pdb/`.
In addition, the PDB also defines the well-known PDB file format, which contains all information required for a 3D structure and is used as input in molecular modelling software as well as 3D visualization tools (see Section 2.8).

2.4 DrugBank - Open Data Drug and Drug Target Database

DrugBank is a freely available bioinformatics and cheminformatics resource (`http://www.drugbank.ca`), which links chemical, pharmacological and pharmaceutical data to comprehensive sequence, structure and pathway information of their corresponding drug targets [28].
The current available version 4.2 contains 7,759 drug entries including 1,600 small molecule drugs, 160 biotech (protein/peptide) drugs, 89 nutraceuticals and over 6000 experimental drugs [29]. Since we focus on the identification of candidate nsSNVs for computational diagnostics, information about already specified drug targets can critically contribute to this process.

2.5 Pathogenicity prediction

Due to the advent of high-throughput variant detection techniques, the amount of identified nsSNVs is growing rapidly [11]. To gain knowledge concerning the pathogenicity of nsSNVs via experimental analysis is laborious and time-consuming, and often even not possible. To solve this intricate problem, various computational methods have been developed over the past decade using different algorithms and features to predict the biological impact of nsSNVs on a protein's function *in silico* and to assess whether an nsSNV is associated with a specific disease.
Most of these prediction methods are based on evolutionary information and/or combine functional and structural parameters as well as multiple sequence alignment derived information. Conservation information usually is obtained from alignments of homologous or somehow related sequences, including position-specific profiles. Some prediction methods also incorporate available annotations, e.g. Gene Ontology (GO) [30] or protein family information (Pfam) [31]. Because protein structure encodes protein function, information concerning the three-dimensional (3D) structural environment, such as solvent accessibility, electrostatics and hydrophobicity, is also a crucial criterion to assume a variant's functional impact [11]. Finally, the *in silico*-derived information about protein structure and function, including essential properties of both, the original and the substituted residues, is combined into features. According to these features, nsSNVs are classified into benign or pathogenic using different machine learning methods such as neural networks, random forests, support vector machines (SVMs) or Bayesian methods and mathematical operations. Figure 2.4 summarizes this general scheme of pathogenicity prediction.

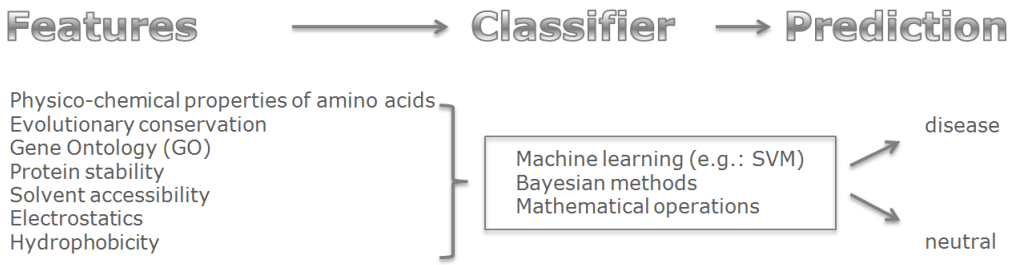


Abbildung 2.4: Overview of the general SNV pathogenicity prediction strategy.

Most techniques were trained on data deposited in databases such as dbSNP or UniProtKB, and with artifically constructed test sets.
In the following, we shortly summarize the underlying principles of prediction approaches selected for integration in the developed software tool BALL-SNP (see Chapter 4).

2.5.1 PANTHER - Evolutionary analysis of coding SNVs

PANTHER computes substitution position-specific evolutionary conservation (subPSEC) scores on alignments [32]. These alignments of evolutionarily related proteins are generated by Hidden Markov Models (HMMs) and collected in the PANTHER library. The subPSEC score defines amino acid probabilities for the occurrence at certain positions among evolutionarily related sequences. Classification is done via scoring a query protein against the entire built PANTHER library, currently comprising over 38,000 HMMs [33].
The PANTHER library as well as the corresponding software tool is available at `http://pantherdb.org/tools/`.

2.5.2 PhD-SNP

PhD-SNP can be divided into two SVM-based strategies: the single-sequence and the sequence profile SVM [34]. The single-sequence SVM classifies the amino acid substitution as neutral or disease-linked based on the nature of the substitution and the properties of the neighboring sequence environment. In addition, the profile SVM calculates sequence profile information derived from multiple sequence alignments and classifies the functional impact according to the ratio of wild type and mutant frequencies of amino acids. Both SVM strategies combined account for the PhD-SNP prediction result.
The software tool is available at `http://snps.biofold.org/phd-snp/phd-snp.html`.

2.5.3 Polymorphism Phenotyping 2 - PolyPhen2

Polymorphism Phenotyping 2 (PolyPhen2) predicts the mutational effect by a naive Bayesian classifier [35]. The used features include both, sequence information derived from multiple sequence alignments and structure information such as, e.g., solvent accessibility.
Moreover, PolyPhen2 computes position-specific independent counts (PSIC) profile scores of two amino acid variants. The PSIC profile is a logarithmic ratio of the likelihood that a given amino acid occurs at a particular site to the background probability of the amino acid occurring at random at a given position within the protein sequence [36].
PolyPhen2 is available at `http://genetics.bwh.harvard.edu/pph2/`.

2.5.4 PROVEAN - Protein Variation Effect Analyzer

PROVEAN calculates a pairwise sequence alignment score based on homologous sequences [37]. This score measures the sequence similarity change of a query sequence to a homologous protein sequence before and after an amino acid substitution within the query sequence. In contrast to most of the other available prediction approaches, PROVEAN is able to predict the functional influence for all types of protein sequence

variations, such as insertions, deletions and substitutions.
Additionally, precomputed PROVEAN scores generated in November 2012 for all possible single amino acid substitutions and single amino acid deletions of human proteins from Ensembl 66 are freely available at `http://provean.jcvi.org`.

Figure 2.5 lists the main state-of-the-art pathogenicity prediction tools currently available. Besides, there are computational methods based on potential energy functions, force fields and molecular dynamics, which analyze the change in a proteins stability, dynamics and interactions to consequently derive the impact of an amino acid substitution [12][13]. These methods, however, can also be time-consuming, since they are computationally expensive, and are generally used for small-scale investigations [14].

2.6 I-Mutant: Protein stability prediction

Since structural information such as protein stability is important for the assessment of nsSNVs, we embedded I-Mutant within the BALL-SNP approach developed in this thesis (Chapter 4) as well as applied this tool within our multi-scale analysis in Chapter 5.
I-Mutant 2.0 is an SVM-predictor based on protein sequence or structure for the automatic assessment of protein stability [38]. It predicts both, the sign and the value of a protein's stability change upon an amino acid substitution.
The sign of the free energy difference, calculated by the substraction of the wild type free energy from that of the mutant, specifies protein stabilization (negative sign) or destabilization (positive sign). The main advantage of I-Mutant compared to other existing protein stability methods is the ability to predict from protein sequence, not requiring structural information.

2.7 The FASTA format

The FASTA format is a text-based format to represent either the sequence of nucleotides or peptides, where nucleotides or amino acids are defined as single-letter codes, e.g. H for Histidine [39]. The FASTA sequence consequently refers to the primary sequence of the residues on a protein backbone.
The first line of a FASTA file corresponds to the so-called header with description information such as the protein name and so on. This description line is labeled with a ">" at the beginning and is optional. A simple example for a FASTA file refers to:

Prediction method	Input	Classifier	Evolutionary analysis	Structural attributes	Annotations	Running modality
MutPred	Fasta sequence	Random forest	SIFT, Pfam, PSI-BLAST	Predictions of: secondary structure, solvent accessibility, transmembrane helices, stability, etc.	/	Web server
PMut	Fasta sequence, UniProt ID	Neural network	PSI-BLAST, multiple sequence alignments (MSA), Pfam	Homolog mapping/predictions	/	Web server
PROVEAN	Ensembl ID, NCBI RefSeq ID, UniProt ID	Alignment scores	BLAST	/	/	Web server, stand-alone
SNPs&GO	Uniprot ID	Support vector machines	Sequence environment, sequence profiles, PANTHER	/	GO	Web server
SNAP	Fasta sequence	Neural network	PSI-BLAST, position-specific independent counts (PSIC) profiles, Pfam	Predictions: secondary structure, solvent accessibility, chain flexibility	SwissProt	Web server
SIFT	Comma separated: chromosome, coordinate, orientation, alleles	Alignment scores	MSA	/	/	Web server, stand-alone
PANTHER	Fasta sequence	Alignment scores	PANTHER library, hidden Markov Models	/	GO	Web server, stand-alone
PhD-SNP	Fasta sequence	Support vector machines	Sequence environment, sequence profiles, MSA	/	/	Web server, stand-alone
SNPs3D	SNP ID, sequence ID	Support vector machines	PSI-BLAST, position-specific scoring matrix, MSA	Structure stability model (solvent accessibility, electrostatics, hydrophobicity...)		Web server
PolyPhen2 (PPh2)	Fasta sequence, SNP ID, UniProt ID	Bayesian classification	PSIC profiles	Homolog mapping/predictions	Pfam	Web server, stand-alone
MutationAssessor	Uniprot ID	Alignment scores	MSA	/	/	Web server
PredictSNP	Fasta sequence	Confidence-based random forest consensus	MAPP, nsSNPAnalyzer, PANTHER, PhD-SNP, PolyPhen, PolyPhen2, SIFT, SNAP	Predictions: secondary structure, solvent accessibility, chain flexibility (SNAP, nsSNPAnalyzer), homolog mapping/ predictions(PolyPhen2)	SwissProt (SNAP), GO (PANTHER), Pfam (PolyPhen2)	Web server, batch script available with many dependencies
Condel	Uniprot ID, chromosome + start + end + mutant	Weighted average score	SIFT, PolyPhen2, MutationAssessor	Homolog mapping/predictions (PolyPhen2)	Pfam (PolyPhen2)	Web server

Abbildung 2.5: Overview of the main state-of-the-art pathogenicity prediction tools [11].

```
>sp|P08588|ADRB1_HUMAN Beta-1 adrenergic receptor
MGAGVLVLGASEPGNLSSAAPLPDGAATAARLLVPASPPASLLPPASESPEPLSQ
QWTAGMGLLMALIVLLIVAGNVLVIVAIAKTPRLQTLTNLFIMSLASADLVMGLL
VVPFGATIVVWGRWEYGSFFCELWTSVDVLCVTASIETLCVIALDRYLA ...
```

FASTA files are the standard format to represent a protein sequence and thus, are generally the input for alignment tools and further methods dealing with protein sequences. The approaches developed in this thesis also require protein sequences and in consequence, rely on FASTA file information.

2.8 The Protein Data Bank (PDB) File

For the assessment of nsSNVs, 3D structural information of the corresponding protein with amino acid substitutions introduced by these nsSNVs, is essential. In general, 3D information derived from X-ray diffraction and Nuclear Magnetic Resonance (NMR) studies is represented via the Protein Data Bank (PDB) file format [27], created in the 1970's, and available at the PDB (Section 2.3).
The formatting is defined by the character positions in a line, where each line starts with left-justified six characters, which denote an identifier for the line type, the so-called record name. There are various record names specifying available information such as SEQRES for the protein FASTA sequence (Section 2.7) or HELIX and SHEET for the secondary structure elements, for example. We concentrate, however, on the ATOM record, which is the most important one used in this thesis.
The ATOM record corresponds to the atomic coordinates for standard amino acids and nucleotides. Additional information besides atom name and coordinates such as the corresponding residue, the residue index and so on are also listed in ATOM lines.
The PDB is a collection of 3D information collected world-wide and although sustained efforts to maintain high-quality data exist, there are unfortunately inconsistent and informal PDB files, not fully checked for errors.

Above, we summarized important methods used in the majority of the developed approaches presented in this thesis. In the following, we describe applied strategies according to the corresponding approach, which incorporates these.

2.9 Pathogenicity prediction of nsSNV sets: material and methods

In Chapter 3, we present the results of an extensive evaluation study of current-state-of-the-art pathogenicity prediction methods. To evaluate the prediction performance of the selected tools on a created test set (see Section 2.12.1), we calculated standard performance measures defined below.

2.9.1 Measures to evaluate statistical performance

Based on the prediction results for the annotated test set (Section 2.12.1), we studied the prediction performance of the pathogenicity prediction methods by calculating *specificity, sensitivity* and *accuracy*. Specificity measures the proportion of negatives, that are correctly identified, whereas the sensitivity quantifies the proportion of actual positives [40], that are correctly detected as such:

$$specificity = \frac{TN}{TN + FP}$$

$$sensitivity = \frac{TP}{TP + FN}$$

where TP = true positives, FP = false positives, FN = false negatives and TN = true negatives.
Besides, the accuracy of a measurement system is the degree of closeness of a quantity to its actual true value and can be seen as the degree of veracity. It identifies the proportion of the true results [40]:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

where TP = true positives, FP = false positives, FN = false negatives and TN = true negatives.

Handling unbalanced data Since the created test set is slightly unbalanced (192 neutral to 147 disease-annotated nsSNVs, details in Section 2.12.1), we also computed the *balanced accuracy* and *Matthews correlation coefficient* (MCC), defined as:

$$balanced\ accuracy = \frac{specificity\ +\ sensitivity}{2}$$

$$MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

where an MCC value of 1 corresponds to a perfect correlation.

Based on the evaluation study of current pathogenicity prediction tools, we depict in Section 3.2 the attempt to adapt single pathogenicity prediction strategies to the assessment of nsSNV sets. Important methods used for the definition of scoring methods to measure the functional effect of nsSNV sets are summarized in the following.

2.9.2 Amino acid distribution probabilities

The amino acid sequence of a protein is characteristic for a protein's folding and function. Minor changes within this sequence may alter or prevent function [41]. To be able to assess the functional effect of amino acid substitutions introduced by nsSNVs, we studied the amino acid spectrum as well as the combination of wild type and mutant residues in neutral and disease-annotated DCM data (see Section 2.12.1). In particular, we compared the observed frequencies with the expected values determined based on the distribution of all amino acids in the test set. The expected values are also computed with regard to codon diversity and the probability to generate a certain substitution referring to all possibilities for a translation from a specific amino acid to another by only one triplet base mutation of the wild type codon.

Single amino acid probabilities

In general, the expected and observed frequencies of an amino acid X within a query protein of a given sequence seq are defined as:

$$\begin{aligned} f_{exp}(X) &= \frac{\#codons(X)}{all\_codons} * length(seq) \\ f_{obs}(X) &= \#X \end{aligned} \tag{2.1}$$

where '#' refers to 'number of', $\#codons(X)$ denotes the number of available codons for amino acid X and all_codons codes for all codons in the genetic code, namely 64.

Since we are mainly interested in the analysis of amino acid substitutions introduced by nsSNVs, we explicitly studied the amino acid distributions at wild type and mutant positions, respectively. At the wild type position of an amino acid sequence the calculated frequencies are defined as:

$$\begin{aligned} f_{exp}(X) &= \frac{\#codons(X)}{all\_codons} * \#mutations \\ f_{obs}(X) &= \#X \end{aligned} \tag{2.2}$$

where '#' refers to 'number of', $\#codons(X)$ denotes the number of available codons for amino acid X, all_codons codes for all codons in the genetic code (64) and $\#mutations$ refers to the number of mutations detected within the particular amino acid sequence.

A nsSNV changes one base within a codon and hence, we calculated the amino acid distributions at mutant positions in a protein's sequence with regard to codon diversity and the probability to generate a certain substitution referring to all possibilities for a translation from a specific amino acid to another by only one triplet base mutation of the wild type codon. For the distributions at mutant positions we define:

$$f_{exp}(X) = codonDiversity(X) * \#mutations$$
$$codonDiversity(X) = \frac{\#subs(Z^* \longrightarrow X)}{\#subs(all)} \tag{2.3}$$
$$f_{obs}(X) = \#X$$

where $\#subs(Z^* \longrightarrow X)$ denotes the number of substitutions from any amino acid to X by one base change and $\#subs(all)$ refers to the number of all possible amino acid substitutions resulting from one base change in the wild type codon.

Amino acid substitution probabilities

Beyond the single amino acid distributions, we also studied the spectrum of combinations of wild type and mutant residues. An amino acid substitution is defined as $X \longrightarrow Y$, meaning amino acid X is substituted by amino acid Y. Based on all mutant residues MT_i, which can result from the observed wild type WT by one base change within its triplet, the probability to generate certain substitutions $WT \longrightarrow MT_i$ is calculated. The expected and observed frequencies of amino acid substitutions in the used data set are determined with respect to the codon diversity by:

$$f_{exp}(X \longrightarrow Y) = codonDiversity(X \longrightarrow Y) * \#mutations(X \longrightarrow Y)$$
$$codonDiversity(X \longrightarrow Y) = \frac{\#subs(X \longrightarrow Y, by\ 1\ base\ change)}{\#subs(X \longrightarrow Z | Z \neq Y, by\ 1\ base\ change)} \tag{2.4}$$
$$f_{obs}(X \longrightarrow Y) = \#subs(X \longrightarrow Y)$$

where $\#subs(X \longrightarrow Y, by\ 1\ base\ change)$ defines the number of substitutions from amino acid X to Y by only one base change in the triplet of X. In addition, $\#subs(X \longrightarrow Z | Z \neq Y, by\ 1\ base\ change)$ denotes the number of all possible substitutions from X to any other amino acid Z unequal Y by one base change in the triplet of X.

2.9.3 The BLOcks SUbstitution Matrix (BLOSUM)

Since evolutionarily related proteins are supposed to form protein families inhering similar functions, the ability to measure the similarity of protein sequences is essential in molecular biology. Typically, alignment methods for protein sequences are applied to measure this similarity. The BLOcks SUbstitution Matrix (BLOSUM) refers to a substitution matrix for local sequence alignments of proteins [42]. BLOSUM comprises blocks of similar sequences and counts the relative frequencies of amino acids and their substitution probabilities.
To date, several different BLOSUM matrices exist based on the underlying similarity of the used protein sequences. Matrices labeled with higher values were built on closely

related sequences, whereas low matrix numbers denote more divergent sequences within the applied alignments. BLOSUM80, for example, was defined on more than 80% identical protein sequences. In contrast, BLOSUM62 relies on sequences with more than 62% similarity. BLOSUM62 refers to the default matrix used in the protein BLAST algorithm (see next subsection).

2.9.4 Position-Specific Iterated BLAST (PSI-BLAST)

The Basic Local Alignment Search Tool (BLAST) compares the biological query sequence of proteins or DNA with a library or database of sequences to detect similar sequences based on a specific threshold [43]. According to the query sequence, different BLAST versions exist, collected in the BLAST package. Among these, the Position-Specific Iterated BLAST (PSI-BLAST) identifies distant relatives of a protein [44].
PSI-BLAST first combines closely related proteins to a profile sequence based on a position-specific scoring matrix (PSSM). Next it queries a protein database based on the generated profile to determine relatives to the query protein with distant evolutionary relationship.
PSI-BLAST is available within the BLAST package at `http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE=Proteins&PROGRAM=blastp&RUN_PSIBLAST=on`.

2.9.5 Position-Specific Scoring Matrix (PSSM)

Sequence similarity search methods using profiles such as PSI-BLAST (Section 2.9.4) have been recognized to more sensitively detect also weak relationships compared to strategies based on simple sequence search queries [45]. An applied profile refers to a so-called Position-Specific Scoring Matrix (PSSM) and encodes which residues are observed at each position in a sequence alignment of evolutionary related sequences [44]. A PSSM measures the amino acid substitution frequencies amongst a protein family, and in consequence, enables the assessment of probable amino acid substitutions at each sequence position of each protein within the family. In addition, it allows the identification of remote homologues of a protein.
The quality of a PSSM, however, relies on the availability of protein families and the amount of available evolutionary related sequences.

2.9.6 PSIPRED

Protein structure encodes protein function and thus, the knowledge about 3D structure elements such as the secondary structure building blocks essentially contributes to the analysis of proteins, their interactions and putative dysfunctions. To date, the prediction of secondary structure elements such as helices or beta sheets has become highly accurate using, for example, the state-of-the-art prediction tool PSIPRED [46]. PSIPRED predicts the secondary structure of a protein via a two-stage neural network, which is based on PSSMs (Section 2.9.5) generated by PSI-BLAST (Section 2.9.4).

The prediction process comprises first a generation of a PSSM, next a prediction of an initial secondary structure sequence and finally a filtering procedure.
PSIPRED is available at `http://bioinf.cs.ucl.ac.uk/psipred/`.

2.10 BALL-SNP: material and methods

In this thesis, we present the implemented tool BALL-SNP to identify candidate nsSNVs for computational diagnostics. Details about BALL-SNP are explained in Chapter 4. Important concepts used within BALL-SNP, are summarized below.

2.10.1 The Biochemical Algorithms Library - BALL

The developed software tool BALL-SNP, presented in Chapter 4, is based on the *Biochemical Algorithms Library* (BALL), an application framework for rapid software prototyping in molecular modelling research and drug design [47]. The overall structure of BALL consists of several layers, each providing functionality for a well-defined field, as can be seen in Figure 2.6. In this stratified architecture, top layers depend on lower ones for their implementation.

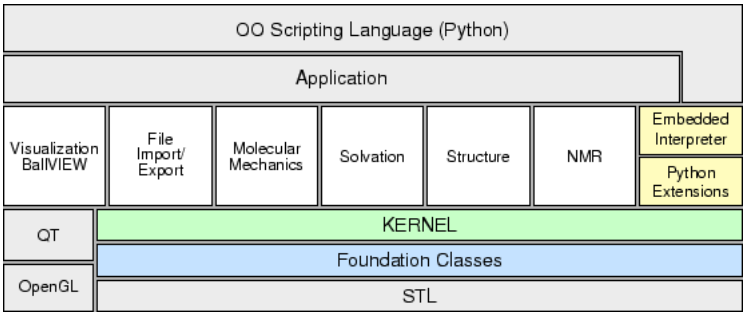


Abbildung 2.6: Overview of the BALL framework structure [48].

The basis of all BALL classes is a comprehensive set of *Foundation Classes* providing general implementations for advanced data structures (e.g. trees, hash maps), mathematical objects (e.g. matrices, vectors), system classes (e.g. file I/O, networking), or design patterns of the C++ standard template library [49]. The BALL *Kernel* contains the data structures for the representation of molecular entities as for example atoms, bonds, molecules and proteins. The third layer consists of several basic components, including molecular mechanics methods, molecular editing, as well as calculation and visualization of electrostatic properties.

Besides this algorithmic fundament, BALL also incorporates a graphical front-end for the visualization of molecular structures, called BALLView [50]. BALLView enables the availability of BALL's broad functionality via an integrated user-friendly graphical user interface (GUI) based on the cross-platform application framework Qt.
In addition, BALL offers a Python scripting interface for rapid prototyping.

2.10.2 Input formats in BALL-SNP

To ensure straightforward usability, we currently offer two different input formats: an *ANNOVAR-based input* [51], as well as a simple tab-separated format, the so-called *BALLformat*. Hence, users are enabled to adopt the output obtained from standard SNP annotation software such as ANNOVAR without substantial re-formatting, as well as use SNV information from different sources compiled in a simple tab-separated input file (BALLformat). Both input formats are below defined in detail.
BALL-SNP focuses on the analysis of the pathogenic relevance of amino acid substitutions introduced by nsSNVs. The SNP calling and annotation process, however, may have great influence on the results of the BALL-SNP analysis. In consequence, the user should carefully adopt the SNP calling and annotation parameters to his application purpose.
Since 3D structure information is essential for the analysis of amino acid substitutions aggregating in one single protein and introduced by nsSNVs in the encoding gene, we automatically extract the PDB identifier of the largest available 3D structure from the UniProtKB [26]. The chosen PDB structure, then is automatically loaded from the Protein Data Bank (PDB) [27].
To maintain flexibility, we also provide the possibility to state a preferred PDB identifier within both input formats or to specify a file name with a user-built 3D model of the query protein in the first line of the input file. The PDB identifier is marked via the flag **PDB:**, whereas an available 3D model file is indicated with the flag **FILE:**.

ANNOVAR-based input

The used parameters for the ANNOVAR call should be carefully chosen by the user based on his application purpose. BALL-SNP does not validate the ANNOVAR output. The file name should end with **_annovar.txt**. Gray-colored text lines in the example are optional. A single point in columns refers to missing information.

PDB:	**2LSQ**					
Line	**Effect**	**Gene**	**Chr**	**Pos**	**rs ID**	**...**
13	nsSNV	ADRB1:... :p.G389R,	chr10	115805056	rs1801253	...
...						

BALLformat input

This tab-separated format can be manually created by the user and may include information gained from different annotation sources. The file name should contain **_BALLformat.txt** at the end. Gray-colored text lines in the example are optional. A single point in columns refers to missing information.

PDB:	2LSQ				
Gene Symbol	**Transcript**	**SNV**	**Chr**	**Pos**	**rs ID**
ADRB1	NM_000684	G389R	chr10	115805056	rs1801253
ADRB1	NM_000684	S475A	chr10	.	.
...					

In addition, further input formats can easily be added.

2.10.3 3D molecular modelling

Since the gap between known protein sequences and available 3D protein structures is still huge and the experimental determination of 3D structures is difficult, techniques to model the missing 3D information of proteins via computational methods have been developed. 3D molecular modelling refers to a template-based protein structure modelling *in silico* [52].
First, a protein with an available 3D structure and a high sequence identity to the query/target protein is searched to serve as scaffold on which the 3D model is built on. In general, this search is performed via multiple alignments of homologous proteins deposited in databases and also includes external data such as secondary structure information, known motifs and conserved features.
The template selection usually relies on specific parameter such as sequence identity, relative alignment length or resolution, while the resolution of an experimentally resolved structure implies the accuracy of atomic coordinates. The sequence identity is calculated by the ratio of matching residues in the pairwise alignment of target and template sequence. Generally, proteins with more than 35% sequence identity reveal homology, while below 35% detailed investigations are required.
Next, based on the selected template, a 3D structural model for the target protein is built. To evaluate the quality of a generated 3D model, various methods are applied, among these the Discrete Optimized Protein Energy (DOPE) score is calculated [53]. DOPE is a pseudo-energetic score to assess the quality of a structure model. In particular, negative DOPE scores correlate with native-like models. The quality of the constructed model, however, critically depends on the quality of the selected template.

Figure 2.7 summarizes the schematic 3D molecular modelling workflow.

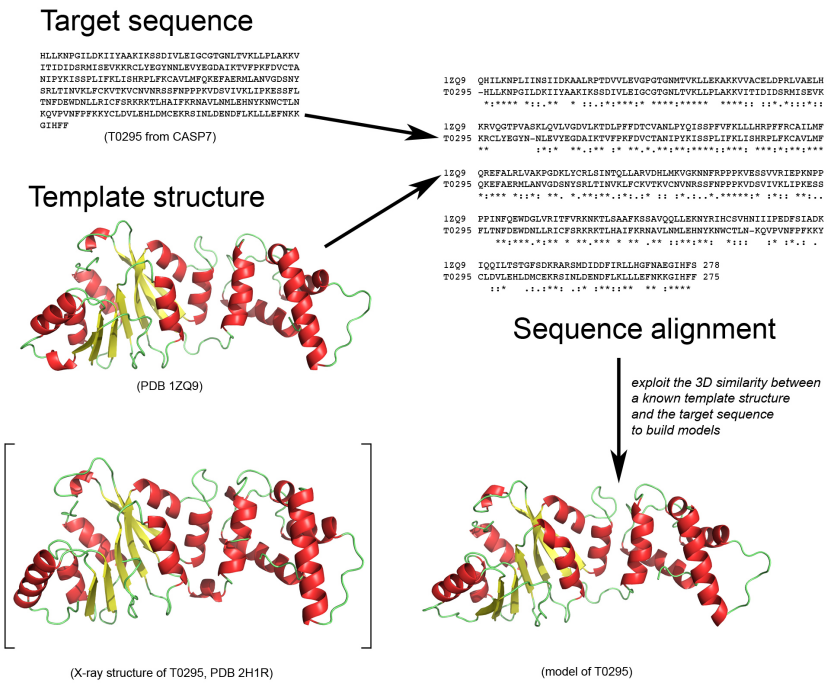


Abbildung 2.7: A schematic overview of 3D molecular modelling [54].

Currently, several software packages for 3D structural modelling are available. One of the most used approaches is the license-restricted toolkit MODELLER [55].

2.10.4 A Database of Comparative Protein Structure Models (ModBase)

Along with computational methods for 3D molecular modelling (see Section 2.10.3), databases to collect already generated 3D models have been built. The most prominent representative refers to ModBase, a database comprising annotated comparative protein structure models built with the automated modelling pipeling ModPipe [56]. ModPipe primarily is based on the toolkit MODELLER [55] in terms of fold assignment, sequence-structure alignment, model building and model assessment.
ModBase currently contains almost 30 million models for about 4.7 million unique protein sequences and is available at `http://salilab.org/modbase`.

2.10.5 Hierarchical cluster analysis

To be able to detect nsSNV close to each other and thus, putative additive effects, we implemented a cluster analysis strategy for nsSNVs.

The goal of a cluster analysis is to partition an amount of objects into groups (cluster) in a way, that the pairwise distances between those assigned into one cluster tend to be smaller than those distributed in different clusters [40].
Hierarchical clustering can be divided into two main strategies: agglomerative (bottom-up) and divisive (top-down). Divisive methods start at the top with one single cluster including all data objects and at each level recursively split one of the existing clusters at the current level into two new cluster. The split is chosen to produce two new groups with the largest between-group distance.
Agglomerative clustering algorithms start with every object representing a singleton cluster. At each step, the two clusters with the smallest distance are merged into one single cluster, sequentially producing one less cluster at the next higher level until all elements end up in one final cluster.
The criteria to link specific clusters are defined based on the pairwise distances between the objects in the cluster. They can be divided into three main concepts (see also Figure 2.8):

- **Single linkage**, also known as nearest neighbor clustering, merges groups according to the distance of the nearest members.
- **Complete linkage** refers to the opposite of the single linkage method, since groups are merged based on the distance of the most remote pair of member objects.
- **Average linkage** defines the distance between the groups as the average of the distances between all pairs of individual objects in the two groups. It can be interpreted as a compromise between single and complete linkage methods, and tends to produce compact cluster.

To measure the distance of two objects $a = (x_a, y_a, z_a)$ and $b = (x_b, y_b, z_b)$ in a 3D space, several common measures have been defined:

- Euclidean distance: $d_E(a,b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2}$
- Euclidean squared distance: $d_E^2(a,b) = (x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2$
- Manhattan distance: $d_M(a,b) = |x_a - x_b| + |y_a - y_b| + |z_a - z_b|$

2.10.6 The Database of Protein interaction SITEs (PiSITE)

Since the biological functions of proteins are driven by their interactions, the knowledge about these are fundamental for the analysis of dysfunctions introduced by genetic variants. The Database of Protein interaction SITEs (PiSITE) collects information concerning protein interaction sites, multiple binding states of a protein or different interaction partner [57]. The identification of interaction sites is based on available PDB

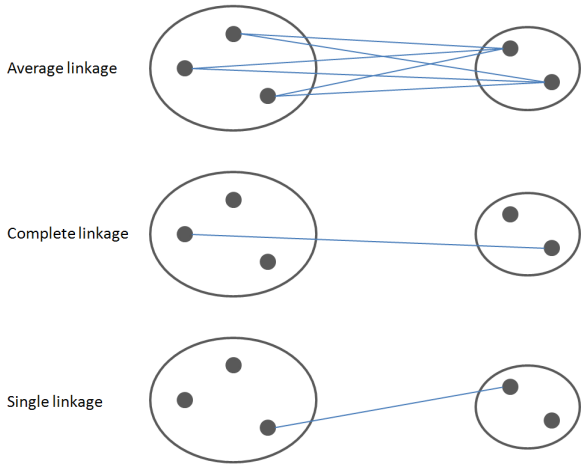


Abbildung 2.8: Example scenarios for the three linkage strategies.

