TECHNISCHE UNIVERSITÄT MÜNCHEN

FAKULTÄT FÜR INFORMATIK

DEPARTMENT FOR BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

Next-Generation Sequencing Data Analysis

Thomas Wieland

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:

Univ.-Prof. Dr. Nassir Navab

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Burkhard Rost
2. Univ.-Prof. Dr. Dr. Fabian Theis
3. Priv.-Doz. Dr. Tim M. Strom

Die Dissertation wurde am 24.03.2015 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 11.10.2015 angenommen.

Acknowledgments

There are many people who helped me throughout this PhD project that I want to thank. First of all I want to thank my supervisor Tim Strom for all his advice and guidance over the last couple of years and for giving me the freedom to work on many interesting topics.

I would also like to thank Thomas Meitinger for giving me the opportunity to work on my project at his institute and for providing me with a lot of helpful adivce.

Next I would like to thank my PhD advisor Burkhard Rost for all the work he had to do in order to help me completing this project.

My sincere thanks also goes to my colleagues at the Institute of Human Genetics for a great working atmosphere and for helping me with all the big and small problems that are an unavoidable part of each PhD project. I want to especially thank Alice and Thomas for correcting this thesis.

And last but not least I want to thank my friends and family, without whom I couldn't have done this whole project. Especially I want to thank Franziska who had to be very patient with me in the last couple of month.

Abstract

Next-Generation Sequencing (NGS) has become one of the most important tools in the field of human genetics. Targeted resequencing of the coding part of the human genome (exome sequencing) has been performed on more than 4,500 samples from over 80 different projects in the course of this PhD project. The samples have been sequenced to identify pathogenic variants and disease associated genes in rare and common diseases. The aim of this PhD project was to investigate and develop methods and parameters to identify such pathogenic variants and genes from large amounts of exome sequencing data.

An existing analysis pipeline has been modified on a large scale in order to reduce runtime, memory usage, required disk space and hands-on time, as well as to increase flexibility and allow easier adaptation and extension. Additionally, new features have been implemented to allow the analysis of other features of the data, such as Structural Variants (SVs) or Copy Number Variations (CNVs), and to allow multiple users to analyze large projects collaboratively.

The data produced during this PhD project has been used to evaluate requirements on study design and certain key quality metrics of exome sequencing data.

Several programs and strategies for variant calling have been benchmarked. Influences of different variant calling procedures and variant quality metrics on sensitivity and specificity have been evaluated and used to draw conclusions on best-practice variant calling. Additionaly, variant calling in RNA sequencing data for detection of RNA editing is discussed.

Variant callers detect on average approximately 23,000 high quality coding variants per exome. Guidelines on filtering and selecting these variants in order to identify those that are disease causing, have been developed and are illustrated by examples, if applicable.

Zusammenfassung

Next-Generation Sequencing (NGS) wurde in den letzten Jahren zu einer der wichtigsten Technologien im Bereich der Humangenetik. Während dieses PhD Projekts wurde das Exom, also der kodierende Bereich des menschlichen Genoms, von mehr als 4.500 Proben aus 80 verschiedenen Projekten sequenziert. Diese Proben wurden sequenziert um mögliche krankheitsverursachende Varianten und Gene in seltenen und häufigen Erberkrankungen zu identifizieren. Das Ziel dieses PhD Projekts war es Methoden und Parameter zu entwickeln und zu untersuchen die das Identifizieren von solchen krankheitsverursachende Varianten und Genen in einer großen Anzahl von Exom Sequenzdaten erlauben.

Große Teile einer bereits existierenden Analyse Pipeline wurden modifiziert um Laufzeit, Arbeits- und Festspeicherverbrauch und benötigte Arbeitszeit zu reduzieren und Flexibilität und Anpassungsmöglichkeiten zu erhöhen. Außerdem wurden neue Module implementiert die die Analyse von anderen Gesichtspunkten der Daten erlauben, wie zum Beispiel die Detektion von strukturellen Varianten und Copy Number Variations (CNVs). Module die die Zusammenarbeit verschiedener Partner in größeren Projekten erlauben wurden ebenfalls entwickelt.

Die Daten die während dieses PhD Projekts entstanden, wurden verwendet um die Anforderungen an das Design von Exom Sequenz Studien und grundsätzliche Qualitätskriterien zu evaluieren.

Mehrere Programme und Strategien zur Identifikation von Varianten wurden hinsichtlich ihrer Leistung überprüft. Einflüsse verschiedener Prozeduren und Qualitätswerte der Daten auf Sensitivität und Spezifität der identifizierten Varianten wurden evaluiert um bestmögliche Strategien zu entwickeln. Zusätzlich wurden Methoden zur Identifizierung von Varianten in RNA Sequenzdaten diskutiert. Diese Methoden wurden Verwendet um Positionen an denen so genannte RNA Editierung stattfindet, zu identifizieren.

Programme zur Variantenidentifikation entdecken durchschnittlich ca. 23.000 kodierende Varianten mit guter Qualität pro Exom Datensatz. Es wurden Richtlinien zur Selektion und Filterung dieser Varianten entwickelt, die angewendet werden können um krankheitsverursachende Varianten zu identifizieren. Diese Richtlinien werden, so weit möglich, an Hand von Datensätzen die während des PhD Projekts analysiert wurden, dargestellt.

Contents

List of Figures

List of Tables

Part I.

Introduction

1. Introduction

1.1. Genetic Disorders

1.1.1. Mendelian Diseases

Mendelian, or monogenic, diseases are diseases caused by mutations in single genes. They can be divided based on their *inheritance pattern*.

Autosomal Recessive Diseases

Two mutant alleles of a disease associated gene are required to cause an autosomal recessive disease. In other words, autosomal recessive diseases are caused either by a *homozygous* mutation or by two *compound heterozygous* mutations in a disease associated gene. People with only a single heterozygous mutation are called *carriers*. Due to selective pressure, single heterozygous mutations are usually rare. Hence, autosomal recessive diseases are more likely to be caused by compound heterozygous mutations with the exception of *consanguineous* families, where homozygous mutations are more likely. The most common lethal autosomal recessive disease in caucasians is *cystic fibrosis*. It occurs in about 1 in 3,000-4,000 Germans[94] and is caused by mutations in the gene *CFTR*.

Autosomal Dominant Diseases

In autosomal dominant diseases, a single heterozygous mutation is sufficient to cause the disease. For instance, the progressive neurodegenerative disorder *Huntington's Disease (HD)* is caused by a CAG repeat in the gene *Huntingtin (HTT)*[81]. If this repeat occurs less than 36 times, an individual does not develop the disorder[130]. If it occurs more than 40 times, an individual will develop the disorder, i.e. the disease is fully *penetrant*. Penetrance is incomplete for repeat counts between 36 and 40.

HD is a late onset disease. It does not reduce reproductive fitness in affected individuals, so it is likely that the causal mutation is passed on to offsprings. Other autosomal dominant diseases, such as *intellectual disability (ID)*, affect patients in their early childhood and severely reduce reproductive fitness, so the disease causing mutations are rarely passed to offsprings. Such diseases are likely to occur due to *de novo* mutations, i.e. novel mutations that occurred in the germline of the parents.

X Chromosome Linked Diseases

X chromosome linked or *X-linked* diseases can be both recessive and dominant. However, recessive X-linked disorders have a different inheritance pattern than autosomal recessive disorders, because male individuals carry only one copy of the X-chromosome. If this copy carries a disease causing mutation, it is sufficient to cause the disease in males. Females

with a single disease causing mutation are carriers and often show a mild phenotype, due to the inactivation of one copy of the X chromosome. Sons of a female carrier have a 50% chance of being affected by the disease and daughters have a 50% chance of being also a carrier. Due to this inheritance pattern, almost exclusively male individuals are affected by X-linked recessive diseases. *Duchenne muscular dystrophy* is an example for a X-linked recessive disease.

Also dominant X-linked diseases have a special inheritance pattern: daughters can inherit a dominant X-linked disease from both their mother and their father whereas sons can only inherit it from their mother. Dominant X-linked disorders are rare. One example are X-linked dominant hypophosphatemic rickets caused by mutations in the *phosphate-regulating endopeptidase gene (PHEX)*.

Y-linked and Mitochondrial Diseases

Diseases linked to the Y chromsome occur only in male individuals. There are not many examples for Y-linked diseases, but some types of infertility are linked to the Y chromosome.

In addition to DNA in the nucleus, mitochondria also carry DNA. These so called "MT chromosomes" contain 37 genes coding for mitochondrial rRNAs, tRNAs and subunits of enzyme complexes of the oxidative phosphorylation system[126]. Mutations in these genes cause mitochondrial disorders (see also Chapter 3.3.4). In humans, mitochondria are passed from the mother to the offspring. In contrast to nuclear chromosomes, several hundreds of MT chromosomes are present in every cell. If and how severely an individual is affected by a disease depends on the proportion of MT chromosomes that carry a mutation.

1.1.2. Complex Diseases

Complex diseases are not caused by mutations in single genes, but by a combination of different genetic and environmental factors. They do not follow a mendelian inheritance pattern. *Genome Wide Association Studies (GWAS)* are used to identify genetic risk loci that play a role in complex diseases. These studies try to identify *Single Nucleotide Polymorphisms (SNPs)* that are significantly associated with a trait. A trait can either be binary, e.g. cases vs. controls, or continuous, e.g. height. If a SNP can be identified as associated with a trait depends on the *effect size* of the SNP and on the number of samples in the case and control groups. At the time of writing, more than 1,700 GWAS have been published including approximately 12,000 significantly associated SNPs[133]. For example, a consortium performed a GWAS with 34,840 type II diabetes (T2D) cases and 114,981 controls[88]. They identified 8 novel disease associated loci in addition to 55 already known loci. Together, these 63 loci account for 5.7% of variance in disease susceptibility.

1.2. DNA Sequencing

The process of determining the order of the four different nucleotides **A**denosine, **G**uanine, **C**ytosine and **T**hymine of a *Deoxyribonucleic acid (DNA)* molecule is called *DNA sequencing*. In 1977 Frederick Sanger and colleagues published a sequencing technique called

chain-termination method which is today known as *Sanger sequencing*[111]. For this method the DNA to sequence is denaturated into single stranded DNA and divided into four reactions. DNA polmymerase and the four different *deoxynucleotides (dNTPs)* (dATP, dGTP, dCTP and dTTP) are added to all four reactions. One of four *dideoxynucleotides (ddNTPs)* is added to each reaction, i.e. ddATP to the first reaction, ddGTP to the second, ddCTP to the third and ddTTP to the fourth. The DNA polymerase synthesizes new DNA strands complementary to the input single stranded DNA by incorporating the appropriate dNTPs. It randomly also incorporates the ddNTPs specific to each reaction. If a ddNTP is incorporated the elongation of the strand stops because ddNTP molecules lack a 3'-OH group which prevents further dNTPs from binding. This leads to DNA molecules with different lengths. The molecules are then sorted by their weight which corresponds to their length. Because only one type of ddNTP was present in each reaction, the last nucleotide of each molecule is known, which allows the reconstruction of the DNA sequence.

For 30 years, Sanger sequencing was the most used sequencing technology and evolved gradually. It has been parallelized and automatized leading to better quality and lower sequencing costs. For example, automated Sanger sequencing was used to sequence the first complete human genome, a task that required 13 years and 2.7 billion US dollars[1]. Today Sanger sequencing is still widely used in diagnostics and small projects.

1.2.1. Next-Generation Sequencing

Next-Generation Sequencing (NGS), or Second-Generation Sequencing, technologies are methods for the massively parallel sequencing of short DNA fragments[114][87]. These technologies reduced sequencing costs per sequenced base pair (bp) by approximately five orders of magnitude compared to automated Sanger sequencing, i.e. First-Generation Sequencing[2]. From 2005 to 2007 the first NGS instruments were introduced by Roche/454[134], Illumina/Solexa[6] and LifeTechnologies/ABI[86]. These three technologies use different sequencing biochemistries and methods for the amplification of the input DNA, which leads to different advantages and disadvantages in terms of read length, quality and throughput. However, they share a basic workflow[114]:

- First, the input DNA is randomly fragmented followed by ligation of common adapter sequences. This process is called *library preparation*.
- In a next step, the single library molecules are amplified in a way that the original molecules and all their copies stay clustered in the same position.
- Actual sequencing is then performed by alternating cycles of addition of fluorescently marked nucleotides and imaging.

The data presented in this PhD thesis has been produced using the Illumina *Sequencing by Synthesis (SBS)* technology, which is described in the next paragraphs.

Library preparation for Illumina NGS starts with random fragmentation of genomic DNA (Figure 1.1). Subsequently, adapters are ligated to the DNA fragments. The adapters allow covalent binding of the DNA to the *flowcell*. Flowcells are glass slides that contain 8 channels, the so-called *lanes*, in which the sequencing takes place.

[1] `http://www.genome.gov/11006929` - Last accessed: 25.08.2014
[2] `http://www.genome.gov/sequencingcosts/` - Last accessed: 21.07.2014

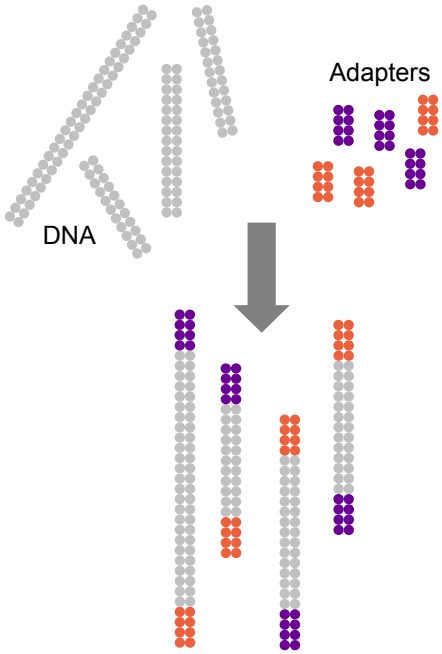


Figure 1.1.: Schema of Illumina library preparation. Adapters are ligated to randomly fragmented DNA molecules.

Single-stranded DNA fragments are then bound randomly to the surface inside the lanes (Figure 1.2). The actual sequencing is performed by measuring the fluorescence of incorporated labeled dNTPs. Since the signal of a single label would be too weak to detect, so-called bridged PCR is performed to generate clusters of identical copies around the initial DNA fragments. After multiple steps of bridged PCR, up to 1,000 copies of each fragment are present.

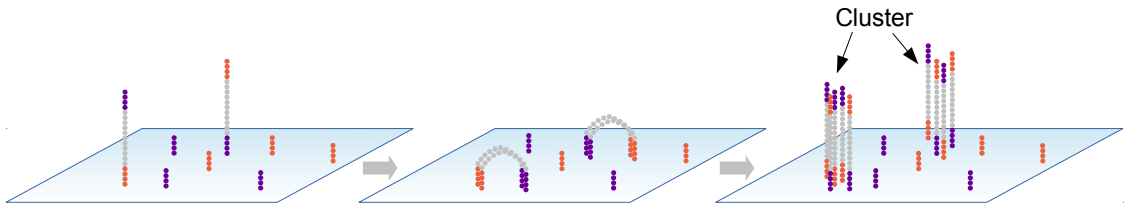


Figure 1.2.: Schema of cluster generation. Single-stranded fragments bind to the flowcell. Clusters are then generated by bridged amplification. Final clusters contain up to 1,000 copies of the initial fragment.

Sequencing is performed in cycles (Figure 1.3). In each cycle four differently labeled dNTPs, primers and polymerase are added to the flowcell. The appropriate dNTPs bind to the nucleic acid chains. The fluorescent label serves as a terminator. Thus, only one dNTP is added to each nucleic acid chain per cycle. Abundant dNTPs, primers and polymerase are then washed off the flowcell. After laser excitation, the fluorescent signals are

measured for each cluster. Then the fluorescent label is cleaved off and the next cycle starts. After each cycle, the Illumina software performs base calling, i.e. it assigns A, C, G or T to each cluster based on the measured fluorescent signals.

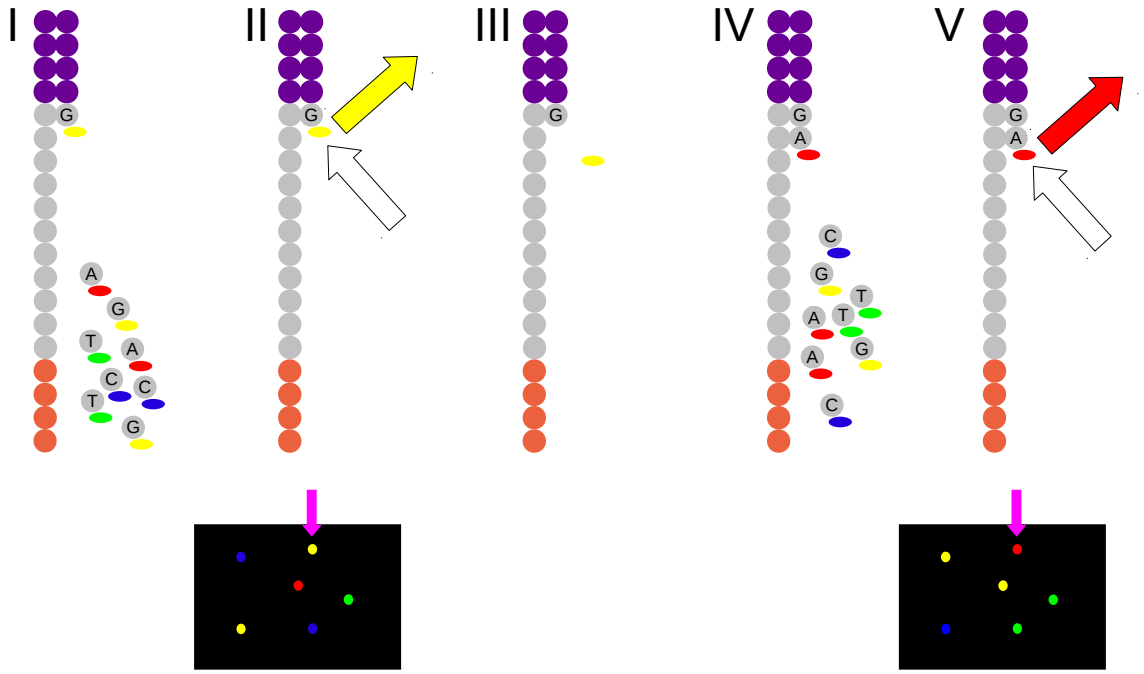


Figure 1.3.: Schema of Illumina Sequencing By Synthesis (SBS). Four differently labeled deoxynucleoside triphosphates (dNTPs), primers and DNA polymerase are added and the appropriate dNTP is added to the nucleic acid chain *(I)*. The fluorescent label serves as a terminator. Thus, only one dNTP is added per cycle. Unused dNTPs, primers and polymerases are washed off. After laser excitation, the fluorescent signals are measured for each cluster *(II)*. Then the fluorescent label is cleaved off *(III)* and the next cycle starts *(IV-V)*.

The number of cycles per Illumina NGS run increased from 35 bp in the first experiments[6] to 300 bp with the latest chemistry version of the MiSeq instrument. Modern Illumina instruments also offer the possibility of performing *paired-end* sequencing, i.e. sequencing from both ends of each fragment (Figure 1.4).

In addition to extending the read length, Illumina increased the throughput also by increasing the cluster density on the flowcell surface. This could be achieved by enhancing the sequencing chemistry as well as the optical system. HiSeq 2500 instruments have now an output of up to 1 terabase (TB) per run (chemistry version 4, two flowcells in 125 bp paired-end mode).

The maximum read length of Illumina SBS is limited mainly by a process called *dephasing*: not all molecules of a cluster incorporate a dNTP in each cycle. These molecules then incorporate the missed dNTP in the next cycle while all other molecules incorporate already the next dNTP. This effect accumulates over time, leading to lower signal intensities

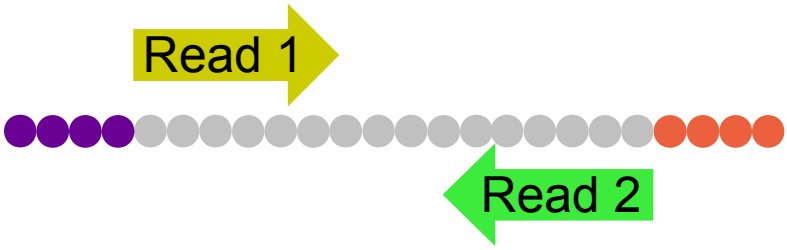


Figure 1.4.: Schema of paired-end sequencing. In paired-end mode, each DNA fragment is read from both ends.

and a higher signal to noise ratio. Thus, the average quality of sequenced bases decreases with the read length.

NGS Applications

In the field of human genetics, NGS is mainly used to identify putative disease causing variants in mendelian diseases and risk factors in complex diseases. Whole genome sequencing of affected individuals is now technically possible for the purpose of variant detection. However, sequencing a whole human genome at high average coverage, i.e. $\geq 30x$ (see Chapter 3.1.1), is still expensive. Thus, various *targeted sequencing* approaches are used for variant detection: *Amplicon sequencing* is used to identify variants in small regions, such as single genes, in large numbers of samples. For the detection of variants in a group of known disease associated genes, disease specific *gene panels* are used.

For the identification of novel disease causing variants and disease associated genes or if the list of candidate genes is too long, enrichment of all coding exons can be used[4]. This approach is called *exome sequencing*. Currently, three major exome enrichment platforms exist: Agilent's SureSelect Human AllExon Kit, Roche/Nimblegen's SeqCap EZ Exome Library and Illumina's TruSeq Exome Enrichment Kit[17]. They all apply the same principle technique: randomly fragmented DNA is hybridized to oligonucleotide baits complementary to the exome targets. Targeted regions are then captured using magnetic streptavidin beads. The three competitors differ in their target region, length and number of oligonucleotide baits and the type of molecule used for capture (Illumina and Nimblegen use DNA, Agilent uses RNA). The manufacturers constantly try to improve their kits by adding more baits for better coverage and a more complete target region. For example, the Agilent Sure Select kit initially targeted approximately 38 Mb of sequence, whereas the newest version of the kit (v5) targets approximately 50 Mb.

In addition to the sequencing of genomic DNA, NGS is also used for sequencing of mRNA, i.e. *RNA-Seq*[132]. mRNA is captured via polyA enrichment and translated into cDNA which is then treated similar to genomic DNA, as described above. RNA-Seq data can be used for *differential expression* analysis between groups of samples, similar to expression data from microarrays. However, unlike microarray data, RNA-Seq data also allows the identification of novel transcripts, alternative splicing and allele specific expression. Also variants can be detected in RNA-Seq data, which allows the investigation of *RNA-editing* (see Chapter 3.4).

NGS is also used in epigenetics. For instance, *chromatin immunoprecipitation followed by sequencing (ChIP-Seq)*[98] is used to identify positions in the genome where a protein of interest, e.g. a transcription factor, binds. *Bisulfite sequencing* can be used to identify the methylation pattern of genomic DNA[59]. Also *micro RNAs (miRNAs)* can be sequenced using NGS. miRNAs are approximately 22 bp long non-coding RNAs that play a role in post-transcriptional gene regulation[13].

1.3. Variant Detection in NGS Data

The detection of *Single Nucleotide Variants (SNVs)* and *small insertions and deletions (indels)* from raw NGS data can be split into three parts:

1. **Alignment** - First, the short NGS reads must be aligned to a reference genome.
2. **Variant Calling** - After alignment, variants, i.e. sites in the genome of the sequenced individual that are different to the reference genome, can be identified.
3. **Variant Filtering And Annotation** - Called variants can then be filtered to remove low quality variants and annotated with additional information such as effect of a variant within a gene.

Basic concepts of these three tasks are introduced in the following paragraphs.

1.3.1. Alignment

Sequencing a whole genome or exome sample typically results in one or more text files in FASTQ format (see Chapter 1.3.4) containing millions of short reads together with quality values for each base. To make sense of these reads, the genome (or exome) of the sequenced individual must be constructed. Theoretically, this can be achieved solely based on the NGS reads without additional information, i.e. generating a *de novo* assembly. Unfortunately, even the best performing[10] *de novo* assemblers, such as *Velvet*[138], *ABySS*[118] or *SOAPdenovo*[78], can not completely assemble complex mammalian genomes using only short NGS reads. Additionally, computational costs are very high, which also limits practical usage.

If a reference genome of the sequenced species is available, NGS reads can be aligned to this reference. Alignment tools from the pre-NGS era, such as BLAT[48], are generally too slow for the alignment of millions of reads per sample. Thus, new alignment tools for NGS data have been developed. Modern NGS aligners, such as Bowtie[62][61], GEM[82] or the Burrows-Wheeler Alignment tool (BWA)[67], use string searching data structures to store the reference genome. For instance, BWA uses the Burrows-Wheeler Transform (BWT) to efficiently store the genome in memory. BWT resorts a given input string, such that equal letters tend to occur in groups, which allows efficient compression. BWA uses BWT to store a compressed prefix trie of the reference genome in memory. A prefix trie is a data structure that stores every prefix of a string such that every exactly repeated substring is only stored once. This allows to search for a string, e.g. a part of a NGS read, in a prefix trie in linear time. The BWA algorithm extends the standard search algorithm of prefix tries to allow mismatches and gaps in the NGS reads. For paired-end reads it first aligns

both reads of a pair separately and then joins these alignments. If the reads of a pair could be mapped to different positions in the reference genome, positions where the two reads are close to each other are preferred.

The alignment of RNA-Seq reads adds an additional problem: the sequenced mature mRNA has already been spliced[28]. Thus, introns are missing and there are reads spanning exon boundaries. These reads can not be aligned to the reference genome using standard alignment programs. However, they are of special interest because they enable the detection of alternative splicing and putative novel transcripts. One strategy to overcome this problem is to align RNA-Seq reads not to the full reference genome but to the reference of the spliced transcriptome. Another strategy applied by modern aligners, such as GEM, is to perform a split alignment, i.e. a separate alignment of the start and end of each read, against the full reference genome. These aligners use a two step process: in a first round a normal alignment as for genomic DNA is performed. This first alignment is then used to define exon boundaries for the split alignment. To further improve accuracy, many of these aligners can be provided with a list of known exon boundaries.

1.3.2. Variant Calling

Variant calling is the process of identifying positions in the genome (or parts of it) of an individual that are different compared to the reference genome. The simplest way to call the genotype at a position is to create a pileup, i.e. a list of all sequenced bases aligned to the position, and calculate the proportion of bases that are different to the reference genome. Two cutoffs for heterozygous and homozygous variants, e.g. 30% and 80%, can then be used to call variants. However, this approach does not take properties such as base or mapping quality into account. Especially at low read depth, such a simple approach is error prone. Thus, more sophisticated variant callers have been developed.

Bayesian models are used by many modern variant callers, such as SAMtools mpileup[68] or GATK UnifiedGenotyper[85]. A simple Bayesian genotyper for SNVs was described by McKenna et al.[85]:

At each position in a diploid genome, 10 possible genotypes G exist. The probability of each G can be calculated using the Bayesian formulation:

$$p(G|D) = \frac{p(G)p(D|G)}{p(D)}$$

where D is the pileup of bases at this position. $p(D)$ can be ignored because it is constant over all genotypes. $p(G)$ is the prior probability of each genotype. Usually, the homozygous reference genotype has the highest prior probability. However, most variant callers allow the definition of prior probabilities for each position. For instance allele frequencies of common variants, e.g. from dbSNP[115], can be used for this purpose. If the genotypes are called for more than one sample simultaneously, i.e. *multi sample calling* is performed, the proportions of reads showing a specific base over all samples can be used as prior probability.

$$p(D|G) = \prod_{b \in pileup} p(b|G)$$

where $p(b|G)$ is the probability of observing the current base given the genotype G. The genotype G is split up into its two alleles A_1 and A_2, such that

$$p(b|G) = \frac{1}{2}p(b|A_1) + \frac{1}{2}p(b|A_2)$$

with

$$p(b|A) = \begin{cases} \frac{e}{3} : b \neq A \\ 1 - e : b = A \end{cases}$$

where e is the phred scaled quality score of b. The genotype at each site is then the one with the largest probability $p(G|D)$. Calling of indels is similar but more complex, because adjacent base pairs must be analyzed jointly.

GATK HaplotypeCaller[85] uses a different approach for variant calling. It first looks for regions that are potentially variable by searching for a significant amount of mismatches in aligned reads. For every such region it constructs a local *de novo* assembly. *de novo* assemblies are represented as graphs where different paths in the graph represent different haplotypes. GATK HaplotypeCaller identifies the most likely haplotype in each graph and if this haplotype contains a variant, i.e. is not representing the reference haplotype, it calls the variant.

1.3.3. Variant Filtering and Annotation

After variant calling, variants can be filtered by fixed thresholds for properties of the variants, such as a minimum read depth or a minimum average mapping quality of the underlying reads. SAMtools varFilter and GATK VariantFiltration are tools that apply such filters on called variants. However, fixed thresholds may not be suitable for different datasets. For instance the read depth at a position depends on the amount of total sequence. Thus, applying the same read depth threshold for samples with different amounts of sequence is not optimal. GATK VariantRecalibrator provides a more sophisticated method for filtering. It uses a set of known, high confidence variants, e.g. from the HapMap project[32], and searches for these variants in the set of called variants. It then models the distribution of these variants relative to annotations such as read depth or mapping quality and clusters them. After that, scores are assigned to all variants based on their distance to the center of these clusters. If a variant is too far away from the center of a cluster, i.e. its score is too low, it is filtered out. The threshold for the score is based on the set of known variants: typically, the threshold is defined such that 99.9% of known variants in the dataset have a higher threshold and are therefore not filtered. The key assumption for this filter method is that known variants that occur at high frequency in a population are more likely to be true than novel variants that have not been seen before (see also Chapter 3.2.5).

Variants can be annotated with a variety of additional information, such as the effect of the variant within a gene, the accession number of the variant in a public database or a conservation score of the affected position. For instance the tools ANNOVAR[131] or SnpEff[16] can be used for this purpose. They both provide a large number of databases for annotation of variant files in the Variant Calling Format (VCF; see next chapter)

1.3.4. File Formats

Over the last few years, several file formats have been developed which are now the de facto standard in NGS data analysis:

- **FASTQ** - FASTQ is a text format used to represent NGS reads. It consists of a unique identifier, the sequenced bases and a Phred-scaled quality value for each base.
- **SAM/BAM** - The *Sequence Alignment/Map format (SAM)*[68] and its binary version, *BAM*, includes the same information as the FASTQ format. Additionaly, it includes information on the alignment, such as the genomic position(s) the read aligns to as well as quality information. It also includes a header where information on the reference genome, the program(s) used to generate the file and the sequenced sample can be stored.
- **VCF** - The *Variant Calling Format (VCF)* was developed by the 1000 Genomes Project Consortium[128][20] and is now the standard output file of most variant callers. Every row in a VCF file represents a single variant and consists of the following columns:
  - CHROM - The chromosome the variant lies on.
  - POS - The position on the chromosome.
  - ID - One or more identifiers of the variant. The identifier can be defined by the user/program, but most commonly the dbSNP[115] identifier is used.
  - REF - The base(s) in the reference genome at this position.
  - ALT - The alternative base(s) of the variant.
  - QUAL - A quality value, usually assigned by the variant caller.
  - FILTER - Indicates if the variant passes all applied filters (see also Chapter 1.3.3). Filter texts are specified by the filter programs, but the text should be "PASS" or "." if all filters are passed.
  - INFO - Additional information on the variant. This consists of semicolon separated "KEY=VALUE" pairs. The possible fields should be explained in the header of the VCF file. The information given in the INFO field depends on the programs used to generate the file, but usually contains additional quality information, e.g. the total read depth at the variant position (DP), or annotations.
  - Genotype fields - The VCF file includes one group of genotype fields per sample. The information given in a single genotype field again depends on the program that generated the file, but it usually includes at least the genotype (GT) of the sample at the variant position, the corresponding genotype quality (GQ) and the read depth (DP).
- **BED** - The BED format[3] is used to store simple genomic regions. It consists of at least three columns:
  1. Chromosome name

[3] `http://genome.ucsc.edu/FAQ/FAQformat.html` - Last Accessed: 30.05.2014

2. Start position
3. End position

The BED format is used for instance to represent ChIP-Seq peaks or to define Exome sequencing target regions.

Part II.

Methods

2. Methods

At the beginning of this PhD project, an exome sequencing analysis pipeline was already available (Figure 2.1)[26]. Briefly, this pipeline started with FASTQ files from the Illumina analysis pipeline. Alignment was performed using BWA[67]. Variants were then called, filtered and annotated using SAMtools[68] and custom Perl scripts. The pipeline was controlled by a config file which had to be created manually for every analyzed sample. This file included the input and output folders and the location of the reference genome.

The annotated variants were then inserted into a relational database. This database had three key tables: (i) the *sample* table that stored information on the sample, such as gender, pedigree name or diagnosed disease; (ii) the *variant* table that stored information on the variant, such as position, variant type or dbSNP identifier; (iii) the *variant to sample* table that represented the occurrence of a variant in a sample and stored additional information on the genotype and quality of the variant in the specific sample. The database could be queried via a custom web interface to identify putative disease causing variants. The key queries allowed to search for variants that were present in different affected members of a pedigree or genes that harbored variants in multiple unrelated, affected samples. These queries allowed to filter for certain properties of the variants, such as variant quality or variant type. By default, all samples diagnosed with a different disease were used as controls to remove common variants. Also basic run statistics, such as coverage or duplicate rate, were inserted into the in-house database.

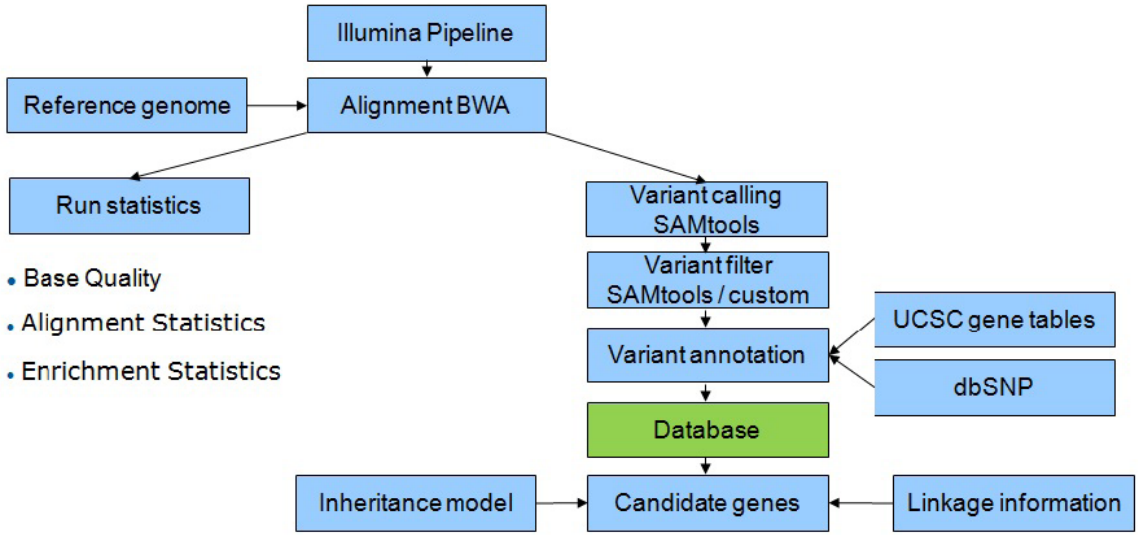


Figure 2.1.: Overview of the exome sequencing analysis pipeline as it was available at the beginning of this project. Picture taken from *Eck,2014*[26]

The ever increasing amount of sequenced samples during this PhD project made large scale changes to the initial pipeline and in-house database necessary. Key goals of these changes were reduced runtime, memory usage, required disk space and hands-on time, as well as increased flexibility and easier adaptation and extension of the pipeline. Also the database design had to be changed in order to handle higher amounts of samples and variants.

Additionally, new features have been implemented to allow the analysis of other features of the data, such as Structural Variants (SVs) or Copy Number Variations (CNVs), and to allow multiple users to analyze large projects collaboratively. These changes are described in the following chapters.

2.1. Improvements

2.1.1. Standard Formats and APIs

As described in Chapter 1.3.4, several de facto standard file formats for NGS data have been developed over the last couple of years. The pipeline has been adapted to use these formats wherever possible. For instance the scripts for variant filtering, annotation and database import now use VCF files. Thus, annotated and filtered output files from the pipeline can easily be used for further analysis with other programs that support VCF files. Furthermore, output of other variant callers, such as GATK[85], or SV/CNV callers (see Chapter 2.2.1) can be easily integrated into the pipeline.

If applicable, standard file formats are processed using APIs, which reduces errors and improves performance.

- VCF files are processed using the VCF Perl API coming with *VCFtools*[20]. This API allows reading and writing of VCF files through standard Perl data structures and provides methods for adding own annotations to a VCF file.

- BAM files are processed using the Bio-SAMtools Perl API[68] which provides methods to access read and coverage information for random regions without converting the BAM file into the SAM format. This is faster, requires less memory and no additional disk space.

2.1.2. Parallelization

Analyzing NGS data is time consuming due to the large data sets and complex tasks. However, many of the tasks, such as read alignment, are independent of each other and can be easily parallelized. Thus, many alignment programs, such as BWA[67], support multithreaded execution. In addition to the parallelization of a single task, tasks that do not depend on each other can run in parallel. For instance, the single FASTQ files of a sample that has been sequenced on different flowcells or lanes can be aligned in parallel (Figure 2.2). Or, after all reads have been aligned, variant calling and calculation of coverage statistics can run in parallel.

The pipeline facilitates parallelization of different tasks using the batch-queuing system

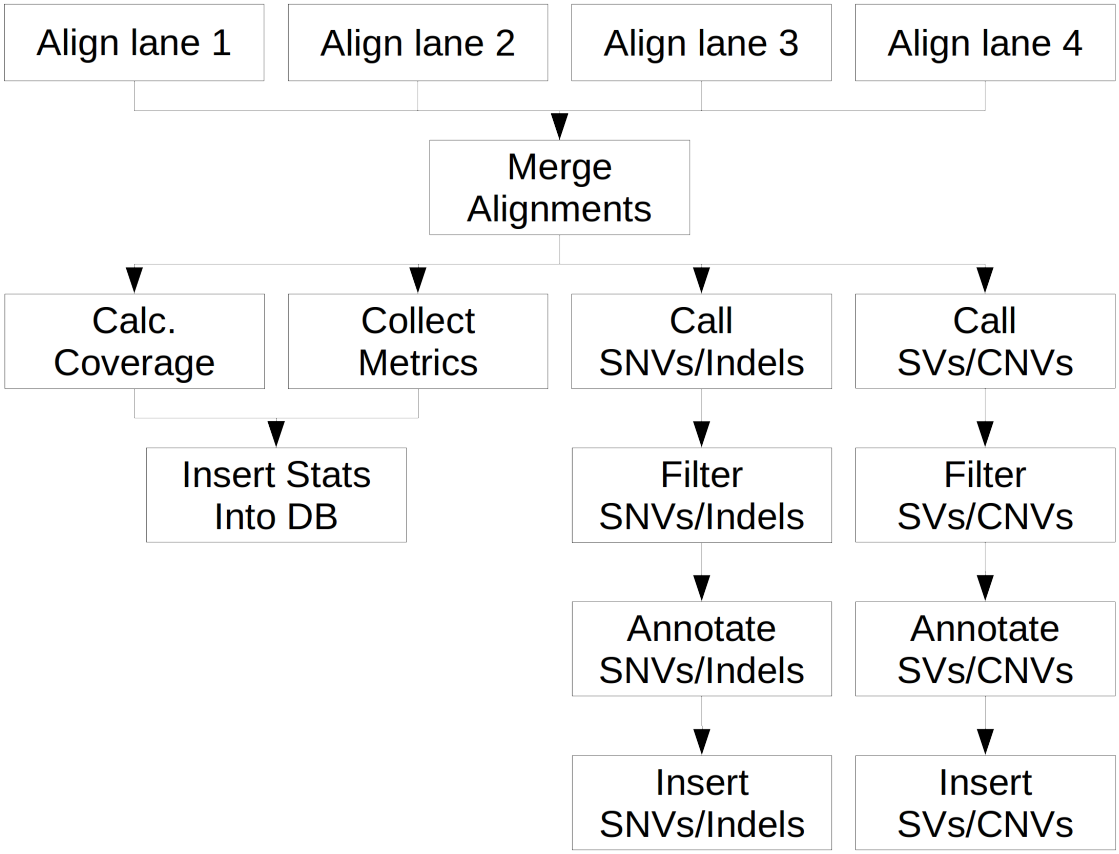


Figure 2.2.: Flowdiagram of the pipeline. Many tasks are independent from each other, e.g. alignment of different lanes, and are therefore parallelized. Some tasks depend on other tasks, e.g. merging of the single lane files depends on alignment of the lanes, and therefore have to wait for those tasks to finish before they can start.

Open Grid Scheduler (OGS)[1]. A batch-queuing system executes jobs on so called *execution hosts* based on the available resources and dependencies of the jobs. Every task of the pipeline is submitted to OGS as a single job. OGS then executes the single jobs when free slots are available and all the predecessors of the jobs have finished. The dependency structure of the tasks is defined in the pipeline. The typical analysis of a single exome sequencing sample can be seen in Figure 2.2. Alignment jobs for the four lanes start at the same time. When they are all finished, the job that merges the BAM files from single lanes begins. When this job is finished the statistics, SNV/indel calling and the SV/CNV calling parts of the pipeline run independently.

[1] `http://gridscheduler.sourceforge.net/` - Last accessed: 30.05.2014

2.1.3. Automatization

To reduce hands-on time, the analysis of standard exome or whole genome samples has been automatized. The pipeline can be started for a single sample, a list of samples or a flowcell by calling a Perl script with only the sample/flowcell name and the name of the OGS queue to which the jobs should be submitted, as arguments. The Perl script then queries the databsae of the in-house *Laboratory Information Management System (LIMS)* to retrieve all necessary information to start the pipeline. This information includes the folder in which the Illumina pipeline stores the FASTQ files, the type of experiment that has been performed (exome sequencing/whole genome sequencing/...) and the version of the enrichment kit. Additionally, the script uses information from a global configuration file, such as the path to the reference genome and to the BED file storing the target regions of the specific exome enrichment kit. All parameters, the pipeline version and the versions of the used programs are stored in the database to reproduce the results at a later time.

2.1.4. Database Changes

Two major changes have been made to the database layout:

1. In addition to diseases, so called *disease groups* have been introduced to represent groups of related diseases. For instance the disease group "tumor" consists of different cancer and adenoma types. In the standard queries of the web interface, samples from different disease groups are now used as controls instead of samples with different diseases. The reason for this change is that variants or genes can play a role in different, related diseases. For example the gene *TP53* is known to play a role in several types of cancer.

2. Relational databases are usually designed to avoid redundancy in the stored data. For instance, the in-house database stores information such as position or type of each variant only once and refers to the stored variant whenever it is called in a new sample. However, for performance reasons this paradigm can be dropped. A table that stores the number of occurrences of a variant for each disease group was introduced, in order to enable faster filtering of variants that are present in controls.

2.2. New Features

2.2.1. Structural Variants and Copy Number Variations

There are four different approaches to detect *Structural Variants (SVs)* or *Copy Number Variations (CNVs)* from (paired-end) NGS data

1. **Read depth approach** - Figure 2.3a - If the coverage at a region is lower or higher than in the surrounding regions, this information can be used to call a deletion or duplication. For whole genome data, a sliding window approach can be used. Briefly, these methods divide the genome in small parts, i.e. windows, and calculate the average read depth of these windows. If a window, or some adjacent windows, has a significantly higher or lower read depth, a duplication or deletion can be called.

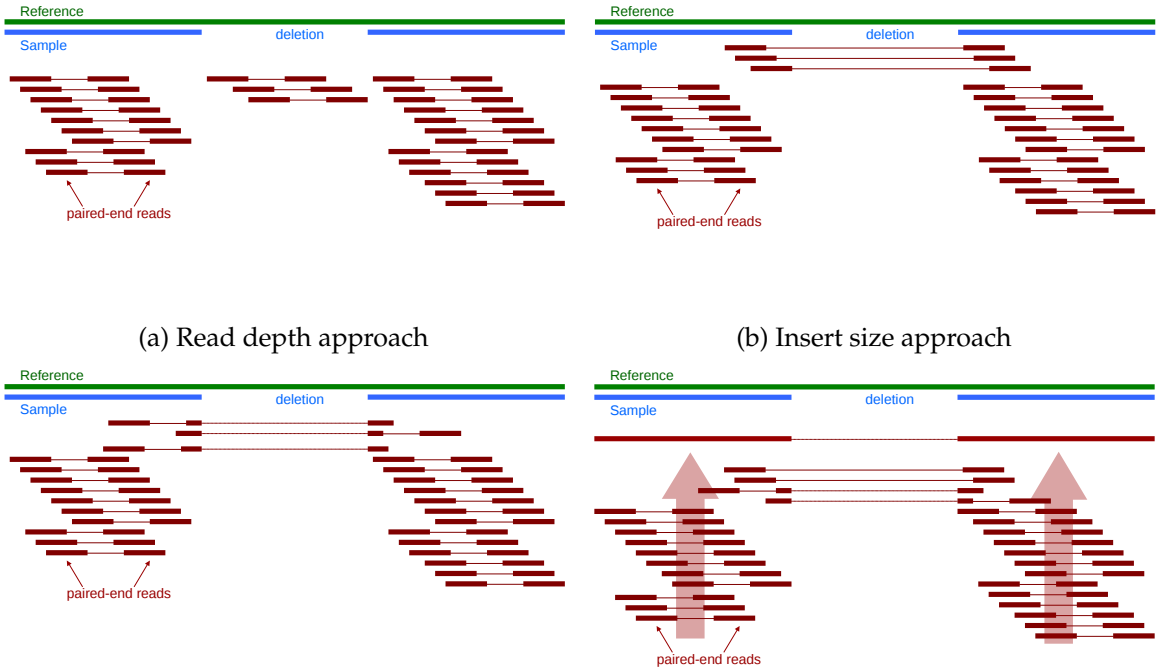
An example for such a program is *CNVnator*[1]. For exome data, read depth approaches are more difficult, because the read depth distribution is not as uniform as for whole genome data due to the capturing process. An example for a read depth based program for exome data is *ExomeDepth*[102]. ExomeDepth takes the average read depth per targeted exon as an input. After normalization, it compares the average read depth of each exon to the average read depth of the same exon in about 10 control samples. It joins significantly different adjacent exons to single deletion or duplication events using a Hidden Markov Model.

2. **Insert size approach** - Figure 2.3b - For paired-end NGS data, the insert size, i.e. the distance between the mapped first and second read of a pair, can be used for the detection of SVs. For instance, if for a significant number of read pairs at a position the insert size is significantly longer than the average insert size (Figure 2.3b), it can be assumed that a deletion has occurred at this position. *Breakdancer*[12] uses insert size information to detect structural variants. In addition to deletions and insertions, breakdancer can also detect translocations, i.e. one of the reads of a pair maps to a different chromosome, and inversions, i.e. one of the reads maps with the wrong orientation. The insert size approach is mainly used for whole genome data, because for exome data both breakpoints of a SV must lie in exons to be recognized, which is rarely the case.

3. **Split read approach** - Figure 2.3c - Reads that overlap with the breakpoints of a SV can be used for SV detection. For deletions this means that one part of a read maps before the deletion and the other part maps after the deletion. For insertions this means that only one part of the read maps to the reference and the part that does not map, maps to the putative insertion. The advantage of this approach is that it can be used to exactly identify the breakpoints and to detect deletions of all different sizes. The size of directly detectable insertions is limited by the read length. The exact sequence of longer insertions can not be determined directly. A tool that uses the split read approach is *Pindel*[137]. It uses the so called pattern growth method to map both ends of the read separately. In the case of a SV a split mapping occurs, i.e. the two ends of the read do not map to the same position. However, this approach is computationally too expensive to perform it for all reads of a sample. Pindel therefore extracts read pairs from a BAM file where only one read of the pair could be mapped and tries to map the other read in proximity to the mapped read using the pattern growth approach.

4. ***De novo* assembly approach** - Figure 2.3d - The most complete and accurate method to detect SVs of all sizes and types would be to perform *de novo* assembly of the NGS reads. Unfortunately, even the best performing[10] *de novo* assemblers, such as *Velvet*[138], *ABySS*[118] or *SOAPdenovo*[78], can not completely assemble complex mammalian genomes solely with short NGS reads. Additionally, computational costs are very high, which also limits practical usage.

For exome data, *Pindel*[137] and *ExomeDepth*[102] have been implemented into the pipeline.



(a) Read depth approach

(b) Insert size approach

(c) Split read approach

(d) *De novo* assembly approach

Figure 2.3.: Methods to detect SVs and CNVs.

On average around 400 SVs and 200 CNVs per sample have been discovered by Pindel and ExomeDepth, respectively (Figure 2.4). Some key properties of these variants will be discussed below.

Pindel

Running Pindel on exome data is useful to detect indels of intermediate size, i.e. indels with a size between about 20 and 1,000 bp. Shorter indels are typically detected by standard variant callers and longer SVs can be detected using read depth based tools. Figure 2.5 shows the size distribution of indels discovered by SAMtools and Pindel. SAMtools is able to detect indels up to a size of approximately 50 bp, which is half of the read length. Longer indels can not be detected using standard variant callers, because reads containing such indels can not be aligned by typical alignment programs. Due to its split mapping approach, Pindel is able to detect also longer indels. Please note that shorter indels detected by Pindel are underrepresented in Figure 2.5, because if an indel is discovered by both SAMtools and Pindel, only the "SAMtools indel" is imported into the database. Importing the same variant twice for a single sample would possibly spoil downstream analysis. The drop in the Pindel graph at a length of approximately 100 bp is due to the lack of longer insertions. Currently only insertions with a length up to the read length are inserted into the database. Longer insertions can also be detected by Pindel, but the quality is lower and the sequence of the insertion can not be reconstructed.

Indels detected by Pindel are often hard to see in the raw data. Figure 2.6 shows a 866

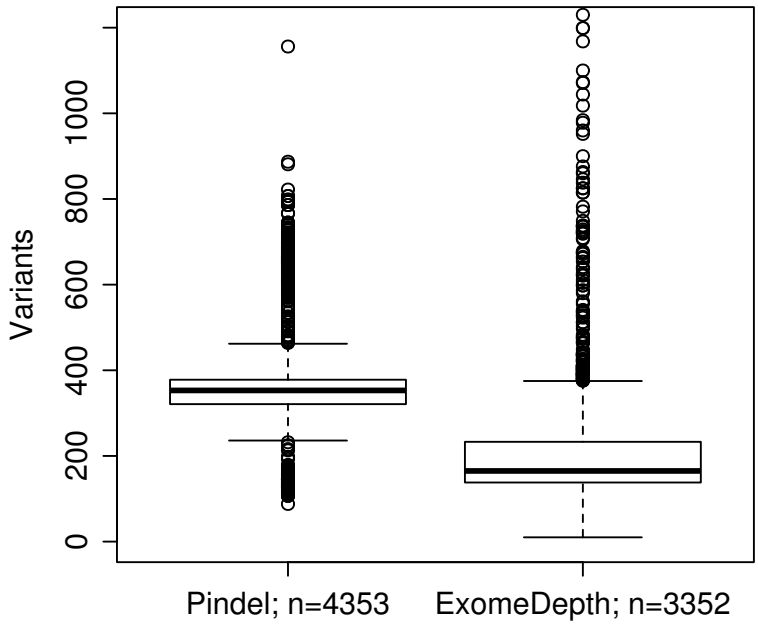


Figure 2.4.: Number of SVs/CNVs detected by Pindel/ExomeDepth per sample.

bp deletion detected by Pindel using the *Integrative Genomics Viewer (IGV)*[108][129] As described above, Pindel uses read pairs where only one of the two reads could be mapped by the standard alignment program and tries to split-align the other read in the proximity of the mapped "anchor" read. However, the split-aligned reads can not be seen in the raw data. Some of the anchor reads are shown in red rectangles in Figure 2.6. They are marked by IGV because the second read in the pair is not mapped. An indication of the detected deletion are the parts of surrounding reads that can not be aligned at the breakpoints, which are represented by the colored mismatches in Figure 2.6.

ExomeDepth

Theoretically, ExomeDepth is able to detect heterozygous one exon deletions, but in practice both sensitivity and specificity are too low to detect such small CNVs. However, about 22% of the detected CNVs consist of only one exon and 89% are shorter than 10 exons (Figure 2.7).

The number of CNVs detected for each sample depends on the similarity of the per exon read depth of the sample to the set of control samples it is compared to. ExomeDepth provides a metric called R_s that illustrates the similarity of the sample to the set of controls. A R_s value of 1 means that the sample is exactly the same as the set of controls and larger values represent bigger differences. The developers of ExomeDepth recommend using only

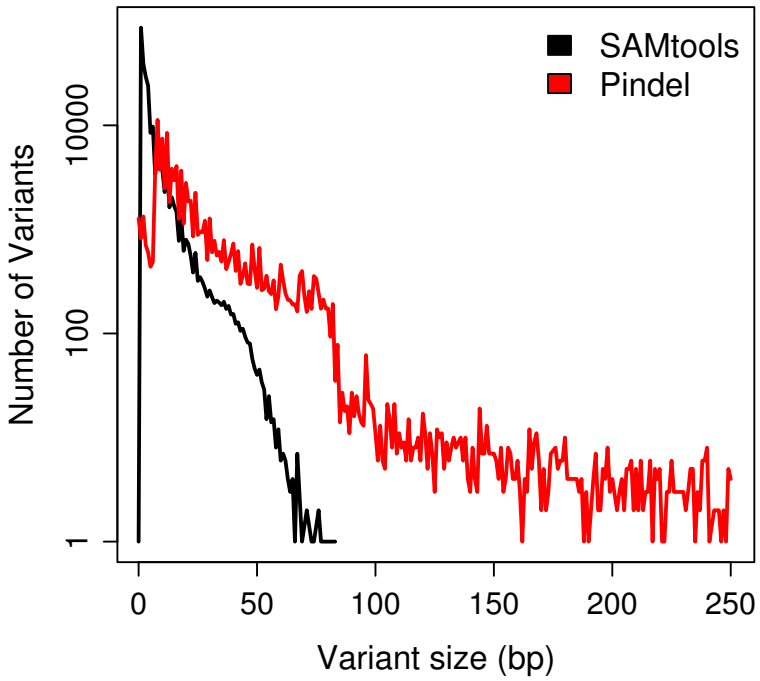


Figure 2.5.: Length distribution of SAMtools and Pindel indels. Note that the y-axis is in log scale.

samples with R_s values below 2.5. From the samples sequenced during this PhD project only those prepared with SureSelect v4 and v5 kits could be analyzed with ExomeDepth, because samples prepared with other kits had in general too high R_s values. From the 2,751 samples that were analyzed with ExomeDepth, 2,522 (92%) had R_s values below 2.5 and 1,245 (45%) had values below 1.5.