(Section 2.3) protein structures of the same protein, whose binding sites are mapped. In particular, the collected information is provided at the protein residue level, enabling the analysis of functional changes caused by amino acid substitutions.
PiSITE offers a web-based interface as well as downloadable flat files. It is available at `http://pisite.hgc.jp`.

2.10.7 The Exome Aggregation Consortium (ExAc)

These days, due to the advent of NGS, a lot of sequencing projects are performed to get insights into the genotype-phenotype relationship. To be able to compare their results despite of different used calling methods and parameters, a coalition of investigators founded the Exome Aggregation Consortium (ExAC) [58]. Their goal is to curate and harmonize exome sequencing data from large sequencing projects and to consequently provide these data to the scientific community. The raw data of the included sequencing projects has been preprocessed using the same pipeline, and in particular jointly variant-called to increase consistency. We also include this information within BALL-SNP (Chapter 4).

2.11 Multi-scale analysis pipeline: material and methods

In Chapter 5, we present a pipeline for the multi-scale analysis of nsSNV sets. Basic concepts and databases included within this pipeline are summarized in this section.

2.11.1 Association rule learning

To discover strong and/or hidden relations between variables in large data sources, association rule learning is generally applied. Given a set I of n binary items or variables, association rule learning aims to find joint values of items, that are most frequently in the data source [40]. These joint values are defined as implications and are formulated as so-called association rules. A rule r_i comprises two different sets of items X and Y:

$$r_i = X \longrightarrow Y, \tag{2.5}$$

where $X, Y \subseteq I$ and $X \cap Y = \emptyset$.

Different measures of significance and interest are applied to quantify the quality of the generated rules [59]. The most important measures are known as support and confidence. Support refers to a frequency constraint determining the quantitative applicability of a rule, while confidence measures its reliability. Support and confidence are mathematically defined as:

$$support(X) = |\{r_i | X \subseteq r_i, r_i \in R\}|, \tag{2.6}$$

where $|.|$ denotes the cardinality of a set and R refers to the set of all generated association rules.

$$confidence(X \longrightarrow Y) = \frac{support(X \cup Y)}{support(X)}, \tag{2.7}$$

where $support(X \cup Y)$ denotes the support of the union of items in X and Y.

Many algorithms to mine association rules are available, where the methods typically can be split into two parts:

1. Find all frequent item sets, that satisfy a minimum support threshold specified by the user.

2. Define association rules based on the frequent item sets, that also fulfill a minimum confidence constraint defined by the user.

Since association rule learning is a state-of-the-art method in market basket analysis, we made use of this powerful strategy and transfered the method to genomic analysis for the identification of associated nsSNV sets in one gene or related nsSNVs in multiple genes (see Figure 2.9).
In the analysis pipeline presented in Chapter 5, we applied the R package arules [60] using the implemented apriori algorithm [61]. The confidence threshold was set to 0.8 and different levels of support, starting with at least 0.5, were tested.

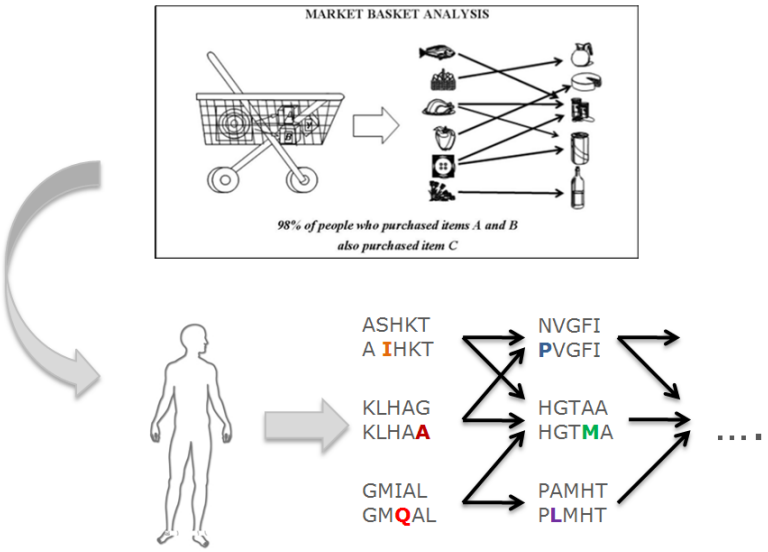


Abbildung 2.9: We transfered the application of association rule learning from the market basket analysis to genomic analysis to identify associated nsSNVs.

2.11.2 The STRING database

The STRING database is a comprehensive resource of known and predicted protein interactions [62]. These interactions comprise direct (physical) and indirect (functional) associations, and are derived from four sources:

- genomic context
- high-throughput experiments
- conserved coexpression
- previous knowledge (e.g. pubmed)

STRING currently covers 9,643,763 proteins [63]. The database is freely available at `http://string-db.org/`.

2.11.3 The Gene Ontology Annotation (GOA) database

The Gene Ontology Annotation (GOA) database provides high-quality Gene Ontology (GO) annotations for proteins deposited in the UniProtKB (Section 2.2.2). The annotations are generated by automatic predictions and manual curation [64]. The applied algorithms for automatic annotation are based on sequence similarity, orthology or domain information as well as existing cross-references and keywords.

Currently, GOA comprises 368 million GO annotations for almost 54 million proteins [65].
It is available at `http://www.ebi.ac.uk/GOA`.

2.11.4 The KEGG database

The use of expert knowledge from e.g. metabolic pathways reveals essential information for the analysis of gene-gene and gene-environment interactions. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database refers to a comprehensive resource of metabolic and regulatory pathways. The KEGG PATHWAY database [66] is a collection of manually created pathway maps, which represent knowledge on molecular interaction and reaction networks for

- metabolism
- genetic information processing
- environmental information processing
- cellular processes
- organismal systems
- human diseases.

These informations promote the biological interpretation of higher-level systemic functions. Currently, about 405,927 pathway maps are deposited.
The KEGG PATHWAY database is freely available at `http://www.genome.jp/kegg/pathway.html`.

2.11.5 naccess

Solvent accessibilities provide an intuitive and quantitatively reasonable idea of the complexity of the molecular interaction network an amino acid residue is involved within a protein [67]. Thus, we also computed solvent accessibilities via naccess [68]. Naccess calculates the solvent accessible area of a protein in a PDB file by defining a probe sphere rolled around the protein's surface. The resulting path of the probe sphere center refers to the solvent accessible surface of the protein. In general, the probe sphere radius corresponds to the radius of the water molecule (1.4 Å), however, naccess also allows user-defined sizes of atoms and the applied probe sphere.
It is freely available at `http://www.bioinf.manchester.ac.uk/naccess/`.

2.11.6 LIGSITEcsc

The identification of binding pockets on protein surfaces is crucial for structure-based drug design applications and protein-ligand docking studies [69]. The LIGSITEcsc is an extended implementation of the well-known LIGSITE algorithm. Instead of using atom coordinates, however, LIGSITEcsc relies on the Connolly surface of a protein to capture surface-solvent-surface events. The identified pockets then are ranked according to the conservation of the involved surface residues.
It is freely available at `http://www.projects.biotec.tu-dresden.de/pocket`.

2.11.7 CELLmicrocosmos

CELLmicrocosmos is an integrative cell modelling and stereoscopic 3D visualization project [70]. To support and visualize the subcellular localization prediction of proteins, CELLmicrocosmos incorporates a PathwayIntegration (CmPI) component (version: CELLmicrocosmos 4.2 PathwayIntegration).
CmPI is connected to DAWIS-M.D., a data warehouse containing a number of databases [71]. It applies a context-based localization prioritization, since each protein usually can obtain different localization entries from databases.
The CELLmicrocosmos including the CmPI is available at `https://www.cellmicrocosmos.org`.

2.12 Used data sets

To be able to analyze existing pathogenicity prediction approaches and develop new strategies to assess disease susceptibility of nsSNV sets, a reliable data set is highly required. The data sets used in this thesis split into patient samples diagnosed with dilated cardiomyopathy (DCM)[72] and healthy control samples from the 1000 genomes project [73]. Details are presented below.

2.12.1 Cardiomyopathy data set

Due to the valuable contribution of the INHERITANCE Project Group, we were able to study a data set containing 842 nsSNVs in 76 genes that are clinically relevant for DCM (known causes and likely candidate genes for DCM) found by studying the genetics in 639 patients with DCM [72]. DCM refers to a frequent disease of the heart muscle (myocardium) and as such belongs to the class of cardiomyopathies [74]. Briefly, the heart's main pumping chamber in DCM patients thins introducing an impaired systolic pump function. Recently, genetic variations have been identified to substantially contribute to DCM [75]. Despite already detected genetic mechanisms contributing

to a DCM cause, there are still unexplained causes [76]. In particular, the observed phenotypes vary with respect to severity and prognosis.
The sequencing of the DCM samples was performed on IlluminaHiSeq instruments [11]. Per patient, roughly 2 billion bases have been sequenced. To ensure diagnostic quality for clinical application, circa 99.1% of the targeted genomic region is covered at least 50-fold. In average, each patient carried 32 nsSNVs in the investigated target region. In consequence, the coverage rate of our data set exceeds those used in classical exome capture studies by several orders of magnitude [11].
Next, we collected available information concerning the data set nsSNVs deposited in available databases [11]. To avoid bias, we collected information available from three different databases: SwissProt/UniProtKB (especially using the HUMSAVAR collection), dbSNP including ClinVar and the HGMD. Further details about the used databases can be found in Section 2.2. When information from more than one of the sources was available (only 5% have information in all three databases), we built a majority-vote-based consensus. While circa 60% are deposited in dbSNP with an rs ID, only 45% have pathogenicity information available [11]. Figure 2.10 represents the detected pathogenicity annotations in ClinVar, HUMSAVAR and the HGMD for the DCM data set.

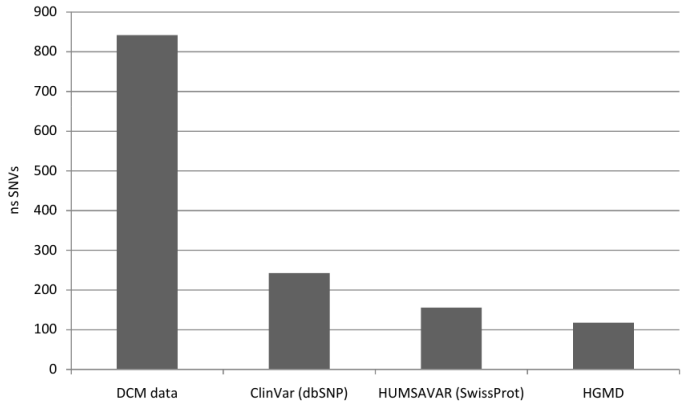


Abbildung 2.10: Detected pathogenicity annotations [11].

The neutral labeled set and the disease-associated set comprise 192 and 147 nsSNVs, respectively [11]. This annotated cardio test set, however, refers to only circa 45% of our whole data set. A total of 55% of the nsSNVs in the DCM data set have no available clinical significance information, and circa 40% have neither an rs ID nor other known identifiers and annotations. This impressively demonstrates the lack of currently available annotation concerning clinical information and phenotypic association of genetic variants deposited in freely accessible databases.

2.12.2 Control data from the 1000 Genomes Project

To be able to evaluate detected putatively DCM-related nsSNV patterns, Dr. Jan Haas from the INHERITANCE Project Group generated a control set based on the general population of the 1000 Genomes Project [73] (1000 genomes). The 1000 Genomes Project sequences multiple genomes to provide a comprehensive resource on human genetic variation.
Jan Haas downloaded the binary files (BAM) including the sequence data from 445 samples and applied the same variant calling and filtering algorithms, as described for the analyzed DCM cohort [72]. In order to match the European INHERITANCE cohort, we only considered individuals with a European descent: Utah residents with northern and western European ancestry (CEU), Finnish in Finland (FIN), British in England and Scotland (GBR), Iberian populations in Spain (IBS) and Toscani in Italy (TSI) [73].

3 Current pathogenicity prediction of nsSNVs - benefits and drawbacks

The invent of Next-Generation Sequencing (NGS) techniques led to a substantial amount of detected genetic variations, still rapidly growing. The vast amount of genetic variant data consequently requires development of automatic procedures to predict the functional and phenotypic effects of nsSNVs [11]. In the previous Chapter, Section 2.5, we listed and briefly described state-of-the-art prediction tools to assess the pathogenic influence of nsSNVs. Since in recent years, a number of approaches dealing with the functional impact of genetic variants on protein function have been developed, the choice of the best performing predictor has become difficult.
In consequence, evaluation studies have been performed to compare the available pathogenicity prediction methods [11]. In 2010, Thusberg et al. analyzed the performance of nine prediction tools on neutral (from dbSNP) and disease-associated (from PhenCode database [77] and IDbases [78]) variant data sets [79]. Castellana and Mazza further studied the uniformity of the predictions of six methods for whole-exome sequencing data [80]. Moreover, Frousios et al. evaluated the prediction performance of nine methods on data from the HGMD and the 1000 Genomes Project, and developed a consensus tool integrating four available prediction methods [81].
In this chapter, we describe and present the results of our study focusing on comprehensive evaluation of prediction concordance and prediction quality of existing tools as well as discuss novel approaches to predict sets of nsSNVs. The following evaluation study of pathogenicity prediction tools for single nsSNVs refers to already published work [11] and constitutes the starting point for our analyses on nsSNV sets.

3.1 Concordance and performance of current state-of-the-art pathogenicity prediction approaches

We systematically explored both, the concordance and performance of existing state-of-the-art nsSNV pathogenicity prediction tools on panel sequencing results of 639 DCM samples [11]. This NGS data set was screened for the full sequence of 76 genes, clinically relevant for DCM and involves 842 nsSNVs. Due to the high coverage rate of at least 50-fold for 99.1% of the target region, the used DCM data set is of high clinical quality (details see Section 2.12.1).
In contrast to previous studies, we extended the number of tested tools and applied

these to the mentioned data set of high-quality. This study was already published in [11]. Figure 3.1 outlines the underlying workflow.
In the following, we state the results most important for the presented thesis.

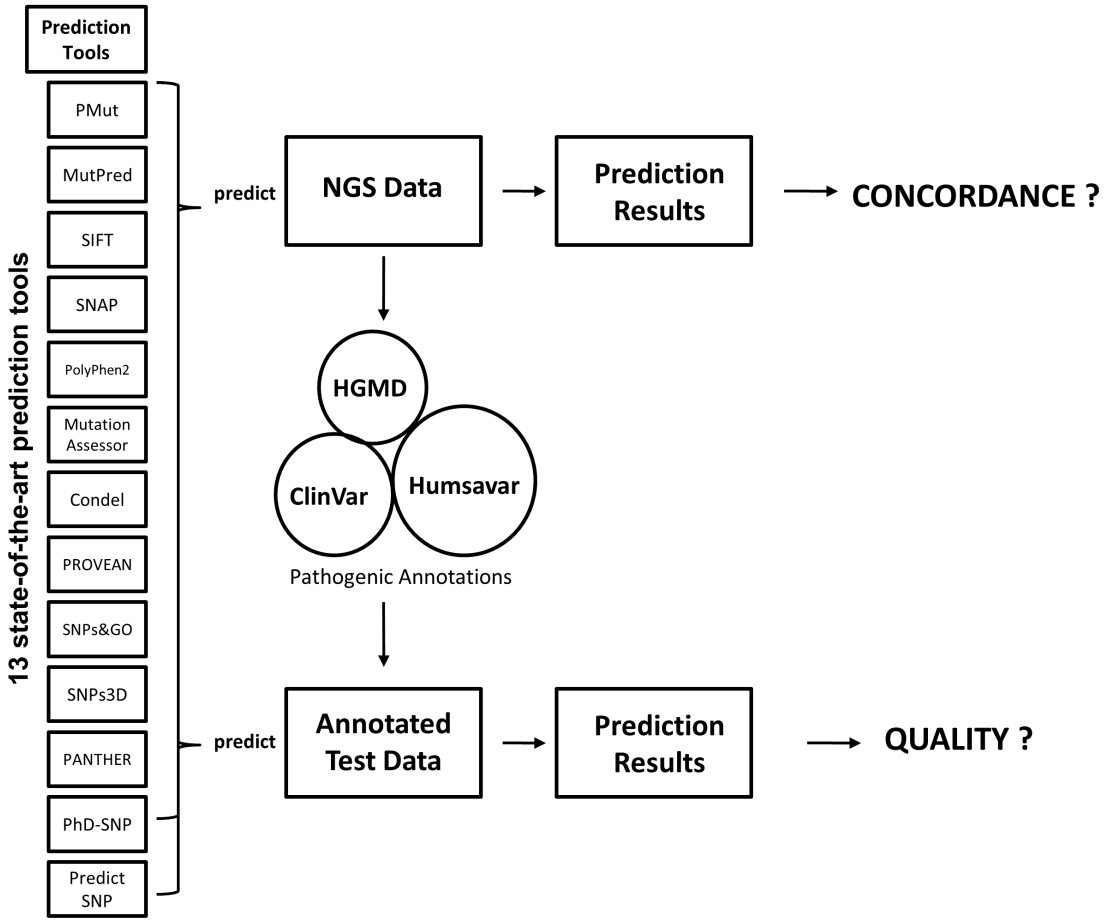


Abbildung 3.1: Graphical overview of the applied test workflow [11].

The selection of the 13 tested tools depended on characteristics important for the majority of pathogenicity analyses: We only considered tools suitable for large-scale studies without prerequisites that are not generally available, such as a homologous 3D structure or a dbSNP ID for prediction. Furthermore, we concentrated on tools, frequently used in the scientific community.
For the calculations, the default parameters proposed and preset by each tool were applied.
The statistical analysis of the 13 state-of-the-art pathogenicity prediction tools comprised two components, the prediction congruency analysis and the prediction quality analysis.

3.1.1 Concordance of prediction methods

Since no pathogenicity information for the complete data set is available, we first focused on the analysis of congruency and association of the state-of-the-art methods. In a first step, we calculated for all 842 nsSNVs in our data set the pathogenicity prediction of the selected 13 tools. There are a dozen of computational tools aiming at functional prediction of nsSNVs, thus, there are also approaches trying to build an unified consensus classification score from them [82][83]. To avoid adding to the complexity, we evaluated the straightfoward majority vote to build a consensus: for each nsSNV, we determined the most frequent prediction result among the single predicted ones. To evaluate the concordance, we computed for each single nsSNV a consensus prediction out of all 13 prediction results and compared the consensus with the prediction of each method, respectively. In addition to overall concordance, we also analyzed the mutual concordance of the tested prediction tools. For each pair of prediction methods, we calculated the similarity of their results concerning all 842 nsSNVs. The similarity score is defined as:

$$Score_{Similarity} = \sum_{n \in nsSNVs} comparison(n) \tag{3.1}$$

$$comparison(n) = \begin{cases} 1 & \text{if } prediction\ of\ Tool\ A\ and\ B\ for\ n\ are\ equal \\ -1 & \text{otherwise} \end{cases} \tag{3.2}$$

We studied both, the overall concordance of one prediction tool compared to the consensus of all prediction methods and the mutual agreement among all methods.
To evaluate the concordance of the 13 prediction methods, we determined the distribution of obtained prediction results on the complete data set of 842 nsSNVs (Section 3.2).
Although SIFT [84] and MutationAssessor [85] predicted about 50% of the 842 nsSNVs to be disease-associated, the other methods proposed the majority of the nsSNVs to be neutral. SNPs3D [86] (90%), PANTHER [33] (35%), MutationAssessor (29%), and PolyPhen2 [35] (23%) failed to predict all of the 842 nsSNVs. In contrast, MutPred [87], SNPs&GO [88], PhD-SNP [34], SNAP [89], PMut [90], PredictSNP [83] and PROVEAN [37] had a prediction failure rate of less than 4%. Figure 3.3 illustrates the comparison of each prediction tool to the consensus prediction result built using all methods. PredictSNP, PROVEAN, PhD-SNP, SNPs&GO, and Condel achieved the best conformity to the overall consensus.
Next, we created a network via Cytoscape [91] based on pairwise comparison of similarity scores of the prediction results to evaluate the mutual agreement among the tested

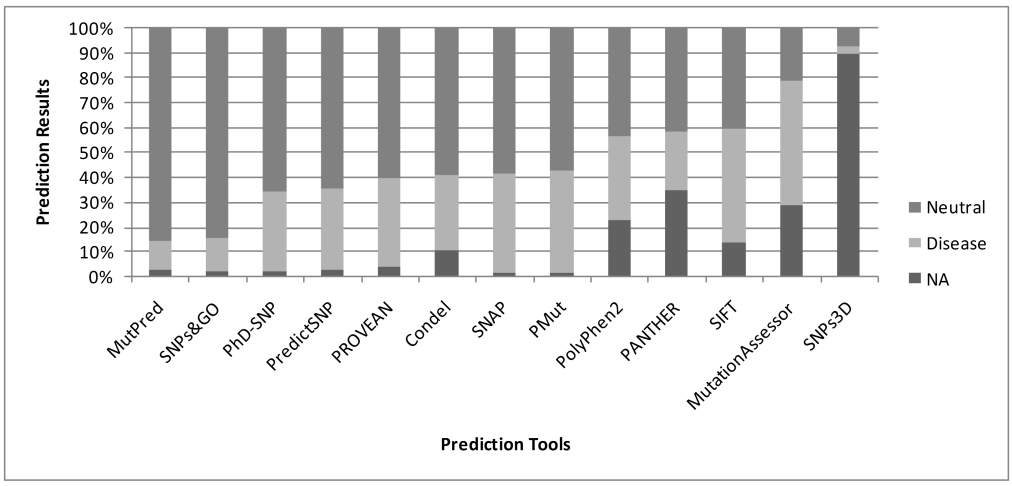


Abbildung 3.2: Distribution of obtained prediction results for the complete data set of 842 nsSNVs [11].

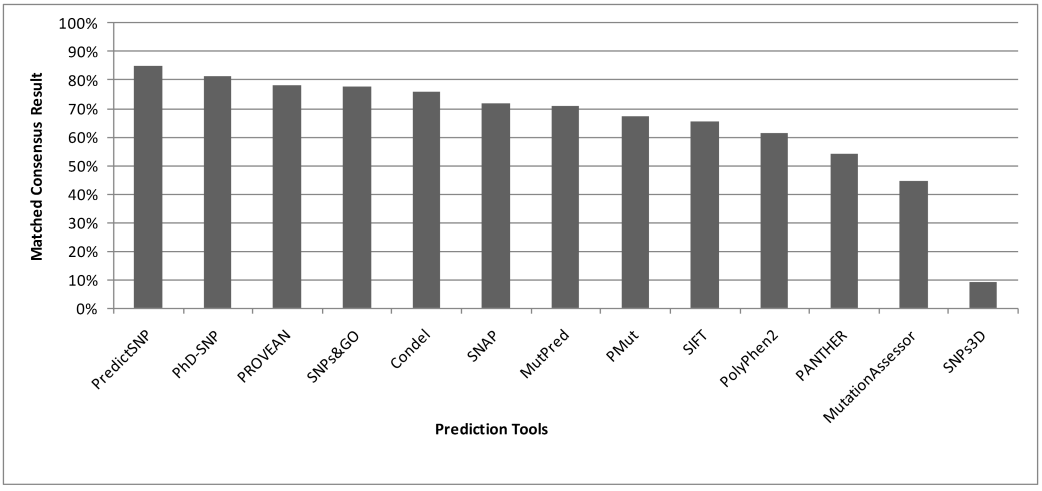


Abbildung 3.3: Comparison of single prediction with consensus prediction results for the complete data set of 842 nsSNVs [11].

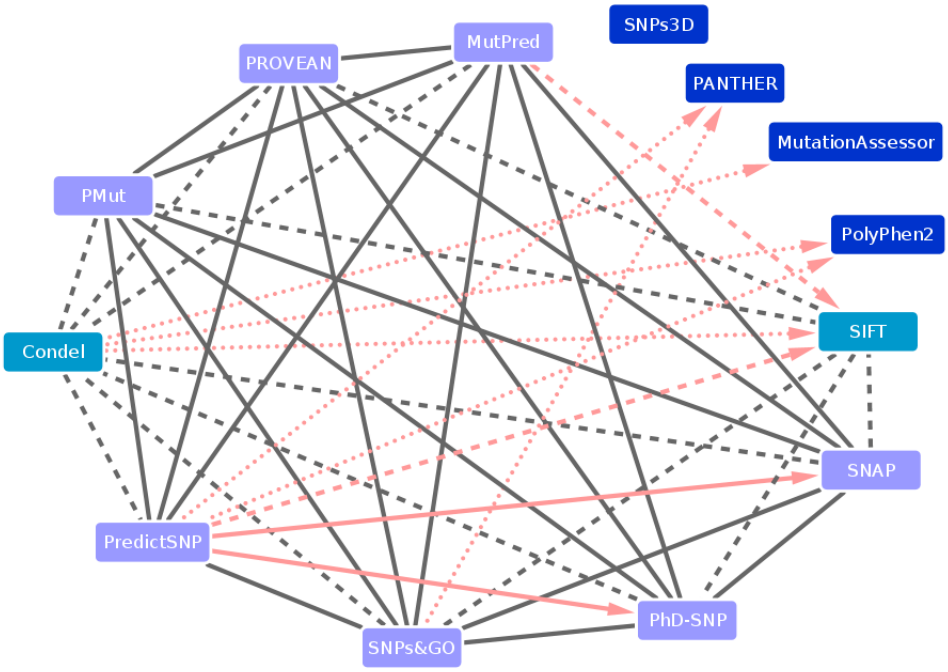


Abbildung 3.4: Network of prediction concordance. Nodes represent prediction tools, while edges mark the pairwise prediction similarity for connected nodes. The dashed edges refer to a similarity of about $70 - 80\%$ and the bold line edges to a similarity of $> 80\%$. For clarity, edges with less similarity value are neglected, except if one method includes another (pointed edges). Pink arrow edges point to included tools. The color of the nodes codes for concordance: purple = high, turquoise = moderate and blue = low [11].

methods (see Figure 3.4). For clarity, only edges marking at least 70% pairwise similarity or connecting methods that incorporate other tools were recognized.
SNPs3D, PANTHER, MutationAssessor, and PolyPhen2 revealed the worst conformity with all other tested tools. In contrast, PredictSNP, PROVEAN, MutPred, SNAP, PMut, SNPs&GO, and PhD-SNP obtained the best concordance values with more than 80%. Except PROVEAN, these tools use machine learning classifiers for pathogenicity prediction, and at least four include structural annotations.
Condel and PredictSNP build consensus predictions based on the single predictions of other methods. Interestingly, machine learning-based applications cluster very well, indicating the chosen classifier method to be essential for the prediction outcome (see Figure 3.5). The underlying classification method even reveals greater influence on the overall concordance than methods incorporating others. Moreover, tools including others show not necessarily equal performance. These findings agree with previous studies [79].

3.1.2 Performance of prediction methods

For 45% of the DCM data set, we were able to extract available pathogenicity annotations from SwissProt/UniProtKB (HUMSAVAR collection), dbSNP including ClinVar and the HGMD to create test sets for the performance evaluation of the selected prediction tools. The neutral labeled set and the disease-associated set comprise 192 and 147 nsSNVs, respectively (details see Section 2.12.1).
In consequence, we were able to measure the prediction quality of the tested 13 prediction tools on the generated test set of 339 annotated nsSNVs. We calculated the confusion matrices [true positives (TP), true negatives (TN), false positives (FP), false negatives (FN)] and consequently accuracy, specificity and sensitivity for the results of each single prediction tool (detail see Section 2.9.1). We also computed the balanced accuracy and the Matthews correlation coefficient (MCC) because the distribution of available annotations concerning neutral and disease-associated nsSNVs is slightly imbalanced (192 to 147).
The results of the prediction performance of the 13 tested tools in our study are shown in Figure 3.6.
The best performance concerning balanced accuracy and sensitivity in combination with prediction ability (NA ratio) was reached by MutPred with 66% accuracy, 96% specificity, 28% sensitivity, 62% balanced accuracy and 0.32 MCC. Despite promising balanced accuracy values, SNPs3D could classify only about 5% of our generated test set. The remaining tools frequently showed a low hit ratio. In addition to SNPs3D and MutationAssessor, most methods revealed much higher specificity values compared to sensitivity. The mean values for accuracy (60%), specificity (69%), sensitivity (49%), balanced accuracy (59%) and MCC (0.20) show the current limitations of nsSNV pathogenicity prediction.
Furthermore, we determined the performance quality of the consensus prediction result,

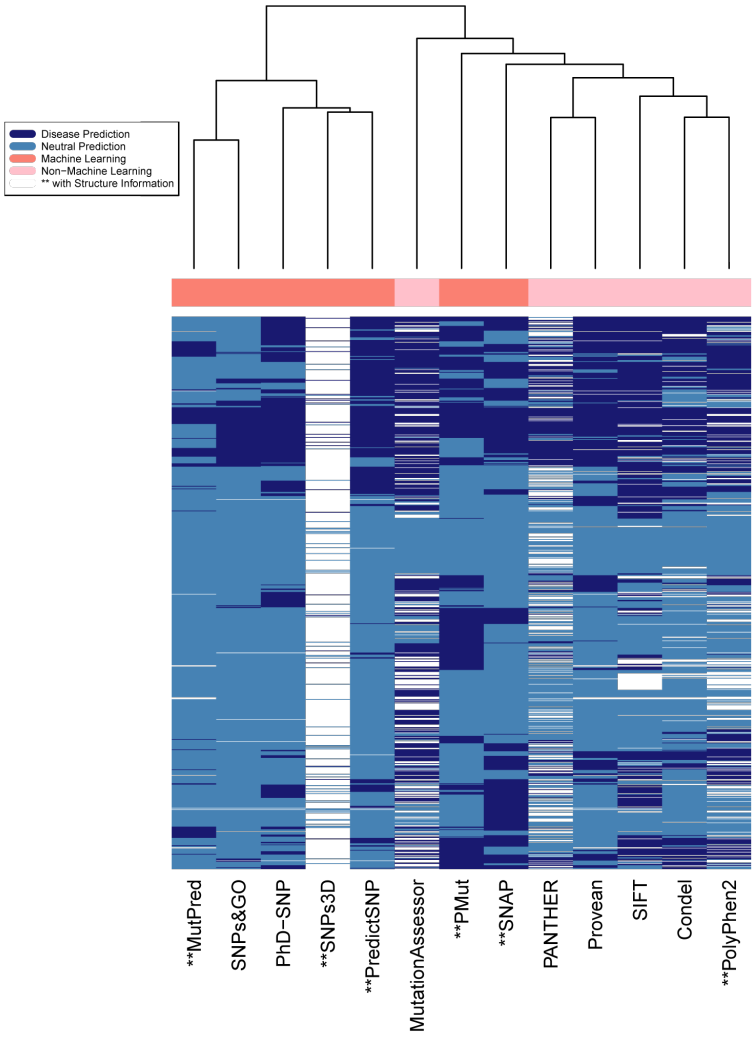


Abbildung 3.5: Heatmap of the prediction results obtained by the 13 state-of-the-art prediction tools. Tools based on machine learning cluster well, indicating that the chosen classifier method is essential for the prediction outcome [11].

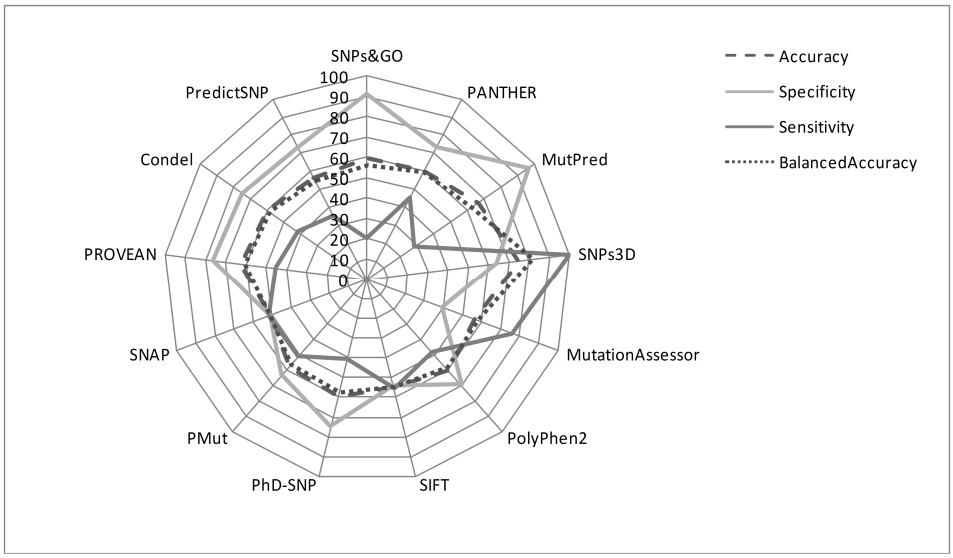


Abbildung 3.6: Prediction results of the 13 tested state-of-the-art tools on 339 nsSNVs with available pathogenicity annotations [11].

computed for each nsSNV during congruency analysis. Cases with balanced 50% vote were excluded within calculation of prediction measures. Referring to the methods' close correlations, we also clustered related methods to build a consensus prediction of non-overlapping tools and calculated a kind of nested majority vote. First, we determined the majority vote for related methods. Next, we determined the majority vote over all these sub-majority votes. In particular, we compared structure-based and sequence-based methods to try to improve the prediction results. We discriminated between the structure-based group (MutPred, PMut, SNAP, SNPs3D, PolyPhen2) and the sequence-based group (PROVEAN, SNPs&GO, SIFT, PANTHER, PhD-SNP, MutationAssessor). We also clustered related methods, namely methods using the same classification method or prediction features: machine learning-based group (MutPred, SNPs&GO, SNPs3D, PhD-SNP, SNAP, PMut) and the non machine learning-based group (PANTHER, PROVEAN, Condel, PolyPhen2, SIFT, MutationAssessor). Since PredictSNP builds a consensus of sequence- and structure-based methods as well as incorporates machine learning and non machine learning-based tools, we excluded this method from our consensus calculation. The resulting quality measures are presented in Table 3.1.

The consensus predictions of machine learning (ML) and structure information-including methods reveal slightly improved accuracy values. We also built a consensus of the results obtained by structure-based and sequence-based methods, which yielded the best consensus prediction result with 65% accuracy (balanced accuracy 63%) and about 63% sensitivity. In addition, the sequence structure consensus as well as the ML consensus

Tabelle 3.1: Prediction results of clustered prediction methods. Sequence consensus: consensus prediction of sequence-based group; structure-consensus: consensus prediction of structure-based group; sequence structure consensus: consensus prediction from sequence and structure consensus; Machine Learning (ML) consensus: consensus prediction of related methods divided into the machine learning group and the non-machine learning group [11].

	Accuracy	**Specificity**	**Sensitivity**	**Balanced accuracy**	**MCC**	**NA**
Sequence consensus	57.7	63.44	50.34	56.89	0.14	2.4
Structure consensus	62.73	76.24	45.39	60.82	0.23	5.0
Sequence structure consensus	**65.0**	**66.0**	**62.86**	**64.43**	**0.27**	**0**
ML consensus	63.74	64.43	61.76	63.1	0.24	0

were able to return a result for each nsSNV in the data set.

3.1.3 Discussion

We analyzed the concordance and performance of current state-of-the-art pathogenicity prediction tools to identify the best method to prioritize nsSNVs as disease cause. During our extensive analysis, we identified several critical drawbacks within the current state-of-the-art pathogenicity prediction strategies.
Many existing prediction methods are not well suited for large-scale studies with real-life data. Often only server-based applications are available, and/or the input is restricted to single sequence variants in one query.
In addition, a major problem concerning all computational methods and databases is the maintenance of the developed software. Rare updates lead to obsolete annotation linkages and can even negatively influence classification results. In fact, some of the available supposedly neutral nsSNV data sets used in former studies contain disease-associated mutations according to actual database entries. We identified, for example, some variants in the neutral VariBench data set of Thusberg et al. [79], as disease-associated, with entries in the HGMD [6]. The major problem in general refers to the limited availability of suitable data and, particularly, high-quality data. Often data sets are constructed from information contained in one particular database without cross-checking these information in additional databases. Some information missing in one database might be available in another. Sometimes, even annotations from different

data sources disagree. In particular, the construction of a neutral labeled data set is highly challenging. Hence, we constructed our positive and negative data sets as consensus out of the three most popular databases for nsSNVs. Based on our findings, we recommend using a consensus of different available data sources to avoid biased data sets in future studies.
The choice of the data and the data quality play an essential role in performance and evaluation studies. Especially the data quality in former studies often varies: Frousios et al., for example, based their analysis partly on low-coverage whole-genome sequencing data [81]. In contrast, our analysis focuses mainly on high-coverage data for targeted sequencing with 99.1% of the targeted genomic region covered at least 50-fold.
Furthermore, none of the currently available approaches consider neighboring nsSNVs or the influence of several nsSNVs. A human individual usually carries more than one nsSNV, and from a medical point of view, especially, the individual combination of nsSNVs plays a crucial role in clinical diagnostics concerning e.g. personalized medicine.

3.1.4 Familial study on glioblastoma multiforme (GBM)

In a familial exome sequencing study of unaffected parents and their two siblings diagnosed with glioblastoma multiforme (GBM), we were able to identify genes with accumulations of homozygous and heterozygous germline variants within both siblings [92]. GBM is the most aggressive and malignant subtype of human brain tumors. The identified accumulations of homozygous and heterozygous variants could not be detected in the healthy parents and thus, might contribute additively to the siblings' observed phenotype. Figure 3.7 exemplarily illustrates the encoded protein structure of Chitinase-3-like protein 1 (*CHI3L1*), one of the detected genes with variant accumulations.
CHI3L1 plays an important role in the regulation of malignant transformation and local invasiveness in gliomas, since it is highly expressed in human glioma tissue. Within our samples of the two siblings, *CHI3L1* revealed several homozygous and heterozygous mutations not present in the parents' exome.
Unfortunately, computational methods to analyze putative synergetic effects of multiple variants in one gene on the cause and severity of a disease are currently missing. In consequence, we aimed to improve this unsatisfactory situation by developing novel strategies to promote the assessment of nsSNV sets in one gene and its coding protein.

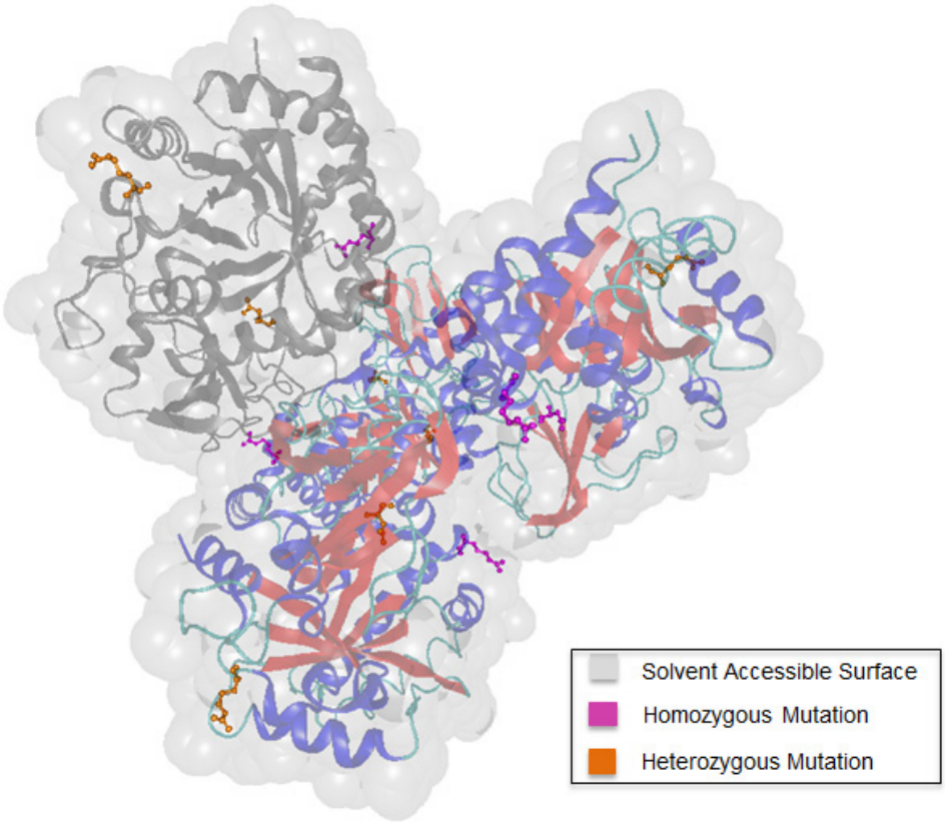


Abbildung 3.7: 3D structure of chitinase-3-like protein 1 (*CHI3L1*) [92]. The four chains of *CHI3L1* are colored according to their secondary structure elements. To highlight the distribution of the detected mutations within one chain, chain C is colored in grey. We distinct between homozygous (pink) and heterozygous (orange) mutations.

3.2 From single nsSNV prediction to the assessment of nsSNV sets

The major drawback of probably all currently available pathogenicity prediction tools is their restriction to one single nsSNV per gene or one single mutation per protein, respectively. This 'one-SNV, one-phenotype'-paradigm, however, is not able to cover the whole range of nsSNV impact on a protein's function regarding diseases such as diabetes or GBM [8]. The computational features suitable for the detection of putative synergetic effects of several nsSNVs in one gene, though, are limited.
In order to develop an approach to assess the effect of multiple nsSNVs, we first tried to detect mutation patterns specific for disease enabling the formulation of classification rules for the pathogenicity prediction of nsSNV sets on the example of the DCM data set of 639 samples (see 2.12.1). To this end, we analyzed amino acid distributions at wild type and mutant positions as well as patterns of amino acid substitutions specific for DCM-linked nsSNVs for the 339 annotated nsSNVs within the DCM samples.
Next, we tried to implement computational scores discriminating between disease-associated and neutral nsSNV sets by adaption of strategies used for the pathogenicity prediction of single nsSNVs. In particular, we focused on features available for the majority of NGS data, such as e.g. the protein sequence, instead of 3D structure information, which is unfortunately often missing.
In the following, we refer to the wild type (WT) amino acid as the unmutated position in the reference genome and the mutant (MT) amino acid as the position in the patient data differing from the reference genome.

3.2.1 Analysis of amino acid distributions in DCM samples

Since we have been able to create a test set of neutral (192) and disease-associated (147) nsSNVs within the DCM data based on annotations in SwissProt/UniProtKB (HUMSAVAR collection), dbSNP including ClinVar and the HGMD, we calculated the amino acid distribution in the neutral and disease labeled DCM test set (see Section 2.12.1) as well as frequencies for WT and MT residues. In particular, we compared the observed frequencies with the expected values determined based on the distribution of all amino acids in the test set. The expected values are also computed with regard to codon diversity and the probability to generate a certain substitution referring to all possibilities for a translation from a specific amino acid to another by only one triplet base mutation of the WT codon. The applied formulas for expected and observed frequencies are defined in Section 2.9.2.
To compare WT and MT residues within substitutions, we computed a BLOSUM62 matrix-based dissimilarity score [93], defined as:

$$Score_{diss} = |BLOSUM62(WT\ aa, MT\ aa) - BLOSUM62(WT\ aa, WT\ aa)|$$

where aa abbreviates amino acid.
Since the identity of BLOSUM62 equals the maximum value and less similar substitutions tend to receive negative values, the higher the score the less similar are the substituted amino acids.
In general, the resulting amino acid distributions at WT and MT residues are quite similar (see Figure 3.8). Except at MT positions in the disease-labeled DCM data, arginine (R) is observed most often, even more frequently as expected (Figure 3.8, part C). R, together with serine (S) and leucine (L), refers to the three amino acids with the most coding triplets (6). Moreover, R was detected as the most frequently mutated WT residue in the considered data set, however, it also reveals the highest occurrence within the analyzed DCM data. Hence, the probability to mutate one of the R residues compared to others is higher.
In contrast to the neutral-labeled DCM set, methionine (M) and tryptophan (W) reveal higher observed frequencies in the disease-annotated DCM data as expected (Figure 3.8, part D). Interestingly, M and W are encoded by only one single codon, and thus, the probability to obtain a M or W residue by one single base mutation is lower compared to the rest of the standard amino acids. Khan et al. analyzed the mutational spectrum of amino acids within protein secondary structure elements such as helices, beta strands and turns. According to their study, M is the only significant MT residue concerning alpha helices [41]. Together with cysteine (C), M is one of two sulfur-containing amino acids. Its thiol side chain also reveals high reactivity [21]. W is the less most occuring amino acid among the 20 standard amino acids. Characteristic for W is its bicyclic structure, consisting of a six-membered benzene ring fused to a five-membered nitrogen-containing pyrrole ring [21]. A substitution resulting in an inserted M or W residue may alter binding affinities within a protein because of the higher reactivity of M or introduce steric clashes due to the size of W modifying a protein's folding. These effects may promote disease-associated dysfunctions of proteins.

Furthermore, we investigated the relationship of mutations correlated with their physicochemical properties. To this end, we distinguished five groups: non-polar (G, A, V, L, M, I), polar (S, T, C, P, N, Q), aromatic (F, Y, W), positively charged (K, R, H) and negatively charged (D, E) amino acids. Explanations for the amino acid abbreviations can be found in 6.1. The results are illustrated in Figure 3.9. In general, non-polar residues represent the majority of WT and MT residues, respectively. Besides these, aromatic amino acids are most frequently detected at neutral-annotated MT positions of the DCM set. In contrast, the majority of disease-associated MT positions in the DCM data is composed of positive-charged residues.

Beyond the single amino acid distributions, we also studied whether the spectrum of combination of WT and MT residues reveals substitutions specific for disease or neutral-annotated DCM data. Based on all MT residues, which can result from the observed WT by one base change within its triplet, we determined the most frequent mutations and calculated the probability to generate certain substitutions (formula

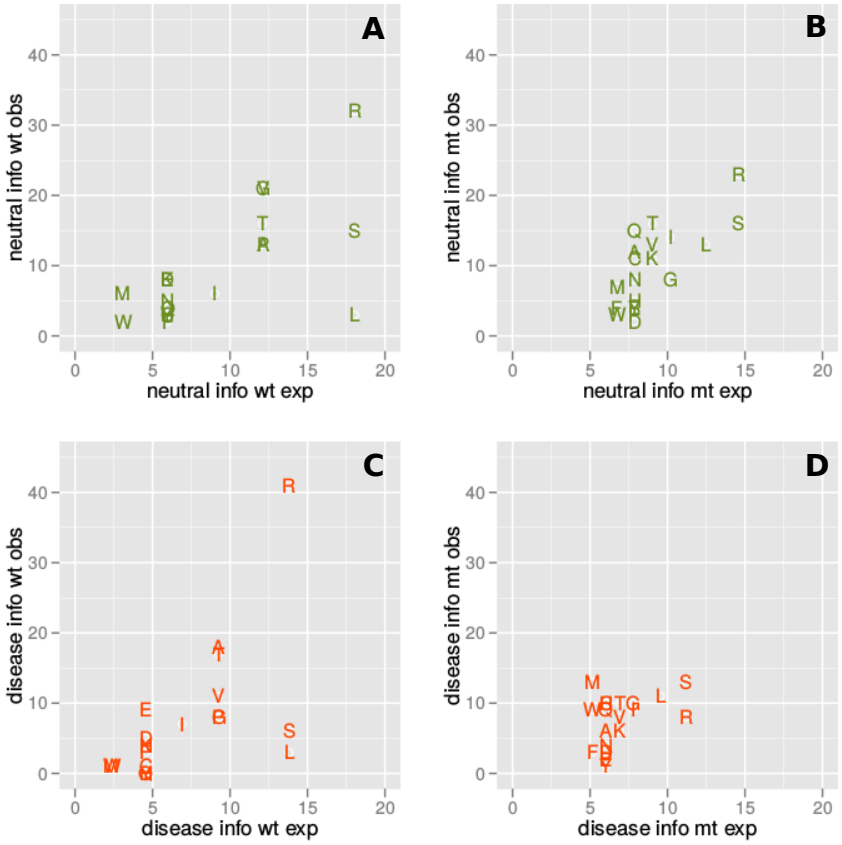


Abbildung 3.8: Distribution of amino acids at WT (wt) and MT (mt) residues in neutral and disease set of the DCM samples. The axes are labeled by the calculated expected versus the observed occurence of the WT and MT residues, respectively. "Neutral info wt obs" e.g. refers to the observed WT occurence within the neutral labeled data set, while "disease info mt exp" denotes the expected occurence of the MT residue within the disease-associated data, and so on.

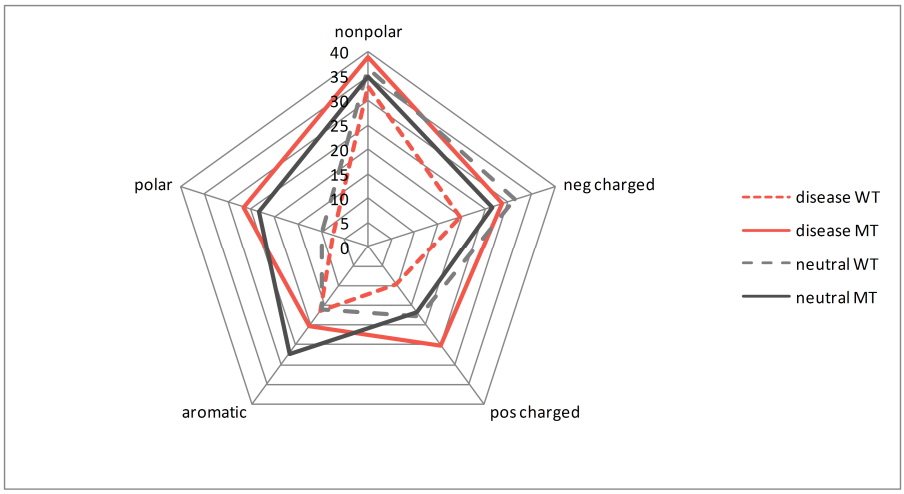


Abbildung 3.9: Distribution of amino acids with certain physico-chemical properties at WT and MT residue positions. The red-colored lines denote the results for disease-labeled DCM data, while the gray printed lines refer to the neutral-annotated DCM set.

definitions see Section 2.9.2). To study WT and MT residues within substitutions, we computed a BLOSUM62 based similarity score as described above. The obtained results are displayed in Figure 3.10. Substitutions with low similarity of the involved WT and MT amino acids in combination with the highest frequency values are R-C, P-L, R-W, G-R, T-M, G-S, and T-I. Interestingly, only P-L and G-R among these substitutions have with 16% and 25% the highest probability values based on codon diversity for a transition from the WT to the detected MT residue by on base change within the WT triplet. In the remaining cases with probability values from 1.2% to 8.3%, a subsitution to a different MT residue would have been more probable according to the genetic code.

In general, no distinct differences in the amino acid substitutions could be determined between the disease and neutral-labeled data sets according to observed and expected frequencies, respectively. Except the substitution T-M could be observed noticeably more often than expected in the disease-annotated data. This at least agrees with the findings of single amino acid distributions, since a substitution to M reveals higher observed frequencies in the disease-annotated DCM data as expected.

The definition of putative mutation rules specific for pathogenic relation of nsSNVs based on the identified amino acid substitution patterns, however, was not possible. The observed results did not allow a distinct classification of amino acid substitutions in DCM patient samples into pathogenic or benign. The identified difference of amino

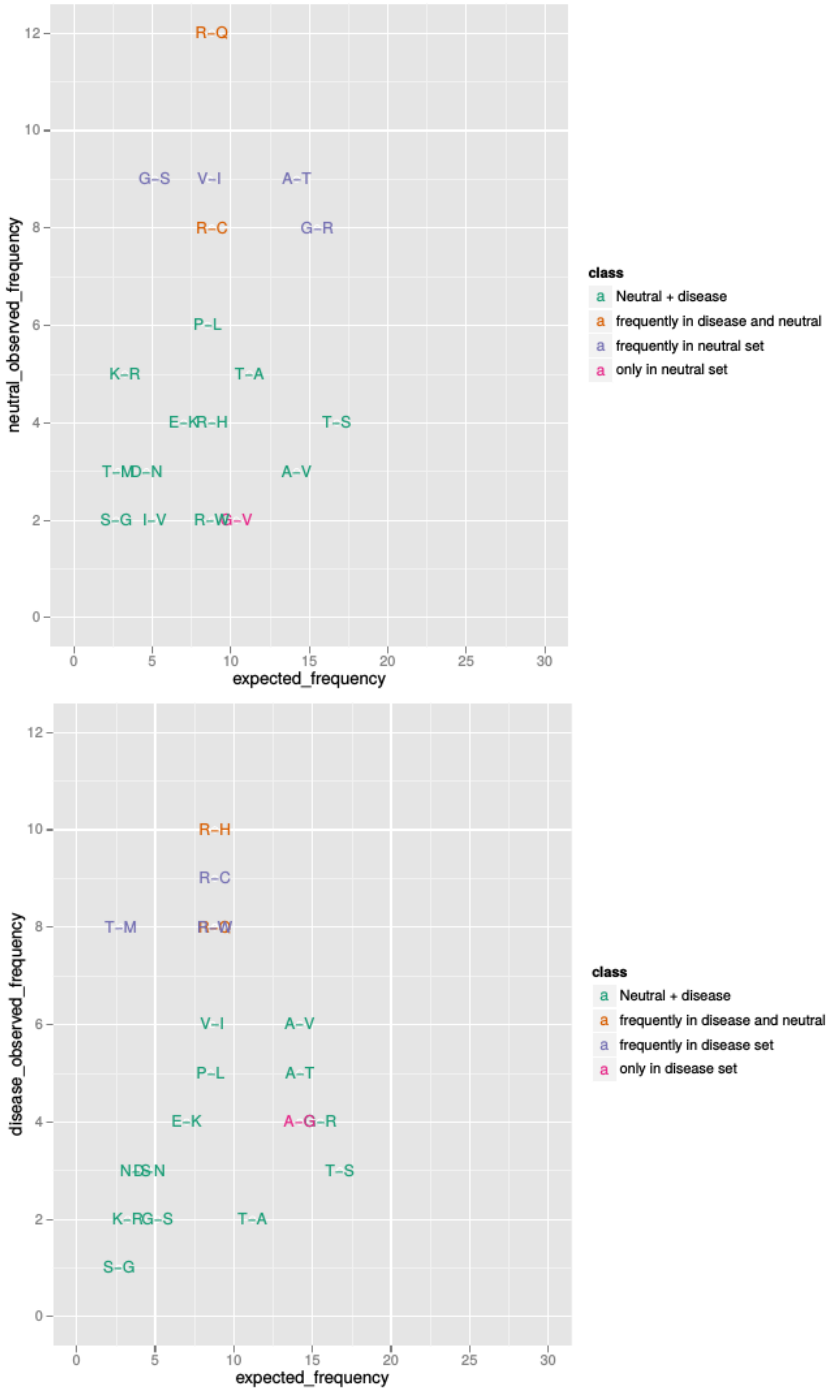


Abbildung 3.10: Amino acid substitution frequencies in both, neutral and disease data set.

acid substitutions between these two groups are only marginal and thus, not significant enough for a classification. Unfortunately, we were not able to determine an exceeding subsitution pattern in one of the analyzed sets in the DCM cohort.
In fact, the distribution of the amino acid substitutions and the substitution pairs in both, disease and neutral-labeled DCM samples are presumably not the major cause for a detectable, functional impact and do not provide the capability to promote nsSNV assessment.
Next, we tried to adapt existing features for single nsSNV prediction to computationally predict the effect of nsSNV sets.

3.2.2 Definition of pathogenicity prediction scores for nsSNV sets

To measure the effect of nsSNV sets including putative additive effects and consequently establish a prediction strategy, we require sophisticated scoring functions. Based on systematic studies of applied scoring methods in single SNV prediction, we developed the following measures:

- BLOSUM Score
- PSSM Score
- Secondary Structure Score

All defined scoring methods compare WT and MT sequence, while the MT sequence includes all amino acid substitutions within a protein detected in one patient sample. In the following, we first present the detailed score definitions and then, we describe the corresponding performance tests and the analysis of the obtained results.

BLOSUM Score

Some of the established approaches for single SNV prediction are homology-based methods and thus, perform nsSNV classification into disease-associated or benign via substitution matrices such as BLOSUM62 to judge the similarity of amino acid substitutions in interchanges between homologous proteins [93] (details in Section 2.9.3).
The formulated BLOSUM Score computes the similarity of WT and MT sequence using the BLOSUM62 substitution matrix:

$$Score_{BLOSUM} = 1 - \frac{\left| \sum_{AA_{wt} \in seq_{wt}} BLOSUM62[AA_{wt}] - \sum_{AA_{mt} \in seq_{mt}} BLOSUM62[AA_{mt}] \right|}{length(seq_{wt})} \tag{3.3}$$

where AA refers to amino acid, wt codes for WT and mt for MT.
Since identical amino acids receive maximum values in the BLOSUM62 matrix, low $Score_{BLOSUM}$-values indicate the substitution of amino acids with completely different properties and hence, may alter protein stability and/or function. A detailed threshold for discrimination of disease and neutral mutations, however, requires systematic tests on high-quality data (see Section 3.2.3).

PSSM Score

Amino acid residues that have been conserved within a protein family more likely play an important role in protein function compared to unconserved ones, since they frequently encode for functional important sites of the protein [94]. To translate the evolutionary information profiles of single mutations to multiple substitutions, we built a conservation-based PSSM score to compare WT and MT protein sequence.
We used the position-specific scoring matrix (PSSM) to create an evolutionary profile score for the mutated amino acid positions (Section 2.9.5). The PSSM takes into account which residues are observed at each position in a sequence alignment of evolutionary related sequences [44]. It measures the amino acid substitution frequencies amongst protein families identified and curated in databases. In consequence, the developed PSSM Score describes the degree of conversation of substituted positions and the evolutionary occurrence of the introduced amino acids at these positions (Section 2.9.5). Amino acid substitutions, which do not comply with the substitution profile of a protein family, indicate destructive influence on the corresponding protein and consequently may contribute to the cause or severity of a disease.
The PSSM score is defined as:

$$Score_{PSSM} = \frac{\sum_{\#AA} PSSM[AA_{wt}] - PSSM[AA_{mt}]}{length(seq_{wt})}, \tag{3.4}$$

where '#' refers to 'number of', AA refers to amino acid, wt codes for WT and mt for MT. For the calculation of the PSSM matrices for WT and MT sequences, we applied the available software tool PSI-BLAST from the BLAST package version 2.2.28 (see Section 2.9.4).

Secondary Structure Score

Structural information is essential when discriminating nsSNVs affecting protein function from functionally neutral ones [11]. Since the arrangement of structural building blocks plays an important role in specific protein folding and hence, modifications may have a tremendous effect on protein stability and function, we compared the sequence of secondary structure elements of WT and MT protein sequences.

The determination of the secondary structure based on sequence information alone is by now highly accurate using efficient computational methods such as the state-of-the-art tool PSIPRED [46]. In consequence, we used the PSIPRED tool version 3.3 (see Section 2.9.6) to perform secondary structure prediction and on the basis of these results as well as the properties of secondary structure elements, we developed the Secondary Structure Score ($Score_{SecStruct}$), defined as:

$$Score_{SecStruct} = 1 - \frac{\#mismatches}{length(seq_{wt})}, \tag{3.5}$$

where '#' refers to 'number of', *wt* codes for WT and a mismatch is valid if at least the two preceding positions in the alignment of WT and MT secondary structure also differ.

Figure 3.11 illustrates a simple example.

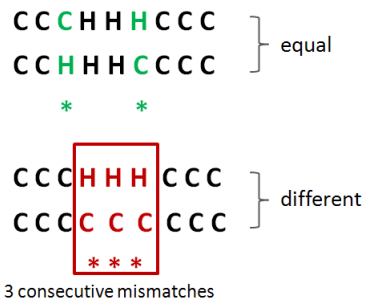


Abbildung 3.11: Example for the defined secondary structure score. The secondary structure elements are denoted as: C = coil, H = helix and E = sheet. A shift by one position may occur due to the prediction method. In particular, we defined three consecutive mismatches as lower boundary for a modified secondary structure sequence.

A shift by one position within the secondary structure may occur due to the prediction method. Moreover, a helical turn comprises approximately four amino acids and the disruption of a helix is supposed to critically influence protein stability [95]. Therefore we account for mismatches only cases where three consecutive alignment positions differ.
In fact, we were able to detect modifications in the secondary structure introduced by amino acid substitutions. Interestingly, these changes not necessarily occurred exactly at the substituted position or close to it indicating a potential influence of tertiary interactions on the mutational impact.

3.2.3 Analysis of the defined prediction scores for nsSNV sets

To test whether the designed scores are suitable to capture putative synergetic effects of nsSNV sets, we selected 58 test samples including 8 genes from the DCM data set (Section 2.12.1). These samples denote different nsSNV sets in one gene detected in DCM patients. Furthermore, the selected test samples comprise minimal overlaps of one or two nsSNVs within one gene. Due to these overlaps, we were also able to analyze the influence of a particular nsSNV in combination of several others. An example for 3 selected test samples refers to:

Gene Symbol	**Transcript**	**nsSNV sets**
ADRB1	NM_000684	S49G, G389R, V5A
ADRB1	NM_000684	S49G, G389R
ADRB1	NM_000684	A187V, G389R

The small size of the data set and the selected samples allowed for a precise analysis of the characteristics and strengths of each defined score in two aspects: the comparison of the scores to each other as well as with respect to the detection capability of accumulation effects.
Referring to this small test set, we calculated $Score_{BLOSUM}$, $Score_{PSSM}$ and $Score_{SecStruct}$ values for the overlapping nsSNV sets. Unfortunately, we were not able to detect significant differences between the related nsSNV sets. Though the most promising of the scores refers to the Secondary Structure Score, its validation requires extensive studies to analyze which modifications of the secondary building blocks in fact have an impact on protein folding and function.
The validation of the calculated scores, however, emerged to be more difficult than anticipated. To the best of our knowledge, currently no data on nsSNV sets and their corresponding functional impact exist, and thus, neither reference scores nor scores for comparisons enabling a threshold definition are available. The assessment whether a certain value of a specific score indicates disease association or functional neutrality without control data proved elusive.
We computed $Score_{BLOSUM}$, $Score_{PSSM}$ and $Score_{SecStruct}$ for all 339 annotated nsSNVs from the DCM data set (Section 2.12.1) to define thresholds to distinct between disease and neutral predicting scores. Based on the available pathogenicity annotations for the 339 nsSNVs, we were able to rank the nsSNVs due to their calculated scores and compare the distribution of neutral and disease annotated nsSNVs among the top scored results, respectively. Unfortunately, besides disease-associated nsSNVs also neutral-labeled mutations obtained worse scoring results. In addition, Wilcoxon-Mann-Withney tests on the neutral and disease distribution of the nsSNVs and their corresponding calculated scores revealed no statistically significant difference between both test sets for all three scores.