Larger CNVs, i.e. with a size of ≥ 10 exons, can be called reliably using ExomeDepth, if the R_s value of the sample is below 2.5. Figure 2.8 shows a heterozygous 319 exon deletion. Heterozygous CNVs are hard to see in raw exome sequencing data because of varying read depth (bottom track in Figure 2.8; see also Chapter 3.1.1). However, the pipeline also outputs the normalized coverage for each exon (top track in Figure 2.8), which can be used for visual assessment of called CNVs.

Database Representation

Variants detected by Pindel and ExomeDepth are stored in the same variant table as SNVs and short indels called by SAMtools. However, due to their size, the pipeline does not insert the deleted/inserted nucleotides but only the start and end coordinates of each variant. Large SVs/CNVs often do not have the exact same start and end coordinates in different samples. In order to compare such variants between samples, variants with more than

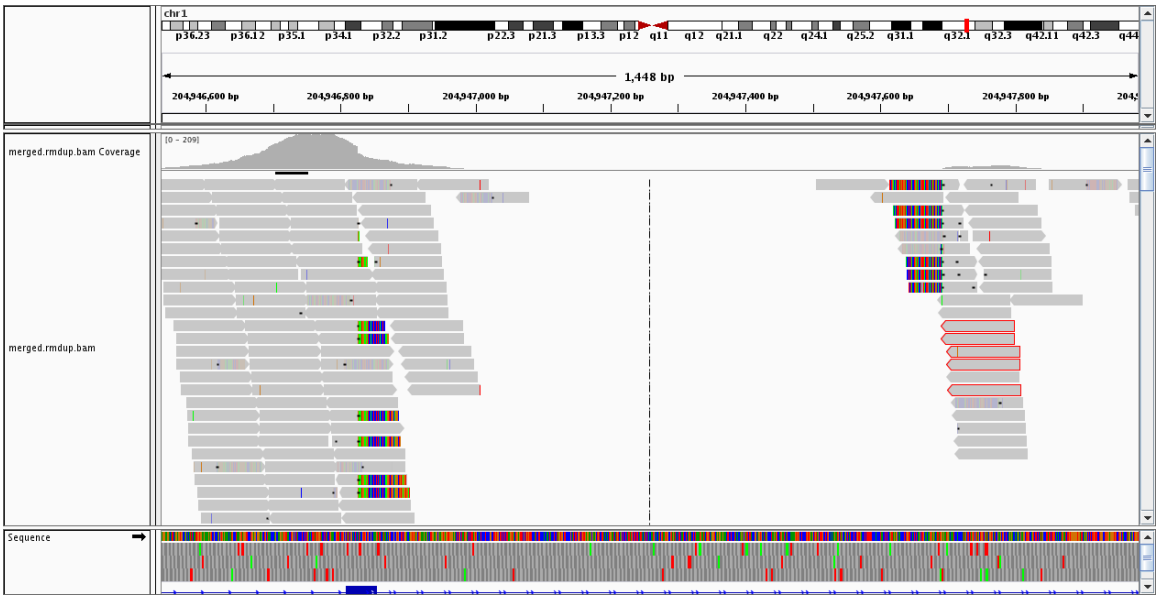


Figure 2.6.: Example of a 866 bp *de novo* deletion detected by Pindel. The raw data is shown using the *Integrative Genomics Viewer (IGV)*[108][129]

90% overlap are joined into a single variant. This allows filtering of SVs/CNVs for variant frequencies with the same queries that are used for SNVs and short indels.

2.2.2. Quality Control

In addition to new analysis methods, new methods for quality control have been developed, in order to detect sample mix-up and contamination and to assess the coverage of genes on a per sample level.

The script *VerifyBamID*[2] has been implemented to detect contamination. It uses known allele frequencies from HapMap[32] variants to calculate the probability that a BAM file contains reads from more than one sample and gives an estimated percentage of contamination. Samples with more than 3% estimated contamination are considered as contaminated. However, this tool can not be used for tumor samples with large scale chromosomal aberrations, because these anomalies cause shifts in allele frequencies of too many variants in the dataset which look similar to contamination with another sample.

To detect sample mix-ups of samples with different sex, the coverage of the gene *SRY* is calculated. This gene is located on chromosome Y and should therefore only be covered if the sample is male. The web-interface also offers a query to calculate the percentage of rare variants shared between two samples. This query can be used if related samples have been sequenced, e.g. children share around 50% of rare variants with each parent. A proportion of shared variants significantly lower than expected points to a sample mix-up.

The coverage of regions targeted by the different exome enrichment kits (see Chapter 3.1.1) is an important metric for the technical quality of an exome sequencing experiment. However, for the detection of putative disease causing variants, the coverage of actual

[2] `http://genome.sph.umich.edu/wiki/VerifyBamID` - Last accessed: 11.07.2014

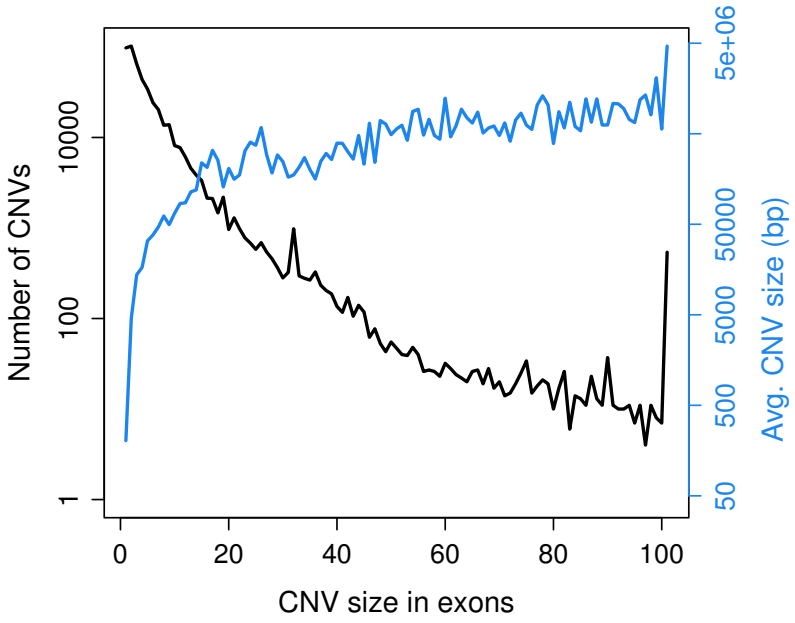


Figure 2.7.: Length distribution of ExomeDepth CNVs. The black line shows the frequency distribution of CNVs with different length (in exons). The peak at 101 exons represents all CNVs with a length ≥ 101. The blue line shows the corresponding average size in base pairs. Note that the y-axis is in log scale.

genes and transcripts is more important than the coverage of target regions defined by manufacturers of enrichment kits. To assess this property the coverage of all RefSeq transcripts of each sample is calculated and inserted into the in-house database. The web-interface provides queries to investigate the coverage of disease candidate genes for single samples, which is especially important for diagnostic samples (see Chapter 3.3.3). Additionally, this information can be used to detect single exon CNVs. For instance the *human growth hormone receptor (GHR)* gene is known to harbor a common deletion of exon 3[120]. Figure 2.9 shows the coverage of this gene in 11 samples. Both homozygous and heterozygous deletions of exon 3 can be seen in the diagram. However, this only works for the comparison of samples prepared with the same version of the exome enrichment kit. Differences between samples prepared with different kits represent technical differences of the kits rather than biological differences of the samples.

2.2.3. Collaborative Features

To jointly analyze large projects, collaborators have been enabled to access their data via the internet. For security reasons all connections are encrypted, access is only granted to clients with known IP addresses and a Yubikey One Time Password generator[3] is required

[3] `http://www.yubico.com/` - Last accessed: 11.07.2014

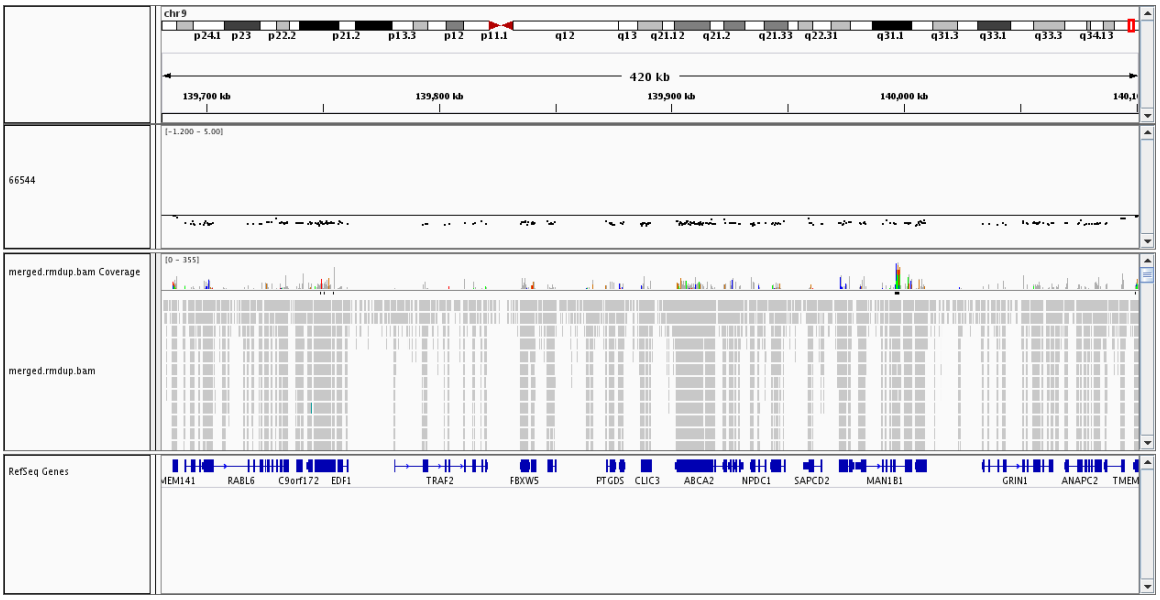


Figure 2.8.: Example of a 319 exon (396 kbp) heterozygous deletion detected by ExomeDepth. The raw data is shown using the *Integrative Genomics Viewer (IGV)*[108][129]

for login. All collaborators are provided with a specific user name and password that grants only access to data from their own projects.

The web-interface now also provides a form to store comments on each variant, in order to allow collaborators to keep track on already analyzed samples and putative disease causing variants (Figure 2.10). It allows the addition of comments such as correctness, mode of inheritance or results from Sanger sequencing to each variant. Also information on the affected gene and free text notes can be added. These comments can then be used for filtering in further analysis, e.g. to search for all correct variants in known disease genes (see also Chapter 3.3.2).

													1	2	3	4	5	6	7	8	9	10
10	PARIS0013	713	713011	GHR	NM_001242399	exomic	paired-end	SureSelect50Mbv5	11.6	230	100	59	73	159	195	134	306	254	285	114	119	245
													100	100	100	100	100	100	100	100	100	100
													59	60	59	59	59	59	59	60	60	59
													1	2	3	4	5	6	7	8	9	10
11	PARIS0014	55	055003	GHR	NM_001242399	exomic	paired-end	SureSelect50Mbv5	11.7	214	100	60	77	172	110	143	297	239	275	128	151	218
													100	100	100	100	100	100	100	100	100	100
													60	60	60	59	60	60	60	60	59	60

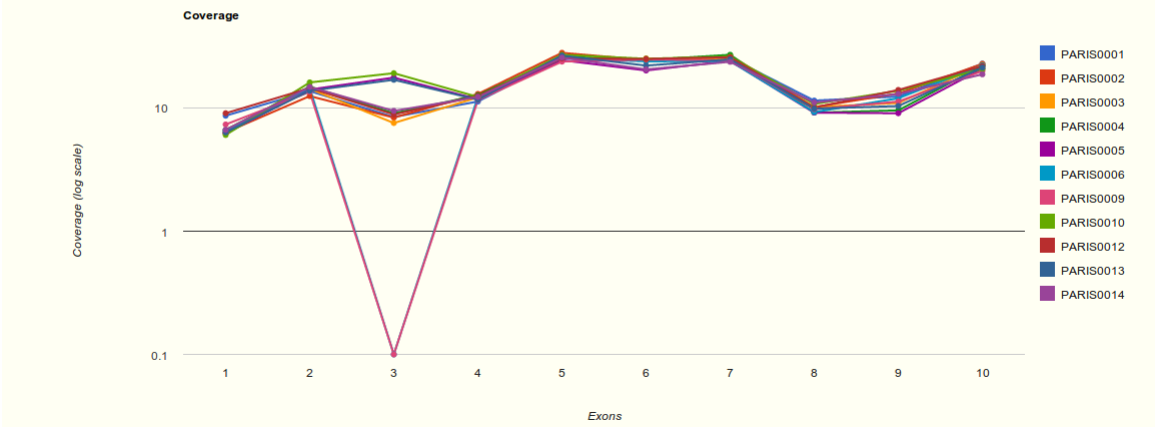


Figure 2.9.: Coverage of the gene *GHR* in 11 samples. Homozygous and heterozygous deletions of exon 3 can be seen. This is a common polymorphism[120]. The diagram is generated dynamically using the Google Charts API.

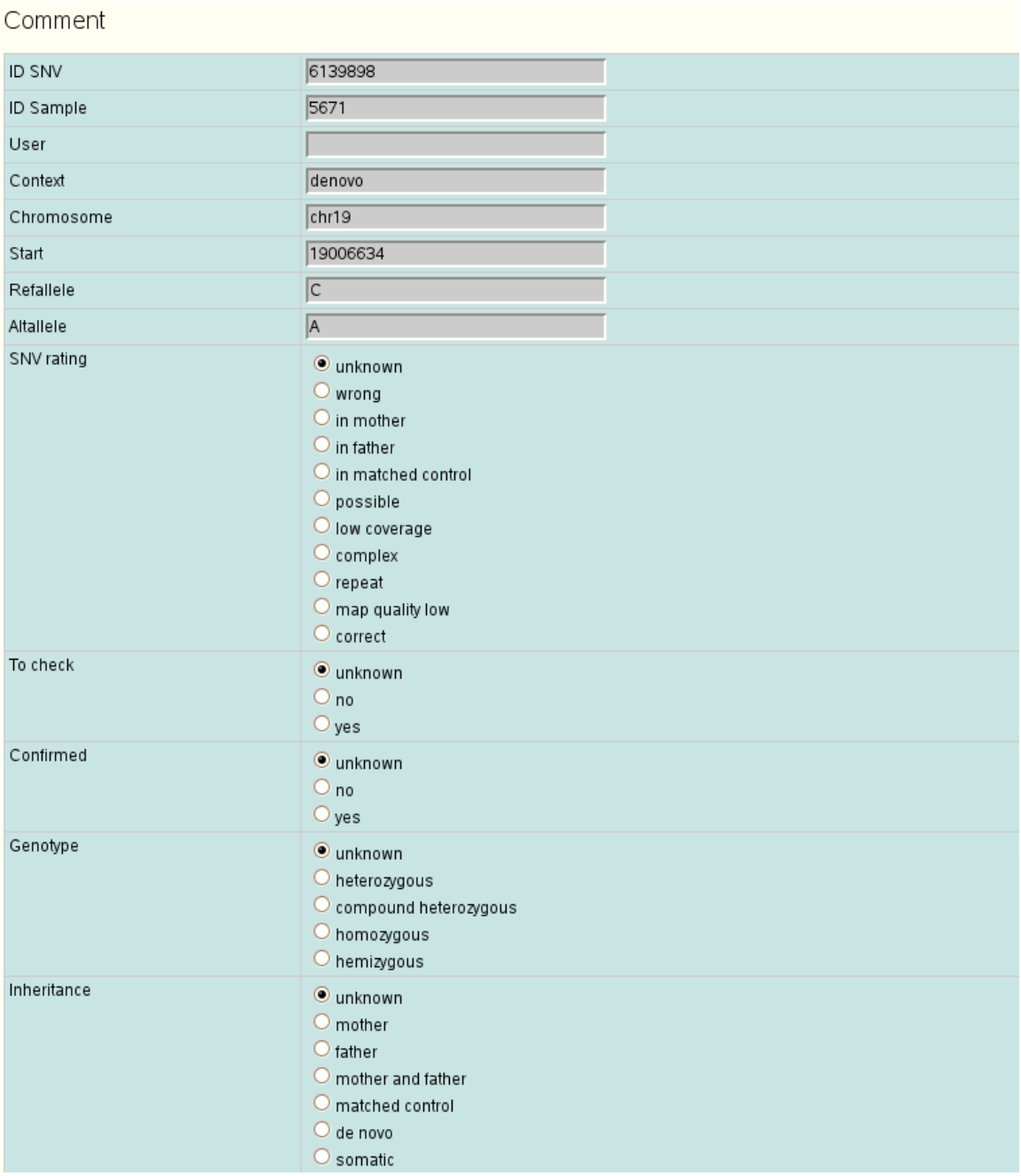


Figure 2.10.: Screenshot of the comments form of the web-interface.

Part III.

Results

3. Results

Over the course of this PhD project 4,567 exomes from 87 different projects have been sequenced on Illumina GAIIx, HiSeq2000 and HiSeq2500 machines. The samples were sequenced to identify pathogenic variants and disease associated genes in rare and common diseases. One third of the samples have been sequenced in the context of developmental disorders, such as *Intellectual Disability (ID)* (Figure 3.1). Neurological disorders and mitochondrial diseases together are responsible for another third of the total sample number.

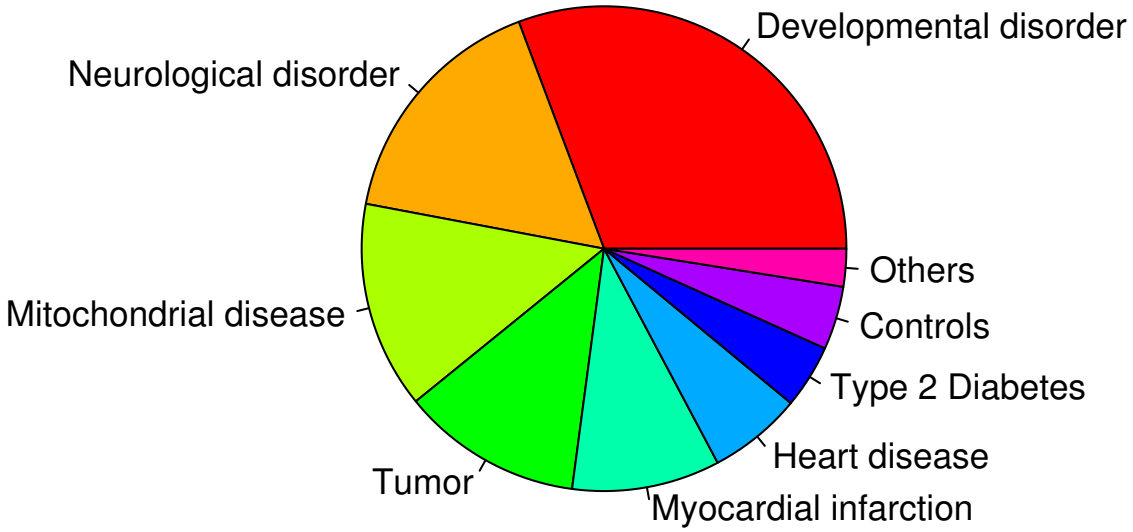


Figure 3.1.: Distribution of samples among disease groups.

The sequencing libraries have been prepared using four different versions of the Agilent SureSelect Human AllExon Kit (Table 3.1). Average coverage, i.e. how many reads overlap a targeted base pair on average, has been above 120x for all kits except for the oldest version (Table 3.2). Around 11,500 high quality synonymous as well as non-synonymous variants have been detected using SAMtools[68].

Kit	Sample
Agilent SureSelect 38Mb kits	91
Agilent SureSelect 50Mb kits (v3)	1,324
Agilent SureSelect 50Mb kits (v4)	881
Agilent SureSelect 50Mb kits (v5)	2,271

Table 3.1.: Number of samples by enrichment kit

Kit	Sequence in Gb (±s.d.)	Coverage (±s.d.)	Synonymous variants (±s.d.)	Non-synonymous variants (±s.d.)
38Mb kits	7.3(±1.8)	82x(±22)	8,533(±256)	7,266(±247)
50Mb kits (v3)	10.2(±2.4)	121x(±27)	10,776(±420)	10,335(±408)
50Mb kits (v4)	9.8(±1.7)	120x(±26)	11,316(±515)	11,344(±499)
50Mb kits (v5)	10.1(±4.0)	123x(±24)	11,507(±317)	11,689(±369)

Table 3.2.: Basic sequencing metrics by enrichment kit

The aim of this PhD project was to investigate and develop methods and parameters to identify candidate pathogenic variants and genes from exome sequencing data. Specifically, three subjects have been investigated and are discussed in the following chapters:

(i) Exome sequencing data must fulfill certain criteria in order to call variants with sufficient quality. Chapter 3.1 evaluates requirements on study design and certain key quality metrics of exome sequencing data.

(ii) Several programs and strategies for variant calling are available. Chapter 3.2 discusses benchmarks for variant callers. Influences of different variant calling procedures and variant quality metrics on sensitivity and specificity are evaluated and used to draw conclusions on best-practice variant calling.

(iii) On average approximately 23,000 high quality coding variants are called per sample. Guidelines on filtering and selecting these variants in order to identify those that are disease causing are discussed in Chapter 3.3.

Additionally, Chapter 3.4 shows results from variant calling in RNA-Seq data with a focus on the identification of RNA editing sites.

3.1. Technical Requirements for Accurate Variant Detection in Exome Sequencing Data

To call variants with sufficient certainty and quality, the quality of the underlying data is crucial. Here, three key quality metrics and their impact on variant calling are discussed: *coverage*, *PCR duplicate rate* and *DNA fragment size*.

3.1.1. Coverage

Sufficient *coverage* or *read depth* of targeted regions is required for variant calling. A minimum read depth of 20x to 40x is assumed to give sufficient power to detect heterozygous variants (see Chapter 3.2). Although the average coverage of targeted regions in exome samples presented here is high (around 120x for newer samples; Table 3.3), there are still approximately 5% and 14% targeted bases covered below 20x and 40x, respectively. This has technical reasons. Some genomic regions can not be captured sufficiently using a hybridization approach because of their sequence composition, e.g. GC rich regions, or because they are not unique.

Kit	Coverage	>20x	>40x
38Mb kits	81.1	81.5%	65.3%
50Mb kits (v3)	120.9	91.4%	81.2%
50Mb kits (v4)	120.5	94.4%	83.5%
50Mb kits (v5)	120.3	95.6%	86.5%

Table 3.3.: Average coverage and % of targeted bases covered more than 20x and 40x per kit. Targeted bases in this case are genomic regions that are in the official target descriptions of the respective Agilent SureSelect kits.

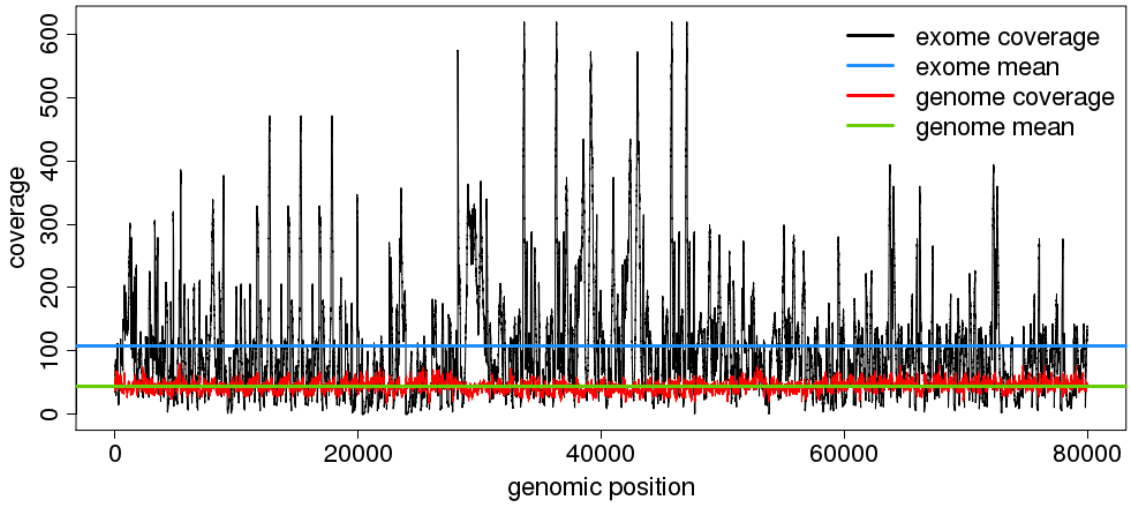
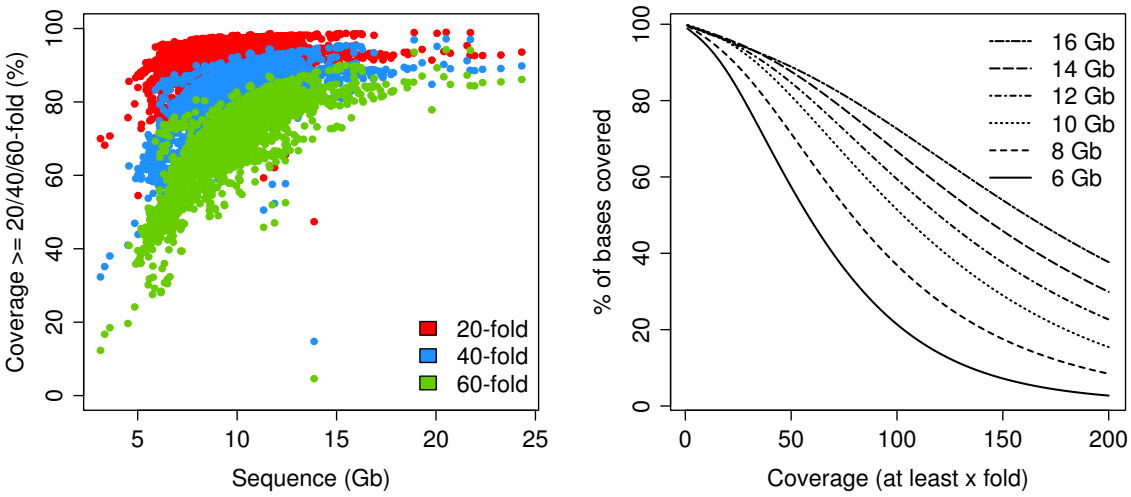


Figure 3.2.: Distribution of read depth for the first 80,000 base pairs of RefSeq transcripts on chromosome 1 for exome sequence (black) and whole genome sequence (red) from the same sample. Horizontal lines show the mean coverage of the exome (blue) and whole genome (green) library, respectively.

Figure 3.2 shows the per base coverage of an exome and a whole genome library of the

same sample. The average coverage of the whole genome library is significantly lower than the average coverage of the exome library. However, the coverage distribution of the whole genome library is more uniform and almost all bases are sufficiently covered whereas there are regions that are not sufficiently covered by the exome sequencing data.

An approach to increase the percentage of sufficiently covered bases is to simply increase the amount of produced sequence per sample. Figure 3.3a shows the percentage of targeted bases covered more than 20/40/60-fold relative to the amount of produced sequence per sample. For every level of coverage it is evident that more than a certain amount of sequence (i.e. around 8 gigabases(Gb) for 20x, 10 Gb for 40x and 12 Gb for 60x) does not increase the amount of sufficiently covered bases. This can also be seen in Figure 3.3b, where the graphs for samples with at least 10 Gb are at the same level up to a coverage of around 50x. In other words, sequencing more than 8-12 Gb per sample only increases the coverage in regions that are already sufficiently covered but does not reduce the amount of regions that are not sufficiently covered.



(a) Percent of target region covered more than 20/40/60-fold by produced sequence.

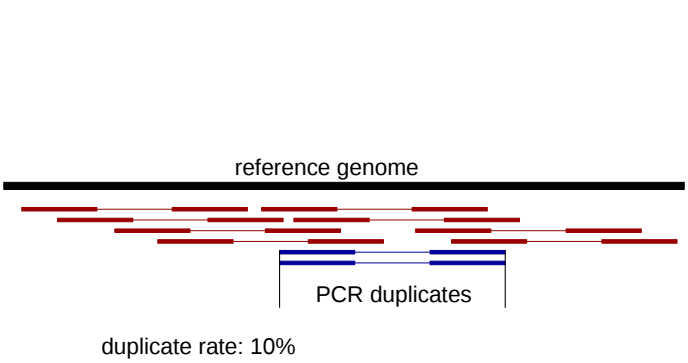
(b) Coverage distribution for different amounts of sequence.

Figure 3.3.: Coverage distribution relative to amount of sequence.

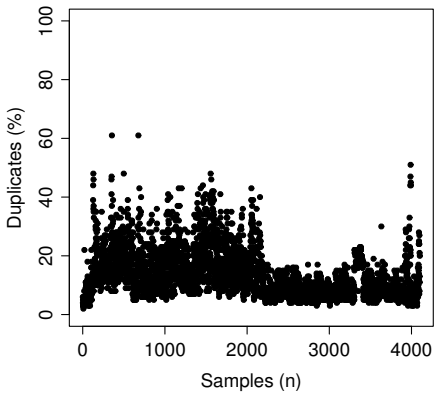
3.1.2. PCR Duplicates

Another important quality metric is the proportion of *Polymerase Chain Reaction (PCR) duplicates*. In the case of paired-end NGS reads, PCR duplicates are usually defined as read pairs that share the same genomic start and end coordinates after read alignment (blue read pairs in Figure 3.4a). These duplicate reads are mainly produced by PCR amplification during the preparation of NGS libraries.

PCR duplicates are removed before variant calling, because they are copies from the same DNA fragment and therefore contain the exact same information which can lead to



(a) PCR duplicates (blue) are read pairs that share the same genomic start and end coordinates. The duplicate rate in this example is 10% since one out of 10 reads is marked as a duplicate.



(b) PCR duplicate rate per sample. Samples are ordered by date of sequencing.

Figure 3.4.: PCR duplicates

problems during variant calling:

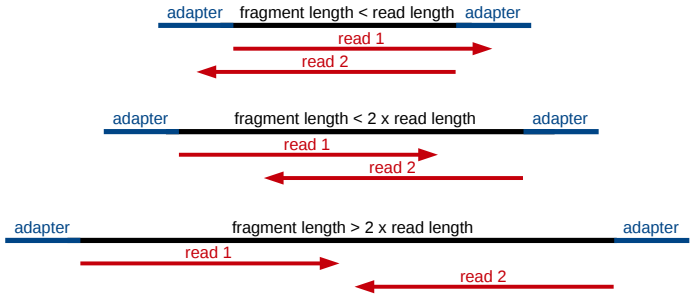
- From a statistical point of view, a perfect NGS experiment can be viewed as randomly sampling DNA molecules. For a diploid organism, this means that at each sequenced position the proportion of reads originating from one of the two alleles follows a poisson distribution with a mean of 0.5. Since PCR duplicates are duplicates from the same DNA molecule, they can lead to a skewed distribution of alleles which in turn can lead to false positive homozygous variant calls. For RNA-Seq experiments removing duplicate read pairs is even more crucial, because the proportion of reads showing a variant is often used to assess allele specific expression or efficiency of RNA editing (see Chapter 3.4).
- PCR can introduce base mismatches into DNA fragments. If such a mistake happens in an early round of PCR it is propagated in the following rounds. Thus, PCR duplicates can lead to false positive variant calls due to propagation of errors.

The amount of PCR duplicates per sample can be seen in Figure 3.4b. Especially in the first samples, the duplicate rate varied strongly. Recent exome library preparation protocols require less rounds of PCR and have therefore a lower, more constant duplicate rate of around 15%. Novel whole genome library preparation protocols require no PCR at all.

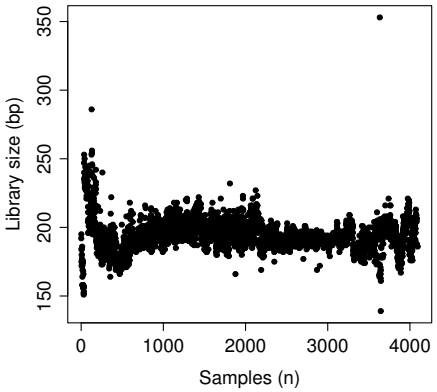
There are also other sources for read pairs with the same start and end coordinates:

- Sometimes one cluster on a flowcell is split into two clusters by the base calling software. These duplicates are called *optical duplicates* and can be distinguished from PCR duplicates by the physical position of the clusters on the flowcell.
- Bad quality and low complexity of the input DNA also influences the duplicate rate.
- With increasing amounts of produced sequence, duplicate read pairs occur by chance.

3.1.3. DNA Fragment Size



(a) Top: If a DNA fragment is shorter than the read length, the two reads of a read pair will contain sequence from the adapters. Middle: If a DNA fragment is shorter than two times the read length, the two reads of a read pair will overlap, i.e. a part of the fragment will be read twice. Bottom: If a DNA fragment is longer than two times the read length, there will be a gap between the two reads of a read pair.



(b) Library size per sample. The average length of the DNA fragments lies around 200 bp in most of the samples.

Figure 3.5.: DNA fragment size

The average *DNA fragment size* or *insert size* can be derived after read alignment. If the average insert size is too short compared to the length of the paired-end reads, problems may arise (Figure 3.5a):

1. If the insert size is shorter than the read length, parts of the sequencing adapters flanking the DNA fragment will also be sequenced. This can cause problems with read alignment and in some cases also leads to incorrect variant calls. To circumvent problems, the known adapter sequence can be clipped before alignment using tools such as *cutadapt*[1]. However, some alignment programs, such as BWA[67][66], are able to align most of those reads also without clipping the adapter beforehand.

2. If the insert size is shorter than two times the read length, the adapter does not get sequenced, but paired-end reads overlap at the ends. The overlapping parts of a read pair can be viewed as partial duplicates (see Chapter 3.1.2). They are not problematic in terms of variant calling, but they minimize the effective sequence yield of the experiment since the overlapping portion of the reads does not contain additional information.

3. If the insert size is longer than two times the read length, there is a gap between the two reads of the read pair. This is the desired situation. For optimal analysis the fragment size should be as large as possible, because longer insert sizes help at spanning repetitive sequences and are also beneficial for structural variant detection.

Figure 3.5b shows the development of the average insert size of exome samples over time. The current average fragment size lies around 200 bp and the paired-end read length

[1] `https://code.google.com/p/cutadapt/` - Last accessed: 06.06.2014

is 100 bp, which can cause some of the problems described above. Unfortunately, in exome sequencing experiments insert size can not be increased easily, because the binding affinity of DNA fragments to the exome capturing beads decreases if the size of the fragments is too large.

3.1.4. Conclusions

Based on the observations above, quality guidelines for exome sequencing have been developed.

The amount of produced sequence per exome sample should be between 8 and 12 Gb. With this amount of sequence approximately 97% of target regions of the SureSelect Human All Exon v5 kit are covered at least 20 times. This is sufficient for variant calling. Adding more sequence does not improve this value.

Using modern, standardized and automated library preparation protocols and high quality input DNA leads to relatively constant duplicate rates of approximately 15%. For exome sequencing experiments, duplicates should be removed or marked using, for instance, SAMtools[68] or Picard Tools[2].

Current enrichment protocols produce exome libraries with an average DNA fragment size of 200 bp. This insert size is long enough for sequencing with 100 bp paired-end reads without a significant amount of reads that contain the adapter sequence. Hence, for this setting, clipping adapters before read alignment is not necessary.

3.2. Benchmarks for Variant Calling

The identification of putatively disease causing variants requires a pipeline that delivers high quality variants, where quality is defined by the amount of true variants that can be called (*sensitivity*) and the amount of called variants that are true (*specificity*). Ideally one wants to call *all and only* true variants, but in practice variant calling and filtering is always a tradeoff between sensitivity and specificity.

Five benchmarks for variant calling and filtering pipelines are discussed in this chapter:

- **Comparing to a gold standard** - To calculate sensitivity and specificity, a set of true variants can be used to compare them to the called variants. Recently, the *Genome In A Bottle Consortium* published a set of gold standard variants for the HapMap/1000 Genome individual NA12878[140]. This individual has been sequenced in many different projects with different technologies and has been analyzed with several different alignment and variant calling tools to ensure that the resulting gold standard variants are not biased. Raw sequencing data can be downloaded to test own pipelines and DNA from this individual can be purchased to also test the sequencing facility.

- **Comparing to arrays** - Especially at the beginning of the NGS era, data from microarrays was used as a gold standard for comparison with variant calls from sequencing. However, comparing to microarrays is not comprehensive since arrays only contain a limited amount of prespecified SNPs. These SNPs are located in genomic regions

[2] `http://picard.sourceforge.net/` - Last Accessed: 06.06.2014

that can be bound uniquely by the probes of the array and can therefore usually be also enriched by exome capture kits. This leads to an overestimation of sensitivity, because regions that are problematic for capturing, sequencing and variant calling are underrepresented. Also genotypes from microarrays contain errors, mainly because of variants on the same allele close to the targeted SNPs that prevent the probes from binding to the DNA.

- **Comparing to *in silico* datasets** - Another way to generate gold standard variant calls is to simulate NGS reads including variants *in silico*[51]. The advantage of this approach is that a true gold standard of variants is available since the variants are known and explicitly generated. A disadvantage is that the *in silico* model might not be able to perfectly mimic the characteristics of true NGS data and therefore the value of quality metrics using this data is limited.

- **Using subsets of data** - Subsets of sequenced data can be used to assess sensitivity and specificity[15]:
  - If confirmed variants are available, one can create subsets of desired read depths of the dataset by randomly drawing reads from the original sequencing files at the variant positions. By repeating these random drawing and calling processes and counting the numbers of successful variant calls, sensitivity of variant calling at certain read depths can be calculated.
  - If a sample has been sequenced with sufficient depth or in more than one experiment, e.g. exome and whole genome sequencing from the same sample, variant calls of the joined data can be used as a *de facto* gold standard. Variants that are called in subsets of the whole dataset but not in the whole dataset can be assumed false positive and so specificity can be calculated.

- **Using novelty of variants** - To assess the influence of certain filters on the false positive rate of variants, annotation of known variants can be used. Variants that have been seen before, e.g. are in dbSNP[115], are assumed to be more likely true than novel variants. Therefore the proportion of novel variants within a dataset can be used as a quality measure. If, for instance, a variant filter removes variants that contain 90% novel variants one can assume that it is more efficient in removing false positives than a filter that removes only 60% novel variants. The same principle is applied by GATK VariantRecalibrator (see Chapter 1.3.3).

In this chapter these benchmarking methods are used to evaluate three different variant callers (with standard parameters, if not stated otherwise): SAMtools mpileup[68] (v.0.1.19), GATK UnifiedGenotyper and GATK HaplotypeCaller[85][24] (v.2.7). These variant callers are benchmarked with regard to read depth and quality scores. Additionally the influence of preprocessing BAM files using GATK Indel Realignment and Base Quality Score Recalibration as well as differences between single sample calling and multi sample calling are investigated.

3.2.1. Comparing to a Gold Standard

Most of the datasets used by the Genome in a Bottle Consortium[3] are freely available. Here, a whole genome sequencing dataset sequenced by Illumina[4] has been used to assess the sensitivity and specificity of the variant callers. This library has been prepared using the Illumina PCR free preparation kit and has been sequenced as 100 bp paired-end run on an Illumina HiSeq, resulting in about 1.67 billion reads. 96% of these reads have been mapped using BWA, leading to an average coverage of about 51x for RefSeq genes. The duplicate rate has been 1.8%.

The Genome in a Bottle Consortium provides a VCF file[5] with variants which is considered as the current gold standard for this dataset. Additionally a BED file[6] containing genomic regions in which variants could be called confidently is provided. This file contains 2.195 billion base pairs on chr1-21 and chrX. These regions have been used as targets for variant calling. Please note that due to the restriction of the analysis to these "confidence regions" it is very likely that sensitivity and specificity obtained in this benchmark are overestimated when compared to actual whole genome variant calls, because "problematic regions", e.g. around centromeres, are not investigated.

Between 2.90 and 2.95 million variants have been called using the variant callers SAMtools mpileup[68], GATK UnifiedGenotyper and GATK HaplotypeCaller[85][24] (Table 3.4).

		SAMtools			UnifiedGenotyper			HaplotypeCaller		
	Gold standard	TP	FN sens	FP spec	TP	FN sens	FP spec	TP	FN sens	FP spec
SNVs	2,742,170	2,734,360	7,769 99.7%	8,454 99.7%	2,739,111	4,601 99.8%	42,924 98.5%	2,739,986	2,707 99.9%	12,719 99.5%
Indels	173,561	147,613	25,518 85.3%	5,444 96.4%	142,024	31,502 81.8%	10,992 92.8%	166,807	7,296 95.8%	2,126 98.7%

Table 3.4.: Comparison of variant calls to the gold standard from the Genome in a Bottle Consortium. The table shows counts of *true positive (TP)*, *false negative (FN)* and *false positive (FP)* variant calls and the respective sensitivities (sens) and specificities (spec) for three different variant callers.

The called variant files were then compared to the gold standard file to calculate numbers of *true positive (TP)*, *false negative (FN)* and *false positive (FP)* variant calls (Table 3.4). TP, FN

[3] `http://www.genomeinabottle.org/` - Last accessed: 11.04.2014

[4] `ftp://ftp-trace.ncbi.nih.gov/giab/ftp/technical/NA12878_data_other_projects/sequence_read/ERP001229/ILLUMINA/sequence_read` - Last accessed: 11.04.2014

[5] `ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/NIST/NISTIntegratedCalls_14datasets_131103_allcall_UGHapMerge_HetHomVarPASS_VQSRv2.18_all_nouncert_excludesimplerep_excludesegdups_excludedecoy_excludeRepSeqSTRs_noCNVs.vcf.gz` - Last accessed: 11.04.2014

[6] `ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/NIST/union13callableMQonlymerged_addcert_nouncert_excludesimplerep_excludesegdups_excludedecoy_excludeRepSeqSTRs_noCNVs_v2.18_2mindatasets_5minYesNoRatio.bed.gz` - Last accessed: 11.04.2014

and FP variant calls are defined as follows:

- TP - A variant is present in the gold standard and the variant file, and has the same genotype in both files.
- FN - A variant is present in the gold standard file but not the variant file, or it is present in both files and has a homozygous genotype in the gold standard file and a heterozygous genotype in the variant file.
- FP - A variant is not present in the gold standard file but is present in the variant file or it is present in both files and has a heterozygous genotype in the gold standard file and a homozygous genotype in the variant file.

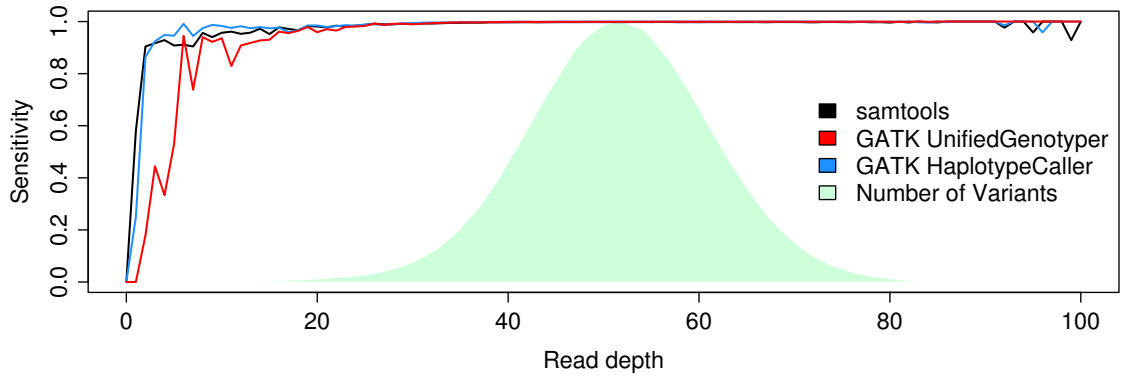
Table 3.4 shows that all three callers have relatively high sensitivity and specificity levels for SNVs, although one SNV every 300 to 800 kbp is missed and there is one wrong SNV every 50 to 250 kbp in a whole genome sequencing experiment. Sensitivity and specificity of variant calling depends on the quality of the underlying data, especially the read depth. Figure 3.6 shows the distribution of sensitivity compared to read depth. As expected, sensitivity is worse for regions with low read depth and the maximum sensitivity is reached at a read depth of about 40. The same applies to specificity (Figure 3.7). Calculations for low and high read depth are less meaningful, because the vast majority of variants are located in regions with a coverage between 40 and 60x. This is due to the generally uniform coverage distribution of whole genome sequencing (see also Chapter 3.1.1).

Variant callers usually provide metrics reflecting the quality of each variant. One such metric that is defined in the specification of the *Variant Calling Format (VCF)*(Chapter 1.3.4) and is therefore reported by most modern variant callers, is the so called *Genotype Quality (GQ)*. It represents the phred scaled probability that the reported genotype is wrong. The GQ can be used to filter variants which should in theory allow to control for specificity. However, for this dataset filtering by GQ is only suitable to a limited extent, since the vast majority of variants has the highest GQ value (Figure 3.8). Moreover, the distribution of the GQ values is not equal, but rather there are distinct peaks.

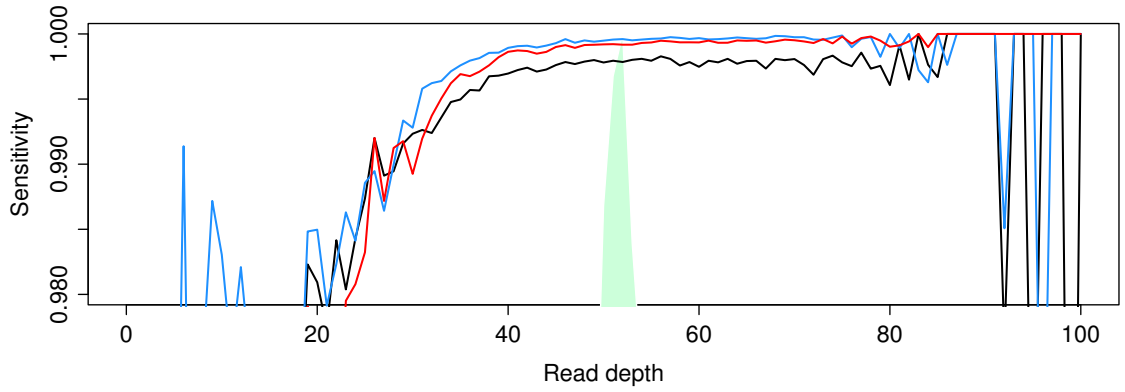
Both sensitivity and specificity values are lower for indels (Table 3.4; Figure 3.9). This is due to the more difficult task of indel calling. However, the GATK HaplotypeCaller performs significantly better than the other two callers, probably because it uses local *de novo* assembly for variant calling which has advantages in repetitive regions. Also the Genome in a Bottle Consortium mainly used the GATK HaplotypeCaller for indel calling, which might introduce some bias .

Despite of insufficient read depth, two major sources for false positive and false negative variant calls can be identified by manually investigating the data:

1. The vast majority of false negative and false positive calls is due to alignment errors or missed calls around short repeats and homopolymers and around indels. Especially SAMtools mpileup (with standard parameters) seems to undercall such variants. This can be problematic in some use cases. For instance the genes *BRCA1* and *BRCA2*, two breast cancer susceptibility genes, contain several homopolymer stretches. Mutations at these homopolymers are often disease causing and therefore of special interest in diagnostics. Figure 3.10 shows an example of such a variant.



(a) Scaled from 0 to 1.



(b) Scaled from 0.98 to 1.

Figure 3.6.: Sensitivity of SNV calls by read depth. Solid lines show sensitivity of the different variant callers (black: SAMtools mpileup; red: GATK UnifiedGenotyper; blue: GATK HaplotypeCaller). The green area shows the distribution of all variants across the read depth spectrum.

This frameshift insertion has been called by SAMtools, but the quality values assigned to it have been very low (Variant Quality=3; Genotype Quality=38). It would have been filtered out if standard filter criteria were applied (i.e. Variant Quality $\geq$ 30; see Chapter 3.2.5). Both GATK callers assign the maximum Genotype Quality (GQ=99) to this variant. The behavior of SAMtools mpileup around homopolymers can be adjusted by increasing the coefficient for homopolymer errors (parameter "-h"). Increasing this parameter from 100 to 150 results in a Variant Quality of 78 and a Genotype Quality of 99 for the variant in the example. Thus, this parameter should be increased if sensitivity around hompolymers is of concern.

2. Some false calls are not false calls but rather reflect different representations at bial-

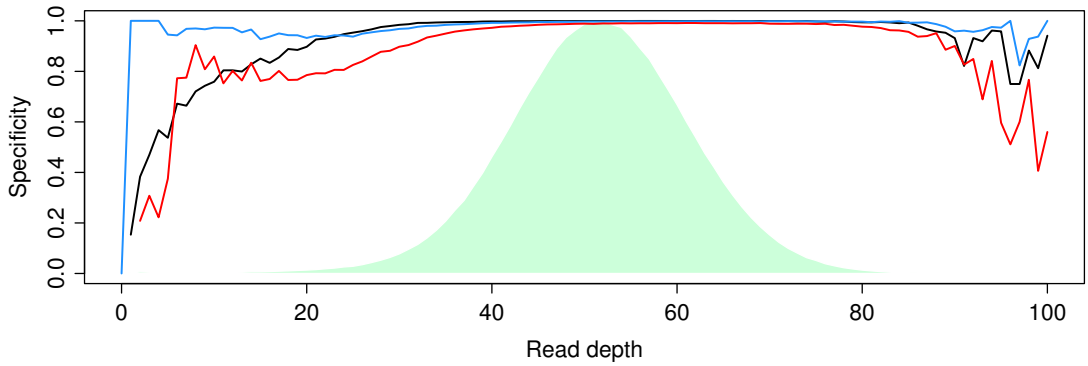


Figure 3.7.: Specificity of SNV calls by read depth. Solid lines show sensitivity of the different variant callers (black: SAMtools mpileup; red: GATK UnifiedGenotyper; blue: GATK HaplotypeCaller). The green area shows the distribution of all variants across the read depth spectrum.

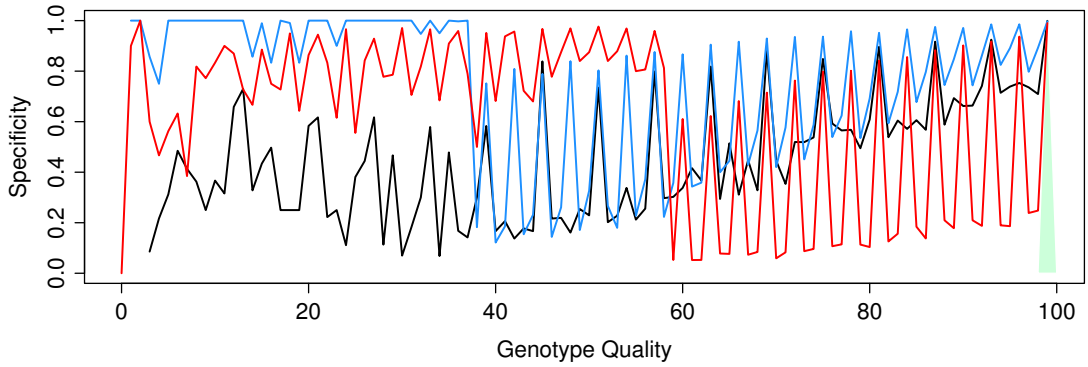


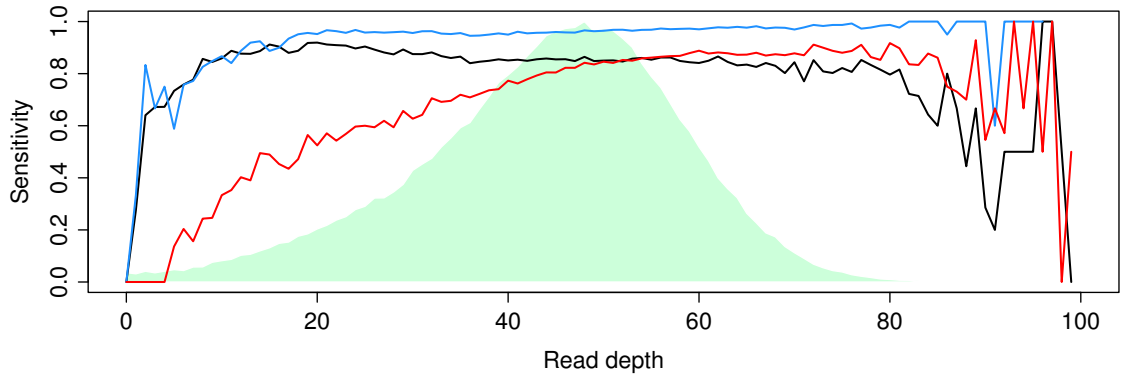
Figure 3.8.: Specificity of SNV calls by Genotype Quality (GQ). Solid lines show sensitivity of the different variant callers (black: SAMtools mpileup; red: GATK UnifiedGenotyper; blue: GATK HaplotypeCaller). The green area shows the distribution of all variants across the read depth spectrum.

lelic sites, i.e. there is more than one variant allele, which make comparisons between two datasets difficult.