3.3 Conclusion

The comprehensive analysis of evaluation and prediction performance of 13 existing tools, which aim to predict the functional impact of nsSNVs (see Section 3.1), revealed major drawbacks concerning current state-of-the-art pathogenicity prediction.
The prediction of single nsSNVs is not able to substantially promote the improvement of computational diagnostics, since common diseases such as diabetes and cancer are caused and influenced by a varying number of genetic variants [96]. In particular, previous studies revealed variants involved in common diseases do not occur at highly conserved regions, however, current prediction methods often rely on conserved features and hence, fail to assess their pathogenic influence [97].
In a familial study of healthy parents and their GBM diagnosed children, we could identify variant accumulations detected in specific genes of the children but not present in their parents [92]. Single variants revealed no pathogenic phenotype in the parents, but their accumulations in the childrens' genome might have additively contributed to their observed disease. Computational methods to predict the disease association of nsSNV sets or to analyze putative synergetic effects of several nsSNVs within one gene, however, are currently missing.
In a simple approach, we tried to adapt existing measures to capture the functional influence of single nsSNVs to the prediction of nsSNV sets. At present, the effects triggering synergetic events introduced by nsSNV sets have not been deciphered so far, and therefore, a definition of computational features for these events is highly challenging. Furthermore, their validation is laborious and time-consuming or even not possible. Our analysis, however, suggests the limitations of current state-of-the-art prediction features for single nsSNV assessment to be adapted in prediction approaches for multiple nsSNVs.
In particular, the presented studies on congruency and performance of available pathogenicity prediction tools as well as the trial to adapt their prediction strategies for multiple nsSNVs, revealed the importance of 3D structure information within the process of detecting pathogenic nsSNVs. Though this information is often limited, approaches to predict the functional impact of nsSNVs are proposed to include these whenever possible. Furthermore, our analysis with respect to the secondary structure score indicates a crucial influence of tertiary interactions, since the mutational effect of nsSNVs not necessarily resides local at substituted positions within a 3D structure. A putative synergetic effect of nsSNV sets underlies complex correlations and the identification of these requires the integration of available information - regardless if genetic or structural - within computational assessment.
Moreover, to improve clinical treatment, developed approaches have to be designed to meet the requirements on clinical application. Experience has demonstrated the importance of imaging techniques in clinics, such as e.g. magnetic resonance tomography. Visual inspection of the effect introduced by several amino acid substitutions within a 3D protein structure might reveal crucial insights in the mechanisms promoting

dysfunction.
In conclusion, we consequently combined genetic and structural information to implement a software tool - BALL-SNP - to identify candidate nsSNVs for computational diagnostics. BALL-SNP is presented in detail in the next chapter.

4 BALL-SNP: A tool to identify candidate nsSNVs

The experimental analysis of the growing amount of detected genetic variations in NGS studies is too cost- and time intensive, while current *in silico* nsSNV pathogenicity prediction tools are not able to contribute to the improvement of clinical and treatment prognoses.
To identify candidate nsSNVs and hence, to promote the application of NGS in clinics, we developed the software tool BALL-SNP, freely available at `http://www.ccb.uni-saarland.de/BALL-SNP` [98]. BALL-SNP serves as a new pipeline for the assessment of multiple nsSNVs in NGS data. It is based on the Biochemical Algorithms Library (BALL) [47] and integrated in BALL's visualization front-end BALLView [50]. BALL is a comprehensive application framework for rapid software prototyping, which offers a large number of molecular data structures and algorithms allowing for sophisticated development of new approaches [98]. Details about the framework are summarized in Section 2.10.1.
Since we aim to combine genetic and structural information, while ensuring intuitive usability, we take advantage of BALL's rich functionality. We extended the versatile C++ class library by adding functionality to import and process Variant Call Format (VCF) based file formats used in DNA sequencing, SNP calling and SNP annotation. We furthermore embedded the currently most important SNV annotation databases (see Section 2.2) and corresponding parsing methods [98]. Since synergetic effects of several mutations aggregating in a protein structure may additively contribute to an observed dysfunction, we implemented a hierarchical bottom-up clustering for mutated residues. In addition, we introduced a compute server and the associated request/response functionality allowing for straightforward integration of available prediction tools [98]. Figure 4.1 outlines the BALL-SNP workflow along with all incorporated data sources.

Besides the 3D visualization, we display additional generated information in an accessible HTML-based interface, facilitating a clearly arranged presentation [98]. Moreover, as BALL-SNP is implemented on top of the standard molecular modelling tool BALLView, an intuitive and direct interaction of the user with the visualized 3D structure representations is possible.
In the following, we point out the main functionality extensions and give an overview about the implementation details developed within this thesis.

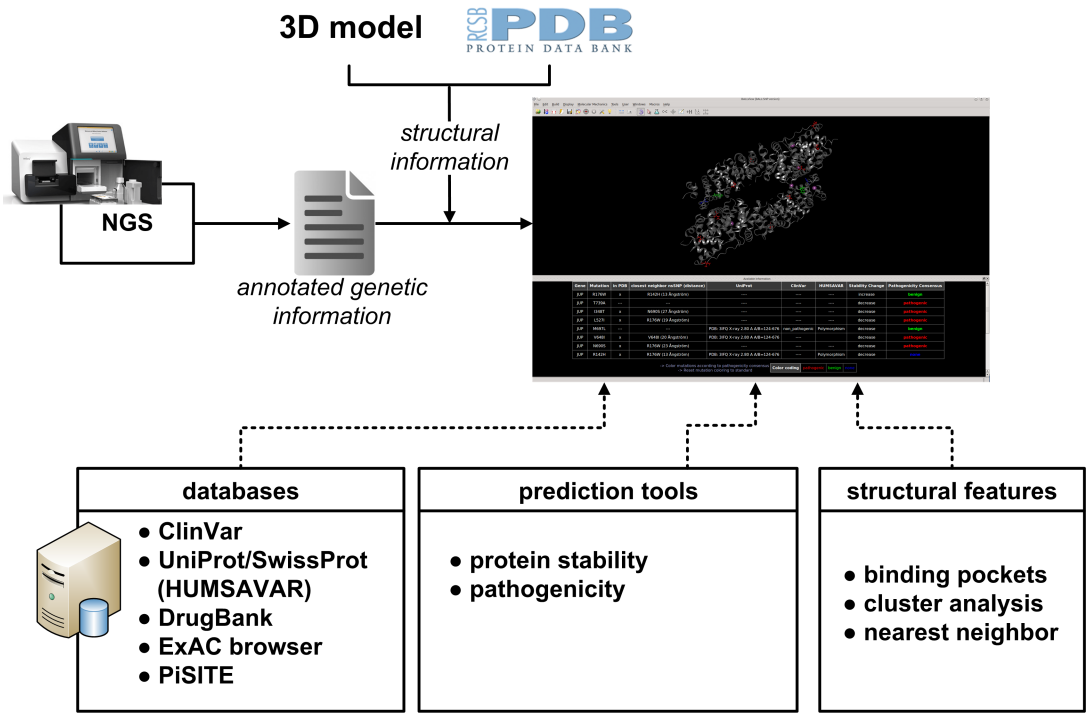


Abbildung 4.1: General BALL-SNP workflow (adapted from [98]).

4.1 Design and Implementation

The BALL implementation is structured into **BALL** core functionality and classes responsible for the visualization front-end BALLView, the **VIEW** component. A major part of the VIEW implementation refers to Qt derived classes creating and handling widgets, dialogs and so on.
The required extensions for BALL-SNP reside in both, the core of BALL and the VIEW component. Figure 4.2 gives an overview of the implemented classes and functionality necessary for BALL-SNP.
Here, we briefly summarize the implemented C++ classes and their tasks within BALL-SNP.

`SNPFile` incorporates all attributes and member functions necessary to read and process the content of the allowed input format files (see Section 2.10.2). The `SNPFileDialog` refers to a central class offering access to diverse functionality. Since we focus on the development of a tool, which is easy to use, in particular to non-experts of molecular modelling or further specialized software, such a class delegating the analysis pipeline without extensive user interaction was necessary. The `SNPFileDialog` connects and triggers certain instances to generate SNV specific information from e.g. annotation databases, available prediction tools and from structural features (see workflow Figure 4.1).
We also constructed a compute server to allow out-sourced, cost-intensive calculations, such as for example SNV pathogenicity predictions. The interface classes to communicate with the constructed compute server refer to `DownloadModelFile` and `DownloadPredictions`. Both include QtNetwork functionality with QtNetworkReply and QtNetworkAccessManager to process server requests and the corresponding server responses.
Since experimentally gained knowledge about nsSNVs is curated in freely available databases (Section 2.2), we implemented a class to parse and make use of these available information, the `DatabaseParser` class.
Structural features such amino acid substitutions aggregating within subunits of a protein can shed light on the mechanisms steering dysfunction. Hence, we introduced methods to analyze the spatial relations of amino acid substitutions. The `SNPClustering` class conducts SNV cluster analysis and stores required cluster properties in instances of the `SNPCluster` class.
The `DatabaseInterface` class is responsible for the representation of the information content prepared and generated by, for example, the `DatabaseParser` and the `SNPClustering`, as well as for the intuitive visualization of these information on an information page widget.

Besides, further add-ons had been necessary for e.g. notification purposes, and were not specific for a certain functionality than rather crucial for all made extensions.

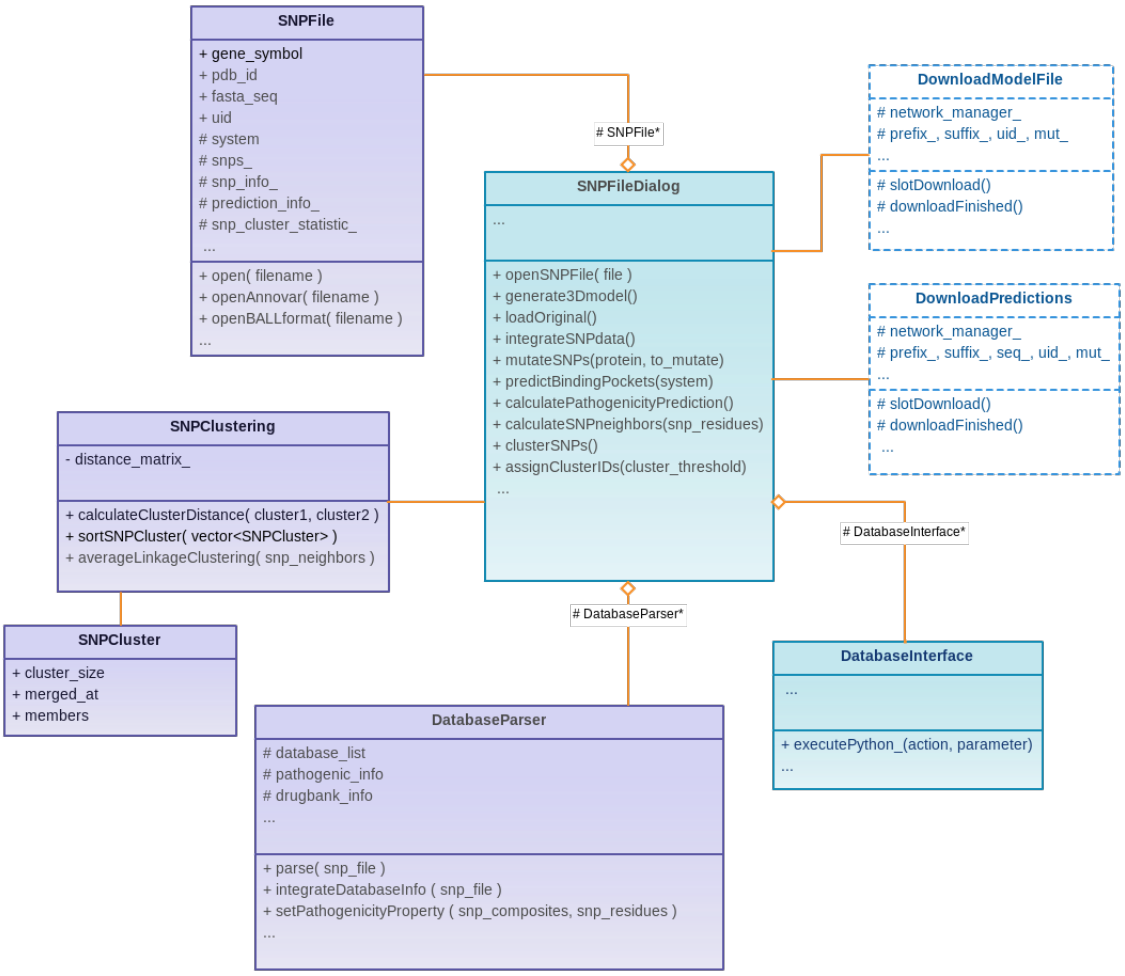


Abbildung 4.2: UML diagram of the implemented extensions and the most important components in BALL-SNP. Purple-colored classes refer to the BALL core, while turquois ones belong to the VIEW component. The blue dash-lined classes specify the interface to the constructed compute server.

BALL handles communication of all widgets and dialogs in the GUI of BALLView through message posting to inform about modifications in a running BALLView instance. To integrate also SNV-relevant messages and their handling, we added a new message type called `SNPdataMessage` to inform notifying components about LOAD, RESET and CLUSTER events concerning SNV data.
Since we generate a large amount of information, we introduced new coloring schemes to enable information-based coloring of the 3D content:

- `COLORING_PATHOGENICITY`: colors mutated residues according to their available pathogenicity information
- `COLORING_INTERACTION`: labels mutated residues with available interaction site information
- `COLORING_CLUSTER`: highlights mutated residues based on the determined cluster affiliation

A molecular structure is partitioned into subcomponents such as e.g. atoms, residues and proteins. To be able to create representations colored according to specific nsSNV characteristics, we added new properties in the existing `residue` class to assign these:

- `PROPERTY_MUTATED`: labels a mutated residue
- `PROPERTY_PATHOGENIC`: assigned to a residue with pathogenic information
- `PROPERTY_BENIGN`: assigned to a residue with benign information
- `PROPERTY_WITH_INTERACTION`: tags a residue with available interaction site information

Furthermore, the additional attribute `cluster_id` to maintain the cluster affiliation and thus, enable the ad hoc selection of cluster distance thresholds, was required.

For the database parsing functionality, we included several python scripts in BALL-SNP to preprocess the different formats of the available data sources and to enable a straightforward update of these without intricately modifying C++ code.
In the following, we describe the functionality of the implemented classes in detail.

4.2 Adjustment of PDB residue information

BALL-SNP critically depends on 3D structures and is consequently connected to the Protein Data Bank (PDB), the most comprehensive archive for 3D structural data of biological macromolecules world-wide [27] (see Section 2.3 for details). Since the standard BALL-SNP input file does neither contain any information about encoded proteins nor information on available 3D content, we parse the required information from

the UniProtKB, which also links available PDB ids. Unfortunately, the PDB information concerning length as well as start and end indices deposited in the UniProtKB are partially inconsistent with the correct information within the PDB file. Moreover, one protein entry may have several PDB identifiers listed in the UniProtKB. To automatically identify the largest structure with the best quality available, we first adjusted the PDB residue information in the UniProtKB and then, compared length and quality of each listed PDB structure. The index adjustment was performed by pairwise alignments of the UniProtKB FASTA sequence (see Section 2.7) and the PDB file ATOM entries within a PDB, since the FASTA sequence within the header of a PDB file not necessarily matches the actual structure sequence of the protein given by the ATOM entries in the PDB file (see Section 2.8 for more details). A Python script preprocesses the parsed UniProtKB in the described form and returns a file mapping UniProtKB identifier to the best available 3D structure in the PDB to allow fast access and direct use in BALL-SNP.
Besides the index information deposited in the UniProtKB, the range of the residue indices within the ATOM records of a PDB file additionally not necessarily corresponds to the actual sequence indices. Since we have retrieved the correct residue indices during the correction of the UniProtKB information, we are able to check whether the residue index within the PDB file is correct or not and adjust it if necessary. The adjustment is performed when the PDB file has been automatically downloaded by BALL-SNP and a visualization is created.

4.3 3D modelling information to overcome missing PDB information

Unfortunately, the gap between known protein sequences and available 3D protein structures is still huge. Since BALL-SNP, however, relies on the 3D information of a protein, we added the possibility to automatically search for templates in ModBase, a well-established database of comparative protein structure models (details in Section 2.10.4).
Therefore, we setup a compute server including the required request and response functionality. The request is created based on the UniProtKB identifier and a list of introduced amino acid substitutions, where a substitution refers to *wild type amino acid + protein residue position + mutant amino acid*, e.g. G65S for glycine is mutated at position 65 to a serine. The built compute server is assigned different functionality purposes of BALL-SNP and hence, the 3D model search pipeline requires the flag "MODEL" to specify which implemented pipeline to trigger. The request syntax using standard HTTP query strings matches:

```
http://www.ccb.uni-saarland.de/ballsnp/cgi-bin/index.py?
seq= MODEL
&uid= UniProtKB id
&mut= mutation list separated with ‘‘_’’
```

The python pipeline called by the constructed compute-server in BALL-SNP to automatically search for a 3D model basically consists of the following steps:

- search ModBase file for the best available 3D model for the given UniProtKB identifier
- check if at least one of the provided amino acid substitutions is comprised in the selected 3D model
- download the coordinate file (PDB) for the identified best 3D model of the given UniProtKB identifier from ModBase server

The quality of the available 3D models is scored based on sequence identity, e-value and DOPE score. We only consider models with more than 60% sequence identity and low values for e-value and DOPE score. These specifications refer to standard parameters used in molecular modelling. The resulting 3D structure of the target sequence is returned to the compute server and BALL-SNP, respectively.

4.4 Integration of available approaches on nsSNV assessment

In recent years, software tools to predict the impact of single nsSNVs on a protein's structure and function have been developed. Details about current state-of-the-art prediction tools are outlined in Section 2.5. Since often experimentally gained information deposited in databases is not available, we selected promising methods for integration into BALL-SNP to make use of their functionality.
To be independent of the software maintenance by a third party and to guarantee stable performance, we currently only focus on the integration of available stand-alone software tools [98]. The underlying databases and resources of all these available approaches, however, exceed the portable size of a downloadable, freely available software tool with a comprehensive molecular modelling library, such as BALL-SNP. Furthermore, the required input formats as well as the obtained output are often incompatible among the tools and thus, a combination of different prediction tools requires additional analysis. In consequence, a straightforward integration was achieved by out-sourcing the calculation of source-expensive approaches on a compute server.

4.4.1 The compute server functionality

As already mentioned, we setup a compute server for the 3D model search in ModBase when no 3D information is available in the PDB. This compute server was extended to

conduct the prediction of pathogenicity and protein stability changes, respectively. The stand-alone versions of the selected methods were installed on the created compute server and accessed by an implemented server script. This script decides based on specified flags (e.g. "MODEL", see Section 4.3) which pipeline to trigger.
BALL-SNP offers the possibility to send a request to this server and to process the corresponding response, accordingly. The BALL-SNP compute server interface corresponds to the implemented classes `DownloadModelFile` and `DownloadPredictions`, illustrated in Figure 4.2.
A general server request based on standard HTTP query strings currently refers to:

```
http://www.ccb.uni-saarland.de/ballsnp/cgi-bin/index.py?
seq= FASTA sequence
&uid= UniProtKB id
&mut= mutation list separated with ‘‘_’’
```

where the `seq` parameter serves as flag in the 3D model construction pipeline (Section 4.3).
BALL-SNP waits and locks the remaining processes until it receives the response to the made request. Since the computation time increases with the number of nsSNVs in the input file, users can decide whether to generate and include this information or just focus on the remaining information [98].
The subsections below elaborate on the integrated functionality.

4.4.2 Protein stability change

Proteins properly folded have minimal potential energy and are usually stable [98]. Amino acid substitutions introducing a change in the protein sequence can have a significant impact on the potential energy of the protein structure, and thus its folding and stability. Consequently, the analysis to which extent a mutation affects protein stability with respect to the wild type, extends the understanding of the mutation impact on protein function and the genotype-phenotype relationship, accordingly [98]. Several methods to predict the change of a protein's binding free energy exist [12][99]. Among these, I-Mutant was shown to have better performance compared to the other available tools [100]. The implemented compute server runs the freely accessible I-Mutant 2.0 code. I-Mutant 2.0 [38] uses SVMs to automatically predict protein stability changes caused by single point mutations in protein sequence (Section 2.6).

4.4.3 Pathogenicity prediction

Since the experimental analysis to gain knowledge concerning the pathogenicity of nsSNVs is laborious and time-consuming, computational approaches have been developed to predict the impact of an amino acid substitution on protein function *in*

silico. In Section 2.5, we list the available prediction tools integrated in BALL-SNP: PANTHER [32], PhD-SNP [34], PolyPhen2 [35] and PROVEAN [37]. The tool selection depended on stand-alone functionality and prediction strategy. We aimed to cover a broad range of different strategies to assess the pathogenic effect of an nsSNV. Beyond sequence-related features, structural features to analyze the genotype - phenotype relationship are essential (PolyPhen2).
Additional to the PROVEAN prediction tool, we include precomputed PROVEAN scores available at `http://provean.jcvi.org`, because PROVEAN emerged to be generally too time-consuming for an interactive workflow in BALL-SNP (see Section 2.5.4).
In a recent study, we were able to show, that prediction accuracy and sensitivity can be further improved by calculating a consensus score for each single nsSNV [11]. Hence, we define an intuitive majority-based consensus score built on the single prediction results:

$$Score_m = \sum_{t \in tools} P_t \begin{cases} 1 & \text{if} \quad P_t = neutral \\ -1 & \text{if} \quad P_t = disease \end{cases} \tag{4.1}$$

where P_t refers to the prediction result of tool t.

$$consensus = \begin{cases} neutral & \text{if} \quad Score_m > 0 \\ disease & \text{if} \quad Score_m < 0 \\ none & \text{otherwise} \end{cases} \tag{4.2}$$

The majority-based score treats every vote identically and selects a unique winner if all included prediction tools are able to make a prediction. Since we also include the I-Mutant2.0 prediction results, we obtain an odd number of tools. In cases, where tools fail to return a prediction result and thus, no consensus decision is possible, we annotate the amino acid substitution with *no consensus*.
Within the 3D structure visualization, the mutated residues can be colored according to their pathogenicity consensus score via the defined coloring method `COLORING_PATHOGENICITY` (see Figure 4.3).

The update and maintenance of the installed prediction tools is straightforward, since this only implies the installation of the newer software version on the compute server as well as a potential server script adaption without substantial re-implementation in BALL-SNP. In consequence, users are not forced to update their local BALL-SNP versions to receive updated predictions.
Moreover, the constructed compute server as well as the implemented interfaces to

request and to process the corresponding response in BALL-SNP, can easily be extended to incorporate further available, stand-alone software tools for the disease-association analysis of nsSNVs.

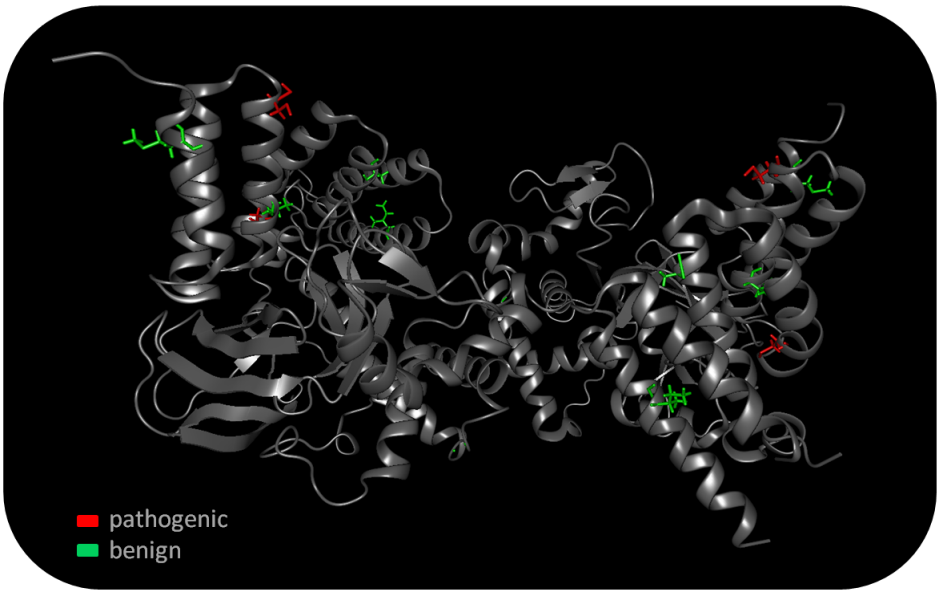


Abbildung 4.3: 3D structure of the protein encoded by *SMYD2*. The amino acid substitutions introduced by nsSNVs are visualized in ball-and-stick representation and colored due to their calculated pathogenicity prediction consensus.

4.5 Integration of available database information

Experimentally gained knowledge about nsSNVs is deposited and curated in different databases [98]. Some of these databases provide additional information concerning the pathogenicity and clinical significance of a nsSNV (details see Section 2.2 and the following). To make use of this knowledge, we include information from SwissProt/UniProtKB [26] and ClinVar (dbSNP) [23] within BALL-SNP. In particular, SwissProt/UniProtKB collects human polymorphisms and disease mutations (annotated in the HUMSAVAR document) assigned according to literature reports on probable disease association. ClinVar is based on the dbSNP [5] and reports human variations while providing clinical significance information.
Besides, we also incorporate data on drug targets curated in DrugBank [28]. The knowledge whether the corresponding protein of a query gene is already a target for medical treatment may provide users with helpful information concerning probable effects of mutations within the target protein, such as e.g. the loss of drug binding.

Information concerning protein interaction sites and in particular, residues participating in these, additionally provide elementary information on protein functions and putative dysfunctions. We therefore integrated The Database of Protein interaction SITEs (PiSITE) (Section 2.10.6). Since PiSITE assigns the interaction information on the residue level, we are able to label mutated residues via `COLORING_INTERACTION`, accordingly. Since interaction residues close to substituted amino acids might be influenced by these, we additionally offer the possibility to color the complete protein representation according to the available interaction information (see Figure 4.4).
Moreover, the Exome Aggregation Consortium (ExAC) provides summary data of exome sequencing projects, freely available for the scientific community [58] (details in Section 2.10.7). We also include these summary data, since the comprised nsSNVs may indicate common polymorphisms inherited in the human population or variations called dependent on the SNP calling technique.
Figure 4.5 exemplarily displays the table representing the integrated information from databases, I-Mutant2.0 (stability prediction) and pathogenicity prediction tools.

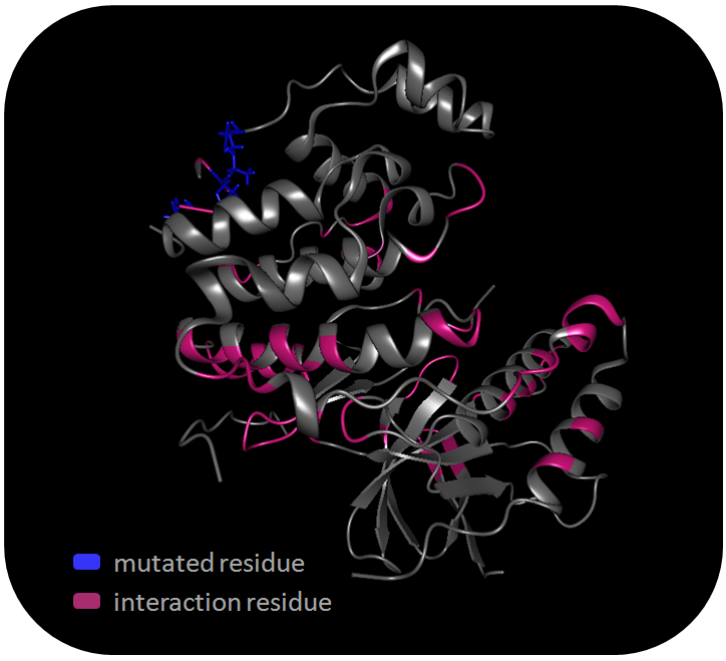
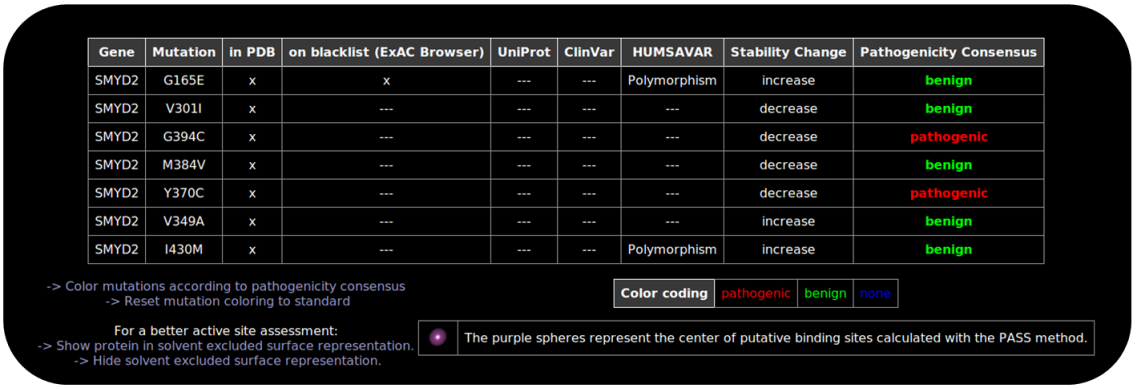


Abbildung 4.4: 3D representation of the protein encoded by *MEF2A* with residues participating in interaction sites labeled in pink.

In cases where database information concerning pathogenicity and/or clinical relevance is available for a given nsSNV, these information is also incorporated into the calculated pathogenicity consensus score for this particular nsSNV (see Section 4.4.3). Since experimentally gained knowledge on the functional impact of an nsSNV is usually more



Gene	Mutation	in PDB	on blacklist (ExAC Browser)	UniProt	ClinVar	HUMSAVAR	Stability Change	Pathogenicity Consensus
SMYD2	G165E	x	x	---	---	Polymorphism	increase	benign
SMYD2	V301I	x	---	---	---	---	decrease	benign
SMYD2	G394C	x	---	---	---	---	decrease	pathogenic
SMYD2	M384V	x	---	---	---	---	decrease	benign
SMYD2	Y370C	x	---	---	---	---	decrease	pathogenic
SMYD2	V349A	x	---	---	---	---	increase	benign
SMYD2	I430M	x	---	---	---	Polymorphism	increase	benign

Abbildung 4.5: Table with generated information from different databases, I-Mutant2.0 and pathogenicity prediction tools for *SMYD2*.

reliable than the predicted impact, the pathogenicity consensus prioritizes pathogenic database information. Benign annotated nsSNVs, however, are treated with equal priority to generated prediction results. In the past, these nsSNVs particularly received re-annotations with disease association in some cases, since experimental studies and available expert knowledge increase. We identified, for example, some variants in the neutral VariBench data set of Thusberg et al. [79] created in 2011, as disease-associated, with entries in the HGMD [6] available in 2014.

To enable a straightforward update and maintenance policy, the integrated data sources are parsed and preprocessed via sophisticated Python scripts included in BALL-SNP's additional data directory. The generated, predefined data formats then are parsed in the C++ implementation via the `DatabaseParser` class. In consequence, a change in a data source will not affect the C++ implementation of BALL-SNP.
Currently, we are focusing on selected important databases that report nsSNV pathogenicity. The embedded database module, however, can easily be extended to include further databases and annotation sources.

4.6 Predicting binding pockets

In addition to data mining information on pathogenicity from databases and *in silico* prediction, further information may provide clinicians essential input. Among these, the proximity of nsSNVs to functional sites such as binding pockets for ligands plays a crucial role. BALL-SNP predicts active sites, which often are located in the largest surface cleft, based on the Putative Active Sites with Spheres (PASS) method [101], that uses probe spheres to characterize regions of buried volume on a protein surface [98]. Based on size, shape, and burial extent of these volumes, positions, which putatively

represent binding sites, are identified.
The basic algorithm was already implemented in BALL, however, not used in the current version. We extended the existing code to visualize the predicted active sites as spheres in BALL-SNP and consequently represent the center of the detected binding pockets (see Figure 4.6).