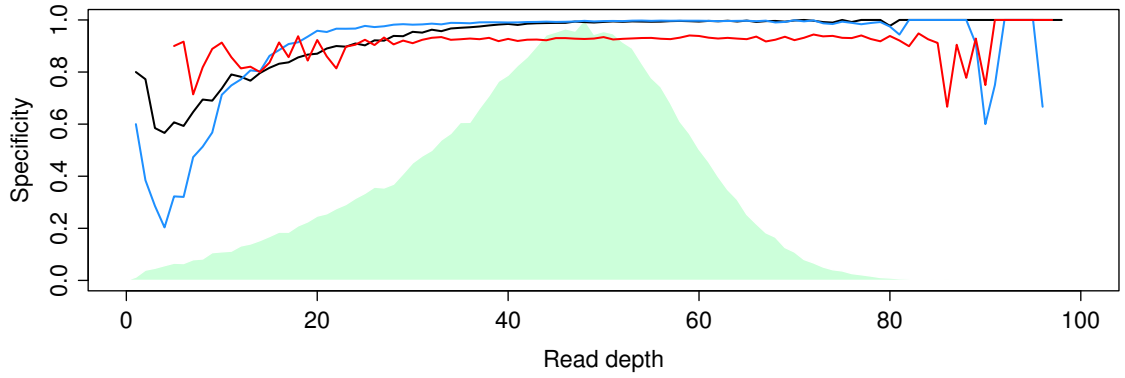
Preprocessing of BAM Files

GATK currently offers and recommends[7] two methods to process aligned BAM files that should increase the quality of subsequent variant calls:

[7] `https://www.broadinstitute.org/gatk/guide/best-practices` - Last accessed: 17.04.2014



(a) Indel sensitivity



(b) Indel specificity

Figure 3.9.: Sensitivity (3.9a) and specificity (3.9b) of Genome in a Bottle indel calls by read depth. Solid lines show sensitivity of the different variant callers (black: SAMtools mpileup; red: GATK UnifiedGenotyper; blue: GATK HaplotypeCaller). The green area shows the distribution of all variants across the read depth spectrum.

- **Indel Realignment** - Mismatched bases are gathering around indels due to alignment errors in such regions. This tool performs local realignment around indels to decrease the amount of incorrect mismatches.
- **Base Quality Score Recalibration** - The quality that is given for each sequenced base by the Illumina base calling software often does not reflect the true probability that the base is wrong. This tool assigns more realistic quality values by incorporating other covariates such as the position of the base in the read or the neighboring bases (i.e. if it is a homopolymer).

These two methods have been applied to the data to assess if they improve the quality

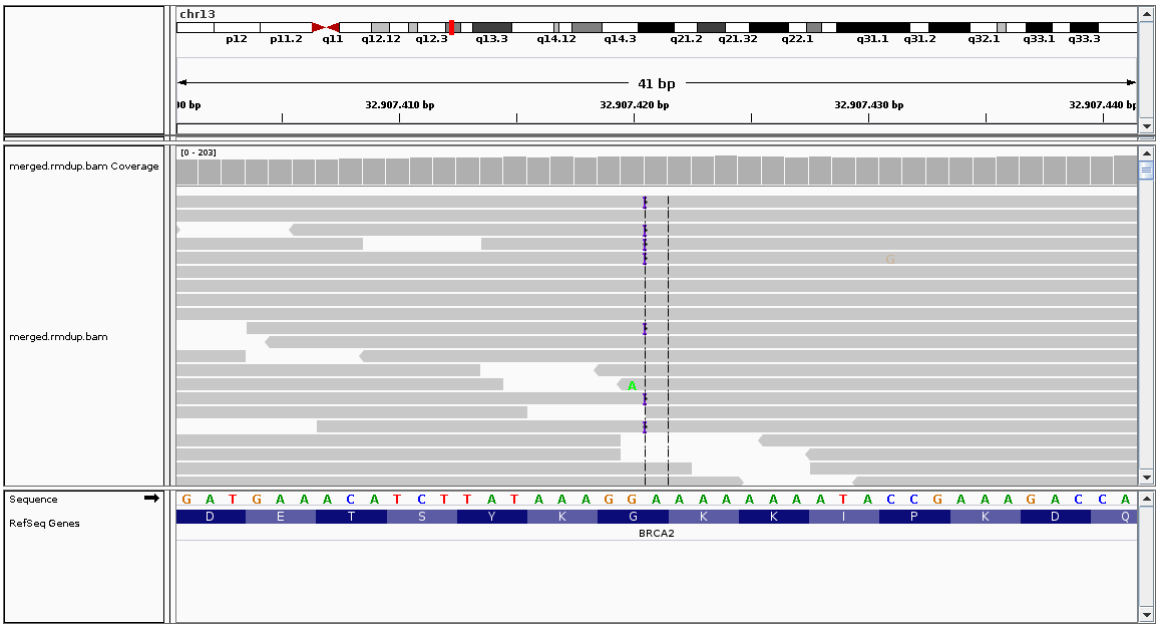


Figure 3.10.: Example of a frameshift insertion in *BRCA2*. SAMtools assigns very low quality values (Variant Quality=3; Genotype Quality=38) whereas GATK Unified Genotyper and GATK Haplotype caller assign high values (Genotype Quality=99).

of variant calls (Table 3.5).

	Gold standard	SAMtools TP	SAMtools FN sens	SAMtools FP spec	UnifiedGenotyper TP	UnifiedGenotyper FN sens	UnifiedGenotyper FP spec	HaplotypeCaller TP	HaplotypeCaller FN sens	HaplotypeCaller FP spec
SNVs	2,742,170	2,734,413	7,735 99.7%	5,333 99.8%	2,739,436	3,750 99.9%	22,135 99.2%	2,740,022	2,665 99.9%	12,916 99.5%
Indels	173,561	148,411	24,635 85.8%	3,275 97.8%	156,134	17,477 89.9%	3,212 98.0%	166,778	7,362 95.8%	2,289 98.6%

Table 3.5.: Comparison of variant calls to the gold standard from the Genome in a Bottle Consortium. The underlying BAM file has been processed with GATK IndelRealigner and Base Quality Score Recalibration. The table shows counts of *true positive (TP)*, *false negative (FN)* and *false positive (FP)* variant calls and the respective sensitivities (sens) and specificities (spec) for three different variant callers.

When comparing the results from optimized to non-optimized files, it is evident that the calls from GATK UnifiedGenotyper become significantly better. However, GATK HaplotypeCaller does not profit from the optimization tools, probably because it performs local *de novo* assembly in order to call variants anyway. SAMtools mpileup variant calls become slightly better after optimization.

3.2.2. Comparing to Arrays

To assess the sensitivity of the variant calling pipeline, heterozygous non-reference SNPs from 26 samples from a trio sequencing project[104] have been compared to results from Affymetrix 6.0 arrays. A total of 66,145 non-reference SNPs from the array were located within the regions targeted by the exome enrichment kit. 64,484 (about 97.5%) of these SNPs could also be found in the corresponding exome sequence when applying the same filters that were used to search for putative *de novo* variants (SAMtools SNV quality ≥ 40).

3.2.3. Comparing to *in silico* Datasets

Methods to generate next generation sequencing data *in silico* have already been developed together with the first alignment and variant calling programs to test their abilities[68]. More sophisticated algorithms try to mimic real NGS data by simulating platform specific errors and biases, such as GC bias for sequencing by synthesis data or indel errors at homopolymer stretches for pyrosequencing data[41][42][84].

Here, a program called WESSIM[51] has been used to simulate an exome sequencing experiment. In addition to simulating platform specific biases, WESSIM also tries to mimic the hybridization step in exome sequencing library preparation by taking the probe sequences from the exome capture kit and looking for positions in the reference genome that might bind to these sequences. WESSIM essentially requires two inputs: the probe sequences, which can be downloaded from the manufacturers homepage and the reference genome from which the data should be generated. To include a set of known variants that act as a gold standard for the assessment of variant calling, a new "personal" genome has been generated by adding a list of variants obtained from a real exome sequencing experiment to the human reference genome. Then WESSIM has been used to generate about 215 million reads. These reads have been used as input for the pipeline.

Interestingly, about 85.3% of the reads mapped to the specified target region, but only 95.7% of the target region has been covered at least once. In real datasets only about 75-80% of the reads are on target, but usually about 99.8% of the target region is covered at least once. However, 95.4% of the target region has been covered more than 20 times, which is in agreement with real data.

		SAMtools			UnifiedGenotyper			HaplotypeCaller		
	Gold standard	TP	FN sens	FP spec	TP	FN sens	FP spec	TP	FN sens	FP spec
SNVs	59,702	50,890	1,250 97.6%	7,778 86.7%	53,177	592 98.9%	6,483 89.1%	53,117	787 98.5%	5,878 90.0%
Indels	13,937	4,047	9,038 30.9%	1,320 75.4%	4,613	8,519 35.1%	1,109 80.6%	5,472	7,678 41.6%	1,006 84.5%

Table 3.6.: Comparison of variant calls obtained from *in silico* data generated to WESSIM to the list of known variants. The table shows counts of *true positive (TP)*, *false negative (FN)* and *false positive (FP)* variant calls and the respective sensitivities (sens) and specificities (spec) for three different variant callers.

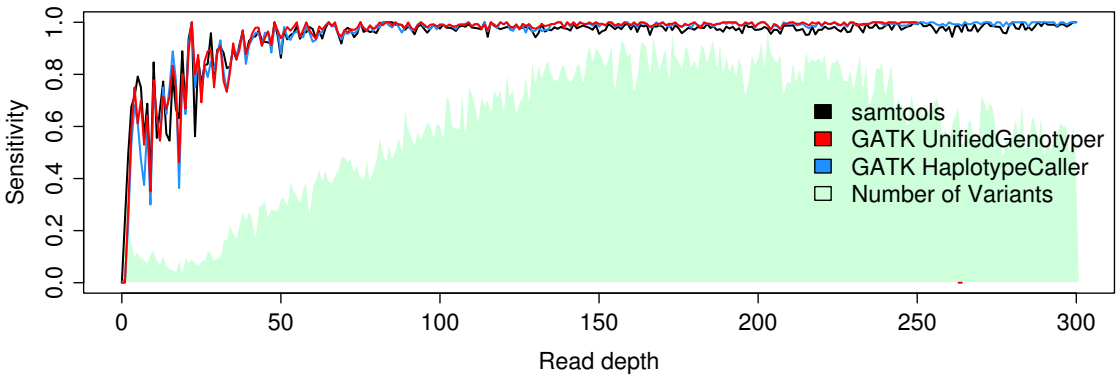


Figure 3.11.: Sensitivity of WESSIM SNV calls by read depth. Solid lines show sensitivity of the different variant callers (black: SAMtools mpileup; red: GATK UnifiedGenotyper; blue: GATK HaplotypeCaller). The green area shows the distribution of all variants across the read depth spectrum.

After filtering for regions that have been covered by the simulated data, approximately 70,000 variants remained in the gold standard file (Table 3.6). For SNVs sensitivity is high and mainly depends on read depth (Figure 3.11). Specificity (Figure 3.12) is much lower compared to specificity from Genome in a Bottle data (Chapter 3.2.1) and also depends largely on readdepth.

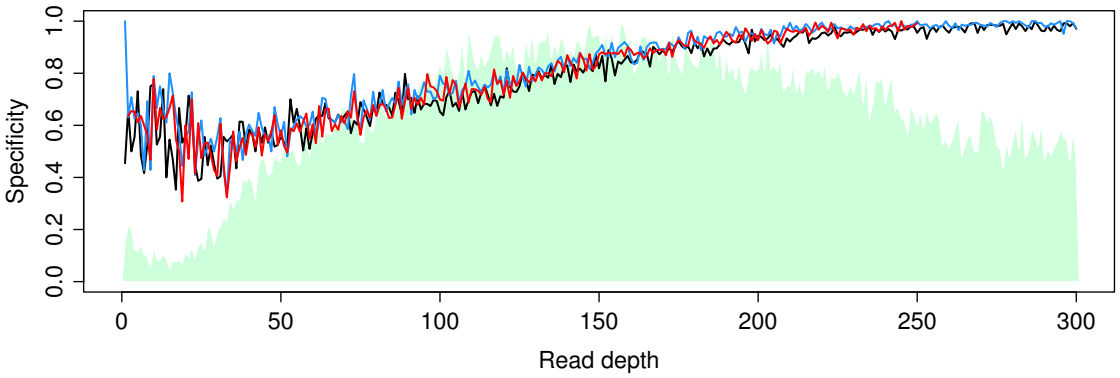


Figure 3.12.: Specificity of WESSIM SNV calls by read depth. Solid lines show specificity of the different variant callers (black: SAMtools mpileup; red: GATK UnifiedGenotyper; blue: GATK HaplotypeCaller). The green area shows the distribution of all variants across the read depth spectrum.

For indels especially the sensitivity levels are low (Figure 3.13), but also specificity is

significantly lower than in the Genome in a Bottle dataset (Figure 3.14).

By looking at the raw data at FN and FP sites, several reasons for the lower sensitivity and specificity values have been identified:

- The coverage distribution of the data is not as uniform as for whole genome sequencing data, but it is also different from real exome sequencing data, from which the original gold standard variants have been obtained. Thus, many of the gold standard variants are not sufficiently covered.

- The generated data often does not reflect the underlying gold standard variant, i.e. there are no/too less reads that show the variant at a given site which leads to false negatives or there are too many reads that show the variant at a given site which leads to false positive homozygous calls.

- As for the Genome in a Bottle data, many FP and FN calls are located around repeats or indels. In contrast to the Genome in a Bottle data, here the gold standard variants have not been filtered as thoroughly which leads to a larger number of variant sites in questionable regions.

- The gold standard variants have been obtained by batched multisample calling using GATK UnifiedGenotyper. This is an explanation why in this test SAMtools mpileup performed worse than the two GATK callers.

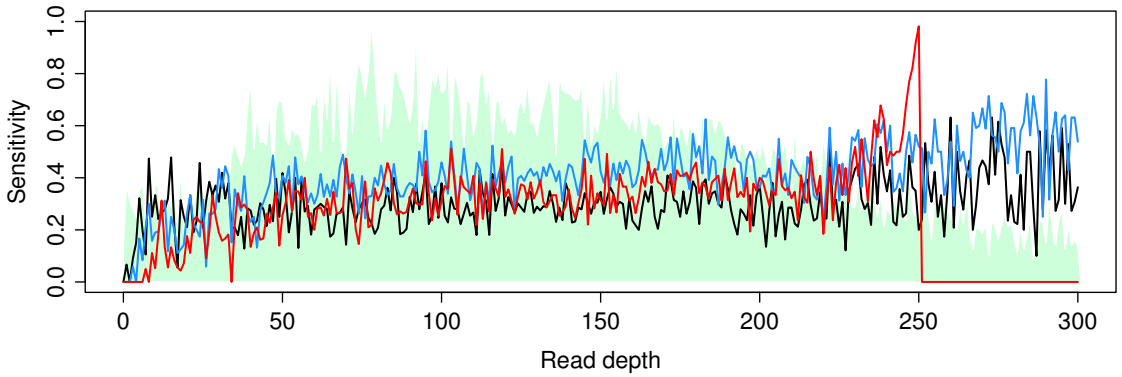


Figure 3.13.: Sensitivity of WESSIM indel calls by read depth. Solid lines show sensitivity of the different variant callers (black: SAMtools mpileup; red: GATK UnifiedGenotyper; blue: GATK HaplotypeCaller). The green area shows the distribution of all variants across the read depth spectrum.

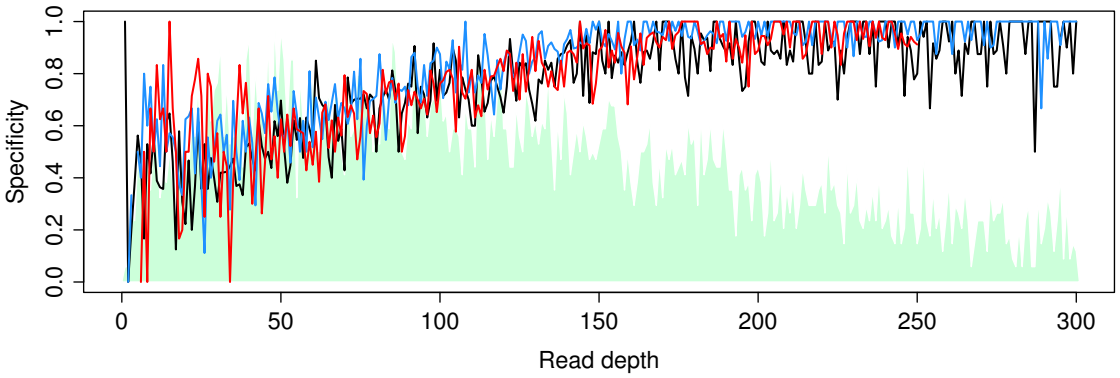


Figure 3.14.: Specificity of WESSIM indel calls by read depth. Solid lines show specificity of the different variant callers (black: SAMtools mpileup; red: GATK UnifiedGenotyper; blue: GATK HaplotypeCaller). The green area shows the distribution of all variants across the read depth spectrum.

3.2.4. Using Subsets of Data

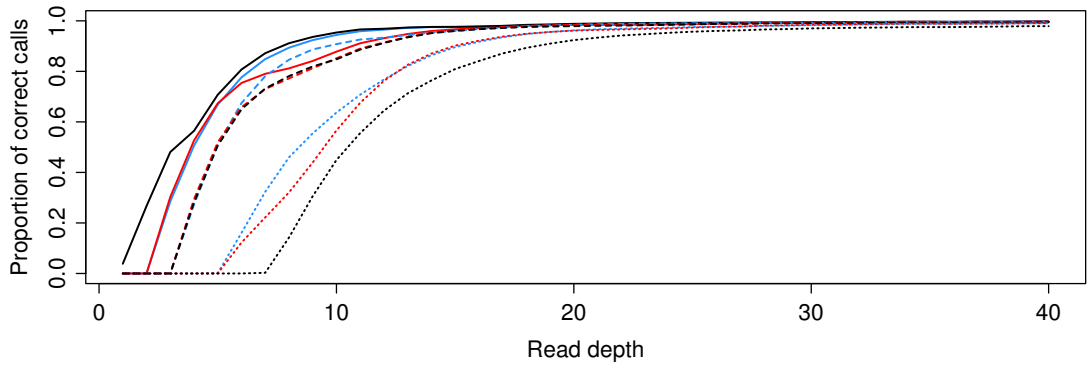
As has already been demonstrated in the last chapters, the most important aspect that influences the quality of variant calling is the read depth. However, comparing to a gold standard as in Chapter 3.2.1 or comparing to *in silico* data as in Chapter 3.2.3 has two major drawbacks for assessing the influence of read depth: (i) variants at positions with sufficient read depth are overrepresented and (ii) other factors, such as alignment problems at repetitive regions and indels, influence the quality of variant calling and are hard to control for.

Downsampling

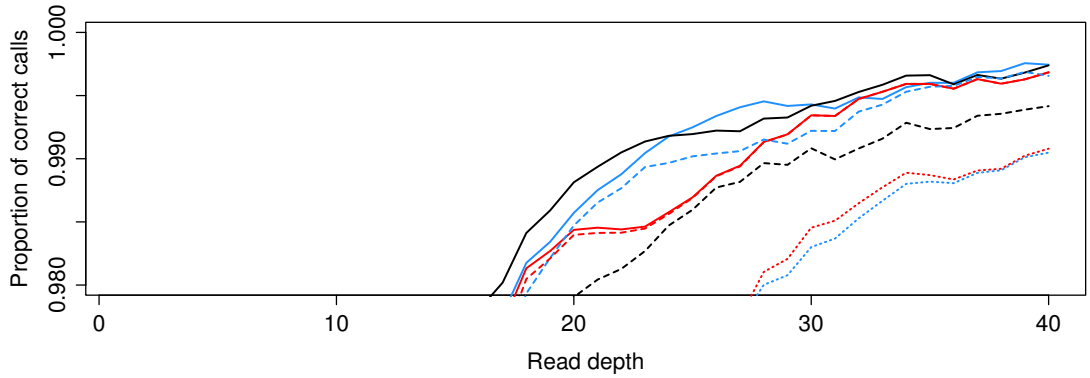
To assess the sensitivity of variant calling at certain read depths, a process called *downsampling* can be used. The general workflow is shown in Algorithm 1:

```
foreach known variant do
    for depth: from 1 to 40 do
        for draws: from 1 to 100 do
            draw depth reads at variant site from original BAM file;
            call variant in drawn reads;
            check if variant has been called correctly;
        end
        sum up all correct/incorrect calls from all draws;
    end
end
sum up all correct/incorrect calls from all variants;
```

Algorithm 1: Schematic of downsampling.



(a) SNV sensitivity - whole range



(b) SNV sensitivity - zoomed to 0.98-1.00

Figure 3.15.: Sensitivity of downsampling SNV calls by read depth. Solid lines show sensitivity of the different variant callers (black: SAMtools mpileup; red: GATK UnifiedGenotyper; blue: GATK HaplotypeCaller) for all variants. Dashed lines show variants with Genotype Quality (GQ) >50 and dotted lines show variants with the maximum Genotype Quality (GQ=99)

This algorithm has been used to analyze 579 known SNVs and 69 indels. 100 draws have been performed at each read depth, leading to 57,900 an 6,900 checked variant calls at each read depth for SNVs (Figure 3.15) and indels (Figure 3.16), respectively. With a read depth of 15-20 a maximum sensitivity is reached for both SNVs and indels. Sensitivity reaches 99.5% for SNVs. The remaining 0.25% are calls that are missed because the proportion of variant reads is too low in the randomly drawn reads. In the case of indels, a sensitivity of approximately 90% is reached. At low read depth GATK HaplotypeCaller performs best. GATK Unified Genotyper performs worse than the other two callers at all read depth levels. Dashed and dotted lines in the plot show how filtering for variants with higher GQ influences sensitivity. While requiring a GQ of at least 50 (dashed lines) only influences sensitivity up to a read depth of 15-20, requiring the maximum GQ of 99 (dotted lines)

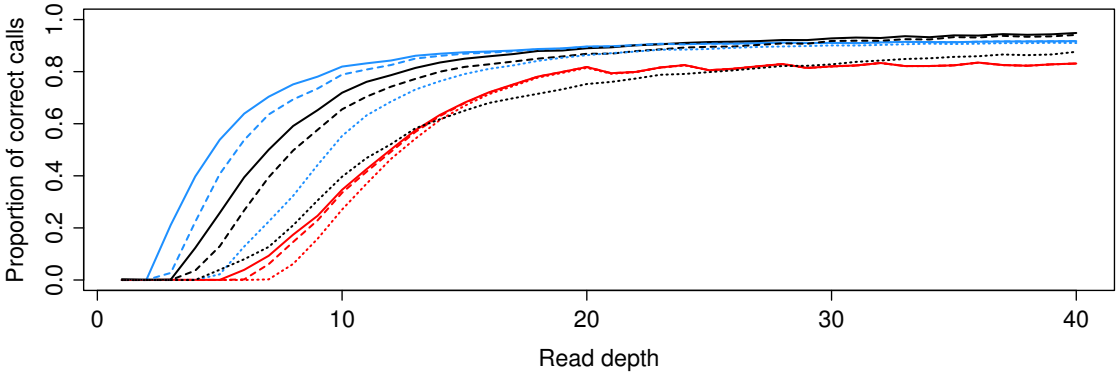


Figure 3.16.: Sensitivity of downsampling Indel calls by read depth. Solid lines show sensitivity of the different variant callers (black: SAMtools mpileup; red: GATK UnifiedGenotyper; blue: GATK HaplotypeCaller) for all variants. Dashed lines show variants with Genotype Quality (GQ) >50 and dotted lines show variants with the maximum Genotype Quality (GQ=99)

lowers sensitivity on the whole read depth spectrum, especially for SAMtools mpileup variants.

Multi Sample Calling

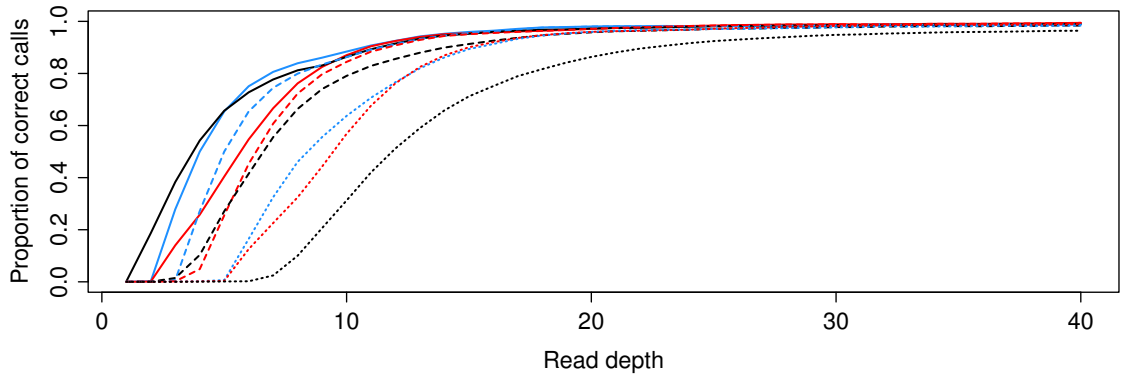
foreach *known variant* **do**
  for *depth: from 1 to 40* **do**
    for *draws: from 1 to 100* **do**
      draw *depth* reads at *variant* site from original BAM file;
      draw *depth* reads at *variant* site from 100 control BAM files that do not harbor the *variant*;
      call *variant* in drawn reads **in multi sample mode**;
      check if *variant* has been called correctly;
    end
    sum up all correct/incorrect calls from all *draws*;
  end
end
sum up all correct/incorrect calls from all *variants*;

Algorithm 2: Schematic of downsampling - multi sample calling.

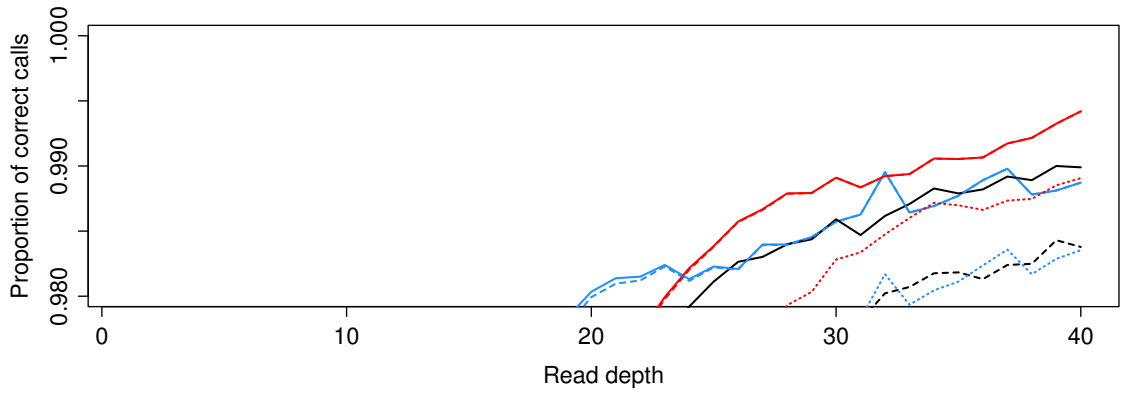
Calling variants from more than one sample together, i.e. *multi sample calling*, is assumed to increase sensitivity especially if the per sample read depth is low. Current multi sample calling alogrithms adapt prior probabilities for variants based on the observed reads of all samples and they apply *Linkage Disequilibrium (LD)* based methods[90]. However, multi sample calling procedures have been developed in the course of population scale sequencing projects, e.g. the 1000 Genome Project[128], where the main purpose was to accurately

detect rare to common variants, but not very rare variants. Most samples sequenced during this PhD project, have been sequenced to detect very rare variants causing rare genetic disorders. To test if sensitivity also increases in the case of very rare variants, the downsampling approach above has been repeated with small modifications (Algorithm 2).

Figures 3.17 and 3.18 show that the detection of singletons using multi sample calling with 101 samples does not improve compared to single sample calling (Figure 3.15). Sensitivity even seems to decrease, when looking at the zoomed Figures 3.15b and 3.17b.



(a) SNV sensitivity - whole range



(b) SNV sensitivity - zoomed to 0.98-1.00

Figure 3.17.: Sensitivity of downsampling SNV calls by read depth. Multi sample calling has been performed using 100 control samples. Solid lines show sensitivity of the different variant callers (black: SAMtools mpileup; red: GATK UnifiedGenotyper; blue: GATK HaplotypeCaller) for all variants. Dashed lines show variants with Genotype Quality (GQ) >50 and dotted lines show variants with the maximum Genotype Quality (GQ=99)

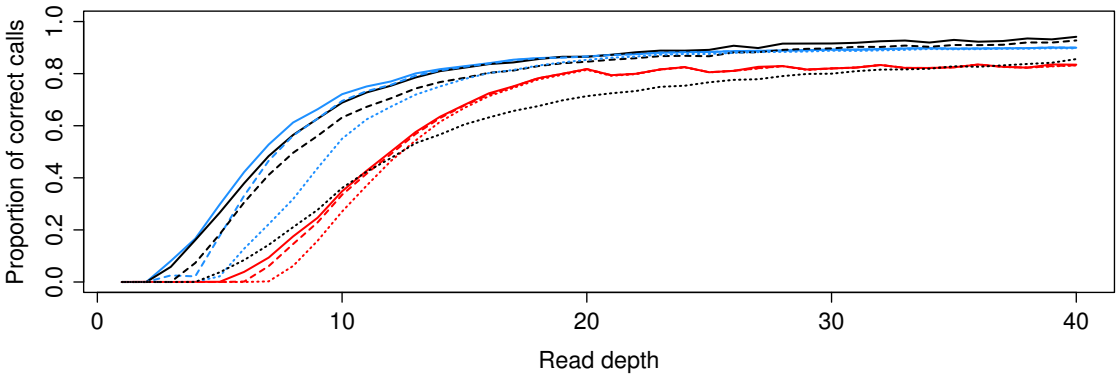


Figure 3.18.: Sensitivity of downsampling Indel calls by read depth. Multi sample calling has been performed using 100 control samples. Solid lines show sensitivity of the different variant callers (black: SAMtools mpileup; red: GATK UnifiedGenotyper; blue: GATK HaplotypeCaller) for all variants. Dashed lines show variants with Genotype Quality (GQ) >50 and dotted lines show variants with the maximum Genotype Quality (GQ=99)

Partitioning of Data

Since the downsampling approach only uses predefined sites of known variants, it can not be used to assess specificity. However, one can use subsets of reads from a sequencing experiment with sufficient read depth or from a sample that has been sequenced in two distinct experiments to call variants and calculate both sensitivity and specificity. Here, subsets from two exome sequencing experiments of two sample from which also whole genome data is available, have been used for this approach. Two different sets of gold standard variants for sensitivity and specificity calculations have been used:

- For sensitivity, the gold standard VCF files include only variants that have been called by all three variant callers in exome and whole genome data.
- For specificity, the gold standard VCF files include variants that have been called by any of the three variant callers in exome or whole genome data.

The two exome sequencing datasets have been split into six subsets each, as depicted in Figure 3.19.

Variant calling has then been performed in each of the 12 subsets, the calls have been compared to the gold standard files described above and TP, FN and FP calls have been summed up. Figure 3.20 shows sensitivity calculations. Due to the splitting of the data, the majority of variants are located in regions with lower coverage. Again, the values for SNVs are higher and more uniform than the values for Indels and a sensitivity plateau is reached at a read depth of about 15 to 20, regardless of the variant caller.

Figure 3.21 shows specificity of variant calling in the 12 subsets relative to the GQ of the calls. The peaks in the green area show that variants are not equally distributed over the

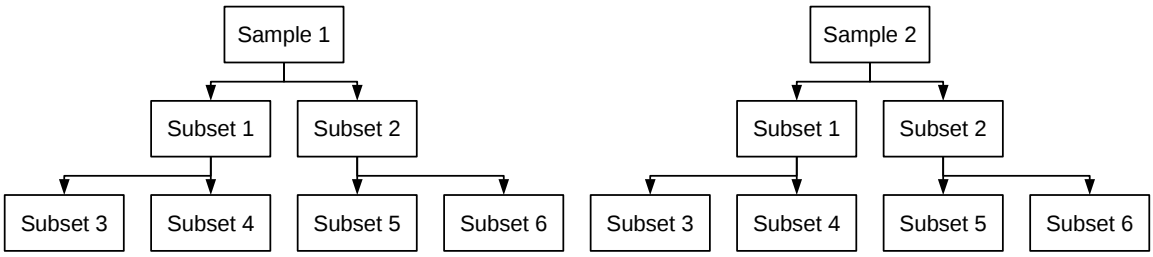
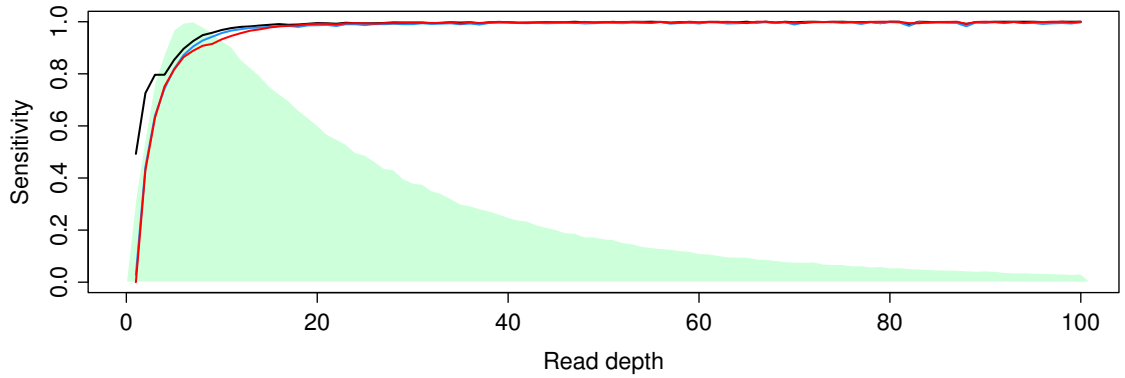
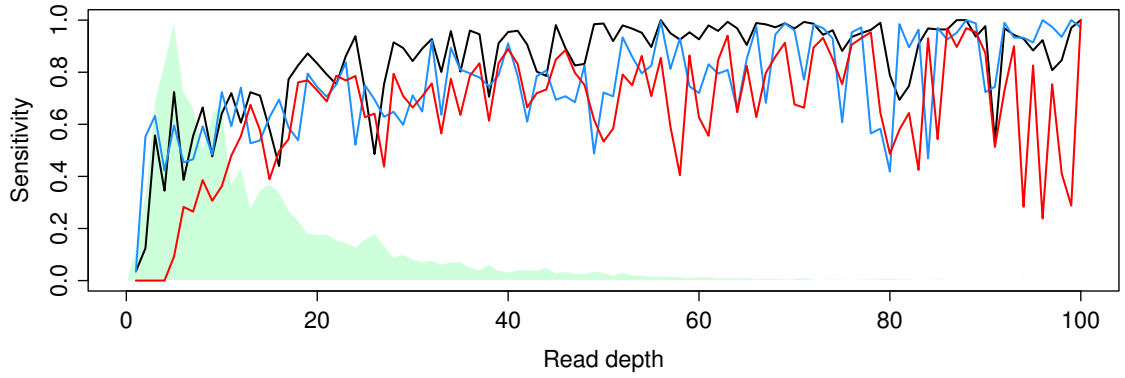


Figure 3.19.: The two samples have been splitted into six subsets, each.



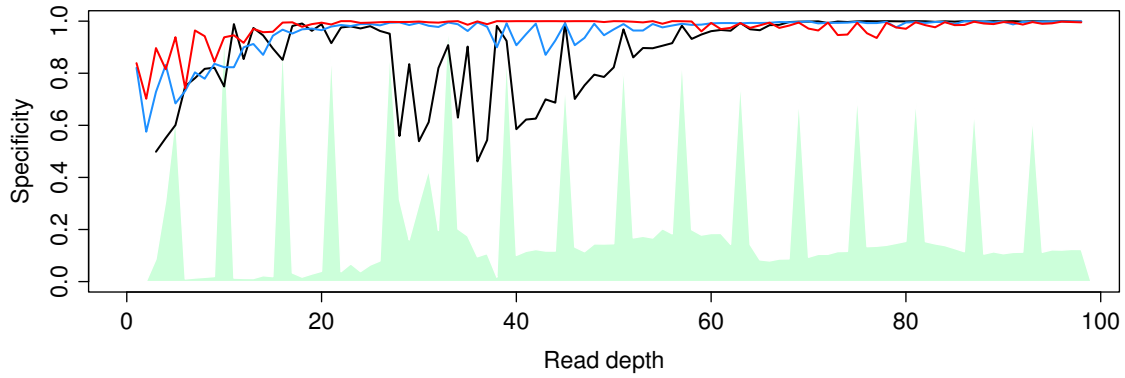
(a) SNV sensitivity - whole range



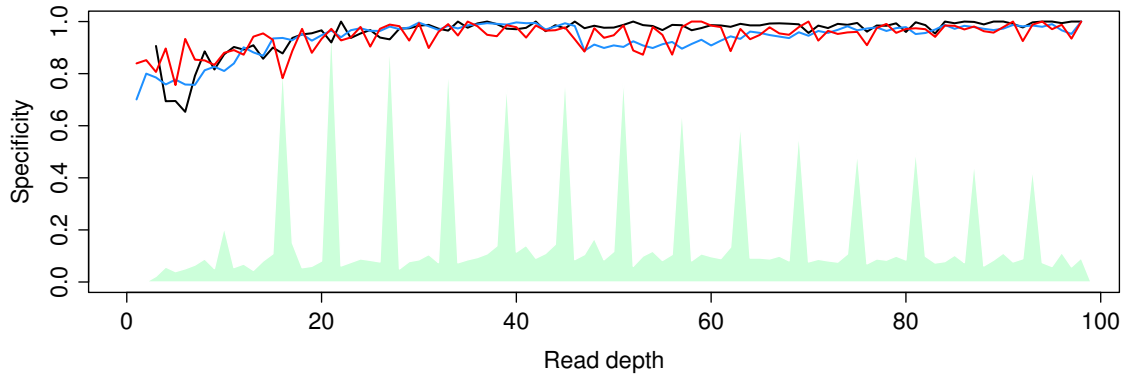
(b) Indel sensitivity - whole range

Figure 3.20.: Sensitivity of SNV and Indel calls by read depth in subsets of two exome samples. Solid lines show sensitivity of the different variant callers (black: SAMtools mpileup; red: GATK UnifiedGenotyper; blue: GATK HaplotypeCaller) for all variants.

whole spectrum of GQ values. Some QG values are overrepresented. Specificity increases slightly with increasing GQ values, but there is no clear value to set a threshold for filtering.



(a) SNV specificity - whole range



(b) Indel specificity - whole range

Figure 3.21.: Specificity of SNV and Indel calls by Genotype Quality in subsets of two exome samples. Solid lines show sensitivity of the different variant callers (black: SAMtools mpileup; red: GATK UnifiedGenotyper; blue: GATK HaplotypeCaller) for all variants.

3.2.5. Using Novelty of Variants

The last chapters showed benchmarking methods to calculate sensitivity and specificity of variants and how read depth and Genotype Quality, as a measure for the reliability of a variant, affect them. This chapter introduces an additional property to assess the influence of read depth and other quality measures on variants: the novelty of a variant. The key assumption here is that true variants are more likely to appear already in public databases, such as dbSNP[115], than false variants. In contrast to previous methods, no gold standard

to compare to is required, but it is not possible to directly identify wrong variants but only to estimate the proportion of false positives in a distinct class of variants. The variants assessed in this chapter are taken from the in-house database and novelty of a variant is defined by its absence from dbSNP v135.

Figure 3.22b shows the scaled proportion of known variants relative to read depth. Interestingly, there is a clear gap at a read depth of 20. This is in line with the findings from the previous chapters where a sensitivity plateau was reached at this coverage. Figure 3.22a shows that the absolute number of variants with a read depth lower than 20 is significantly higher than in the adjacent bins, suggesting that a majority of these variants are false positives.

This method easily allows assessement of the effect of filters. A property often used for filtering in the in-house database is the *variant quality* assigned by SAMtools. The values usually range from 0 to 225 which allows for a more fine-grained filtering than the Genotype Quality, but the correlation between those two values is high (0.89 pearson correlation).

Figure 3.23b shows that there is a clear increase of proportions for variants with a quality of at least 30. Figure 3.23a shows the absolute number of variants in the different bins. It can be seen that there is an excess of variants in the low quality bins, suggesting that the majority of those variants are false positives. However, there is a drop in proportion of novelty in the bin with the highest variant quality, which also holds most variants. This suggests that some of these variants are also false positives. To further investigate this, it is helpful to split the variants in rare and common variants. Rare is in this case defined as 10 or less alleles in the in-house database.

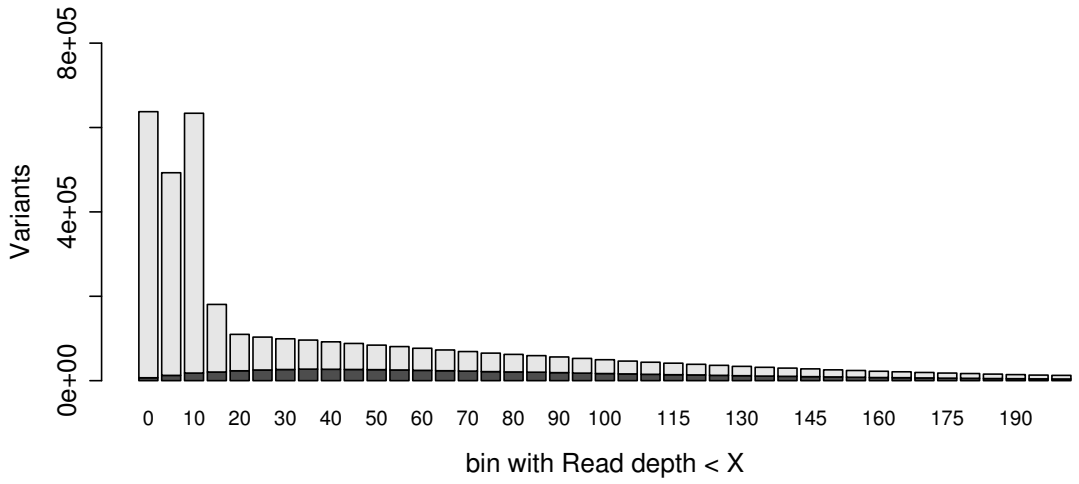
The majority of variants are rare (Figure 3.25a vs. Figure 3.24a; see also Chapter 3.3). As expected, proportions of dbSNP variants are high in variants that are common in the in-house database (Figure 3.24b) and significantly lower in rare variants (Figure 3.25b). The drop in the proportion of novelty in the last bin is only present in rare variants. Assuming that the proportion in the last bin should be similar to the proportion in the second last bin, it can be concluded that up to 10% of rare variants with maximum variant quality are false positives.

3.2.6. Conclusions

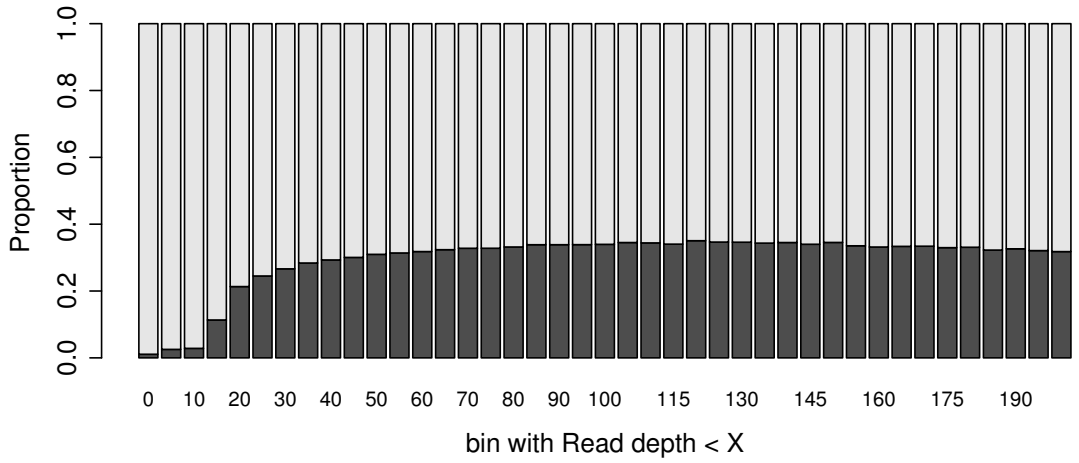
The results from the benchmarks in this chapter can be used to draw general conclusions on variant calling and filtering.

The most important factor is read depth. All tested variant callers perform equally well at SNV calling at positions with more than 15-20x coverage. GATK HaplotypeCaller, the newest of the tested variant callers and still under active development, performs better at indel calling than the other two callers. However, at the beginning of this PhD project, GATK HaplotypeCaller was not yet avaible, thus SAMtools mpileup was chosen as main variant caller. As of today, GATK HaplotypeCaller might be the best choice.

The two GATK programs for improving the quality of BAM files in order to improve variant calling, Indel Realigner and Base Quality Score Recalibrator, only improve variant calls of GATK UnifiedGenotyper. Thus, it should only be used if GATK UnfiedGenotyper is used for variant calling.



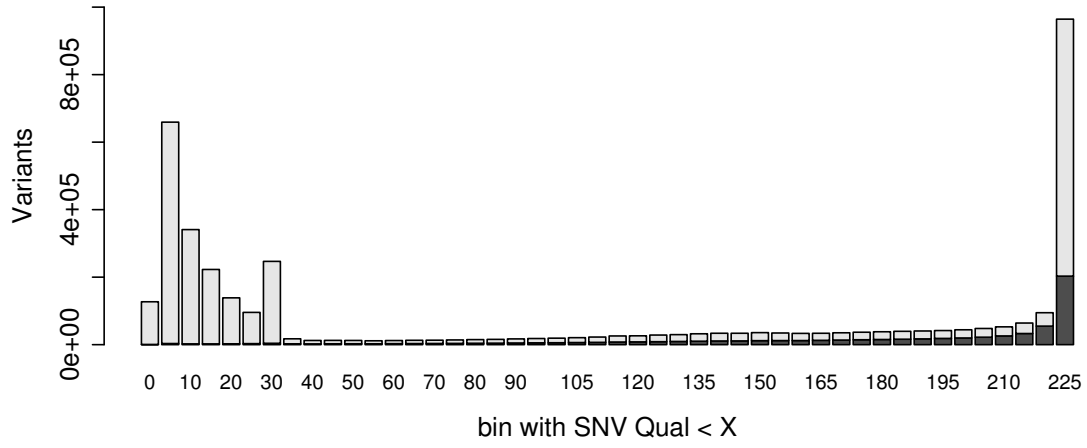
(a) Total variants



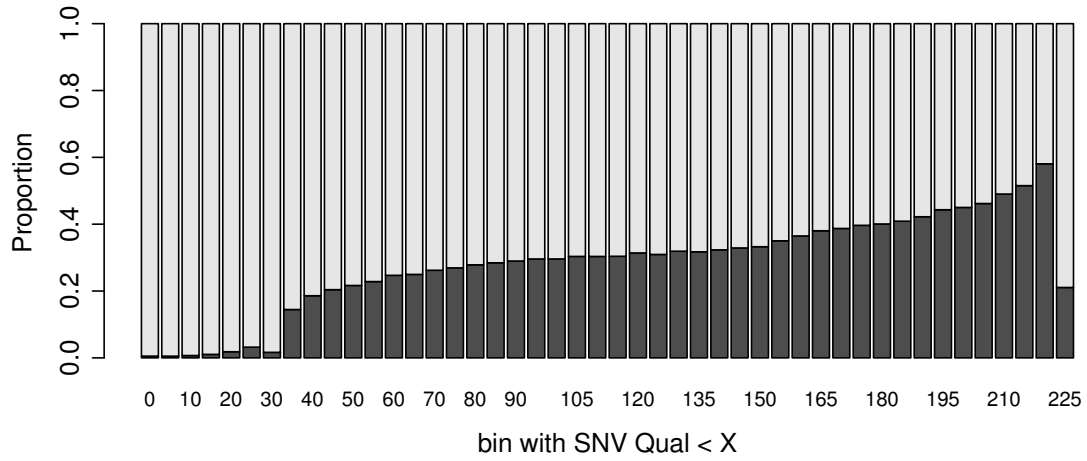
(b) Proportions

Figure 3.22.: Proportions of dbSNP variants for different bins of read depth. dbSNP variants are depicted in dark grey, other variants in light grey.

Multi sample calling does not improve the calling of singletons. However, it has still an advantage compared to single sample calling: if only one sample has a variant at a position, also genotypes for all other samples are calculated if the coverage of the samples is sufficient at the position. If a variant gets called in only one sample using single sample calling, one does not know if the other samples do not have the variant or they do not have enough reads at this position. Thus, multi sample calling allows calculations of allele frequencies whereas single sample calling does not.



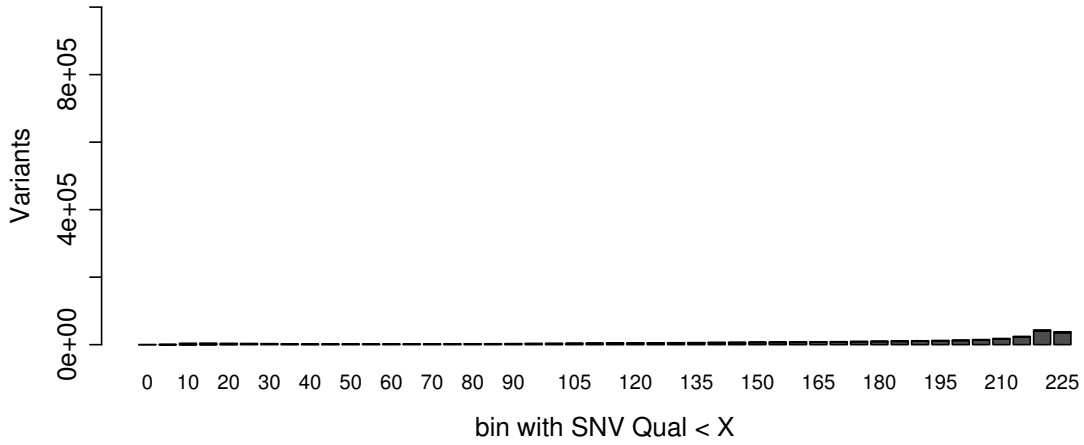
(a) Total variants



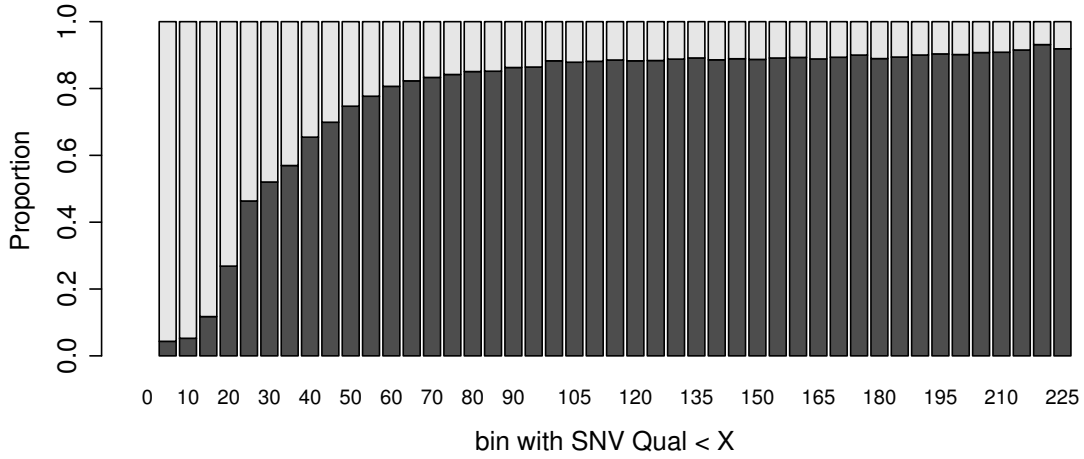
(b) Proportions

Figure 3.23.: Proportions of dbSNP variants for different bins of variant quality. dbSNP variants are depicted in dark grey, other variants in light grey.

In addition to read depth, also the variant quality given by SAMtools can be used for filtering variants. It could be shown that a cutoff of 30 is sensible, with the exception of variants in homopolymer regions where the cutoff should be lower. Genotype Quality is not as useful for filtering, because the values are not assigned continuously over the whole spectrum. No sensible cutoff could be defined. However, if variants are called using GATK UnifiedGenotyper or GATK HaplotypeCaller, the recommended way of filtering is using GATK VariantRecalibrator (see Chapter 1.3.3).



(a) Total common variants

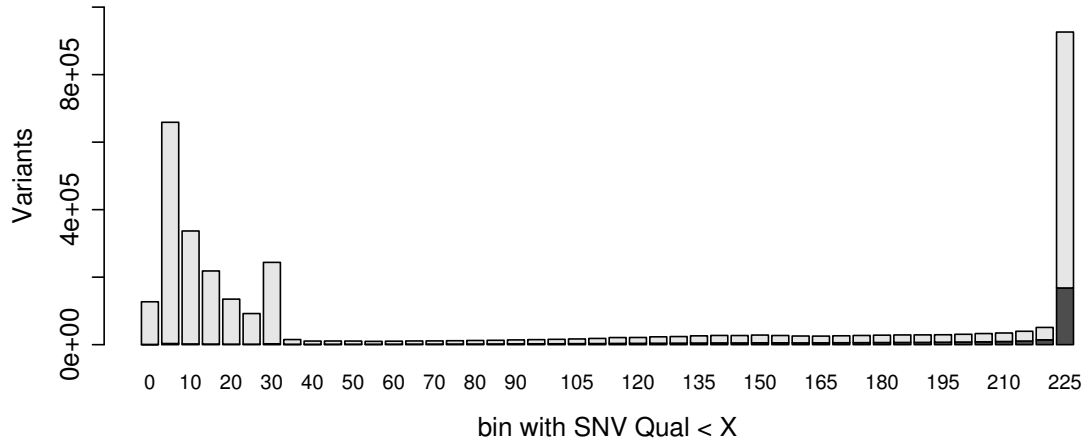


(b) Proportions for common variants

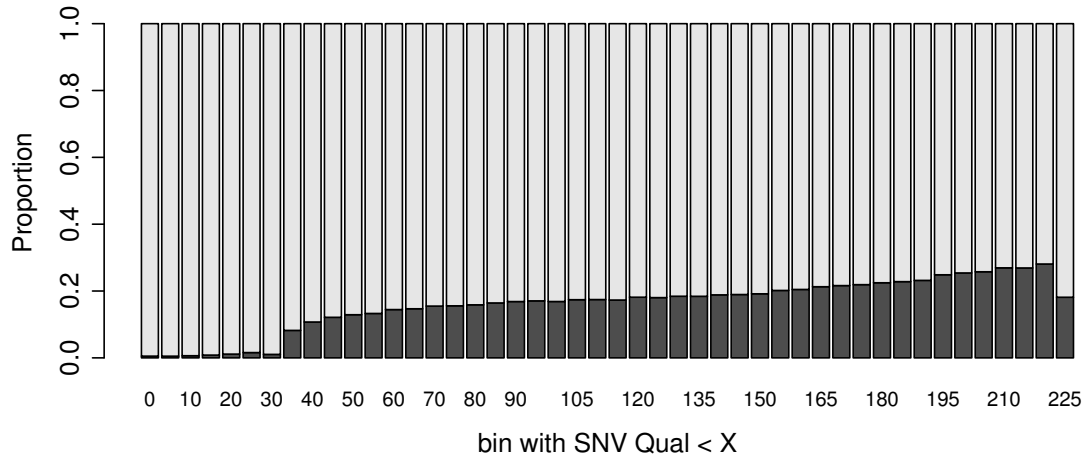
Figure 3.24.: Proportions of dbSNP variants for different bins of variant quality. Variants are common (i.e. >10 alleles) in our in-house database. dbSNP variants are depicted in dark grey, other variants in light grey.