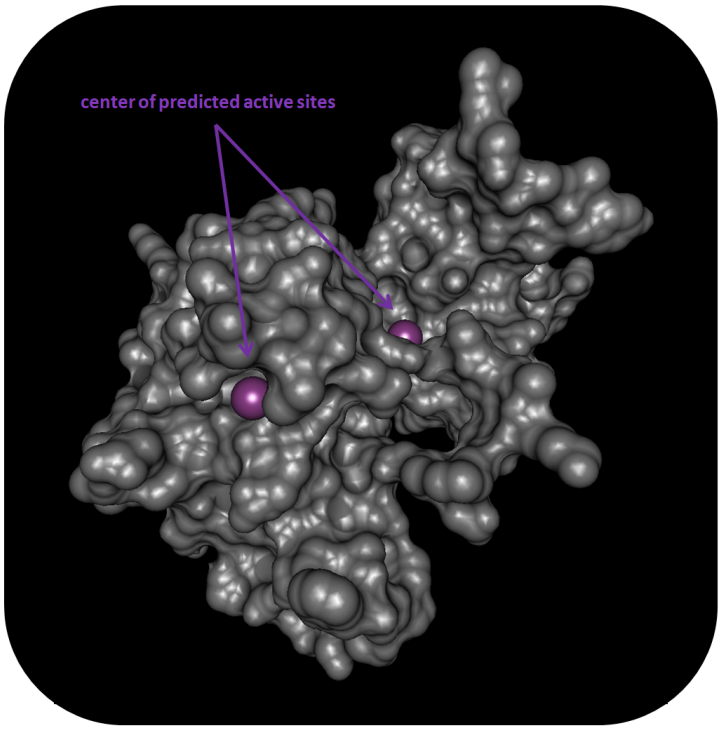


Abbildung 4.6: The *MAP2K3* protein in Solvent-Excluded-Surface (SES) representation including the center of predicted active sites (purple spheres).

4.7 Cluster analysis of nsSNVs

Several mutated residues in one protein may have a synergetic effect on the cause and severity of a disease phenotype [98]. The detection of putative quantitative effects requires 3D structural information and visualization. To support the visual inspection of spatial relations, we implemented a hierarchical bottom-up clustering performed on the 3D structure of the encoded protein and the amino acid substitutions introduced by nsSNVs. The applied distance metric refers to the Euclidean distance of the mutated residues C_α-atoms. The linkage criterion to determine the distance between sets of amino acid substitutions was defined according to the average linkage variant. Details on hierarchical clustering and the linkage strategies are explained in Section 2.10.5.

The cluster analysis is implemented in the `SNPClustering` class within the BALL core. To store required cluster properties such as *members* or the distance at which two clusters were merged into the current, we defined a `SNPCluster` class serving as a data container for one single cluster.
Moreover, proteins comprise polypeptide chains, also known as protein subunits. These subunits may be identical, homologous or dissimilar and dedicated to different tasks. Amino acid substitutions in one subunit can also reside in identical subunits of the protein. Since the naturally occuring form of some proteins not only consists of one but several subunits, nsSNVs can introduce substitutions in several identical subunits, simultaneously.
The spatial analysis of mutated residues within one subunit as well as among all subunits of the protein might add critically to the understanding of synergetic effects introduced by nsSNVs. To be able to differentiate between mutations in identical protein subunits, we introduced the notation *wild type amino acid + residue position + mutant amino acid + _subunit*, e.g. G56S_A for the substitution of glycine at position 56 in subunit A to serine.
Since distance thresholds depend on the size of the protein and its folded 3D structure and thus, are difficult to generalize, we provide the user with all possible distance thresholds and offer the possibility to visualize the resulting clusters, respectively. The results of the cluster analysis are represented in tabular format on BALL-SNP's information page (Figure 4.7).

Cα-atom distance	cluster
13	**cluster1:** 2 nsSNPs **cluster2:** 2 nsSNPs ->show
19	**cluster1:** 2 nsSNPs **cluster2:** 2 nsSNPs **cluster3:** 2 nsSNPs ->show
20	**cluster1:** 2 nsSNPs **cluster2:** 2 nsSNPs **cluster3:** 2 nsSNPs **cluster4:** 2 nsSNPs ->show
21	**cluster1:** 3 nsSNPs **cluster2:** 3 nsSNPs **cluster3:** 2 nsSNPs **cluster4:** 2 nsSNPs ->show
25	**cluster1:** 3 nsSNPs **cluster2:** 3 nsSNPs **cluster3:** 2 nsSNPs **cluster4:** 2 nsSNPs **cluster5:** 2 nsSNPs **cluster6:** 2 nsSNPs ->show
27	**cluster1:** 5 nsSNPs **cluster2:** 3 nsSNPs **cluster3:** 2 nsSNPs **cluster4:** 2 nsSNPs **cluster5:** 2 nsSNPs ->show
28	**cluster1:** 5 nsSNPs **cluster2:** 5 nsSNPs **cluster3:** 2 nsSNPs **cluster4:** 2 nsSNPs ->show
43	**cluster1:** 6 nsSNPs **cluster2:** 6 nsSNPs **cluster3:** 2 nsSNPs **cluster4:** 2 nsSNPs ->show
53	**cluster1:** 8 nsSNPs **cluster2:** 8 nsSNPs ->show
79	**cluster1:** 16 nsSNPs ->show

Abbildung 4.7: The cluster analysis results with all possible distance thresholds are presented in a tabular format on the information page.

Users can choose which threshold to visualize within the 3D structure and can print the corresponding clustering mutated residues via links on the information page (see Section 4.8). Figure 4.8 examplifies the printing of the cluster affiliation of specific amino acid substitutions at a certain distance threshold.

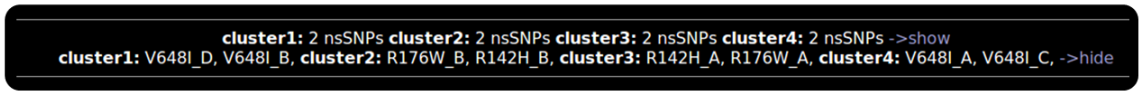


Abbildung 4.8: The members of the clusters at a specific threshold can be printed via links on the information page. The notation of an amino acid substitution is defined as *wild type amino acid - residue position - mutant amino acid - _chain of the protein.*

Within the 3D structure visualization, the clustering mutations can be labeled according to their cluster affiliation via the defined coloring method `COLORING_CLUSTER`, exemplified in Figure 4.9.

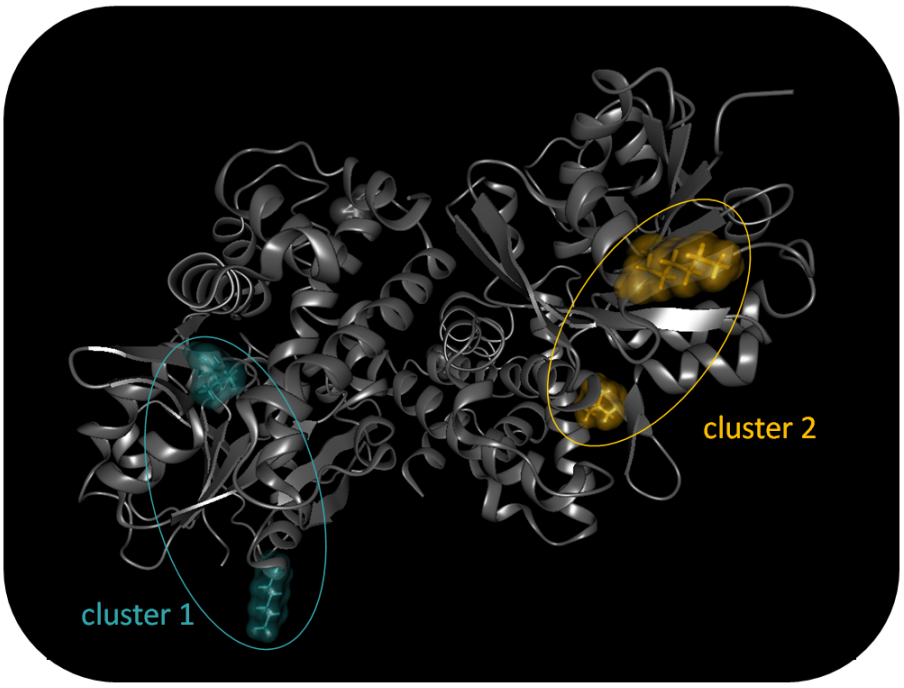


Abbildung 4.9: Glycerol-3-phosphate dehydrogenase 1-like protein encoded by the GPD1L gene in cartoon representation with highlighted results of the cluster analysis at a distance threshold of 26Å.

4.8 Representation of generated information - the information page

In the previous sections, we generated information critical for the assessment of nsSNVs as candidates for clinical studies. To allow a user-friendly, well-structured representation of all these generated information without extensive expert knowledge of the underlying implementation, we use established web technology standards to provide a QtWebKit-based HTML environment. This environment was adopted from PresentaBALL, a package for presentations and lessons in structral biology, embedded as a plugin within BALL/BALLView [102].
Though, we also aim at an intuitive interaction with information content and a clearly arranged representation, the demands BALL-SNP poses on such an interface, differ compared to those of PresentaBALL. In consequence, we implemented a novel interface class called `DatabaseInterface`. This C++ interface is assigned two main tasks: representation of the generated information via an HTML interface widget in BALLView and intuitive techniques to interact with and visualize these information in the 3D structure.
The content of the HTML interface widget is parsed, prepared and generated by the `DatabaseParser` class and includes for example the information from different embedded databases (see Section 4.5) and cluster analysis results (see Section 4.7).
Since the information page is already visualized at the start of a running BALL-SNP instance, when no SNV data was loaded yet, the `DatabaseParser` has to generate an updated instance of this HTML interface widget with the calculated information after SNV input data was loaded.
Besides, the user is able to trigger visualization events via standard HTML hyperlinks providing complete access to the Python interface of BALL. The link syntax based on standard HTTP query strings refers to:

```
BALL-SNP/BALL/data/DatabaseInterface/database_index.html?
action= Python function to call
&module=common_functions
```

We restrict the user interaction, however, to modify the residue coloring according (e.g. `COLORING_PATHOGENICITY`) to specified properties (e.g. `PROPERTY_PATHOGENIC`, `PROPERTY_BENIGN`), creating predefined visual representations such as Solvent-Excluded-Surface (SES), and to the corresponding reset functionality. For this purpose, we added corresponding functions to the collection of predefined Python functions (common_functions), which are called from the HTML interface widget and provided from PresentaBALL:

- `colorDatabaseInfo()`: color mutated residues according to generated database info

- `resetDatabaseInfo()`: reset coloring of mutated residues to standard
- `setClusterThreshold(distance)`: set distance threshold to join two clusters
- `colorClusterInfo(distance)`: color mutated residues according to their cluster affiliation at a specific distance
- `resetClusterInfo()`: reset cluster coloring of mutated residues to standard
- `colorInteractionInfo()`: color mutated residues participating in interaction sites
- `resetInteractionInfo()`: reset coloring of mutated residues to standard
- `colorSystemInteractions()`: color the complete 3D representation of a protein based on available interaction site information
- `resetSystemInteractions()`: reset the protein coloring to standard
- `createSESModelAndHighlightSNPs()`: create SES model representation for protein and highlight mutated residues
- `removeSESModel()`: remove created representation of SES model from 3D view

In contrast to PresentaBALL, the HTML interface is fully integrated within BALL/ BALLView without plugin activation requirements and plugin dependent usage. To avoid complexity, the allowed interactions are restricted to defined use-case scenarios valid in the analysis of nsSNVs.

4.9 Application Scenarios

A valid and high-quality data set is essential when analyzing the phenotypic effect of nsSNVs on human health. To avoid artifacts, that may arise from using artificially generated data sets and to prove the benefits of the developed tool BALL-SNP, we applied it to clinical, high-quality data of DCM patients (details in Section 2.12.1) as well as cancer data from one patient diagnosed with breast cancer.
In general, there are two practical scenarios based on NGS data, which are, to the best of our knowledge, not implemented in the previously existing methods: the assessment of the effect of several nsSNVs within a single protein, and the contribution of nsSNV sets to ligand binding or protein stability. BALL-SNP is able to support the user in selecting candidate nsSNVs for further analysis and finding possible solutions in both scenarios.
In the following, we outline BALL-SNP's rich functionality and ability to assess the effect of nsSNVs in different use-cases.

Figure 4.10 demonstrates the general pipeline BALL-SNP processes when loading a delivered input file.

4.9.1 Analysis of cardiomyopathy data

The results presented in this section have already been published in [98].
Within the high-quality NGS data set of 639 patients suffering from DCM [72], we identified three cases exemplary for the previously mentioned practical scenarios in NGS data analysis [98]. Genes *JUP*, *VCL* and *SMYD2* revealed nsSNV cluster in potentially interesting locations. In particular, the nsSNVs of the DCM data within these genes reveal no pathogenic annotations in available databases.

Analysis of nsSNVs in JUP The gene *JUP* coding for junction plakoglobin is involved in cell junction, which influence the arrangement and function of cells within a tissue. In particular, *JUP* is involved in arrhythmogenic right ventricular dysplasia (ARVD), a congenital heart disease [103].
For the example of *JUP*, we exemplarily process the pipeline of an input file within BALL-SNP step by step.
An exemplary input file for BALL-SNP in *BALLformat* (see Section 2.10.2) is shown below for the gene *JUP*:

PDB:	**3IFQ**				
JUP	NM_002230	T739A	chr17	39912019	.
JUP	NM_002230	N690S	chr17	39912444	.
JUP	NM_002230	V648I	chr17	39913771	.
JUP	NM_002230	M697L	chr17	39912145	rs1126821
JUP	NM_002230	R142H	chr17	39925713	rs41283425
JUP	NM_002230	I348T	chr17	39921186	.
JUP	NM_002230	L527I	chr17	39915041	.
JUP	NM_002230	R176W	chr17	39925402	.
JUP	NM_002230	R203C	chr17	39925321	.

The gray-colored text line is optional.
Based on this input file, BALL-SNP first checks automatically for available 3D structures of the protein encoded by gene *JUP*, if no PDB identifier was provided, and/or downloads the PDB structure. Next, the 3D structure is visualized in the 3D view of BALL-SNP. Since the calculation requires few minutes, the user can then decide whether to compute the pathogenicity consensus and protein stability predictions for each nsSNV in the input file or skip this step. In the case of *JUP*, 4 of the 9 inherited nsSNVs reveal a disease-association in the consensus pathogenicity prediction (details see Section 4.4.3). Table 4.1 lists the introduced amino acid substitutions and the calculated pathogenicity

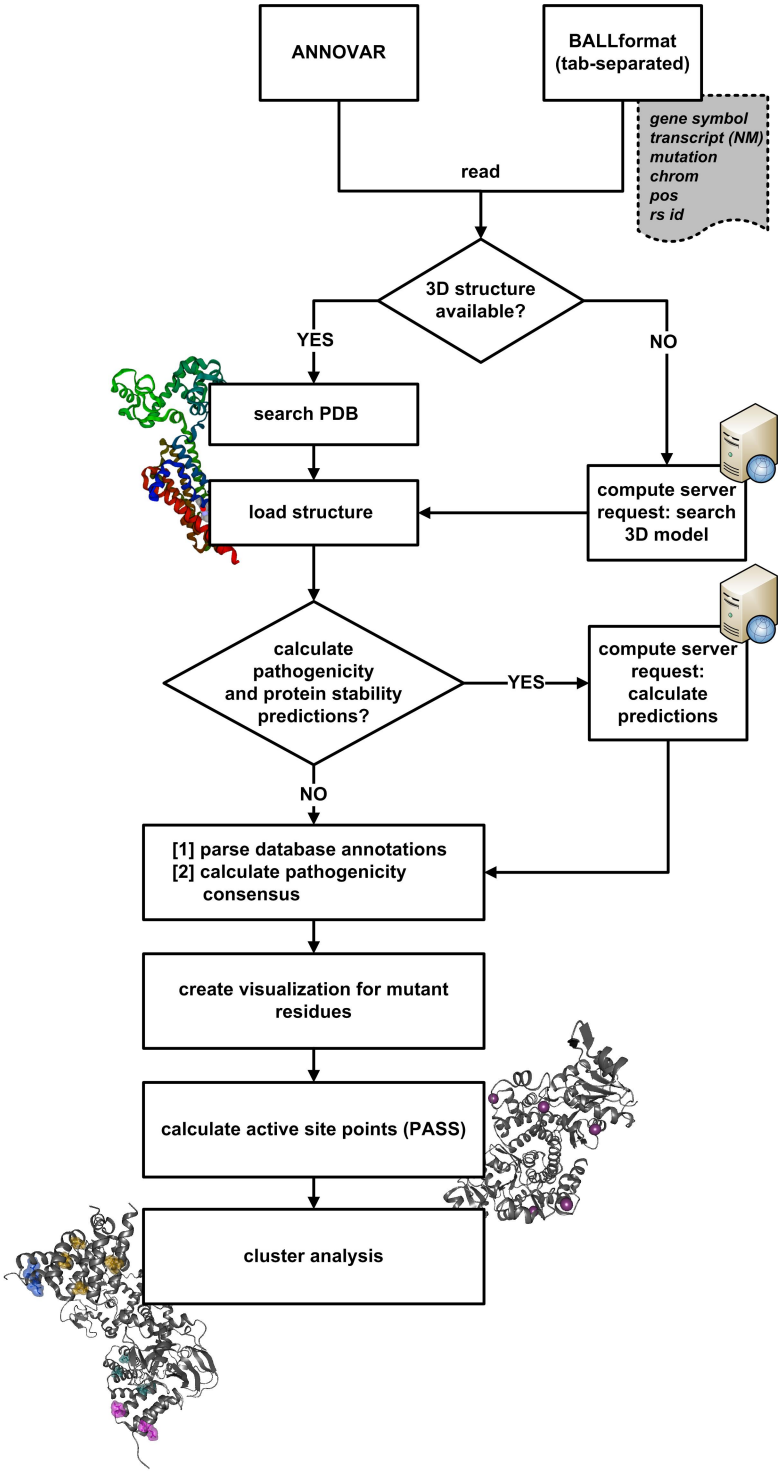


Abbildung 4.10: General pipeline processed in BALL-SNP for a given input file.

consensus as well as the predicted protein stability change. The database info, generated in the next step, however, contains only information for the amino acid substitution M697L, namely a non-pathogenic clinical significance.

Tabelle 4.1: Pathogenicity consensus as well as prediction of the resulting protein stability for the amino acid substitutions introduced by nsSNVs in *JUP*.

amino acid substitution	pathogenicity consensus	protein stability
R142H	pathogenic	decrease
R176W	pathogenic	decrease
R203C	pathogenic	increase
I348T	pathogenic	decrease
V648I	benign	decrease
N690S	benign	decrease
M697L	benign	decrease
T739A	benign	decrease

In further steps, mutated residues are labeled, binding pockets predicted and a cluster analysis on all amino acid substitutions is performed, simultaneously. Based on this information, users can highlight the mutated residues in the visualization, accordingly. Figure 4.11 illustrates the result of the cluster analysis at a threshold of 24Å. The coloring of the amino acid substitutions is defined by their cluster affiliation.
The combination of both, the generated pathogenicity consensus and the cluster analysis, indicate a synergetic influence on the protein's function of several mutated residues. Mutations with a pathogenic prediction are clustering and in particular, benign predicted substitutions also show a close neighborhood.
In addition, the center of putative active sites were labeled with purple spheres within the 3D visualization. Interestingly, the amino acid substitution L527I indicates proximity to a predicted binding site (see Figure 4.12).
The nsSNVs in our data set, identified in the coding region of *JUP* obtained either no annotation or a benign one. Based on the performed analysis with BALL-SNP, however, putative synergetic effects of the introduced amino acid substitutions are uncovered, identifying these nsSNVs as promising candidates for further clinical studies.

Analysis of nsSNVs in VCL *VCL* codes for vinculin, an actin filament-binding protein, involved in both, cell-matrix and cell-cell adhesion. *VCL* has been reported to be associated with DCM, a congestive heart failure [99]. Database search yields no annotations for the nsSNVs in *VCL* from our data set.
In contrast, the calculated pathogenicity consensus associates 6 of 9 nsSNVs as disease-linked and I-Mutant predicts all except one to decrease protein stability (see Table 4.2).
Interestingly, BALL-SNP identifies, that amino acid substitutions corresponding to

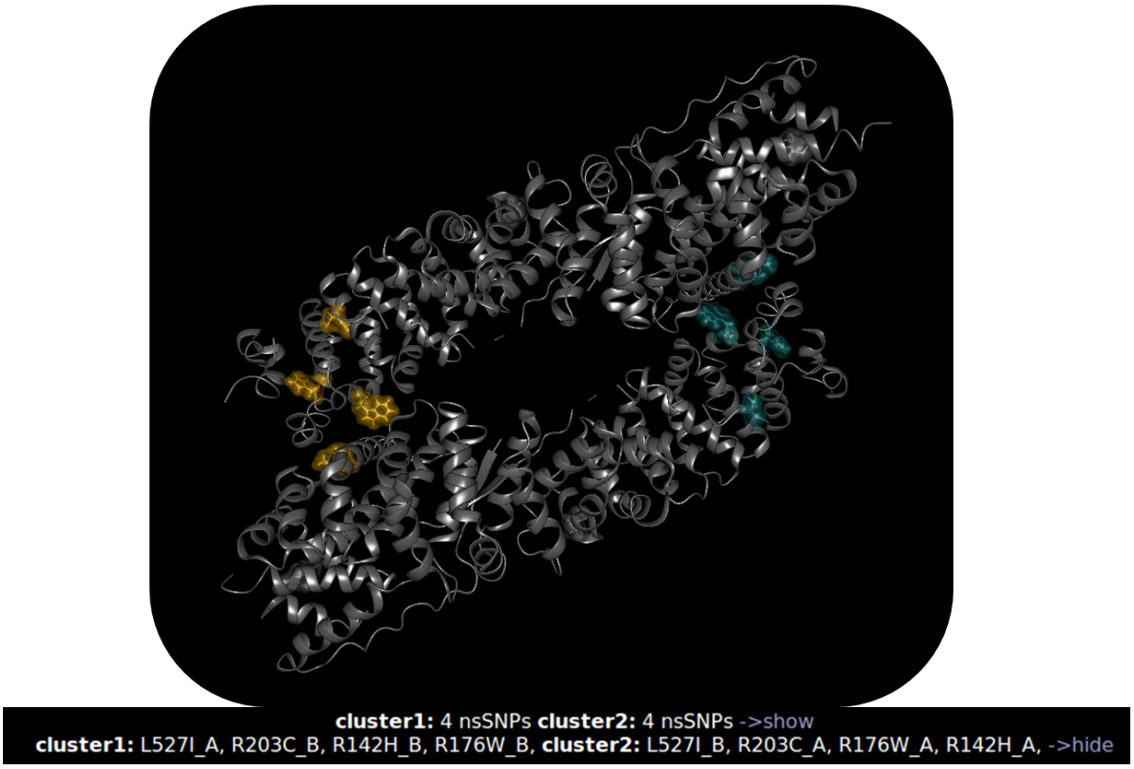


Abbildung 4.11: The wild type structure of the protein encoded by *JUP* is displayed in the cartoon representation (gray), while the inherited amino acid substitutions are colored according to their cluster affiliation.

Tabelle 4.2: Pathogenicity consensus as well as prediction of the resulting protein stability for the amino acid substitutions introduced by nsSNVs in *VCL*.

amino acid substitution	pathogenicity consensus	protein stability
R230H	pathogenic	decrease
H363R	pathogenic	decrease
A413T	benign	decrease
I519L	pathogenic	decrease
R586W	pathogenic	increase
V658A	no consensus	decrease
R759Q	pathogenic	decrease
A922V	benign	increase
N1010K	pathogenic	decrease

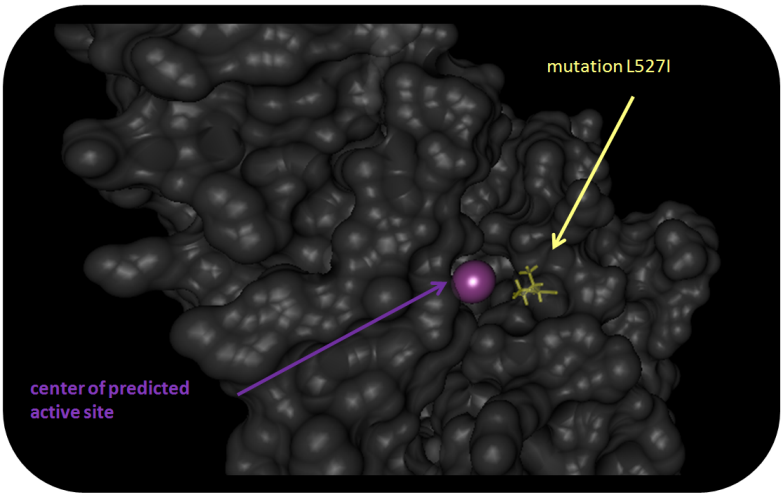


Abbildung 4.12: Cutout of the solvent-excluded surface of *JUP*. The purple sphere represents the center of a predicted binding site. The mutation L527I (highlighted in yellow) is located within the detected, putative binding pocket [98].

nsSNVs, cluster together in the protein structure at a threshold of 19Å(see Figure 4.13). The mutations I519L, R586W and V658A cluster with C_α-atom distances between 15 to 19Å, respectively. Except V658A, the substitutions obtained a disease-association in the pathogenicity consensus. H363R and R759Q are in close neighborhood with a C_α-atom distance around 19Å. In addition, both mutations are predicted to decrease protein stability and are linked to disease.
Since we are also able to detect clusters between the different subunits of a protein, we identified R230H in chain B and A922V in chain A to be located close to each other (C_α-atom distance of 19Å). Mutations at the interface of protein subunits may critically influence the stability of the protein, since they can alter the binding affinity of the subunits among each other.
The BALL-SNP analysis of the nsSNVs of *VCL* again contributed important information not available in state-of-the-art approaches to assess the functional effect of nsSNVs.

Analysis of nsSNVs in SMYD2 *SMYD2* codes for a N-lysine methyltransferase, which methylates both, histones and non-histone proteins. While the database search only returns either no or benign annotations, 2 of the 7 nsSNVs are consensually predicted to be pathogenic (Table 4.3).
Furthermore, BALL-SNP impressively shows that several amino acid substitution pairs introduced by nsSNVs are located next to each other, implying a cumulative effect. The mutations Y370C and M384V (at a C_α-atom distance of 9Å) are adjacent in an opposite

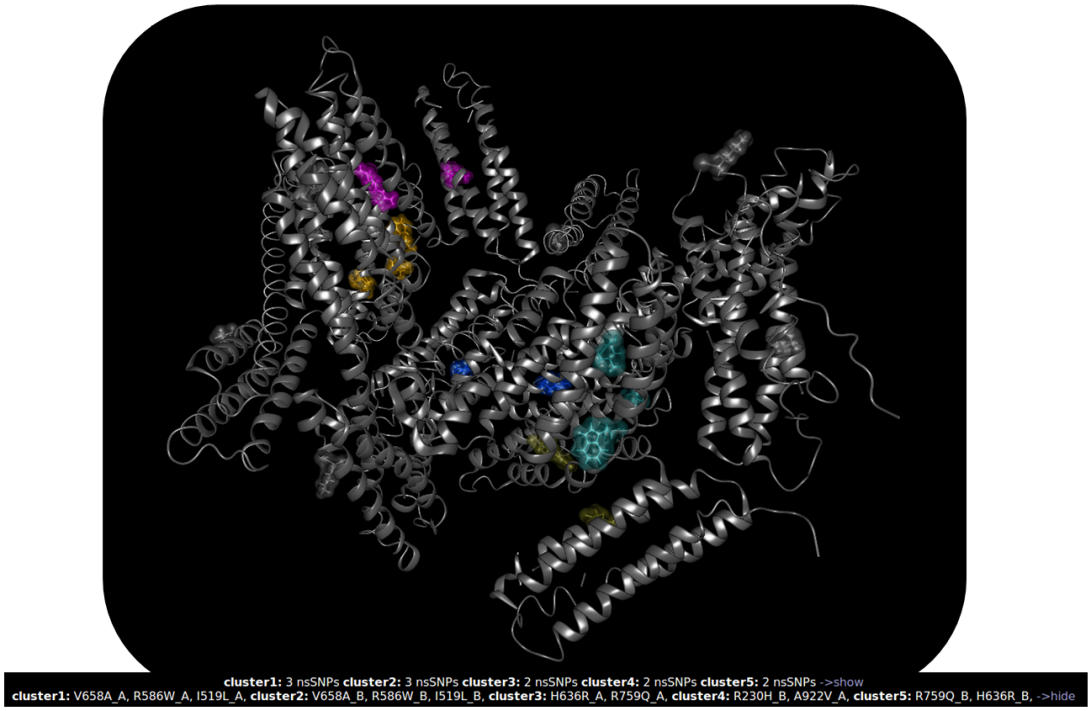


Abbildung 4.13: The 3D structure of the protein encoded by *VCL*. The wild type structure of the protein is displayed in the cartoon representation (gray), while the mutated residues are colored by their cluster affiliation at a threshold of 19Å.

Tabelle 4.3: Pathogenicity consensus as well as prediction of the resulting protein stability for the amino acid substitutions introduced by nsSNVs in *SMYD2*.

amino acid substitution	pathogenicity consensus	protein stability
G165E	benign	increase
V301I	benign	decrease
V349A	benign	decrease
Y370C	pathogenic	decrease
M384V	benign	decrease
G394C	pathogenic	decrease
I430M	benign	decrease

direction, and both are predicted to lead to decreased protein stability. Furthermore, the substitutions G394C and I430M are located close to each other (12Å C_α-atom distance) as well as V301I and V349A (16Å C_α-atom distance). Interestingly, both pairs produce opposite predictions of I-Mutant concerning their impact on protein stability.

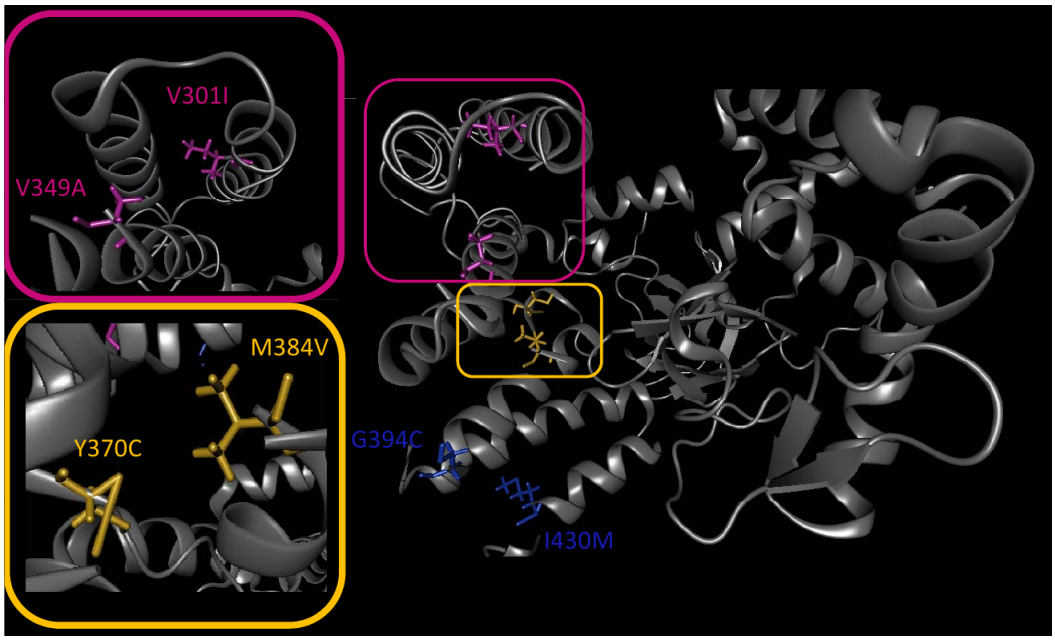


Abbildung 4.14: Cutouts of the 3D structure of the *SMYD2* encoded protein. The clustering pairs of amino acid substitutions are highlighted in different colors. The color framed pictures are close-up views of the correspondingly colored mutation pairs. All pairs are located next to each other, indicating a cumulative effect [98].