3.3. Identifying Disease Causing Variants

Variant calling and filtering as described in the last chapter leads to approximately 23,000 good quality coding variants per sample. These variants must be annotated and filtered in order to identify putative disease causing variants. Figure 3.26 shows the filter strategy that has been used throughout this PhD project. It is roughly based on guidelines for variant detection in research[80] and diagnostic[105][107] settings.



(a) Total rare variants



(b) Proportions for rare variants

Figure 3.25.: Proportions of dbSNP variants for different bins of variant quality. Variants are rare (i.e. $\leq$10 alleles) in our in-house database. dbSNP variants are depicted in dark grey, other variants in light grey.

First, a frequency filter is applied to the variants of a sample (Chapter 3.3.1). Variants that are too frequent with respect to the prevalence and mode of inheritance of the investigated disease, can be filtered out.

To further reduce the amount of putative disease causing variants, the assumed mode of inheritance can be taken into account (Chapter 3.3.2). If, for instance, a recessive mode of inheritance is assumed, all variants that are not homozygous or compound heterozygous

can be filtered out.

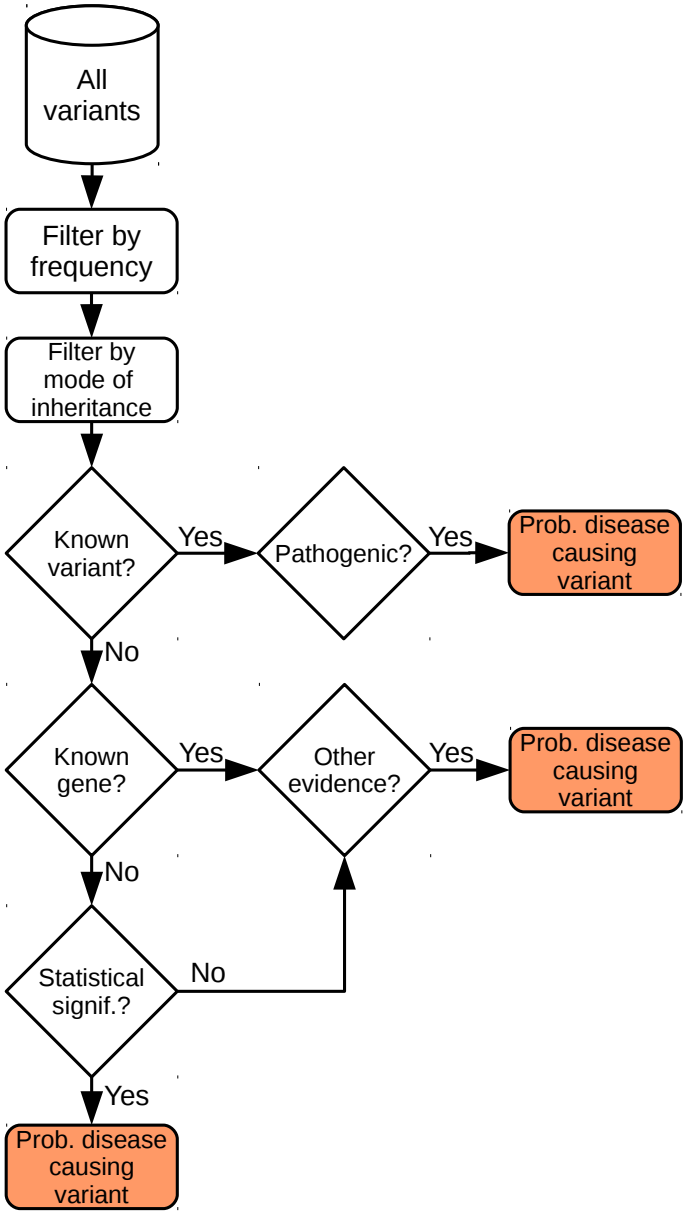


Figure 3.26.: Flowdiagram for the detection of putative disease causing variants from exome data.

The remaining variants are then searched for known disease causing variants from public databases, such as HGMD[121] or ClinVar[60], and for variants in genes known to be associated with the respective disease (Chapter 3.3.3). If a sample has a variant that has been described to be associated with a disease fitting the phenotype of the sample, this

variant is very likely disease causing. Another strong evidence for disease causality of a variant is, if it has not been previously described to be disease causing but is located in a gene that is associated with the respective disease. However, in this case additional evidence for the pathogenicity of the variant is required, such as *in silico* predictions of the deleterious effect of the variant or experimental evidence (Chapter 3.3.5).

If no known variants or variants in known genes can be found, calculations of statistical significance are used to identify variants in genes that are significantly enriched in cases compared to controls (Chapter 3.3.4). In addition to statistical evidence or if statistics can not be used, e.g. if the number of cases is not sufficient, experimental evidence and *in silico* predictions of deleteriousness can be used to identify disease causing variants (Chapter 3.3.5).

Example datasets that have been generated during this PhD project are used to demonstrate the different steps throughout this chapter.

3.3.1. Variant Frequencies

Approximately 11,500 synonymous, 11,000 missense, 300 loss-of-function (nonsense, stoploss, splice) and 300 frameshift variants with sufficient quality (SAMtools variant quality≥30) have been called per sample (Table 3.7).

Kit	Varianttype						
	synonymous (±s.d.)	missense (±s.d.)	nonsense (±s.d.)	stoploss (±s.d.)	splice (±s.d.)	frameshift (±s.d.)	indel (±s.d.)
38Mb kits	8,353 (±256)	7,036 (±228)	48 (±6)	18 (±3)	37 (±4)	60 (±7)	75 (±16)
50Mb kits (v3)	10,785 (±298)	9,765 (±271)	84 (±7)	28 (±3)	146 (±10)	185 (±14)	170 (±25)
50Mb kits (v4)	11,328 (±363)	10,503 (±317)	97 (±8)	34 (±4)	177 (±9)	274 (±14)	313 (±21)
50Mb kits (v5)	11,498 (±331)	10,811 (±334)	100 (±8)	37 (±4)	183 (±10)	278 (±18)	315 (±29)

Table 3.7.: Average number of variants per sample.

	Singletons	AF <1%	AF <5%	All variants
Non-synonymous variants	355,209	691,800	706,407	730,022
Synonymous variants	161,761	338,145	348,823	371,156
All variants	513,572	1,023,092	1,048,182	1,093,770

Table 3.8.: Overview of all called coding variants. About half of the variants in the database are present in only a single sample.

One of the major assumptions for the detection of putative disease causing variants for rare diseases is that these variants are also rare. Hence, the frequency of a variant in all

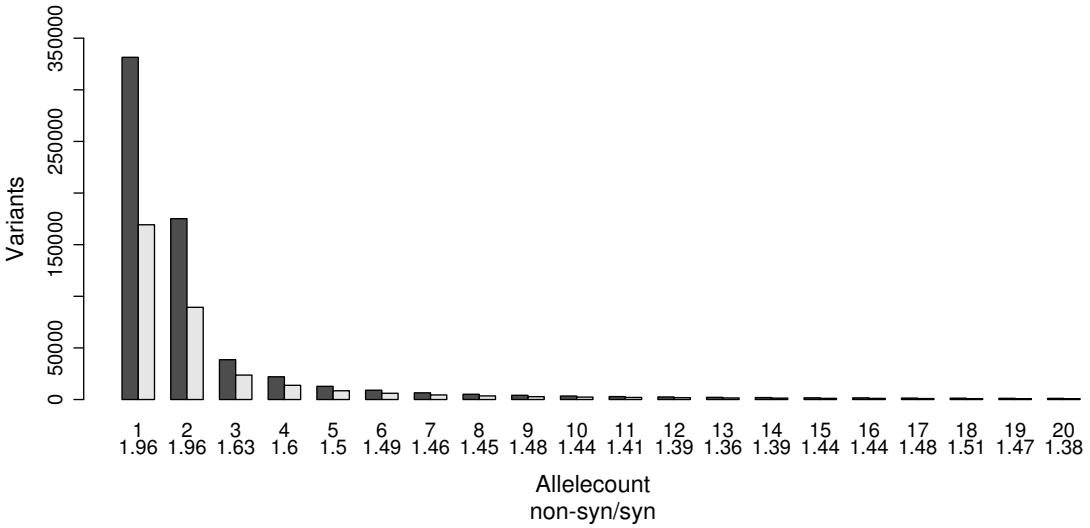


Figure 3.27.: Non-synonymous (dark grey) and synonymous (light grey) variants per allele count. At lower allele counts, the proportion of non-synonymous variants is higher.

samples that do not have the same disease can be used for filtering. The actual cutoff frequency to use for filtering depends on the incidence and mode of inheritance of the investigated disease. Cystic fibrosis is, for instance, the most common lethal genetic disease in caucasian populations and occurs in about 1 in 3,000-4,000 Germans[94]. It is recessive and caused by mutations in the *CFTR* gene. Up to 3/4 of the cases in Europe are caused by a single mutation, a deletion of a single amino acid at position 508 (Phe508del). Currently, 102 samples in the in-house database which are not diagnosed with cystic fibrosis, are carriers of this mutation. Thus, the cutoff frequency for identifying disease causing mutations for cystic fibrosis in the in-house database must be above 1.25%.

313 trios of children with intellectual disability (ID) and their parents have been sequenced in two different projects to detect disease causing *de novo* variants[104]. These variants are not expected to occur in any other samples, despite the prevalence of ID of 1.5-2.5%[65]. This has two reasons: (i) ID is genetically heterogenous. Thus, the putative disease causing mutations spread across many possible loci. (ii) The investigated cases are assumed to be dominant, have severely reduced reproductive fitness and the putative disease causing mutations are assumed to have full penetrance. Thus, true disease causing mutations are not expected to occur in samples that do not suffer from ID.

The vast majority of variants are rare, when looking at all samples together (Table 3.8; Figure 3.27). 47% of the variants are *singletons*, i.e. they have been called in only a single sample, and 92% of the variants have a frequency of below 1% in the in-house database. On a per sample level, variant frequencies are distributed equally (Figure 3.28), with the exception of two peaks: the peak at very low frequencies represents private variants of a specific sample and the peak at very high frequencies represents private variants of the sample from which the reference genome was derived. With approximately 4,500 sequenced sam-

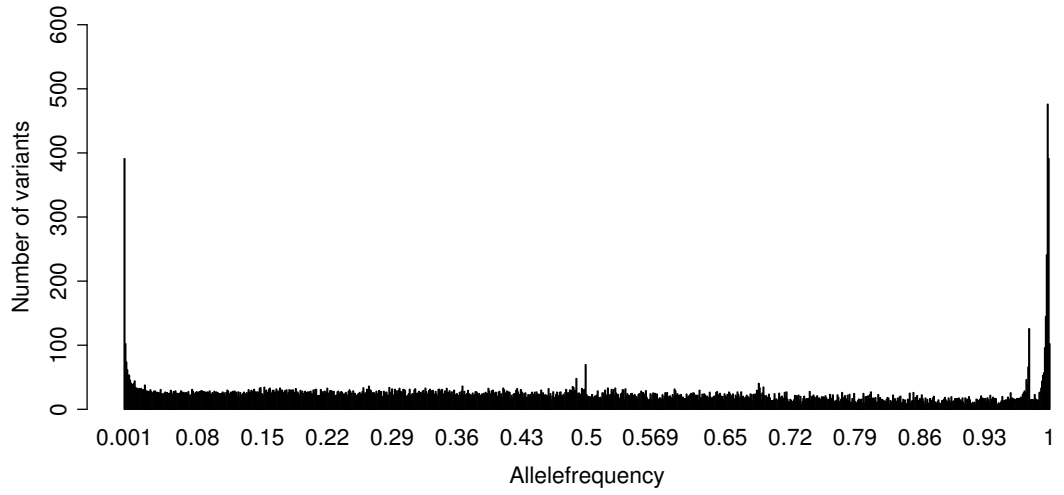


Figure 3.28.: Allele frequency of variants of a single sample (average over all samples).

ples, filtering for variants with an allele frequency below 0.1%, which corresponds to variants that are present in up to eight controls in Figure 3.27, about 350 variants per sample remain. As expected, approximately 2/3 of these variants are non-synonymous. Due to the genetic code, substitutions in the last base of a codon often do not change the amino acid, whereas changes in the other two bases do. Hence, statistically it is expected that around 2/3 of substitutions cause amino acid changes.

To further reduce the number of very rare variants per sample, it would be possible to sequence more and more samples until the complete spectrum of non-lethal coding variants is known. To achieve this, the number of sequenced samples must be significantly higher than 4,500, as can be seen in Figure 3.29. Also with 4,500 sequenced samples, each new sample introduces approximately 100 novel variants. This number decreases only slowly.

Frequencies from Public Databases

In addition to frequencies from the in-house database, frequencies from public databases can also be used for filtering[44]. Two widely used databases for this purpose are the *Exome Variant Server (EVS)*[127] and the *1000 Genome Project*[128]. EVS offers variants from 6,503 exome samples (4,300 European Americans and 2,203 African Americans) sequenced in the course of the *NHLBI GO Exome Sequencing Project (ESP)*. The 1000 Genome Project currently offers variants from 1,092 samples from 26 populations for which exome sequencing and low coverage whole genome sequencing was performed.

These two datasets can be used via the in-house web interface. However, a major drawback of data from public databases compared to data from the in-house database is, that it has been analyzed differently and may therefore be biased differently.

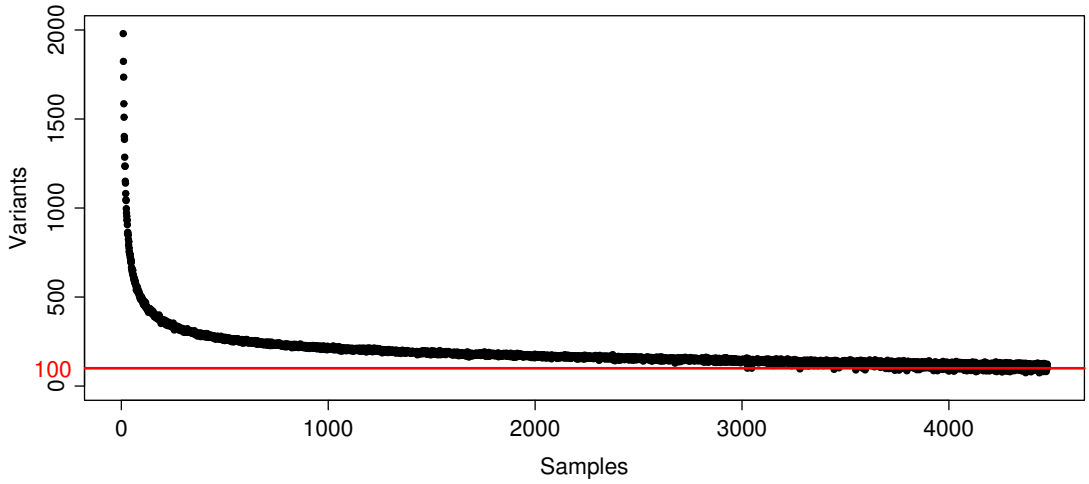


Figure 3.29.: Novel variants in database when adding new samples

3.3.2. Mode of Inheritance

The assumed inheritance pattern, or *mode of inheritance*, of a disease can be used for filtering of candidate variants. If, for instance, a recessive mode of inheritance is assumed for a disease, only homozygous and compound heterozygous variants are of interest. If additionally the parents of the patient are consanguineous, the disease causing variant is most likely homozygous. If a dominant mode of inheritance is assumed, sequencing of distantly related, affected family members is beneficial for filtering of candidate variants.

A special mode of inheritance are *de novo* mutations, i.e. mutations that occur in the germline of a parent and can therefore be only detected in somatic cells of the child but not the parents. *De novo* mutations are assumed to cause severe disorders, such as ID, that massively reduce reproductive fitness and are unlikely to be inherited. The common method to detect *de novo* variants using exome sequencing is to perform *trio* sequencing, i.e. sequencing of the affected child and both parents, and then identifying variants that are present in the child but not the parents. During this PhD project, 313 trios of children with intellectual disability and their parents have been sequenced. Altogether, 618 *de novo* variants have been identified in these trios (Table 3.9). In 71 of the patients *de novo* variants have been identified in genes already associated with ID (see also Chapter 3.3.3).

To assess the general properties of the set of *de novo* variants in cases, the *de novo* mutation rate per sample has been calculated and compared to the mutation rate of 50 control trios (Figure 3.30 and Table 3.10). The mutation rate fitted to the expected poisson distribution (Figure 3.30a) and has been higher in cases than in controls (Figure 3.30b).

Interestingly, people in the case group have a significantly higher number of protein altering variants than people in the control group (Table 3.10). Especially the fraction of loss-of-function variants is strikingly higher in cases compared to controls, whereas the fraction of synonymous variants is lower (Figure 3.31). This does not directly lead to the discovery of single disease causing variants, but it gives some indication of the prescence

Mutationtype	Mutations
frameshift	61
splice	26
stoploss	2
nonsense	32
indel	14
missense	376
syn	106
nearsplice	1
Total	618

Table 3.9.: *de novo* variants in 313 patients with intellectual disability.

of disease causing variants in the case group.

	Protein-altering	**Point**	**Missense**	**Synonymous**
Sequence length (bases)	$2.54x10^{7}$	$3.34x10^{7}$	$2.54x10^{7}$	$7.93x10^{7}$
Cases (n=313)				
Mutations	511	542	376	106
Mutations per person	1.63	1.73	1.20	0.34
Mutation rate	$3.21x10^{-8}$	$2.59x10^{-8}$	$2.36x10^{-8}$	$2.13x10^{-8}$
Controls (n=50)				
Mutations	36	55	31	22
Mutations per person	0.72	1.1	0.62	0.44
Mutation rate	$1.42x10^{-8}$	$1.65x10^{-8}$	$1.23x10^{-8}$	$2.77x10^{-8}$
p-value	$2.68x10^{-7}$	$2.44x10^{-3}$	$6.61x10^{-4}$	0.32

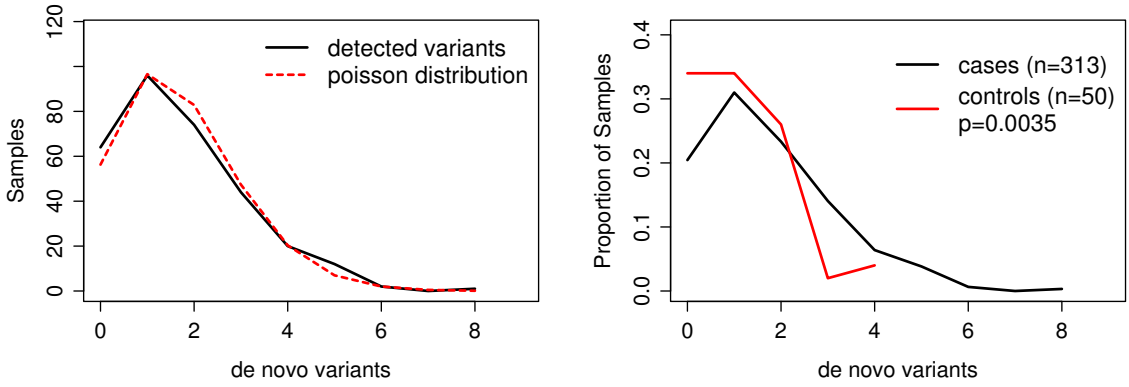
Table 3.10.: *de novo* mutation rate calculations for 313 patients with intellectual disability and 50 controls.

It is assumed that the *de novo* mutation rate is proportional to the age of the father at birth[56]. Indeed, the number of *de novo* variants in the samples presented here is significantly correlated to the age of the father at birth (Pearson's correlation p-value=0.0004; Figure 3.32b) and not significantly correlated to the age of the mother at birth (Pearson's correlation p-value=0.020; Figure 3.32a).

3.3.3. Known Pathogenic Variants and Genes

After filtering for frequency and mode of inheritance, the remaining variants can be looked up in public databases that include known variant to phenotype relationships [44].

One of the most used databases for this purpose is the *Human Gene Mutation Database (HGMD)*[121], which contains variants obtained by literature mining. The professional



(a) *de novo* point mutations per sample compared to the poisson distribution.

(b) *de novo* point mutations of cases vs. controls

Figure 3.30.: de novo point mutations per sample

version of HGMD, which requires a fee-based license, currently contains 148,413 variants. The free, public version contains 105,417 variants[8]. It contains variants from the professional version that are at least three years old.

Another, recently started, public database is *ClinVar*[60]. ClinVar contains variants that were submitted by researchers and clinicians or extracted from public databases (e.g. OMIM[95]) or expert consensus reports. It currently contains 107,784 variants and is growing rapidly[9].

If a detected variant (or a pair of compound heterozygous variants for recessive diseases) can be found in such a database and the variant is marked as causal for a matching disease/phenotype, it is very likely that it is indeed disease causing in the investigated sample. However, public databases also contain uncertain and false positive findings. For instance, Bell et al.[5] identified 460 recessive variants that were marked as disease associated in HGMD. They omitted 122 (27%) of these variants because they were common polymorphisms, sequencing errors or lacked evidence of pathogenicity.

The problem of uncertain or false positive classifications in databases also becomes evident by investigating the overlap of HGMD and ClinVar. At the time of writing, 28,968 of 105,643 disease causing mutations from HGMD can also be found in ClinVar. Of these 28,968 mutations, only 18,100 (62%) are classified as pathogenic in ClinVar. An example for problems due to uncertain classifications can be found by investigating the carrier status for cystic fibrosis in the in-house database. As mentioned above, approximately 1 in 3,000-4,000 Germans suffers from cystic fibrosis[94]. According to the Hardy-Weinberg equilibrium, up to 3.6% of the population are carriers of a cystic fibrosis mutation. However, 551 of 4,297 samples (12.8%) in the in-house database carry a mutation in CFTR that is classified as disease causing in HGMD. When looking at these mutations in detail, the most common mutation (168 of 551) is not the expected deletion (Phe508del; most common mutation in literature[94]), but a synonymous mutation directly upstream of the donor splice

[8] `http://www.hgmd.org/` - Last accessed: 18.06.2014
[9] `http://www.ncbi.nlm.nih.gov/clinvar/submitters/` - Last accessed: 18.06.2014

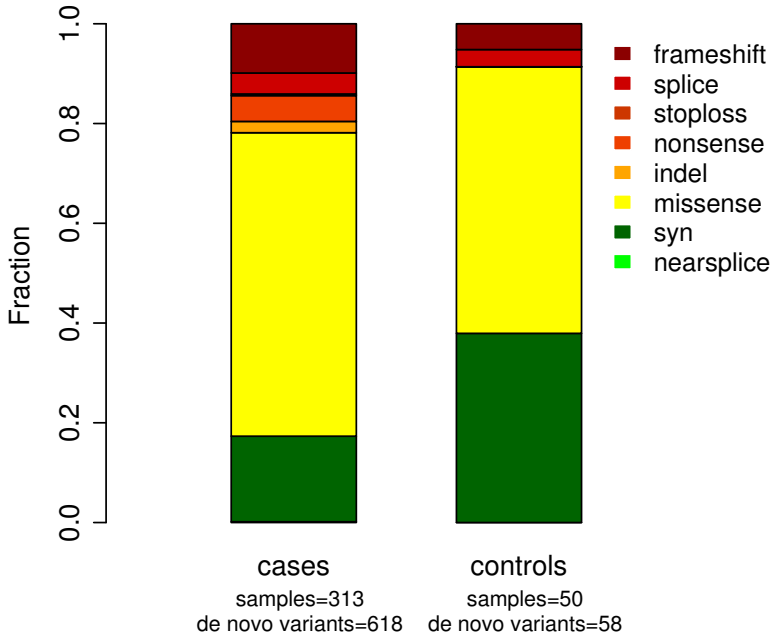
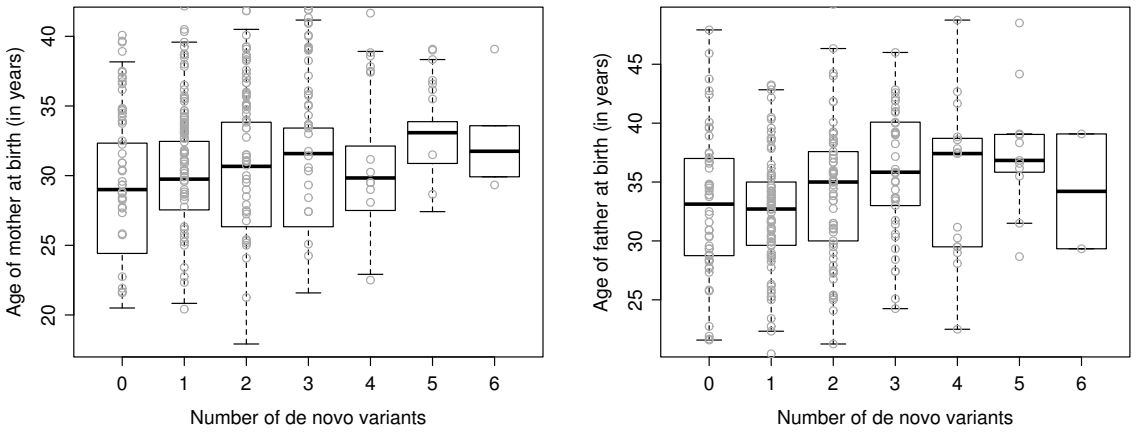


Figure 3.31.: Fraction of functions of de novo mutations



(a) Age of mother at birth

(b) Age of father at birth

Figure 3.32.: Number of de novo mutations vs. age of parents at birth

site of intron 10. This mutation is classified with "Uncertain significance" by ClinVar and also the evidence in the describing literature[92][123] is unclear, suggesting that this variant is rather a common polymorphism than a disease causing variant.