Figure 4.14 illustrates these 3D spatial observations, in detail. The overall results of the hierarchical cluster analysis based on average linkage are shown in Figure 4.15.
In conclusion, BALL-SNP was able to identify promising candidates for further clinical studies and computational diagnostics in all of the three presented analyses. Based on these findings, further analyses on the relation to disease traits of the detected nsSNVs are possible.

Abbildung 4.15: Cluster analysis of the amino acid substitutions in the *SMYD2* encoded protein. The protein consists of two chains A and B. Hence, the mutations are labeled accordingly. Part A: The amino acid substitutions are highlighted within the protein 3D structure according to their cluster affiliation. Part B: The overall cluster analysis results are shown in tabular format on the information page. The red marked distance refers to the highlighted cluster of mutations in the structural visualization [98].

4.9.2 Analysis of breast cancer data

Circulating tumor cells (CTCs) are considered as a valuable liquid tumor biopsy. Especially for metastatic cancers, CTCs have demonstrated a substantial potential for prognostic purposes. In an exome sequencing study of 3 CTCs captured from one breast cancer patient and of the corresponding tumor tissue, we identified genes enriched with mutations, that overlap in all four samples (study currently unpublished). Among these, especially *MAP2K3* and *KCNJ12* revealed several nsSNVs and have been described in the context of various cancers according to literature [104] [105].
For both, however, no 3D structure is available from the PDB. In such cases, BALL-SNP offers an automated 3D model search of the corresponding protein via a compute server and the database of comparative protein structure models, ModBase (Section 4.3). BALL-SNP was able to detect 3D models in ModBase with target-template alignments of about 84% (*MAP2K3* to template 3ENM) and 90% (*KCNJ12* to template 3SPC) sequence identity as well as reliable e-values and DOPE scores, respectively. In consequence, we were able to analyze the spatial relationship of the detected variants in the corresponding proteins of *MAP2K3* and *KCNJ12* with BALL-SNP.

MAP2K3 In the encoded protein of *MAP2K3*, variant R264H resides close to a predicted active site center and thus, probably has an influence on the binding affinity of *MAP2K3* protein (see Figure 4.16). Interestingly, both detected amino acid substitutions additionally revealed a disease-associated pathogenicity prediction consensus (see Table 4.4). Further studies on putatively altered binding affinities due to these mutations are required to guide drug target analyses.

Tabelle 4.4: Pathogenicity consensus as well as prediction of the resulting protein stability for the amino acid substitutions introduced by nsSNVs in *MAP2K3*.

amino acid substitution	pathogenicity consensus	protein stability
R67W	pathogenic	decrease
R264H	pathogenic	decrease

KCNJ12 Within the found 3D model of the ATP-sensitive inward rectifier potassium channel 12, encoded by *KCNJ12*, the inherited nsSNV-introduced amino acid substitutions cluster together, impressively (see Figure 4.17).
Variants D173N, A185V and M302I are located next to each other with pairwise C_α-atom distances between 18Å and 16Å, respectively. Hence, these variants may putatively add to a quantitative effect on a dysfunction of the corresponding protein. Besides, D173N and I100V also cluster with a C_α-atom distance of 21Å. All mentioned mutations are furthermore predicted to decrease protein stability. In addition, 4 of 6 amino acid substitutions are predicted to be pathogenic (see Table 4.5).

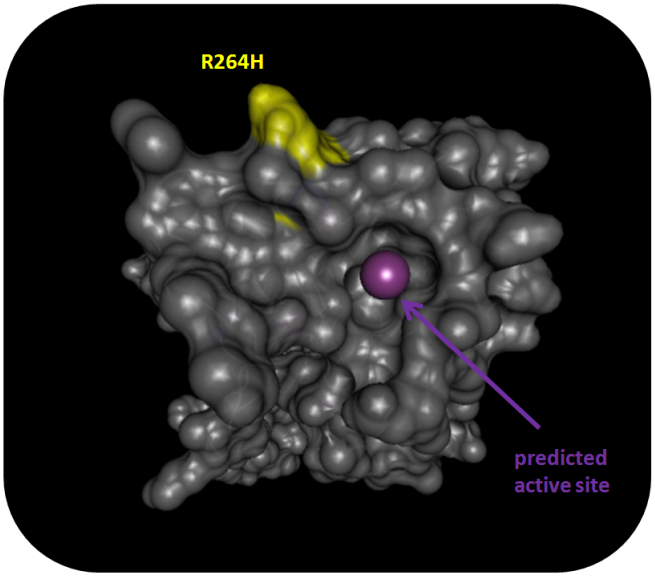


Abbildung 4.16: SES representation of the found 3D model of *MAP2K3* encoded protein. The included amino acid substitution R264H is located close to a putative active site. The purple sphere represents the center of a predicted binding pocket.

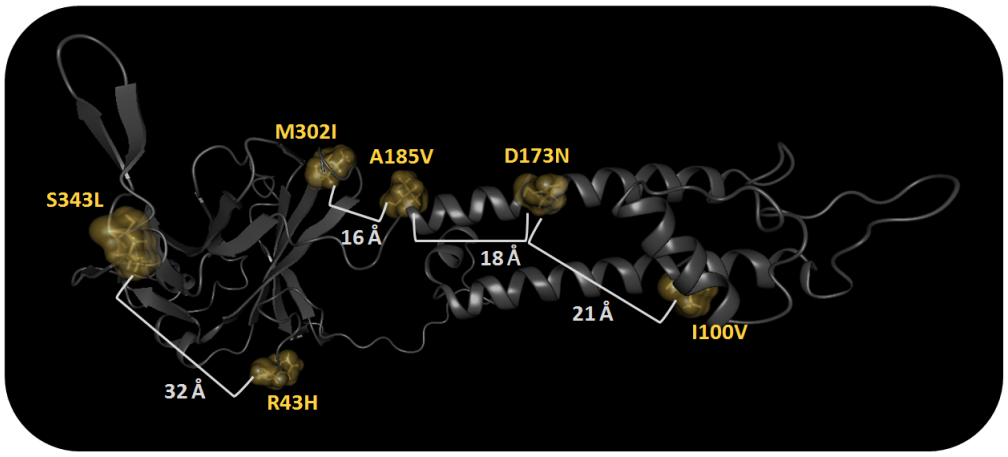


Abbildung 4.17: 3D model of *KCNJ12* encoded protein in cartoon representation. The majority of the nsSNV-introduced amino acid substitutions cluster within the 3D structure. The variants D173N, A185V and M302I are located close to each other with pairwise C_α-atom distances of 18Å and 16Å, respectively. In addition, D173N and I100V are close neighbors with a C_α-atom distance of 21Å. S343L and R43H at least reveal a C_α-atom distance of 32Å.

Tabelle 4.5: Pathogenicity consensus as well as prediction of the resulting protein stability for the amino acid substitutions introduced by nsSNVs in *KCNJ12*.

amino acid substitution	pathogenicity consensus	protein stability
R43H	pathogenic	decrease
I100V	benign	decrease
A185V	pathogenic	decrease
D173N	pathogenic	decrease
M302I	pathogenic	decrease
S343L	benign	increase

In conclusion, the results obtained in the BALL-SNP analyses hint to a disease-relation of the studied nsSNVs. Although further validation studies are required, these results may contribute to the selection of clinically relevant candidate nsSNVs.

4.9.3 Analysis of interaction sites

Unfortunately, annotations of genetic variants and in particular, their introduced amino acid substitutions are often missing. In consequence, we consider multiple data sources and calculate several properties of mutated residues within BALL-SNP. For the applied clinical data sets, however, no interaction site information from PiSITE (Section 2.10.6) was available. To be able to nevertheless demonstrate BALL-SNP's capability to make use of this information, we selected a representative protein with amino acid subsitutions at residues with available PiSITE interaction information from the UniProtKB (Section 2.2.2). The gene *MEF2A* encodes for the myocyte-specific enhancer factor 2A, a transcriptional activator with specific binding to the MEF2 element [106]. It mediates cellular functions in skeletal and cardiac muscle development as well as in neuronal differentiation and survival.
We selected four amino acid substitutions annotated in the UniProtKB to demonstrate BALL-SNP's interaction labeling functionality. In addition, we computed the pathogenicity and protein stability predictions (see Table 4.6).

Tabelle 4.6: Pathogenicity consensus as well as prediction of the resulting protein stability for the amino acid substitutions introduced by nsSNVs in *MEF2A*.

amino acid substitution	pathogenicity consensus	protein stability
R269A	pathogenic	decrease
K270A	pathogenic	decrease
L273A	pathogenic	decrease
V275A	pathogenic	decrease

The chosen nsSNVs are annotated in the UniProtKB to reduce trancriptional activity

when pairwise associated with each other. Figure 4.6 represents the 3D structure of the endcoded protein with residues participating in the protein's interaction sites colored in pink. The amino acid substitutions introduced by the selected nsSNVs are additionally visualized with their SES and colored in blue. The mutated residues at position 273 and 275 participate in an interaction site of the protein and thus, are colored in pink including their SES. In combination with the available annotations in UniProtKB, the pathogenicity consensus and I-Mutant's stability change predictions, the nsSNVs L273A and V275A indicate disease-relevance, which should be further examined in clinical studies.

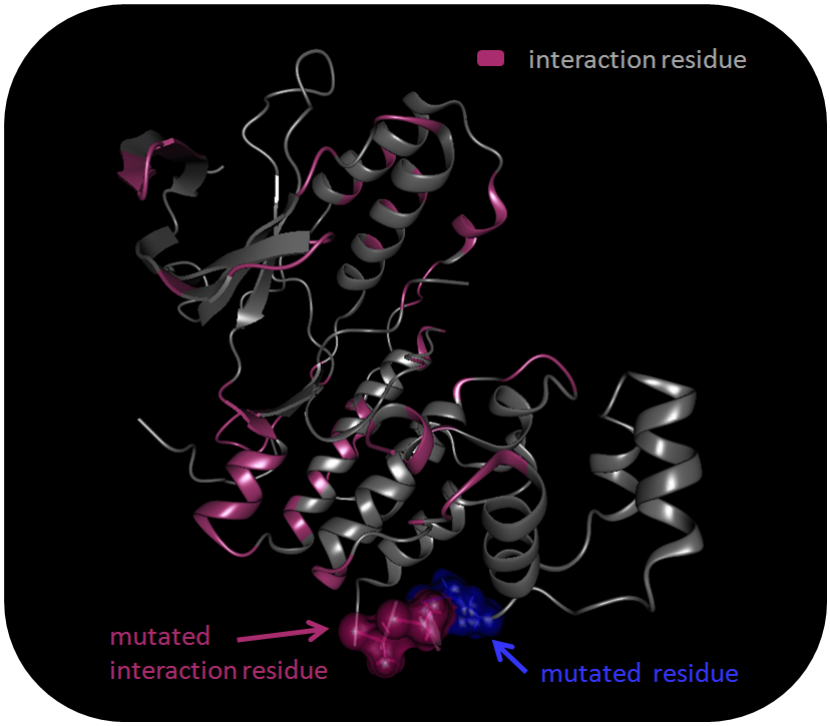


Abbildung 4.18: 3D representation of the protein endoced by *MEF2A*. Residues with available information concerning their participation in protein interaction sites are colored in pink. Mutated residues are colored in blue, except they contribute to an interaction site (pink color).

4.10 Conclusion

The analysis of the genotype-phenotype relation and in particular, of the influence of nsSNVs on protein stability and function, is essential in human health care. In spite of the fact, that the majority of common diseases such as cardiomyopathy are caused by

accumulation of several nsSNVs, computational methods to analyze cumulative nsSNVs and their putative quantitative contribution to an observed pathogenic phenotype are missing. In consequence, the validation of the clinical relevance of nsSNV spatial interactions is limited.
Our tool BALL-SNP combines genetic and structural information to provide scientists the possibility to get deeper insights on the potential effects of accumulated mutations in proteins. BALL-SNP enables the assessment of the impact of nsSNV clusters on protein stability, and consequently assists the selection of candidate nsSNVs for experimental validation. It is based on a standard molecular modelling framework, allows the use of standard NGS output, embeds important nsSNV annotation databases and performs nsSNV cluster analysis.
Although further improvement is needed to meet requirements of the clinical application, BALL-SNP already makes an important contribution to the existing instruments of candidate nsSNV analysis.
Besides nsSNV sets inherited in one gene, nsSNVs in different genes can imply synergetic events [107]. To be able to study these, we developed a multi-scale analysis approach, presented in the next chapter.

5 Multi-scale analysis of nsSNV sets in multiple genes

Traditional computational approaches predict the influence of nsSNVs on pathogenicity for single variants. The impact of nsSNVs on a patient's phenotype, however, can arise from multiple factors such as e.g. gene-gene and gene-environment interactions [20].
In the previous chapter, we presented BALL-SNP, a tool to assess the functional impact of nsSNV sets in one gene for the identification of candidate nsSNVs in NGS data. Since the genetic basis of most common diseases refers to multiple genetic factors working in aggregate [16], sets of genes inhering nsSNVs may also exhibit synergetic effects.
Previous studies analyzed the occurrence and characteristics of compensatory mutations [108], though there might also be cumulative effects of mutations, packing single benign effects together to an observable disease phenotype. More precisely, benign-annotated nsSNVs in combination might be responsible for a pathological effect. Westphal et al., for example, studied congenital disorders of glycosylation and identified a mild polymorphism in *ALG6* putatively exacerbating an already severe pathogenic phenotype caused by *PMM2* dysfunction [109]. This phenomena often is denoted by epistatis, in essence defined as the interaction between genes. In available literature, however, this term is confusingly and even conflictively used according to Heather J. Cordell [110]. In fact, a differentiation of genes and proteins in the context of interaction analysis is exhausting. In this chapter, we refer to the analysis of nsSNVs in multiple genes including the analysis of introduced amino acid substitutions in the encoded proteins avoiding the term epistasis. We study the accumulations of nsSNVs on the genomic level, while examining their interactions on the protein level.
Since biological systems are driven by complex biomolecular interactions [9], the study of putative synergetic events affecting a patient's phenotype requires a multi-scale analysis comprising 3D context, interaction information and functional cascades. To address this issue, we extended the traditional pathogenicity prediction approach by the analysis of nsSNV-affected genes and their mutated proteins in terms of available 3D structures, pathway analysis and subcellular localization on the example of a high-quality clinical data set of cardiomyopathy patients. The conducted multi-scale pipeline is summarized in Figure 5.1.
The work presented in this chapter will be published under [111].

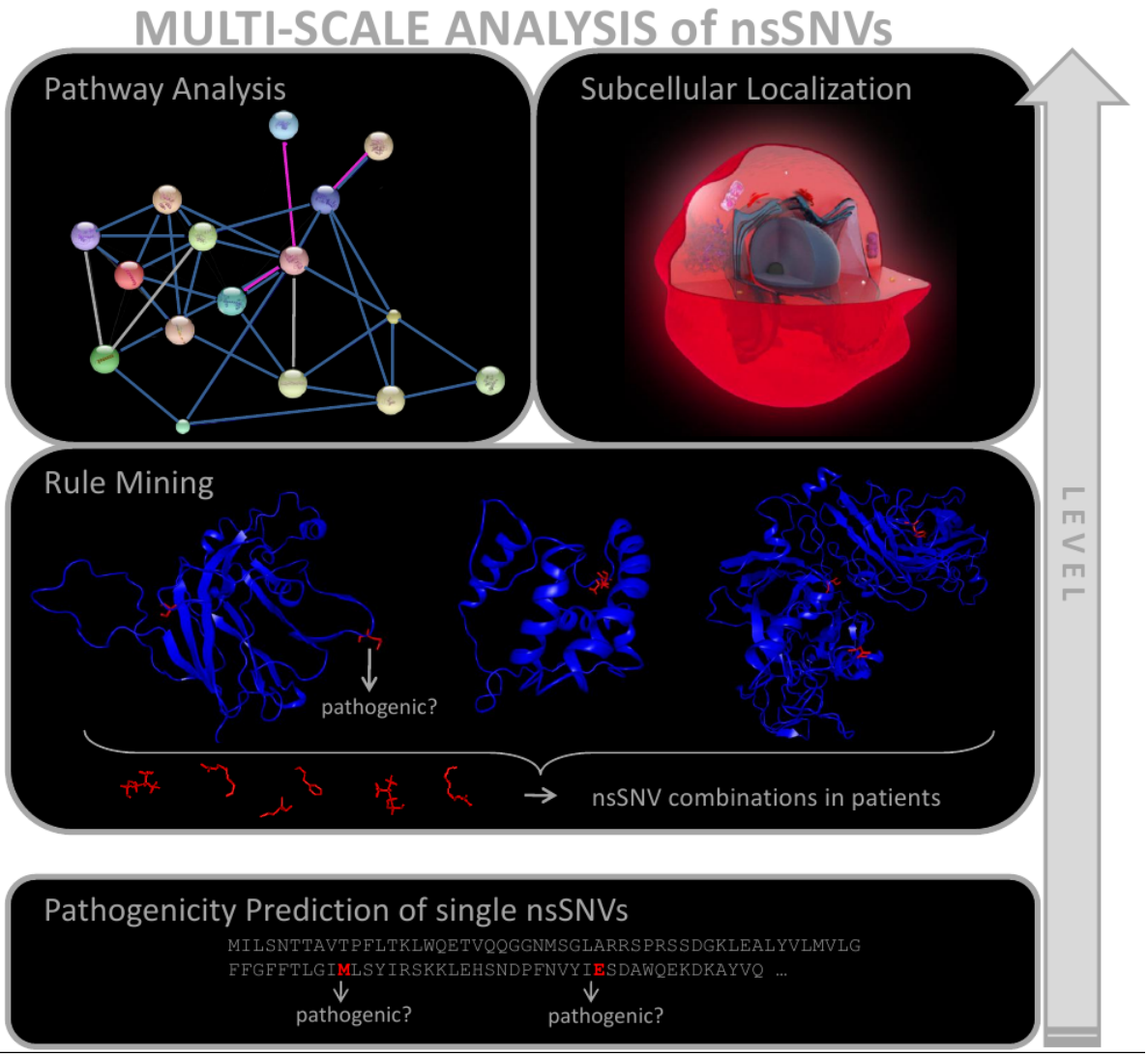


Abbildung 5.1: Schema for multi-scale analysis of nsSNVs. The traditional approach of pathogenicity prediction of single nsSNVs is extended by more complex levels such as rule mining to detect nsSNV sets in patients. On the top level, pathway analysis as well as subcellular localization are applied to study the pathogenic impact of multiple nsSNVs [111].

5.1 Adaption of association rule learning to identify genes with synergetic nsSNVs

Human individuals frequently carry more than one single nsSNV even within one gene. Beyond this, genes and their encoded proteins interact with each other.
To discover strong relations between variables in large data sources, association rule learning is generally applied. Association rule learning uncovers hidden relationships within the tested data by formulating association rules, while using different measures of interestingness to quantify the quality of the generated rules (details see Section 2.11.1).
Via association rule learning, we studied whether there are frequent combinations of nsSNVs within our DCM data set (see Section 2.12.1) and a healthy control cohort based on the general population of the 1000 Genomes Project (see Section 2.12.2). The DCM data comprises 76 genes inhering 842 nsSNVs from 639 DCM patients. We applied the R package arules [60] with confidence threshold 0.8 and different levels of support, starting with at least 0.5.
In fact, we were able to identify associated combinations of nsSNVs with high support and significant confidence values in both, single genes and multiple genes. In the DCM data, we detected frequent combinations of nsSNVs within the single genes *MYPN*, *CACNA1C*, *DMD*, *ADRB2* and *RBM20* with high support and significant confidence values. Table 5.1 lists the detected nsSNVs significantly associated within one gene. The nsSNV combinations in *RBM20* and *CACNA1C* are even found in at least 90% of all patients.

Tabelle 5.1: Significantly associated nsSNVs in single genes with confidence of at least 0.8.

Gene	nsSNVs	DCM patients with nsSNVs
ADRB2	G16R, E27Q	64%
CACNA1C	M869V, K1893R, P1868L	96%
DMD	D878G, R2933Q	72%
MYPN	S691N, S707N	72%
MYPN	F628L, S691N, S707N, P1135T	70%
MYPN	F628L, S803R, S691N, S707N, P1135T	67%
MYPN	S803R, S691N, S707N, P1135T	67%
RBM20	E1223Q, W768S	98%

Since experimental validation is limited, we rely on experimental knowledge deposited in available databases. According to database entries in the HGMD (Section 2.2.3) and UniProtKB (Section 2.2.2), *RBM20* is already related to DCM and *CACNA1C* to the Timothy and Brugada Syndrome, which is a genetic disease characterized by an abnormal electrocardiogram. *MYPN*, that incorporates the most identified

associated nsSNV accumulations, is linked to different forms of cardiomyopathies (familial, hypertrophic, dilated). *ADRB2* participates in signal transduction and namely in the adrenergic signaling in cardiomyocytes. *DMD* is involved in several pathways relevant for cardiac diseases such as DCM, hypertrophic cardiomyopathy (HCM), arrhythmogenic right ventricular cardiomyopathy (ARVC) and viral myocarditis. All identified nsSNV associations, however, are annotated as benign nsSNVs.
Furthermore, we detected nsSNV combinations in different genes (*CACNA1C*, *SMYD2*, *PARVB*, *KCNE1*, *RBM20*, *KCNQ2* and *JUP*) significantly associated with each other in the DCM patients. Table 5.2 lists the detailed nsSNV combinations. Using the corresponding association rule setup, these specific nsSNV - gene combination is not present in the control cohort. Just *SMYD2*, *PARVB*, *KCNE1* and *JUP* revealed an association in the healthy controls.

Tabelle 5.2: These associated nsSNV combinations in seven different genes are detected in more than 70% of all patients with confidence of 0.8 and higher. Interestingly, the 26 patients without identified disease nsSNVs share these combinations.

Gene	Expression	nsSNV set
CACNA1C	Heart, brain, ovary, etc.	M869V, K1893R, P1868L
SMYD2	Heart, brain, etc.	G165E
PARVB	Heart, skeletal muscle	V6A
KCNE1	Heart, lung, etc.	S38G
RBM20	Heart	E1223Q, W768S
JUP	Heart	M697L

The nsSNVs in *CACNA1C* and *RBM20*, could even be detected in 90% of all DCM samples. According to a large-scale analysis of the human transcriptome in 2004, all of the associated genes revealed significant expression in heart [112]. The majority of detected genes with associated combinations is already known in the context of diseases such as Brugada Syndrome, Long QT Syndrome, Naxos Disease and different stages of DCM. In contrast, all association rule detected single nsSNVs within these genes are annotated as benign, except the N749T mutation in *KCNQ2*, which has currently no available annotations. Figure 5.2 compares the information annotations of all genetic variants within the DCM data set with the association rule detected.

Among the 639 DCM patients, we identified 26 without already known or annotated disease-associated nsSNVs. The 26 DCM patients mainly inhere benign and not annotated variants. Interestingly, the intersection of their inherited nsSNVs revealed exactly the detected associated nsSNV combinations in *CACNA1C*, *SMYD2*, *PARVB*, *KCNE1*, *RBM20*, *KCNQ2* and *JUP*.

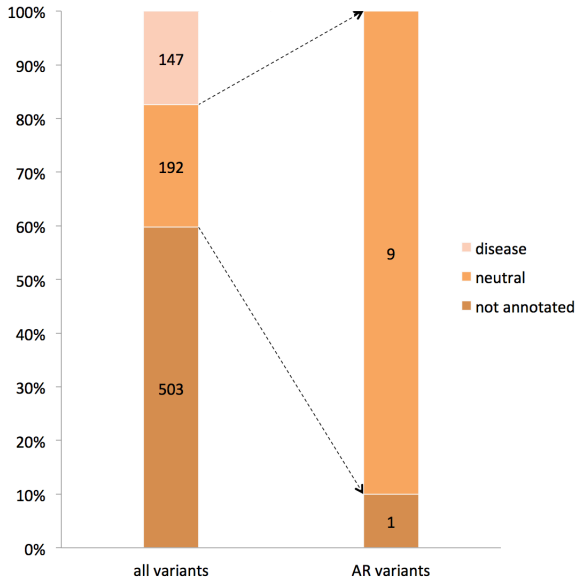


Abbildung 5.2: Comparison of all genetic variants within the DCM data set and the association rule (AR) detected. Except one not annotated AR variant, all other AR variants have neutral annotations available [111].

5.2 Network analysis of genes with associated nsSNVs

Due to the growing availability of high throughput biological data, the analysis of molecular networks gained significant interest. In addition, the incorporation of expert knowledge from gene ontology (GO) and biochemical pathways in e.g. genome-wide association studies (GWAS) has been shown to provide more meaningful results [9]. In consequence, we used several information sources to determine the biological and functional connections of the detected, associated genes with nsSNV combinations. The databases queried are: The STRING database [62], the Gene Ontology Annotation (GOA) database [65] and the KEGG PATHWAY database [66] (details see Section 2.11).
The network analysis splits into functional interaction and biological intersection of the associated nsSNVs in different genes. Referring to the corresponding GO terms of the genes with associated nsSNV combinations, the majority participates in protein binding, voltage-gated ion channel activity and transport. A mutation of residues involved in complex interaction networks can critically influence large interaction cascades by spreading the implemented loss across the network.
Besides the biological connections among the associated nsSNVs, we also investigated their topological characteristics within the human interaction network. To detect putative interaction hubs, we determined betweenness and degree for each node in

the human STRING network using the R package igraph [113]. The degree of a node identifies the number of edges connected to the node, whereas the node betweenness is an indicator of the nodes centrality in the network. The calculations were based on the downloadable version of the human STRING network including scored links between proteins and interaction types for protein links, available at `http://string-db.org/` (version v9_05 and v9_1, November 2014).
To provide more insight into the biological intersection and functional interaction as well as combine the corresponding findings, we visualized these genes with related genes and highlighted GO overlaps within the resulting networks using Cytoscape [91] (see Figure 5.3). The edges in the network refer to available interactions between their nodes. *CACNA1C*, *JUP*, *SMYD2*, *PARVB* and *KCNQ2* are directly connected to large hubs within the STRING human network.
In addition, *KCNE1*, *KCNQ2* and *CACNA1C* interact functionally with each other [114]. *KCNE1* attenuates the current amplitude of the *KCNQ2* channel subunit and slows its gating kinetics [115]. According to the KEGG PATHWAY database, *KCNE1* is part of the adrenergic signaling in cardiomyocytes. A perturbation of its channel function by inherited mutations results in increased susceptibility to cardiac arrhythmias. *KCNQ2* belongs to the cholinergic synapse and *CACNA1C* even takes part in both pathways.
Interestingly, *KCNE1*, *CACNA1C* and *KCNQ2* are already targets of drugs against arrhythmia, atrial fibrillation, congestive heart failure, left ventricular hypertrophy and isolated systolic hypertension according to DrugBank entries [116].

5.3 Structural location of amino acid substitutions

The protein structure reveals interactions between residues which are distant in primary sequence but close in 3D space. The effect of an nsSNV critically depends on the structural location of the mutated residue, especially if it is buried in the hydrophobic core or exposed on the protein surface [117]. In particular, disease-associated variants often affect intrinsic structural features of proteins [118].
First, we selected all proteins within our DCM data set, with a 3D structure available in the PDB (see Section 2.3). Next, we calculated solvent accessibilities using naccess (see Section 2.11.5) for the 8 proteins (comprising 46 amino acid substitutions) in our data set with an available PDB structure to analyze whether disease-associated amino acid substitutions cluster on the protein surface or at buried sites.
Solvent accessibility provides an intuitive and quantitatively reasonable idea of the complexity of the molecular interaction network a residue is involved in [67]. The results confirm the findings of Wang et al. for single nsSNVs [119]: The majority (89%) of disease-linked mutations introduced by nsSNVs is located inside the protein probably affecting stability, whereas benign-annotated substitutions mainly cluster on the protein surface (67%).

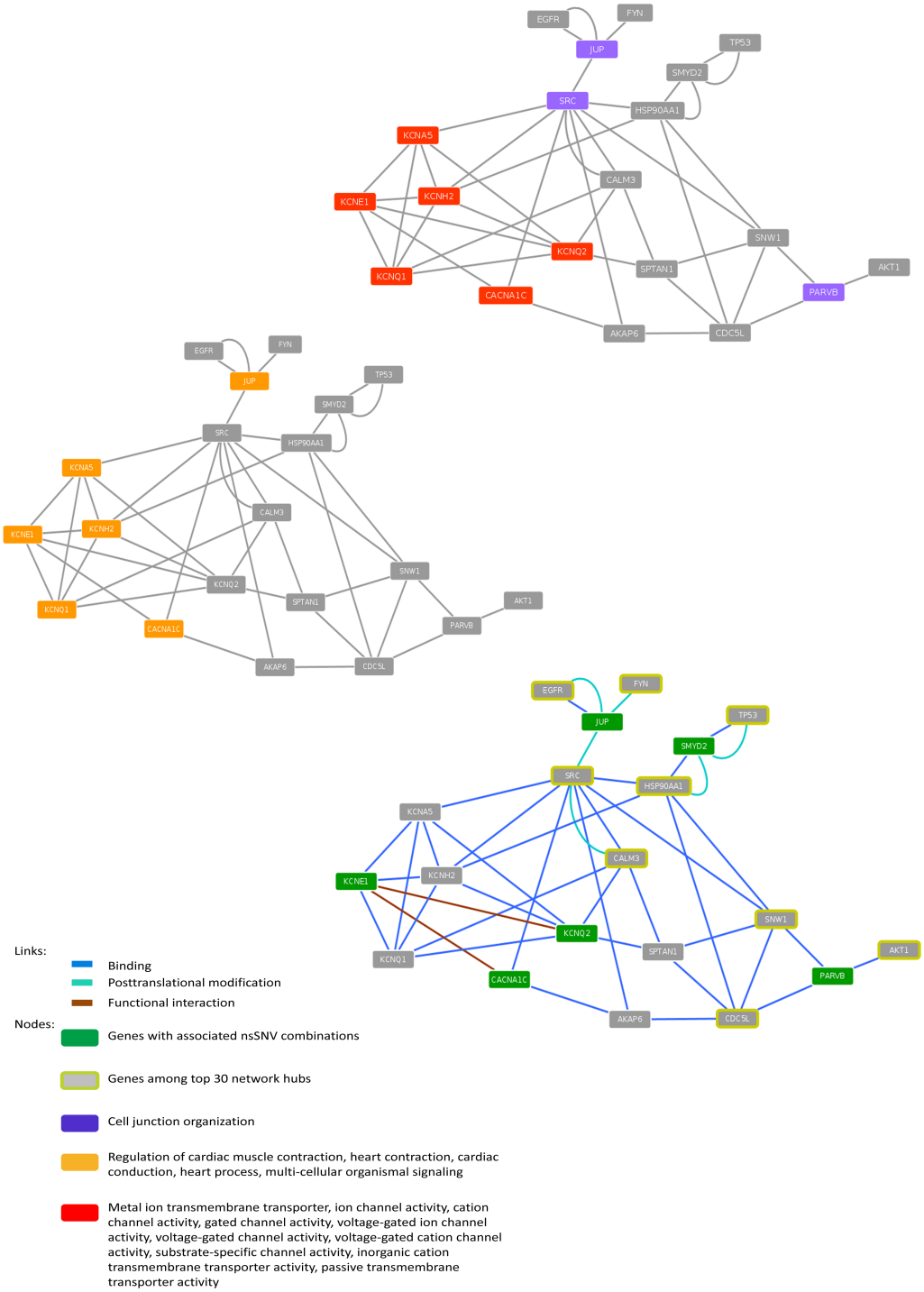


Abbildung 5.3: GO annotations and interactions of associated nsSNVs. The network is based on the STRING human network. The associated nsSNV genes reveal great overlap in their GO annotations. Some of them are also connected to the top ranked hubs within the STRING human network [111].

Since previous studies identified the majority of pathogenic nsSNVs to destabilize a protein's structure [120], we also analyzed protein stability changes upon mutation based on I-Mutant2.0 predictions (see Section 2.6). The majority (81%) of substitutions are predicted to decrease protein stability, independent of their location in the 3D protein structure (see Figure 5.4).

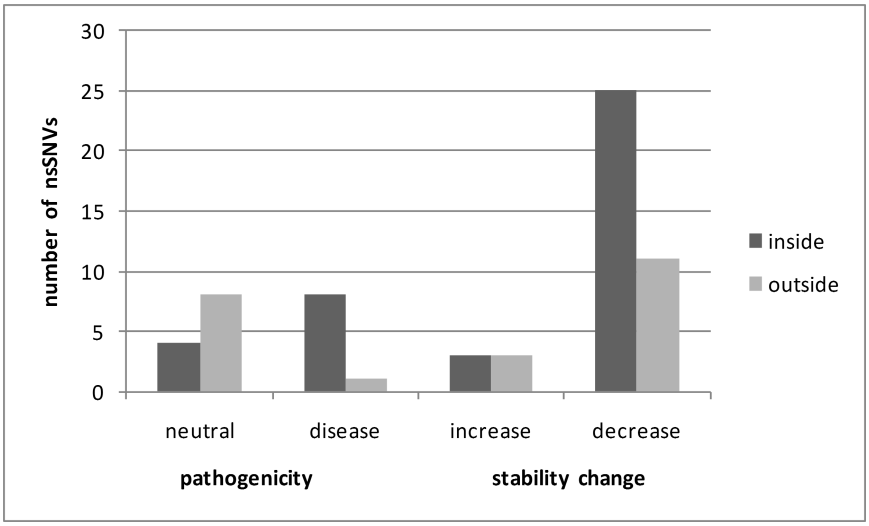


Abbildung 5.4: Information on nsSNV-introduced mutations in proteins of the DCM data set with available 3D structure. We distinguish between amino acid substitutions on the surface of a protein (outside) and buried in the protein structure (inside) [111].

For five protein structures, we were also able to predict possible binding pockets using LIGSITEcsc [69]. 8 of 10 mutations at the proteins surface are found close to a predicted binding pocket using the default parameter settings of 1Å grid space and a probe radius of 5Å. Interestingly, two of the detected significantly associated nsSNV combinations, *KCNE1 S38G* and *SMYD2 G165E*, are also located close to a putative binding pocket (see Figure 5.5) of the corresponding protein.
SMYD2 lysine-methylates the tumor suppressor *TP53*, leading to decreased DNA-binding activity and subsequent transcriptional regulation activity of *TP53* [121]. According to literature, the binding interface of *TP53* and *SMYD2* is located between the catalytic SET domain (residue 1-282) and the C-terminal domain [122].

5.4 Subcellular localization of mutated proteins

For visualizing and analyzing the localization of the proteins encoded by genes with nsSNVs, we used the CELLmicrocosmos 4.2 PathwayIntegration (CmPI) (see Section

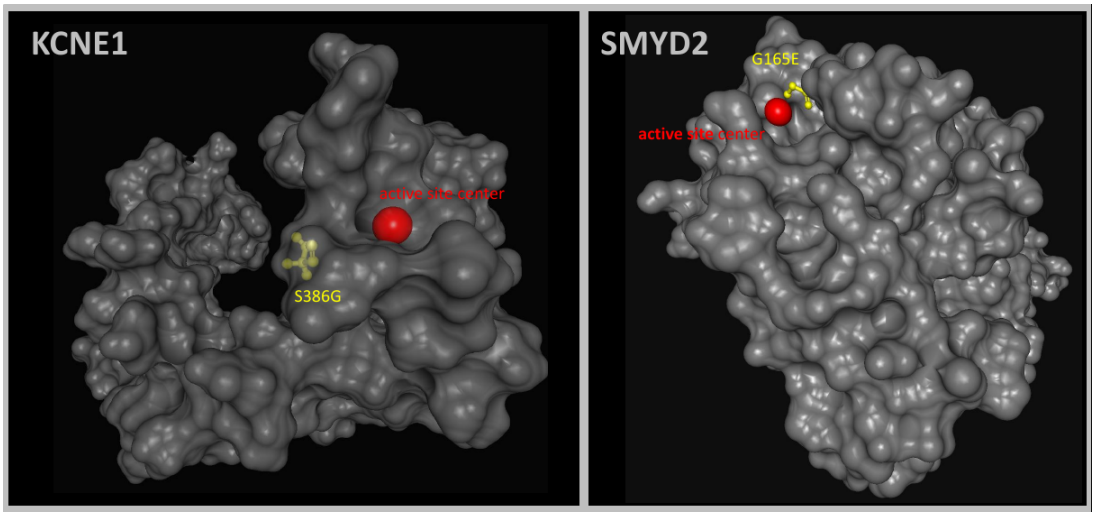


Abbildung 5.5: 3D structures of the encoded proteins of *KCNE1* and *SMYD2*. KCNE1: SES of *KCNE1* encoded protein with the mutated *S386G* in ball-and-stick representation, highlighted in yellow. The red sphere represents the center of a predicted binding pocket. SMYD2: SES of the *SMYD2* associated protein with the mutated *G165E* in ball-and-stick representation, highlighted in yellow. The red sphere represents the center of a predicted binding pocket [111].

2.11.7) [70]. In the context of this work, the following databases were queried: BRENDA [123], GO [124], Reactome [125], and UniProtKB [26].
Based on the previously discussed methods, seven genes were identified showing specific nsSNV combinations in more than 70% of all analyzed patients: *CACNA1C*, *JUP*, *KCNE1*, *KCNQ2*, *PARVB*, *RBM20*, and *SMYD2*. In particular, all of the detected genes have significant expression in heart. Using CmPI, Homo sapiens-related potential localizations for these seven genes were acquired using cell component-gene association data from the aforementioned databases. The associated proteins of these genes show five potential localizations: nucleus, cytosol, cell membrane, lysosome, and the extracellular matrix. Moreover, five of them provide multiple potential localizations. An overview and distribution of these locations can be found in Figure 5.6.

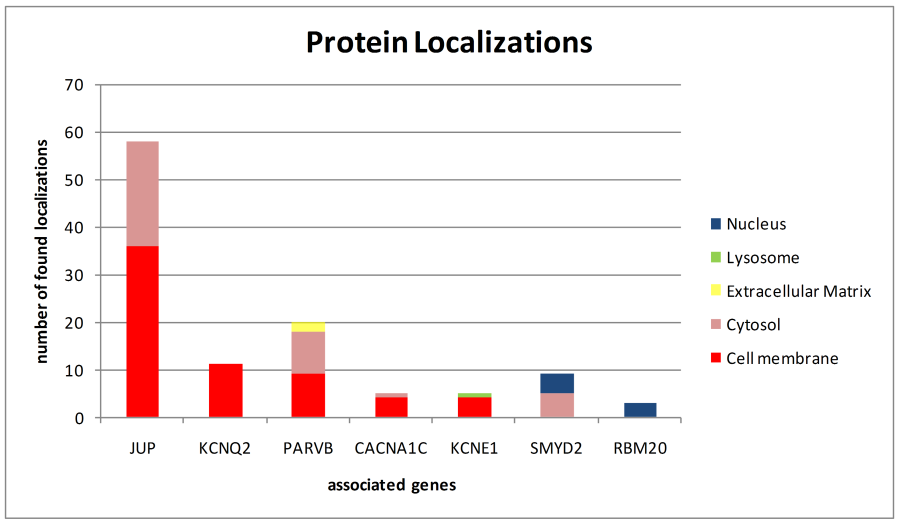


Abbildung 5.6: Subcellular localization chart of all localizations for proteins of the associated genes [111].

Based on the localization data, the hypothesis can be formulated that these proteins are assembled in a potential cascade starting from the nucleus, through the cytosol, entering the cell membrane and proceeding to the extracellular matrix, or vice versa. This theory is supported by the fact that *RBM20* is exclusively localized at the nucleus and *KCNQ2* at the cell membrane, whereas *PARVB* seems to travel between the extracellular matrix, the cell membrane and the cytosol.
We visualized the connections of these proteins including the assigned subcellular locations in Figure 5.7. For the purpose of clarity, Figure 5.7 condenses the detected interactions to the identified potential cascade.

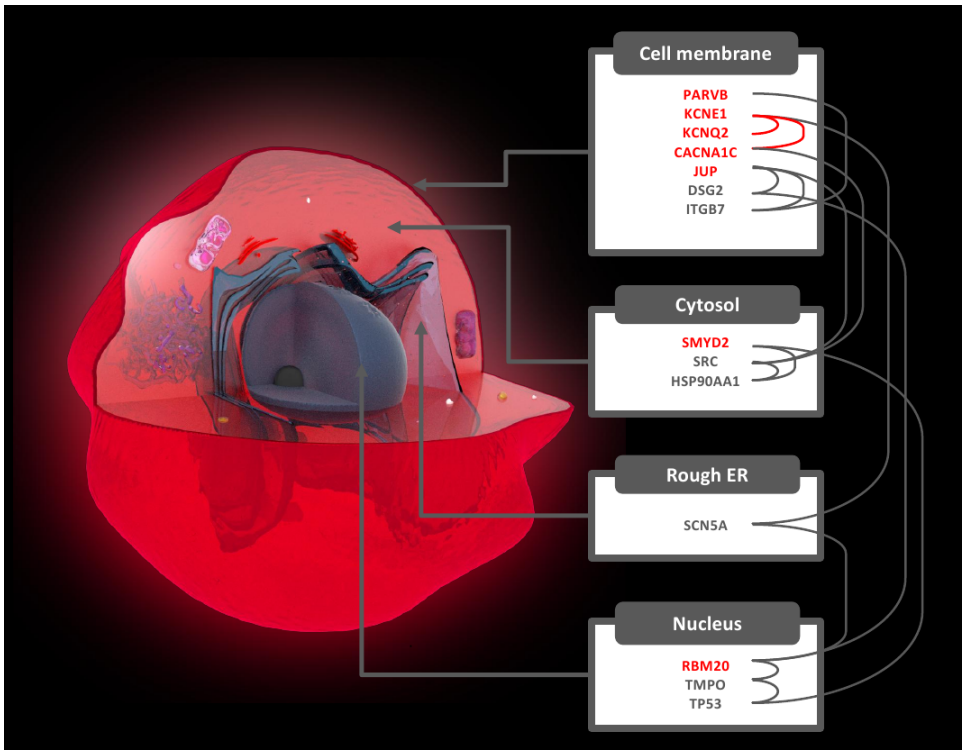


Abbildung 5.7: Schematic visualization of the subcellular localization. The red gene symbols represent the detected associated genes. Besides the subcellular assignment, interactions between the listed genes are visualized (right side). Red-labeled edges mark direct connections of associated genes. In addition to links within one cell compartment, there are also multiple edges crossing different compartments [111].

5.5 Conclusion

In this study, we detected patients suffering from cardiomyopathy without identified disease-associated nsSNVs, but inhering mainly benign-labeled variants. Via association rule learning, we detected associated combinations of nsSNVs within at least 70% of all cardiomyopathy patients in the data set. These specific combinations could not be identified as associated in the control cohort of healthy humans, which hints to disease relevance, however requires further analysis. Due to the lack of prediction tools able to assess a cumulative effect of nsSNVs in multiple genes, a pathogenicity prediction for the identified associated nsSNVs was not possible. Furthermore, only for two of the identified associated nsSNV genes, *KCNE1* and *SMYD2*, a 3D structure of the encoded protein was available. For the remaining associated genes, even an adequate template for structural modelling was missing. Interestingly, the associated nsSNVs in *KCNE1* (*S38G*) and *SMYD2* (*G165E*) are located at the surface of the encoded protein close to predicted binding pockets.
In addition, we studied available pathway information including GO annotations and analyzed interaction networks. Proteins might act at different stages of the same pathway contributing quantitatively to the progressive dysfunction of the pathway until a disease phenotype is observed. Both genes in Westphal's study [109], for example, encode enzymes involved in a different part of the post-translational modification process without a direct interaction. *KCNE1* and *CACNA1C* as well as *KCNQ2* and *CACNA1C* participate in the same pathways and mainly contribute to voltage-gated ion channel activity and transport. Ion channels are key components in a wide variety of biological processes, such as muscle contraction (e.g. cardiac muscle contraction), epithelial transport of nutrients and ions or T-cell activation. A number of genetic disorders (e.g. Long QT syndrome, Brugada syndrome) are related to ion channel dysfunctions.
Groh et al. [126] created a network based on the combination of diseases and associated genes from the Online Mendelian Inheritance in Man (OMIM) database [114]. According to this network, disease genes not necessarily refer to hubs and are even often non-essential. *JUP* and *PARVB* are involved in cell junction organization and interact as well as *SMYD2* directly with the top-ranked hubs within the STRING human network. Cell junctions play a major role in communication between neighboring cells and cell stress reduction.
Furthermore, the subcellular localization revealed the seven proteins to form a potential cascade: *RBM20* is only found in the nucleus, *SMYD2* travels between the nucleus and the cytosol, and the other five genes are mostly associated to the cell membrane, where *PARVB* shows potential localizations between the extracellular matrix, the cell membrane and the cytosol.
Finally, all systematic analyses point to a connection of the detected genes featuring associated nsSNVs - in functionality as well as in their contribution to biological pathways. Moreover, the specific nsSNV combinations identified as significantly associated

in the DCM patients could not be detected associated within the healthy control data. In a next step, further association studies on even larger patient cohorts with cardiomyopathies are required to validate the identified nsSNVs. Additional studies on patients with phenotypes different from cardiomyopathies, in particular, can assess nsSNV specificity.

The constructed multi-scale analysis pipeline for nsSNV sets in multiple genes supports the computational study of synergetic effects and their impact on pathogenicity. We demonstrated this on the example of a cardiac phenotype, however, the analysis can be likewise applied to other diseases such as cancer.

6 Discussion and Conclusion

Advances in high-throughput DNA sequencing techniques have enabled the reliable detection of individual sequence variants in the human genome [1]. Understanding the role of genetics in disease has become a central part of medical research. The decryption of the genotype-phenotype relation and in particular, the study of the pathological effect of genetic variants such as nsSNVs, are increasingly important in human health care. The analysis of NGS data, however, still remains a challenge. In particular, the interpretation of pathogenicity of single variants or combinations of variants is crucial to provide accurate diagnostic information or guide therapies.

Since the amount of identified nsSNVs is growing rapidly, while the experimental analysis to gain knowledge concerning their disease association is laborious and time-consuming, computational approaches have been developed to predict the functional impact of nsSNVs *in silico*.

In a comprehensive evaluation study of available methods to computationally predict the pathogenicity of nsSNVs, we uncovered several drawbacks of currently existing prediction tools with respect to performance, congruency, applicability and clinical relevance. The major limitation, however, denotes to the underlying 'one-SNV, one-phenotype'-paradigm. Contrary to Mendelian disorders, common diseases such as diabetes or cancer, are caused by a varying number of genetic alterations and environmental factors. Typically, a human individual inheres more than one nsSNV, and beyond single variations, these individual combinations of nsSNVs may add to pathogenic processes.

In a familial study of healthy parents and their children diagnosed with glioblastoma multiforme, for example, we could identify variant accumulations detected in specific genes of the children, not present in their parents [92]. Single variants revealed no pathogenic phenotype in the parents, their accumulations in the childrens' genome, though, might have additively contributed to the observed disease. To the best of our knowledge, approaches to assess the functional effect of nsSNV sets are currently limited.

In this thesis, we developed strategies to study both, nsSNVs accumulating in one gene and nsSNV combinations in multiple genes. The developed approaches have been tested on high-quality NGS data from 639 DCM patients inhering 842 nsSNVs in 67 genes [72].

In a first step, we implemented a straight-forward approach to discriminate between disease-associated and neutral nsSNV sets by adaption of existing strategies for single nsSNV pathogenicity prediction. Our analysis, however, revealed the limitations of single nsSNV prediction strategies for the application on nsSNV sets.

In the evaluation study of single pathogenicity prediction approaches, we have been able

to demonstrate the importance of structural information when analyzing the functional impact of nsSNVs [11]. Hence, we combined genetic and structural information to implement a software tool - BALL-SNP - to assess the functional impact of multiple nsSNVs in one protein [98]. BALL-SNP is based on the molecular modelling framework BALL and its visualization front-end BALLView. It promotes genetic variant interpretation and identification of candidate nsSNVs for computational diagnostics via nsSNV cluster and 3D spatial analysis. The input refers to the output of standard SNP annotation software preventing substantial re-formatting efforts. In addition, we offer a simple tab-separated format to allow the inclusion of information from several sources.
Since studies on the positive and negative aspects of GWAS revealed the importance of the integration of already available biological knowledge [9], BALL-SNP incorporates pathogenicity and clinical significance information available in databases such as UniProtKB (HUMSAVAR), HGMD and dbSNP (ClinVar). In fact, the information deposited in different databases or the existing cross-links from one database to others incorporates inconsistencies complicating information retrieval. The PDB indices for start and end residues in the UniProtKB, for example, not necessarily correspond to the actual PDB sequence coverage. But the correct index information is critical to select the best available 3D structure needed for BALL-SNP analyses. In consequence, required strategies to uncover and correct these inconsistencies had to be implemented. Further existing discrepancies in information sources have to be identified and corrected in future studies to improve current information curation and thus, permit a real gain of knowledge.
To make use of available pathogenicity prediction tools for single nsSNV, we compute a pathogenicity consensus score based on the single prediction results for each nsSNV. Although this majority-vote based consensus score is intuitive and machine learning methods may perform better, its dependency on the underlying training data is from our point of view less restrictive in contrast to built models from machine learning techniques. However, some of the prediction tools base their predictions on models trained on database data, which may also explain some of their detected limitations. The applied consensus score particularly improved the statistical performance of single pathogenicity prediction approaches in the evaluation study. While there are numerous pathogenicity prediction methods for single nsSNVs available, however, we currently focus on the integration of available stand-alone software tools to guarantee stable performance and to be independent of the software maintenance by a third party.
Since we aimed to develop a software tool, which is also usable for non-experts, the general pipeline processed in BALL-SNP for a given input file with nsSNV information is designed to automatically delegate the required operations and calculations without extensive user interaction. The generated information and the 3D structural content are visualized in a clearly arranged representation including an intuitive user interface. The availability of a 3D structure, however, represents a critical requirement for an nsSNV analysis with BALL-SNP. If no PDB structure is available, we offer an automatic

search for available 3D models. A pipeline for automated 3D modelling requires toolkits, which are often restricted to non-profit users or even claim a license key for each modelling process. This contradicts the philosophy of BALL-SNP to be open-source and usable for both, non-experts and professionals. In consequence, we decided to make use of databases comprising available 3D models. In fact, the quality of the BALL-SNP analysis depends on the quality of the 3D model whose quality relies on the availability of adequate templates. Hence, different 3D models may receive different analysis results. Experienced users, though, are able to compare different models in BALL-SNP by root-mean-square deviation (RMSD) calculations and select the most appropriate one for their studies. Unfortunately, there are cases where no 3D structure and even no valid 3D model or only small fractions of a protein structure are available. Structural information proved its significant importance when assessing the effects of nsSNVs on a protein's function and stability. In consequence, we prioritize quality results instead of being able to return any result and accept this limitation. In addition, the PDB is a collection of 3D information collected world-wide and though sustained efforts to maintain high-quality data exist, there are unfortunately inconsistent and informal PDB files, not fully checked for errors. Hence, BALL-SNP might not be able to return analysis results for each PDB file deposited in the PDB. We, however, implemented strategies to overcome common inconsistencies such as the correction of the residue index within the ATOM records in PDB files. Whenever further inconsistencies are identified, corresponding correction methods will have to be implemented in the future to avoid failures.

Since BALL-SNP is an open source project and due to its modular architecture, it is easily extendable and adaptable to include further third party tools or retieve additional information from data resources. In future versions of BALL-SNP, additional expert knowledge provided by professional users concerning detected disease-relation of particular nsSNVs may be incorporated to enable user-defined application scenarios. In addition, genetic variations have been recognized to affect drug selection, dosing and adverse events [18]. NsSNV-introduced amino acid substitutions may change existing binding affinities of a protein or a potential drug target. Hence, the integration of a workflow for therapeutic use in BALL-SNP to also study interactions of putative medical substrates and drug targets represents an important extension for furture work. A database with small ligands and known drugs should be included to test whether mutations alter known binding affinities by comparing the binding to the unmodified protein against the binding to the mutated protein. To this end, BALL-SNP may also assist drug development with respect to individual genetic variations.

Besides nsSNV sets in one gene, we also studied nsSNV combinations in multiple genes. According to Yue et al., about 25% of nsSNVs are deleterious but not disease-related [127]. Since proteins with redundant function exist, no noticeable phenotypic change may be detected when only one involved gene inheres an nsSNV introducing a dysfunction of the encoded protein [128]. Several *in silico* prediction tools, however,

penalize deleterious effects of nsSNVs as disease-associated without considering the network environment. Due to overlapping protein functions, produced dysfunctions may be buffered by the environment [129]. In consequence, the analysis of the pathogenic influence of nsSNV sets in multiple genes requires a multi-scale pipeline to capture gene-gene environment effects.

In this thesis, we developed an integrative approach based on state-of-the-art pathogenicity prediction of single nsSNVs, association rule mining, pathway analysis and subcellular localization of the involved genes. We demonstrated its strength on the analysis of the DCM data set and a healthy control cohort based on the general population of the 1000 Genomes Project. Via association rule learning, we detected associated nsSNV sets in seven genes present in at least 70% of all 639 patient samples, but not in the control cohort. Structural analyses of these revealed primarily an influence on the protein stability, which agrees with previous studies on nsSNVs [119]. For amino acid substitutions located at the protein surface, we generally observed a proximity to putative binding pockets. Considering the subcellular localization of the proteins encoded by the genes harboring the associated nsSNVs, we observed a cascade, starting from the nucleus, proceeding through the inner cell body to the extracellular matrix. The performed systematic analysis point to a connection of the identified genes featuring associated nsSNVs - in functionality as well as in their contribution to biological pathways. In conclusion, the detected associated nsSNVs may putatively influence cardiovascular phenotypes. Since statistical relevance not necessarily indicates clinical relevance, further analyses and experimental validation are required. Despite cardiovascular diseases, the developed pipeline can likewise be applied to other diseases such as cancer.

Currently, the data suitable for such multi-scale studies, though, is limited. The majority of the available classical exome capture studies reveal lower coverage rates compared to the DCM data. In addition, control cohorts of healthy humans are limited, particularly in high quality. The interpretation of results obtained by a comparison of data differing in quality consequently poses a great challenge on the scientific community. Furthermore, the lack of experimental information often prevents the detection of clinical significance as well as its validation. Future studies critically rely on advances in high-quality data generation.

In this thesis, we concentrated on genetic variants in coding regions, in particular nsSNVs. Beyond these, intronic SNVs in promoter regions and splicing sites as well as genetic variants in regulatory regions may have an influence on human phenotypes, since they may affect gene splicing, transcription factor binding and messenger RNA degradation. In fact, about 88% of genetic variants with weak trait-association from GWAS represent non-coding variants [130]. The prediction of their pathogenic impact *in silico*, however, is even more complex. Their effect, though, is expected to be weaker compared to coding nsSNVs [131].

Due to missing computational methods to analyze cumulative nsSNVs and to assess their impact on pathogenicity, the validation of their clinical relevance is limited. In fact, genetic testing enables predictive diagnosis and can enhance pre-symptomatic intervention. Future studies, however, focusing on translation of computational findings to applicable mechanisms in clinical routine and capturing diagnostic demands, are highly required.

To improve medical treatment, computational approaches should be designed to address the requirements in clinical application. An intuitive and, in particular, visual inspection of genomic data and genetic information might have a greater chance to reach clinical acceptance, since imaging techniques such as e.g. magnetic resonance tomography (MRT) have demonstrated their great importance in clinics over the last years. In addition, the knowledge of individual nsSNV combinations and their functional impact may pave the way to tailor medical care for patient-specific treatment [132].

In summary, we developed approaches to cover these requirements of computational diagnostics within this thesis. We implemented a non-commercial software tool - BALL-SNP - to promote nsSNV candidate selection by inspection and visualization of combined genetic and structural information. Although further extensions are required to assist clinical use of NGS data, BALL-SNP already decisively contributes to existing instruments of candidate nsSNV analysis. Furthermore, the developed multi-scale analysis pipeline revealed promising results for the study of DCM patients and may serve as template for future approaches to identify complex interactions of genes harboring nsSNVs.

References

[1] G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, 2010.

[2] F. S. Collins, M. S. Guyer, and A. Charkravarti. Variations on a theme: cataloging human DNA sequence variation. *Science*, 278(5343):1580–1581, Nov 1997.

[3] J. N. Bailey, M. A. Pericak-Vance, and J. L. Haines. The impact of the human genome project on complex disease. *Genes (Basel)*, 5(3):518–35, 2014.

[4] M. Choi, U. I. Scholl, W. Ji, T. Liu, I. R. Tikhonova, P. Zumbo, A. Nayir, A. Bakkaloglu, S. Ozen, S. Sanjad, C. Nelson-Williams, A. Farhi, S. Mane, and R. P. Lifton. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A*, 106(45):19096–101, 2009.

[5] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 29(1):308–11, 2001.

[6] P. D. Stenson, E. Ball, K. Howells, A. Phillips, M. Mort, and D. N. Cooper. Human Gene Mutation Database: towards a comprehensive central mutation database. *J Med Genet*, 45(2):124–6, 2008.

[7] P. D. Stenson, E. V. Ball, M. Mort, A. D. Phillips, K. Shaw, and D. N. Cooper. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Curr Protoc Bioinformatics*, Chapter 1:Unit1 13, 2012.

[8] G. H. Fernald, E. Capriotti, R. Daneshjou, K. J. Karczewski, and R. B. Altman. Bioinformatics challenges for personalized medicine. *Bioinformatics*, 27(13):1741–1748, Jul 2011.

[9] J. H. Moore, F. W. Asselbergs, and S. M. Williams. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–455, Feb 2010.

[10] Tugba G. Kucukkal, Ye Yang, Susan C. Chapman, Weiguo Cao, and Emil Alexov. Computational and experimental approaches to reveal the effects of single nucleotide polymorphisms with respect to disease diagnostics. *International Journal of Molecular Sciences*, 15(6):9670, 2014.

[11] S. C. Mueller, C. Backes, J. The Inheritance Study Group Haas, H. A. Katus, B. Meder, E. Meese, and A. Keller. Pathogenicity prediction of non-synonymous single nucleotide variants in dilated cardiomyopathy. *Brief Bioinform*, 2015.

[12] R. Guerois, J. E. Nielsen, and L. Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol*, 320(2):369–87, 2002.

[13] M. Li, M. Petukh, E. Alexov, and A. R. Panchenko. Predicting the impact of missense mutations on protein-protein binding affinity. *J Chem Theory Comput*, 10(4):1770–1780, 2014.

[14] S. Stefl, H. Nishi, M. Petukh, A. R. Panchenko, and E. Alexov. Molecular mechanisms of disease-causing missense mutations. *J Mol Biol*, 425(21):3919–36, 2013.

[15] The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–320, 2005.

[16] N. J. Schork, S. S. Murray, K. A. Frazer, and E. J. Topol. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev*, 19(3):212–9, 2009.

[17] A. H. Burghes, H. E. Vaessin, and A. de La Chapelle. Genetics. the land between Mendelian and multifactorial inheritance. *Science*, 293(5538):2213–2214, Sep 2001.

[18] K. M. Giacomini, C. M. Brett, R. B. Altman, N. L. Benowitz, M. E. Dolan, D. A. Flockhart, J. A. Johnson, D. F. Hayes, T. Klein, R. M. Krauss, D. L. Kroetz, H. L. McLeod, A. T. Nguyen, M. J. Ratain, M. V. Relling, V. Reus, D. M. Roden, C. A. Schaefer, A. R. Shuldiner, T. Skaar, K. Tantisira, R. F. Tyndale, L. Wang, R. M. Weinshilboum, S. T. Weiss, and I. Zineh. The pharmacogenetics research network: from SNP discovery to clinical drug response. *Clin. Pharmacol. Ther.*, 81(3):328–345, Mar 2007.

[19] T. G. Kucukkal, M. Petukh, L. Li, and E. Alexov. Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. *Curr Opin Struct Biol*, 32C:18–24, 2015.

[20] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, Oct 2009.

[21] L. Stryer. *Biochemistry*. Spektrum Akademischer Verlag, 1996.

[22] Barkur S. Shastry. Role of SNPs and Haplotypes in Human Disease and Drug Development. In Mauro Ferrari, Mihrimah Ozkan, and Michael J. Heller, editors, *BioMEMS and Biomedical Nanotechnology*, pages 447–458. Springer US, 2007.

[23] M. J. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, and D. R. Maglott. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*, 42(Database issue):D980–5, 2014.

[24] A. Bateman, M. J. Martin, C. O'Donovan, M. Magrane, R. Apweiler, E. Alpi, R. Antunes, J. Arganiska, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, G. Chavali, E. Cibrian-Uhalte, A. D. Silva, M. De Giorgi, T. Dogan, F. Fazzini, P. Gane, L. G. Castro, P. Garmiri, E. Hatton-Ellis, R. Hieta, R. Huntley, D. Legge, W. Liu, J. Luo, A. MacDougall, P. Mutowo, A. Nightingale, S. Orchard, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Turner, V. Volynkin, T. Wardell, X. Watkins, H. Zellner, A. Cowley, L. Figueira, W. Li, H. McWilliam, R. Lopez, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M. C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. de Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, B. E. Suzek, C. Vinayaka, Q. Wang, Y. Wang, L. S. Yeh, M. S. Yerramalla, and J. Zhang. UniProt: a hub for protein information. *Nucleic Acids Res.*, 43(Database issue):D204–212, Jan 2015.

[25] Y. L. Yip, H. Scheib, A. V. Diemand, A. Gattiker, L. M. Famiglietti, E. Gasteiger, and A. Bairoch. The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Hum Mutat*, 23(5):464–70, 2004.

[26] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*, 34(Database issue):D187–91, 2006.

[27] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–42, 2000.

[28] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, 34(Database issue):D668–672, Jan 2006.

[29] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, 4(Database issue):D1091–1097, Jan 2014.

[30] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1):25–29, May 2000.

[31] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate, and M. Punta. Pfam: the protein families database. *Nucleic Acids Res*, 42(Database issue):D222–30, 2014.

[32] H. Mi, A. Muruganujan, and P. D. Thomas. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res*, 41(Database issue):D377–86, 2013.

[33] P. D. Thomas, M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res*, 13(9):2129–41, 2003.

[34] E. Capriotti, R. Calabrese, and R. Casadio. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, 22(22):2729–34, 2006.

[35] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4):248–9, 2010.

[36] S. R. Sunyaev, F. Eisenhaber, I. V. Rodchenkov, B. Eisenhaber, V. G. Tumanyan, and E. N. Kuznetsov. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.*, 12(5):387–394, May 1999.

[37] Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan. Predicting the functional effect of amino acid substitutions and indels. *PLoS One*, 7(10):e46688, 2012.

[38] E. Capriotti, P. Fariselli, and R. Casadio. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res*, 33(Web Server issue):W306–10, 2005.

[39] FASTA format description of NCBI. http://blast.ncbi.nlm.nih.gov/blastcgihelp.shtml.

[40] R. Tibshirani, T. Hastie, and J. Friedman. *The Elements of Statistical Learning*. Springer Verlag, 2001.

[41] S. Khan and M. Vihinen. Spectrum of disease-causing mutations in protein secondary structures. *BMC Struct Biol*, 7:56, 2007.

[42] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.*, 89(22):10915–10919, Nov 1992.

[43] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, Oct 1990.

[44] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, Sep 1997.

[45] S. Henikoff and J. G. Henikoff. Embedding strategies for effective use of information from multiple sequence alignments. *Protein Sci.*, 6(3):698–705, Mar 1997.

[46] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292(2):195–202, Sep 1999.

[47] A. Hildebrandt, A. K. Dehof, A. Rurainski, A. Bertsch, M. Schumann, N. C. Toussaint, A. Moll, D. Stockel, S. Nickels, S. C. Mueller, H. P. Lenhof, and O. Kohlbacher. BALL–biochemical algorithms library 1.3. *BMC Bioinformatics*, 11:531, 2010.

[48] BALL architecture.
http://www.ballview.org/Overview/Architecture.

[49] B. Stroustrup. *The C++ Programming Language*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2000.

[50] A. Moll, A. Hildebrandt, H. P. Lenhof, and O. Kohlbacher. BALLView: a tool for research and education in molecular modeling. *Bioinformatics*, 22(3):365–6, 2006.

[51] K. Wang, M. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, 38(16):e164, 2010.

[52] A. R. Leach. *Molecular Modelling: Principles and Applications*. Prentice Hall, 2001.

[53] M. Y. Shen and A. Sali. Statistical potential for assessment and prediction of protein structures. *Protein Sci.*, 15(11):2507–2524, Nov 2006.

[54] E. di Luccio and P. Koehl. A quality metric for homology modeling: the H-factor. *BMC Bioinformatics*, 12:48, 2011.

[55] N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper, and A. Sali. Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics*, Chapter 5:Unit 5.6, Oct 2006.

[56] U. Pieper, B. M. Webb, G. Q. Dong, D. Schneidman-Duhovny, H. Fan, S. J. Kim, N. Khuri, Y. G. Spill, P. Weinkam, M. Hammel, J. A. Tainer, M. Nilges, and A. Sali. ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, 42(Database issue):D336–346, Jan 2014.

[57] M. Higurashi, T. Ishida, and K. Kinoshita. PiSite: a database of protein interaction sites using multiple binding states in the PDB. *Nucleic Acids Res.*, 37(Database issue):D360–364, Jan 2009.

[58] Exome Aggregation Consortium (ExAC), Cambridge, MA. http://exac.broadinstitute.org.

[59] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, jun 1993.

[60] Kurt Hornik, Bettina Grün, and Michael Hahsler. arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. *Journal of statistical software*, 14:1–25, 2005.

[61] Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. Algorithms for Association Rule Mining - a General Survey and Comparison. *SIGKDD Explor. Newsl.*, 2(1):58–64, jun 2000.

[62] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering, and L. J. Jensen. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*, 41(Database issue):D808–15, 2013.

[63] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. von Mering. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, 43(Database issue):D447–452, Jan 2015.

[64] E. C. Dimmer, R. P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, M. J. Martin, B. Bely, P. Browne, W. Mun Chan, R. Eberhardt, M. Gardner, K. Laiho, D. Legge, M. Magrane, K. Pichler, D. Poggioli, H. Sehra, A. Auchincloss, K. Axelsen, M. C. Blatter, E. Boutet, S. Braconi-Quintaje, L. Breuza, A. Bridge, E. Coudert, A. Estreicher, L. Famiglietti, S. Ferro-Rojas, M. Feuermann, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, J. James, S. Jimenez, F. Jungo, G. Keller, P. Lemercier, D. Lieberherr, P. Masson, M. Moinat, I. Pedruzzi, S. Poux, C. Rivoire, B. Roechert, M. Schneider, A. Stutz, S. Sundaram, M. Tognolli, L. Bougueleret, G. Argoud-Puy, I. Cusin, P. Duek-Roggli, I. Xenarios, and R. Apweiler. The UniProt-GO annotation database in 2011. *Nucleic Acids Res.*, 40(Database issue):D565–570, Jan 2012.

[65] R. P. Huntley, T. Sawford, P. Mutowo-Meullenet, A. Shypitsyna, C. Bonilla, M. J. Martin, and C. O'Donovan. The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Res*, 2014.

[66] K. F. Aoki-Kinoshita and M. Kanehisa. Gene annotation and pathway mapping in KEGG. *Methods Mol Biol*, 396:71–91, 2007.

[67] X. de La Cruz and M. Calvo. Use of surface area computations to describe atom-atom interactions. *J Comput Aided Mol Des*, 15(6):521–32, 2001.

[68] S.J. Hubbard and J.M. Thornton. NACCESS. *Computer Program London, Department of Biochemistry and Molecular Biology, University College London*, 1993.

[69] B. Huang and M. Schroeder. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol*, 6:19, 2006.

[70] B. Sommer, B. Kormeier, P. S. Demenkov, P. Arrigo, K. Hippe, O. Ates, A. V. Kochetov, V. A. Ivanisenko, N. A. Kolchanov, and R. Hofestadt. Subcellular localization charts: a new visual methodology for the semi-automatic localization of protein-related data sets. *J Bioinform Comput Biol*, 11(1):1340005, 2013.

[71] B. Kormeier. *Data Warehouses in Bioinformatics*, pages 111–130. Springer, 2014.

[72] J. Haas, K. S. Frese, B. Peil, W. Kloos, A. Keller, R. Nietsch, Z. Feng, S. Muller, E. Kayvanpour, B. Vogel, F. Sedaghat-Hamedani, W. K. Lim, X. Zhao, D. Fradkin, D. Kohler, S. Fischer, J. Franke, S. Marquart, I. Barb, D. T. Li, A. Amr,

P. Ehlermann, D. Mereles, T. Weis, S. Hassel, A. Kremer, V. King, E. Wirsz, R. Isnard, M. Komajda, A. Serio, M. Grasso, P. Syrris, E. Wicks, V. Plagnol, L. Lopes, T. Gadgaard, H. Eiskjaer, M. Jorgensen, D. Garcia-Giustiniani, M. Ortiz-Genga, M. G. Crespo-Leiro, R. H. Deprez, I. Christiaans, I. A. van Rijsingen, A. A. Wilde, A. Waldenstrom, M. Bolognesi, R. Bellazzi, S. Morner, J. L. Bermejo, L. Monserrat, E. Villard, J. Mogensen, Y. M. Pinto, P. Charron, P. Elliott, E. Arbustini, H. A. Katus, and B. Meder. Atlas of the clinical genetics of human dilated cardiomyopathy. *Eur Heart J*, 2014.

[73] 1000 genomes project. www.1000genomes.org.

[74] J. Jameson, D. L. Kasper, and T. R. Harrison. *Harrison's: Principles of Internal Medicine*. McGraw-Hill Medical Publishing Division, New York, 16th edition, 2005.

[75] E. Grunig, J. A. Tasman, H. Kucherer, W. Franz, W. Kubler, and H. A. Katus. Frequency and phenotypes of familial dilated cardiomyopathy. *J. Am. Coll. Cardiol.*, 31(1):186–194, Jan 1998.

[76] J. Haas, K. S. Frese, Y. J. Park, A. Keller, B. Vogel, A. M. Lindroth, D. Weichenhan, J. Franke, S. Fischer, A. Bauer, S. Marquart, F. Sedaghat-Hamedani, E. Kayvanpour, D. Kohler, N. M. Wolf, S. Hassel, R. Nietsch, T. Wieland, P. Ehlermann, J. H. Schultz, A. Dosch, D. Mereles, S. Hardt, J. Backs, J. D. Hoheisel, C. Plass, H. A. Katus, and B. Meder. Alterations in cardiac DNA methylation in human dilated cardiomyopathy. *EMBO Mol Med*, 5(3):413–29, 2013.

[77] B. Giardine, C. Riemer, T. Hefferon, D. Thomas, F. Hsu, J. Zielenski, Y. Sang, L. Elnitski, G. Cutting, H. Trumbower, A. Kern, R. Kuhn, G. P. Patrinos, J. Hughes, D. Higgs, D. Chui, C. Scriver, M. Phommarinh, S. K. Patnaik, O. Blumenfeld, B. Gottlieb, M. Vihinen, J. Valiaho, J. Kent, W. Miller, and R. C. Hardison. PhenCode: connecting ENCODE data with mutations and phenotype. *Hum Mutat*, 28(6):554–62, 2007.

[78] H. Piirilä, J. Valiaho, and M. Vihinen. Immunodeficiency mutation databases (IDbases). *Hum Mutat*, 27(12):1200–8, 2006.

[79] J. Thusberg, A. Olatubosun, and M. Vihinen. Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat*, 32(4):358–68, 2011.

[80] S. Castellana and T. Mazza. Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools. *Brief Bioinform*, 14(4):448–59, 2013.

[81] K. Frousios, C. S. Iliopoulos, T. Schlitt, and M. A. Simpson. Predicting the functional consequences of non-synonymous DNA sequence variants–evaluation

of bioinformatics tools and development of a consensus strategy. *Genomics*, 102(4):223–8, 2013.

[82] A. Gonzalez-Perez and N. Lopez-Bigas. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet*, 88(4):440–9, 2011.

[83] J. Bendl, J. Stourac, O. Salanda, A. Pavelka, E. D. Wieben, J. Zendulka, J. Brezovsky, and J. Damborsky. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput Biol*, 10(1):e1003440, 2014.

[84] P. C. Ng and S. Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31(13):3812–4, 2003.

[85] B. Reva, Y. Antipin, and C. Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res*, 39(17):e118, 2011.

[86] P. Yue, E. Melamud, and J. Moult. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, 7:166, 2006.

[87] B. Li, V. G. Krishnan, M. E. Mort, F. Xin, K. K. Kamati, D. N. Cooper, S. D. Mooney, and P. Radivojac. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, 25(21):2744–50, 2009.

[88] E. Capriotti, R. Calabrese, P. Fariselli, P. L. Martelli, R. B. Altman, and R. Casadio. WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. *BMC Genomics*, 14 Suppl 3:S6, 2013.

[89] Y. Bromberg, G. Yachdav, and B. Rost. SNAP predicts effect of mutations on protein function. *Bioinformatics*, 24(20):2397–8, 2008.

[90] C. Ferrer-Costa, J. L. Gelpi, L. Zamakola, I. Parraga, X. de la Cruz, and M. Orozco. PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics*, 21(14):3176–8, 2005.

[91] M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P. L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker, and G. D. Bader. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc*, 2(10):2366–82, 2007.

[92] C. Backes, C. Harz, U. Fischer, J. Schmitt, N. Ludwig, B. S. Petersen, S. C. Mueller, Y. J. Kim, N. M. Wolf, H. A. Katus, B. Meder, R. Furtwangler, A. Franke, R. Bohle, W. Henn, N. Graf, A. Keller, and E. Meese. New insights into the genetics of glioblastoma multiforme by familial exome sequencing. *Oncotarget*, 6(8):5918–31, 2015.

[93] M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, N. Shaw, C. R. Lane, E. P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G. Q. Daley, and E. S. Lander. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.*, 22(3):231–238, Jul 1999.

[94] T. R. Rebbeck, M. Spitz, and X. Wu. Assessing the function of genetic variants in candidate gene association studies. *Nat. Rev. Genet.*, 5(8):589–597, Aug 2004.

[95] B. Kneissl, S. C. Mueller, C. S. Tautermann, and A. Hildebrandt. String kernels and high-quality data set for improved prediction of kinked helices in alpha-helical membrane proteins. *J Chem Inf Model*, 51(11):3017–25, 2011.

[96] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, Sep 1996.

[97] P. D. Thomas and A. Kejariwal. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc. Natl. Acad. Sci. U.S.A.*, 101(43):15398–15403, Oct 2004.

[98] S. C. Mueller, C. Backes, O. V. Kalinina, B. Meder, D. Stockel, H. P. Lenhof, E. Meese, and A. Keller. BALL-SNP: combining genetic and structural information to identify candidate non-synonymous single nucleotide polymorphisms. *Genome Med*, 7(1):65, 2015.

[99] T. M. Olson, S. Illenberger, N. Y. Kishimoto, S. Huttelmaier, M. T. Keating, and B. M. Jockusch. Metavinculin mutations alter actin interaction in dilated cardiomyopathy. *Circulation*, 105(4):431–7, 2002.

[100] E. Capriotti, P. Fariselli, and R. Casadio. A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics*, 20 Suppl 1:i63–8, 2004.

[101] Caetano Traina Jr and Agma Traina. Fast feature selection using fractal dimensions. *Proceedings of the 15th Brazilian Symposium on Databases*, 2000.

[102] S. Nickel, D. Stoeckel, S.C. Mueller, H.-P. Lenhof, A. Hildebrandt, and A.K. Dehof. PresentaBALL - a Powerful Package for Presentations and Lessions in Structural Biology. *BioVis - 3rd IEEE Symposium on Biological Data Visualization*, 2013.

[103] A. Asimaki, P. Syrris, T. Wichter, P. Matthias, J. E. Saffitz, and W. J. McKenna. A novel dominant mutation in plakoglobin causes arrhythmogenic right ventricular cardiomyopathy. *Am J Hum Genet*, 81(5):964–73, 2007.

[104] D. H. Teng, Y. Chen, L. Lian, P. C. Ha, S. V. Tavtigian, and A. K. Wong. Mutation analyses of 268 candidate genes in human tumor cell lines. *Genomics*, 74(3):352–364, Jun 2001.

[105] K. V. Fuentes Fajardo, D. Adams, C. E. Mason, M. Sincan, C. Tifft, C. Toro, C. F. Boerkoel, W. Gahl, and T. Markello. Detecting false-positive signals in exome sequencing. *Hum. Mutat.*, 33(4):609–613, Apr 2012.

[106] B. Zhu, B. Ramachandran, and T. Gulick. Alternative pre-mRNA splicing governs expression of a conserved acidic transactivation domain in myocyte enhancer factor 2 factors of striated muscle and brain. *J. Biol. Chem.*, 280(31):28749–28760, Aug 2005.

[107] N. Katsanis, S. J. Ansley, J. L. Badano, E. R. Eichers, R. A. Lewis, B. E. Hoskins, P. J. Scambler, W. S. Davidson, P. L. Beales, and J. R. Lupski. Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder. *Science*, 293(5538):2256–2259, Sep 2001.

[108] C. Ferrer-Costa, M. Orozco, and X. de la Cruz. Characterization of compensated mutations in terms of structural and physico-chemical properties. *J Mol Biol*, 365(1):249–56, 2007.

[109] V. Westphal, S. Kjaergaard, E. Schollen, K. Martens, S. Grunewald, M. Schwartz, G. Matthijs, and H. H. Freeze. A frequent mild mutation in ALG6 may exacerbate the clinical severity of patients with congenital disorder of glycosylation Ia (CDG-Ia) caused by phosphomannomutase deficiency. *Hum Mol Genet*, 11(5):599–604, 2002.

[110] H. J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.*, 11(20):2463–2468, Oct 2002.

[111] S.C. Mueller, B. Sommer, C. Backes, J. Haas, B. Meder, E. Meese, and A. Keller. From single variants to protein cascades: Multi-scale modeling of SNV sets in genetic disorders.

[112] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101(16):6062–7, 2004.

[113] Gabor Csardi Nepusz and Tamas. The igraph software package for complex network research. *InterJournal*, Complex Systems, 2006.

[114] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33(Database issue):D514–7, 2005.

[115] W. P. Yang, P. C. Levesque, W. A. Little, M. L. Conder, P. Ramakrishnan, M. G. Neubauer, and M. A. Blanar. Functional expression of two KvLQT1-related potassium channels responsible for an inherited idiopathic epilepsy. *J Biol Chem*, 273(31):19419–23, 1998.

[116] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res*, 36(Database issue):D901–6, 2008.

[117] E. Capriotti and R. B. Altman. Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC Bioinformatics*, 12 Suppl 4:S3, 2011.

[118] S. Sunyaev, V. Ramensky, and P. Bork. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet.*, 16(5):198–200, May 2000.

[119] Z. Wang and J. Moult. SNPs, protein structure, and disease. *Hum Mutat*, 17(4):263–70, 2001.

[120] P. Yue, Z. Li, and J. Moult. Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol*, 353(2):459–73, 2005.

[121] J. Huang, L. Perez-Burgos, B. J. Placek, R. Sengupta, M. Richter, J. A. Dorsey, S. Kubicek, S. Opravil, T. Jenuwein, and S. L. Berger. Repression of p53 activity by smyd2-mediated methylation. *Nature*, 444(7119):629–32, 2006.

[122] L. Wang, L. Li, H. Zhang, X. Luo, J. Dai, S. Zhou, J. Gu, J. Zhu, P. Atadja, C. Lu, E. Li, and K. Zhao. Structure of human SMYD2 protein reveals the basis of p53 tumor suppressor methylation. *J Biol Chem*, 286(44):38725–37, 2011.

[123] A. Chang, I. Schomburg, S. Placzek, L. Jeske, M. Ulbrich, M. Xiao, C. W. Sensen, and D. Schomburg. BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Res*, 43(Database issue):D439–46, 2015.

[124] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–9, 2000.

[125] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D'Eustachio. The Reactome pathway knowledgebase. *Nucleic Acids Res*, 42(Database issue):D472–7, 2014.

[126] K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabasi. The human disease network. *Proc. Natl. Acad. Sci. U.S.A.*, 104(21):8685–8690, May 2007.

[127] P. Yue and J. Moult. Identification and analysis of deleterious human SNPs. *J Mol Biol*, 356(5):1263–74, 2006.

[128] M. A. Nowak, M. C. Boerlijst, J. Cooke, and J. M. Smith. Evolution of genetic redundancy. *Nature*, 388(6638):167–171, Jul 1997.

[129] G. Z. Wang, J. Liu, W. Wang, H. Y. Zhang, and M. J. Lercher. A gene's ability to buffer variation is predicted by its fitness contribution and genetic interactions. *PLoS ONE*, 6(3):e17650, 2011.

[130] No authors listed. The International HapMap Project. *Nature*, 426(6968):789–796, Dec 2003.

[131] E. Capriotti, N. L. Nehrt, M. G. Kann, and Y. Bromberg. Bioinformatics for personal genome interpretation. *Brief. Bioinformatics*, 13(4):495–512, Jul 2012.

[132] T. G. Kucukkal, Y. Yang, S. C. Chapman, W. Cao, and E. Alexov. Computational and experimental approaches to reveal the effects of single nucleotide polymorphisms with respect to disease diagnostics. *Int J Mol Sci*, 15(6):9670–717, 2014.

Appendix: Table of Abbreviations

AA	Amino Acid
ATP	Adenosine Triphosphate
BALL	Biochemical Algorithms Library
BLAST	Basic Local Alignment Search Tool
CmPI	CELLmicrocosmos 4.2 PathwayIntegration
CTC	Circulating Tumor Cell
DCM	Dilated Cardiomyopathy
DOPE	Discrete Optimized Protein Energy
ExAC	Exome Aggregation Consortium
FN	False Negatives
FP	False Positives
GO	Gene Ontology
GUI	Graphical User Interface
GWAS	Genome-Wide Association Studies
HGMD	Human Mutation Database
HMM	Hidden Markov Models
KEGG	Kyoto Encyclopedia of Genes and Genomes
MT	Mutant Type (amino acid)
NMR	Nuclear Magnetic Resonance
NIH	National Institutes of Health
nsSNV	non-synonymous Single Nucleotide Variant
PDB	Protein Data Bank
PSIC	Position-Specific Independent Counts
PSSM	Position-Specific Substitution Matrix
RMSD	Root-Mean Square Deviation
SES	Solvent-Excluded Surface
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
subPSEC	substitution Position-Specific Evolutionary Conservation
SVM	Support Vector Machines
SW	Smith-Waterman
TM	Transmembrane
TN	True Negatives
TP	True Positives
UML	Unified Modeling Language
VCF	Variant Call Format
WT	Wild Type (amino acid)

Amino acid	Three-letter abbreviation	One-letter abbreviation	Amino acid	Three-letter abbreviation	One-letter abbreviation
Alanine	Ala	A	Methionine	Met	M
Arginine	Arg	R	Phenylalanine	Phe	F
Asparagine	Asn	N	Proline	Pro	P
Aspartic Acid	Asp	D	Serine	Ser	S
Cysteine	Cys	C	Threonine	Thr	T
Glutamine	Gln	Q	Tryptophan	Trp	W
Glutamic Acid	Glu	E	Tyrosine	Tyr	Y
Glycine	Gly	G	Valine	Val	V
Histidine	His	H			
Isoleucine	Ile	I			
Leucine	Leu	L			
Lysine	Lys	K			

Abbreviations for amino acids [21].

Publication List

Journal Publications

- **2015**
  - Christina Backes, Petra Leidinger, Gabriela Altmann, Maximilian Wuerstle, Benjamin Meder, Valentina Galata, Sabine C Mueller, Daniel Sickert, Cord Sthler, Eckart Meese, Andreas Keller: The influence of Next-Generation Sequencing and storage conditions on miRNA patterns generated from PAXgene blood. Analytical Chemistry 07/2015.
  - Nicole Ludwig, Yoo-Jin Kim, Sabine C. Mueller, Christina Backes, Tamara V. Werner, Valentina Galata, Elke Sartorius, Rainer M. Bohle, Andreas Keller, Eckart Meese: Posttranscriptional deregulation of signaling pathways in meningioma subtypes by differential expression of miRNAs. Neuro-Oncology 02/2015.
  - Sabine C. Mueller, Christina Backes, Jan Haas, Hugo A. Katus, Benjamin Meder, Eckart Meese, Andreas Keller: Pathogenicity prediction of non-synonymous single nucleotide variants in dilated cardiomyopathy. Briefings in Bioinformatics 01/2015.
  - Sabine C. Mueller, Christina Backes, Olga V. Kalinina, Benjamin Meder, Daniel Stoeckel, Hans-Peter Lenhof, Eckart Meese, Andreas Keller: BALL-SNP: combining genetic and structural information to identify candidate non-synonymous single nucleotide polymorphisms. Genome Medicine 01/2015; 7(1):65.
- **2015 - submitted**
  - Sabine C. Mueller, Bjoern Sommer, Christina Backes, Jan Haas, Benjamin Meder, Eckart Meese, Andreas Keller: From single variants to protein cascades: Multi-scale modeling of SNV sets in genetic disorders.
  - Sabine C. Mueller, Christina Backes, Alexander Gress, Nina Baumgarten, Olga V. Kalinina, Andreas Moll, Oliver Kohlbacher, Eckart Meese, Andreas Keller: BALL-SNPgp - from genetic variants towards computational diagnostics.

- **2014**
  - Christina Backes, Christian Harz, Ulrike Fischer, Jana Schmitt, Nicole Ludwig, Britt-Sabina Petersen, Sabine C. Mueller, Yoo-Jin Kim, Nadine M. Wolf, Hugo A. Katus, Benjamin Meder, Rhoikos Furtwaengler, Andre Franke, Rainer Bohle, Wolfram Henn, Norbert Graf, Andreas Keller, Eckart Meese: New insights into the genetics of glioblastoma multiforme by familial exome sequencing.Oncotarget 12/2014.
  - Andreas Keller, Petra Leidinger, Britta Vogel, Christina Backes, Abdou ElSharawy, Valentina Galata, Sabine C. Mueller, Sabine Marquart, Michael G. Schrauder, Reiner Strick, Andrea Bauer, Joerg Wischhusen, Markus Beier, Jochen Kohlhaas, Hugo A. Katus, Joerg Hoheisel, Andre Franke, Benjamin Meder, Eckart Meese: miRNAs can be generally associated with human pathologies as exemplified for miR-144*. BMC Medicine 12/2014; 12(1):224.
  - Jan Haas, Karen S. Frese, Barbara Peil, Wanda Kloos, Andreas Keller, Rouven Nietsch, Zhu Feng, Sabine Mueller, Elham Kayvanpour, Britta Vogel, Farbod Sedaghat-Hamedani, Wei-Keat Lim, Xiaohong Zhao, Dmitriy Fradkin, Doreen Koehler, Simon Fischer, Jennifer Franke, Sabine Marquart, Ioana Barb, Daniel Tian Li, Ali Amr, Philipp Ehlermann, Derliz Mereles, Tanja Weis, Sarah Hassel, Andreas Kremer, Vanessa King, Emil Wirsz, Richard Isnard, Michel Komajda, Alessandra Serio, Maurizia Grasso, Petros Syrris, Eleanor Wicks, Vincent Plagnol, Luis Lopes, Tenna Gadgaard, Hans Eiskjr, Mads Jrgensen, Diego Garcia-Giustiniani, Martin Ortiz-Genga, Maria G. Crespo-Leiro, Rondal H. Lekanne Dit Deprez, Imke Christiaans, Ingrid A. van Rijsingen, Arthur A. Wilde, Anders Waldenstrom, Martino Bolognesi, Riccardo Bellazzi, Stellan Moerner, Justo Lorenzo Bermejo, Lorenzo Monserrat, Eric Villard, Jens Mogensen, Yigal M. Pinto, Philippe Charron, Perry Elliott, Eloisa Arbustini, Hugo A. Katus, Benjamin Meder: Atlas of the clinical genetics of human dilated cardiomyopathy. European Heart Journal 08/2014; 36(18).
  - Tim Seifert, Andreas Lund, Benny Kneissl, Sabine C. Mueller, Christofer S. Tautermann, Andreas Hildebrandt: SKINK: a web server for string kernel based kink prediction in -helices. Bioinformatics 02/2014; 30(12).
- **2013**
  - Petra Leidinger, Christina Backes, Stephanie Deutscher, Katja Schmitt, Sabine C. Mueller, Karen Frese, Jan Haas, Klemens Ruprecht, Friedemann Paul, Cord Stahler, Christoph Jg Lang, Benjamin Meder, Tamas Bartfai, Eckart Meese, Andreas Keller: A blood based 12-miRNA signature of Alzheimer disease patients. Genome biology 07/2013; 14(7):R78.
  - Ina Meiser, Sabine C. Mueller, Friederike Ehrhart, Stephen G. Shirley,

Heiko Zimmermann: A new validation method for clinical grade microencapsulation: quantitative high speed video analysis of alginate capsule. Microsystem Technologies 01/2013; 21(1).

- **2011**
  - Benny Kneissl, Sabine C. Mueller, Christofer S. Tautermann, Andreas Hildebrandt: String Kernels and High-Quality Data Set for Improved Prediction of Kinked Helices in -Helical Membrane Proteins. Journal of Chemical Information and Modeling 11/2011; 51(11):3017-25.
- **2010**
  - Andreas Hildebrandt, Anna Katharina Dehof, Alexander Rurainski, Andreas Bertsch, Marcel Schumann, Nora C. Toussaint, Andreas Moll, Daniel Stoeckel, Stefan Nickels, Sabine C. Mueller, Hans-Peter Lenhof, Oliver Kohlbacher: BALL - biochemical algorithms library 1.3. BMC Bioinformatics 10/2010; 11:531.

Book Chapters

- **2012**
  - I. Meiser, S. C. Mueller, H. Zimmermann, F. Ehrhart: Quantitative 3D High Speed Video Analysis of Capsule Formation during Encapsulation Processes. 03/2010: pages 204-207.
- **2010**
  - I. Meiser, S. C. Mueller, M. M. Gepp, H. Zimmermann, F. Ehrhart: Quantitative High Speed Video Analysis of Biopolymer Encapsulated Cells while Capsule Formation. 12/2008: pages 2255-2258.

Conference Proceedings

- **2013**
  - S. Nickels, D. Stoeckel, S.C. Mueller, H.-P. Lenhof, A. Hildebrandt, A.K. Dehof: PresentaBALL A powerful package for presentations and lessons in structural biology. Biological Data Visualization (BioVis), 2013 IEEE Symposium on; 01/2013.
- **2012**
  - S. Nickels, H. Sminia, S.C. Mueller, B Kools, A. K. Dehof, H. Lenhof, A. Hildebrandt: ProteinScanAR - An Augmented Reality Web Application for High School Education in Biomolecular Life Sciences. Information Visualisation (IV), 2012 16th International Conference on; 01/2012.

Conference Poster

- **2012**
  - S.C. Mueller, S. Nickels, D.Stoeckel, A.K. Dehof, H.P. Lenhof, A. Hildebrandt: PresentaBALL - A Web Technology based Presentation Extension of BALL, Visualizing Biological Data (VIZBI), 2011.
- **2011**
  - S.C. Mueller, S. Nickels, K. Sons, B. Friedrich, D. Stoeckel, A.K. Dehof, H.P. Lenhof, P. Slusallek, A. Hildebrandt: Distributed Collaborative Molecular Modelling. Molecular Modelling Workshop, 2011.
  - Sabine C. Mueller, Benny Kneissl, Christofer S. Tautermann, Andreas Hildebrandt: Kinks in alpha-helical membrane proteins: Manually annotation, extensive analysis and successful prediction, European Conference on Bioinformatics (ECCB), 2011
  - S.C. Mueller, S. Nickels, D.Stoeckel, A.K. Dehof, H.P. Lenhof, A. Hildebrandt: PresentaBALL - A Web Technology based Presentation Extension of BALL, Visualizing Biological Data (VIZBI), 2011.

Acknowledgement

Mein ganz besonderer Dank gilt meinem Doktorvater Prof. Dr. Andreas Keller für die Vergabe des interessanten Themas, für seine Unterstützung und der Betreuung dieser Arbeit. Weiterhin möchte ich mich bei Prof. Dr. Eckart Meese für seine Unterstützung bedanken.

Unseren Kooperationspartnern in Heidelberg, Prof. Dr. Benjamin Meder und Dr. Jan Haas, danke ich für die gute Zusammenarbeit und die Bereitstellung des klinischen Datensatzes. Dr. Björn Sommer möchte ich für die Unterstützung hinsichtlich der Entwicklung des multiskalen Ansatzes, insbesondere für die Lokalisierungskomponente, danken. Des Weiteren gilt mein Dank Dr. Olga Kalinina und Alexander Gress für hilfreiche Diskussionen und die Zusammenarbeit bei der Entwicklung von BALL-SNP. In diesem Zusammenhang bedanke ich mich auch bei Nina Baumgarten für ihren Beitrag an der Integration einer Pathogenitätsvorhersage in BALL-SNP.

Weiterhin möchte ich mich bei allen Mitarbeitern und (ehemaligen) Kollegen der Humangenetik in Homburg sowie dem Lehrstuhl für klinische Bioinformatik in Saarbrücken bedanken, allen voran Christina Backes für ihre Diskussionsbereitschaft und ihre Unterstützung.

Meinen Freunden möchte ich danken, dass sie mich in stressigen Phasen unterstützt und aufgemuntert haben, wenn immer es nötig war.

Ein ganz besonderer Dank gilt meiner Familie für ihre Unterstützung und ihr Vertrauen in mich. Von ganzem Herzen möchte ich Michael Gepp dafür danken, dass er mich stets motiviert, immer an mich geglaubt und nie seine Geduld verloren hat.