Variants in Known Genes

If no previously described disease causing mutation could be identified, novel variants in genes known to be associated with the disease of interest are investigated. Lists of known genes for a disease can be obtained from public databases, such as OMIM[95], or by literature research.

Search results

Date: 20.06.2014

DNA Id:	BRCA0001
Pedigree:	TU2253
Foreign ID:	24804
Libtype:	exomic
Libpair:	paired-end
Assay:	SureSelect50Mbv5
Seqence:	9.200
Average coverage:	114.25
At least 20x coverage:	96.09

n	Gene Symbol	Transcript	Avg depth	Depth > 20	Avg MapQual	Position, Avg depth, Depth > 20, Avg MapQual
1	ATM	NM_000051	118	97	59	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 63 32 77 42 82 100 43 49 99 64 42 186 169 47 33 172 113 50 192 54 163 106 128 81 126 167 59 16 182 40 174 165 47 89 100 100 100 87 100 95 88 100 100 82 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 97 0 100 96 100 100 100 100 60 60 59 60 59 59 59 58 59 58 59 60 59 59 59 60 59 59 59 60 59 59 59 59 59 59 58 59 60 59 59 60 60 60
2	BRCA1	NM_007294	174	100	59	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 118 245 181 46 36 82 120 65 155 155 198 142 124 206 175 65 113 95 156 23 79 47 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 97 100 100 58 59 59 58 59 59 59 60 59 60 60 59 59 59 59 60 59 57 59 60 59 60
3	BRCA1	NM_007297	177	100	59	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 118 245 181 46 36 82 120 65 155 155 198 142 124 206 175 65 113 95 156 23 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 97 58 59 59 58 59 59 59 60 59 60 60 59 59 59 59 60 59 57 59 60
4	BRCA1	NM_007298	126	100	59	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 118 245 181 46 36 82 120 65 155 155 200 142 124 138 175 65 113 95 156 23 79 47 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 97 100 100 58 59 59 58 59 59 59 60 59 60 60 59 59 60 59 60 59 57 59 60 59 60
5	BRCA1	NM_007299	124	100	59	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 152 245 46 36 82 120 65 155 155 200 142 124 138 175 65 113 95 156 23 79 47 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 97 100 100 58 59 58 59 59 59 60 59 60 60 59 59 60 59 60 59 57 59 60 59 60
6	BRCA1	NM_007300	172	99	59	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 118 245 181 46 36 82 120 65 155 155 200 10 142 124 206 175 65 113 95 156 23 79 47 100 100 100 100 100 100 100 100 100 100 100 0 100 100 100 100 100 100 100 100 97 100 100 58 59 59 58 59 59 59 60 59 60 60 59 59 59 59 59 60 59 57 59 60 59 60
7	BRCA2	NM_000059	119	99	59	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 63 110 65 8 41 223 50 64 142 96 60 54 93 187 148 112 157 273 92 121 183 67 142 104 219 218 100 100 100 0 100 100 100 100 100 99 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 60 59 59 60 59 59 60 57 59 59 60 59 59 59 59 59 60 60 59 59 60 60 59 59 59 59

Figure 3.33.: Screenshot of the coverage mask of the web interface for breast cancer candidate genes.

If a list of known genes is available, it is important to know how well they are covered by exome sequencing. This is particularly important if exome sequencing should be used for diagnostic. Six patients with suspected familial breast cancer (BRCA) have been sequenced to test whether exome sequencing is a suitable method for BRCA diagnostics. For this purpose, the detection of heterozygous variants in nine candidate genes is required. Figure 3.33 shows the coverage information for different isoforms of three of these genes. The bottom row of this table shows that exon 4 of the gene *BRCA2* is not sufficiently covered for this sample. In fact this is the case for all samples sequenced with version 5 of the Agilent SureSelect Human All Exon enrichment kit. For a diagnostic purpose, this exon would have to be sequenced additionally using a different technology.

Figure 3.34 shows three variants that have been detected in breast cancer associated genes in three of the six familial breast cancer diagnostic samples. The bottom two variants are classified as disease causing mutations by HGMD. The top variant is not in HGMD. For variants like this, i.e. novel variants in disease associated genes, additional evidence is

Search results

Date: 20.06.2014
Number of samples: 4300

Individuals tested 6

Show All entries

n	idsnv	Chr	Gene symbol	Non syn/ Gene	Omim	Class	Function	pph2	pph2 prob	Sift	dbSNP 135	av Het	Clinical	HGMD	1000 genomes AF	EVS ea	EVS aa	Valid	Count	Disease	DNA id
1	6080077	chr13:32905168-32905168	BRCA2	171 (18)	600185	snp	splice				rs81002846	0			0			by-cluster	1	BRCA	BRCA0003
2	6083251	chr13:32907420-32907420	BRCA2	171 (18)	600185	indel	frameshift					0		CI972557	0				1	BRCA	BRCA0004
3	5040278	chr17:41258504-41258504	BRCA1	107 (6)	113705	snp	missense,5utr,intronic	benign probably damaging	0.026 0.999 0.994 1 0.985 0.996	0	rs28897672	0	1	CM940172	0			by-cluster	2	BRCA	BRCA0002

Figure 3.34.: Screenshot of three variants detected in breast cancer candidate genes.

required in order to show that the variant is disease causing rather than a benign polymorphism (see Chapter 3.3.5). Since the variant in this example is at a splice site, for instance RNA sequencing data that shows if the splice site is disrupted, could be generated.

3.3.4. Statistical Significance

In many samples sequenced in the course of this PhD project, e.g. the first 51 intellectual disability trios[104], known disease causing variants and genes have been ruled out before exome sequencing. Hence, in the majority of samples, no known disease causing variants or variants in disease associated genes were identified. To identify novel disease causing variants and disease associated genes, it is recommended to use formal calculations of statistical significance[80]. How to calculate statistical significance depends on the mode of inheritance of the investigated disease and is limited by the number of available samples.

623 patients with mitochondrial disorders (MD), a group of highly heterogenous conditions characterised by faulty oxidative phosphorylation, have been sequenced [25] [83] [34] [30] [38] [35] [37] [36] [57] [21]. Mitochondrial disorders are recessive. Thus, only homozygous and compound heterozygous variants are of interest in these samples. Fisher's exact test can be used to calculate the burden of each gene. For example mutations in the gene *ACAD9* have been shown to be disease causing in mitochondrial disorders[39]. Querying the in-house database, 15 of 623 samples with a mitochondrial disorder and 7 of 3,969 samples with other diseases have homozygous or compound heterozygous variants in this gene. Fisher's exact test results in a p-value of 7.387×10^{-9} for these values, which is below the genome wide significance threshold of 1.7×10^{-6} recommended by MacArthur et al.[80][10].

For *de novo* variants, significance can be calculated for each gene by calculating the probability that the observed *de novo* variants have not occurred by chance. A method to calculate this probability is *TADA*[40]. It takes the number of samples, the number and type (i.e. missense or loss-of-function) of *de novo* variants per gene and the gene specific mutation rate into account. The mutation rate is a function of the length of the gene and its nucleotide composition, e.g. a C next to a G increases mutation rate approximately by a factor of 10 for this position compared to the genome wide mutation rate. In addition to *de*

[10] 0.05 Bonferroni-corrected for 30,000 genes

novo mutations, TADA also allows the integration of variant frequencies from case-control studies. Applying TADA to *de novo* mutations from intellectual disability trios (see Chapter 3.3.2) results in 15 genes that reach genome wide significance. In addition to 8 genes with known association to ID, 7 genes (*ASXL3, SON, DDX3X, BCL11B, SETD5, TCF20, TRIP12*) with previously unknown association have been identified. Additional investigation of these genes is required, for example by sequencing them in larger cohorts of patients with ID or functional studies as described in the next chapter.

If the number of available cases and controls is high enough, statistical methods originally devoloped for GWAS can be used to detect significant variants. However, these methods have been designed for the analysis of common variants with low to modest effect in large numbers of samples. Their statistical power is often too low to analyse rare variants in relatively few samples. To overcome this problem, researchers have developed methods, such as *burden tests* or *variance-component tests*, to jointly analyze multiple rare variants that are located in the same genomic region, e.g. genes[64]. Which of these methods should be used depends on assumptions on the investigated disease and its genetic architecture, e.g. if it can be assumed that all tested variants in a gene have a negative effect or not. Tools such as *PLINK/SEQ*[11] or *EPACTS*[12] include a multitude of tests and allow to run them on standard multi sample VCF files. They also include methods for filtering of variants and samples to reduce the number of false positives, which is crucial for good results.

3.3.5. Additional Evidence

If no known variants or variants in known genes and also no novel variants or genes could be identified using statistical significance, e.g. because the number of samples was too small, other evidence can be used to identify disease causing variants. Also if a variant has been already identified as disease causing by the steps described in the last chapters, additional evidence is beneficial in order to prevent false positive variant classifications.

The methods to gather additional evidence are divided into two parts in this chapter: the first part describes *in silico* methods that try to predict variant effects and the second part describes experimental methods to get insight on the effect of variants.

Conservation and Prediction Scores

To assess the putative pathogenicity of variants without a known effect, several *in silico* tools that assign scores based on certain properties, have been developed. These scores can be divided into two classes:

1. **Conservation scores** - These scores are usually based on multiple sequence alignments of the human reference genome to groups of other genomes, e.g. mammals or vertebrates. Sites that are conserved in more distantly related genomes have higher conservation scores than sites that are rapidly evolving. It is assumed that sites are conserved because they are functionally important. Thus, mutations at such sites are more likely to be harmful than mutations in unconserved regions. Conservation

[11] `http://atgu.mgh.harvard.edu/plinkseq/` - Last accessed: 03.09.2014
[12] `http://genome.sph.umich.edu/wiki/EPACTS` - Last accessed: 03.09.2014

scores are usually available for the whole genome. Scores assessed in this chapter are *phyloP*[117], *GERP++*[22] and *SiPhy*[31][74].

2. **Prediction scores** - In addition to conservation between different species, prediction scores take additional information of a given variant into account, such as the position in a secondary structural element of the protein or the differences of properties of the wildtype and the mutant amino acid. Since these properties are primarily specific to protein coding regions, prediction scores are often only available for coding variants or, more specific, missense variants. Variants with known effect are used to train a prediction model to separate disease causing/deleterious and benign variants on the basis of the respective properties. Prediction scores assessed in this chapter are *Polyphen2*[3], *SIFT*[58], *MutationTaster*[113], *LRT*[14], *MutationAssessor*[106], *FATHMM*[116] and *CADD*[53]. In contrast to the other tools, CADD offers scores for SNVs as well as small indels for the whole genome. It uses 63 different annotations to calculate its score. These annotations include also GERP, phyloP, Polyphen2 and SIFT predictions. As a training set it uses nearly 15 million high-frequency human-derived alleles and the same amount of simulated variants. The key assumption here is that deleterious variants are depleted in the high-frequency alleles but not in the simulated variants.

The scores used in this chapter were derived from the *database of Human Non-synonymous SNVs and Their Functional Predictions (dbNSFP; v2.4)*[76][77]. This database provides precalculated scores from the tools mentioned above for all possible coding and splice site variants based on the GENCODE 9 annotation. In addition to scores and predictions from the single tools, dbNSFP also provides ranked scores for each tool. The ranked score of a variant is defined as the rank of this variant in the list of all variants ordered by the original score, divided by the total number of variants. Hence, the ranked scores range from 0 to 1 and a higher rank score means that a variant is predicted to be more deleterious. dbNSFP v2.4 also includes two aggregated scores (RadialSVM and LR) that combine the single scores and the maximum frequencies of variants in the 1000 genomes data (manuscript submitted).

The average score of variants per allele count can be used to assess the general properties of conservation and prediction scores (Figure 3.35). It can be assumed that deleterious variants are depleted at higher allele counts due to selective pressure. Hence, a good prediction tool should give higher average scores for rare variants than for frequent variants. Figure 3.35 shows this trend for most tools. Two scores are significantly different compared to the others: MutationTaster and LR

- MutationTaster is most likely different because it does not provide a continuous score but only classifies variants as damaging or neutral and gives a p-value for the reliability of each classification. A continuous score was derived by dbNSFP using the p-value for damaging variants and 1 - p-value for neutral variants, but this score reflects rather the reliability of each classification and not the deleteriousness.

- As mentioned above, LR is an aggregation of other scores and takes also allele frequency data from the 1000 genomes project into account. Thus, variants with higher allele counts receive lower scores.

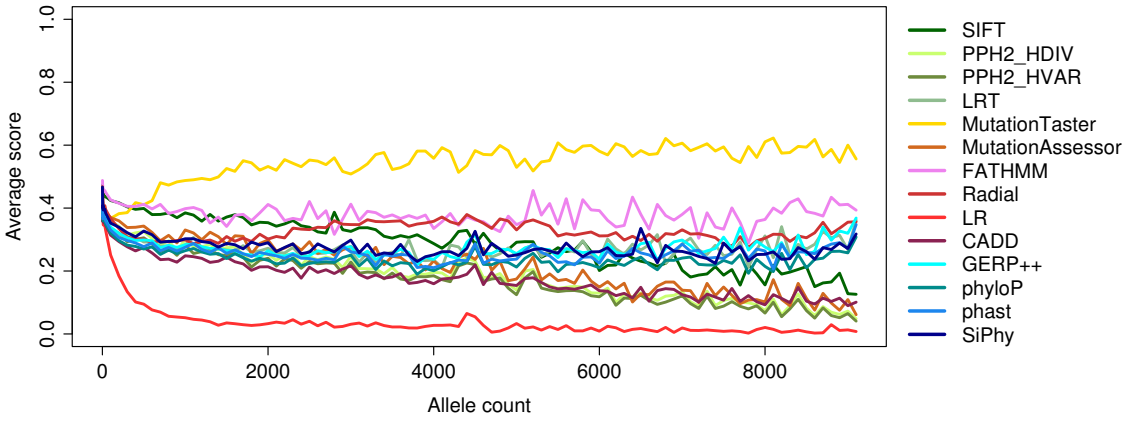
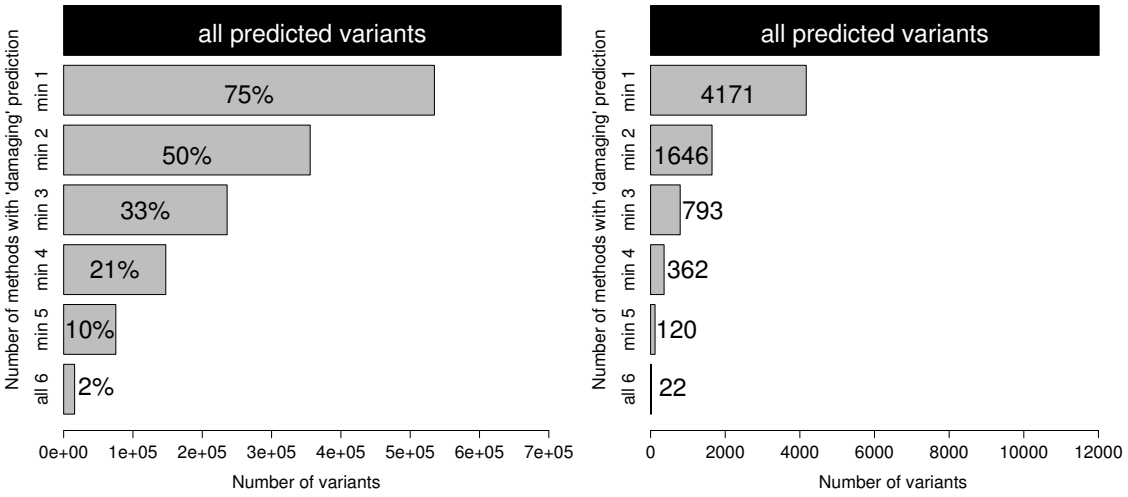


Figure 3.35.: Average predicted function by alternative allele count



(a) Total variants

(b) Per sample

Figure 3.36.: Overlap of six prediction scores (SIFT, PPH2_HVAR, LRT, MutationTaster, MutationAssessor and FATHMM)

Another way to assess the viability of prediction scores is to measure their overlap. Figure 3.36 shows the overlap of six scores: SIFT, PPH2_HVAR, LRT, MutationTaster, MutationAssessor and FATHMM. RadialSVM, LR and CADD were excluded, because they use at least some of the other scores as basis for their predictions. PPH2_HDIV was excluded, because it is the same score as PPH2_HVAR, but with a different training dataset. Thus, the overlap of PPH2_HVAR and PPH2_HDIV is large (89% according to dbNSFP[77]). Figure 3.36a shows that about 700,000 variants from the in-house database have a prediction from at least one of the six tested scores. 75% of these variants have been classified as damaging

by at least one tool, but only 2% of the variants are predicted to be damaging by all six tools. On a per sample level, on average 4,171 variants have been classified as damaging by at least one tool, but only 22 variants are predicted to be damaging by all six tools (Figure 3.36b). A naive way to combine single scores would be to use a majority of predictions, i.e. at least 4 of 6 scores must classify a variant as damaging. However, some scores correlate more than others[77], so a simple majority vote may be biased. More sophisticated aggregations of single scores, such as LR, RadialSVM or CADD, are preferable.

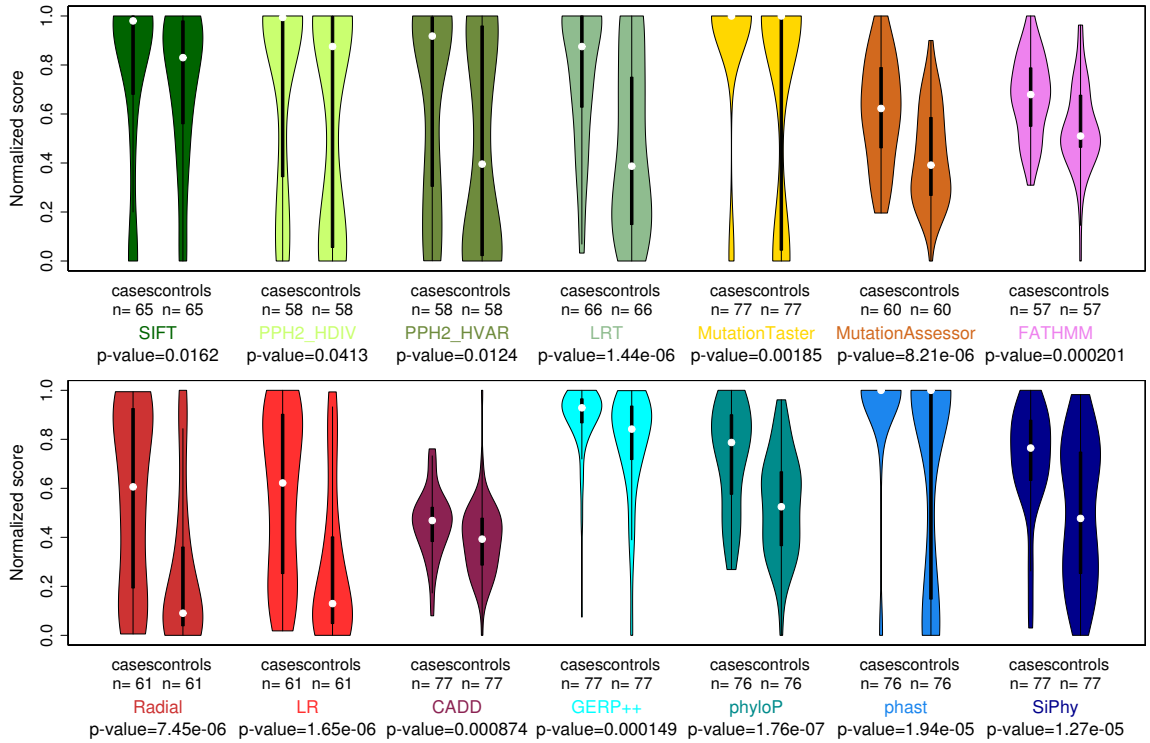


Figure 3.37.: Ranked prediction and conservation scores of 77 *de novo* variants from 71 samples in known disease genes vs. matched variants from controls. Wilcoxon rank-sum test (two sided) is used to test between groups.

To assess the ability of the single prediction and conservation scores to discriminate between putative disease causing and benign variants, one can compare the scores of known disease causing variants and benign variants. Here, a list of *de novo* variants from patients with intellectual disability (ID) (see Chapter 3.3.2) in genes associated with ID were compared to variants with corresponding annotations and allele frequencies from control samples (Figure 3.37). It is not certain that the variants in genes associated with ID are causal in all of these patients, but it can be assumed that they are more likely to be deleterious than random variants from control samples. Notably, all scores predict the variants in known ID genes to be more deleterious than the control variants. The most significant differences (Wilcoxon rank-sum test, two sided) between variant groups were achieved by the conservation scores phyloP, phast and SiPhy, the prediction score MutationAssessor and the aggregating scores Radial and LR.

In addition to providing raw scores, prediction tools also assign an effect, i.e. *damaging*

or *tolerated*, to each variant. Figure 3.38 shows these predictions for the 77 *de novo* variants from Figure 3.37 and for 93 *de novo* variants that have been identified in the same patients. This example represents a typical use case of prediction tools: one is looking at a list of variants from patients and tries to distinguish disease causing from benign variants. Especially SIFT and PPH are not able to distinguish between the two groups of variants. Other tools perform better at separating putative disease causing and benign variants, e.g. LR or Radial, but only half of the putative disease causing variants are predicted to be damaging.

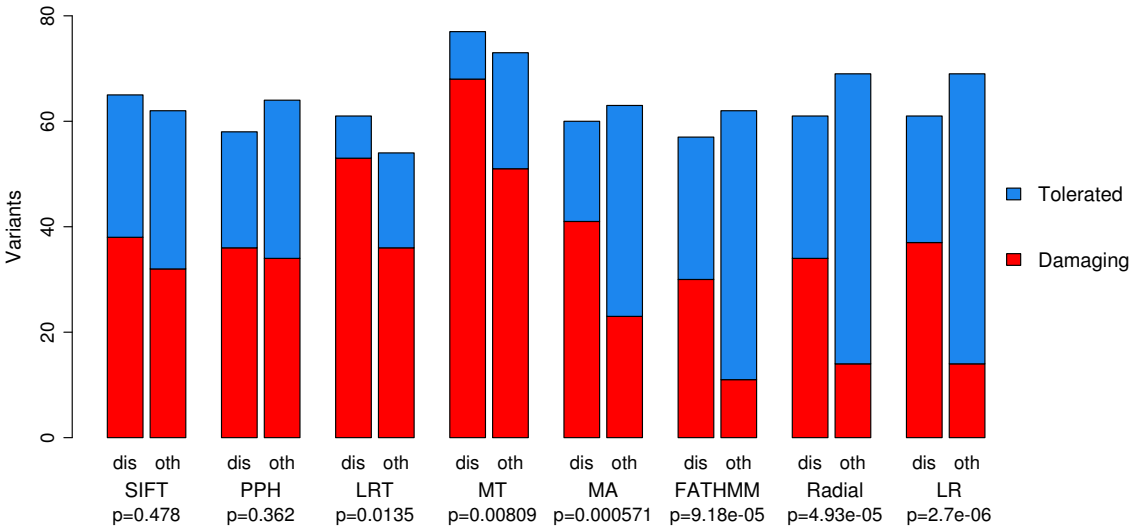


Figure 3.38.: Deleterious vs. tolerated predictions for *de novo* variants from 71 samples. 77 variants are in known disease genes (dis) and 93 variants are in other genes (oth). Fisher's exact test (two sided) is used to test for differences between groups.

Experimental Evidence

Experimental evidence can be further divided into two groups: (i) general evidence, often from high-throughput methods, for instance expression data of genes from tissues of interest. Such information can often be found in public databases and it is therefore possible to easily use it for filtering of many variants. (ii) evidence that is based on the variant itself, for instance if the transcript that carries the variant is expressed lower than the wild type transcript. These experiments have to be performed for every variant separately, which is time consuming and expensive. Hence, such experiments are only performed for very promising candidate variants and genes. However, such specific experiments are the most valuable sources of evidence.

General Experimental Evidence

A widely used resource are *Protein Protein Interactions (PPI)*. PPI can be used to filter for variants that are located in genes encoding for proteins which are known to interact with

disease associated proteins. There are several public databases, such as *MIPS*[97], that include PPIs either mined from literature or from high throughput experiments[55].

Another piece of evidence for the involvement of a certain gene and its variants in a disease is, if the gene is part of a *biological pathway* that is known to play a role in the disease or includes other disease associated genes. For instance the *Kyoto Encyclopedia of Genes and Genomes (KEGG)*[45][46] database includes information on biological pathways.

To assess the role of putative disease causing variants, it is important if the affected gene is expressed in the tissue of interest, e.g. disease causing genes for intellectual disability are usually expressed in the brain. The *GNF Gene Expression Atlas 2*[124] includes microarray expression data of 79 tissues. It can also be queried via the interface of the *UCSC Genome Browser*[49].

Similar to the expression of a gene in a tissue of interest, the presence of the encoded protein is important. Recently a draft map of the human proteome has been published[50]. It includes data of proteins from 17,294 genes in 30 tissues.

Information on the influence of variants in genes on the phenotype of other organisms can give evidence on disease associations. The *Mouse Genome Database*[9] contains information on the influence of heterozygous and homozygous gene knockouts on the phenotype of mice. For instance, according to this database, a homozygous knockout of the ortholog of the gene *STXBP1* that carries *de novo* variants in several patients with intellectual disability[104], leads to neuron apoptosis and degeneration in mice. This gives additional evidence that the *de novo* variants in this gene are disease causing. However, intellectual disability is a heterogenous disorder that is already linked to a large number of genes. Thus, also such seemingly convincing evidence should be critically examined.

Specific Experimental Evidence

The type of functional experiment that can be performed depends on the type of variant and the affected gene. For instance, a common 34 bp deletion in the promoter of *TXNL4A* could be identified by whole genome sequencing in six patients with *Burn-McKeown syndrome*, a rare condition with a characteristic combination of choanal atresia, sensorineural deafness, cardiac defects and craniofacial dysmorphisms. Consequently, reporter gene and *in vivo* assays were performed to assess the expression of *TXNL4A* which was indeed reduced[135].

In another project, somatic mutations in two *ATPases* were detected in nine aldosterone producing adenomas. These ATPases control sodium, potassium and calcium ion homeostasis. Functional *in vitro* studies of the mutants showed loss of pump activity and strongly reduced affinity for potassium[8].

In patients with mitochondrial disorders, oxygen consumption of the cells is decreased. So-called rescue experiments are performed to investigate if a variant is disease causing[35]. Wild type cDNA of the candidate gene is expressed in fibroblast cell lines of the patients. Then the oxygen consumption is measured and compared to the oxygen consumption of cells without the wild type cDNA. If the variant is causing the phenotype, the oxygen consumption of the cell lines with the expressed cDNA should be significantly increased compared to the original patient cell line, i.e. the phenotype is rescued.

3.3.6. Conclusions

This chapter showed strategies to identify a putative disease causing variant in the set of approximately 23,000 coding variants of a single exome sample.

The first step to reduce the number of variants is to filter the original set of variants for variants with an allele frequency below a cutoff that is appropriate for the investigated disease. However, even after analyzing more than 4,500 exomes, every new exome carries approximately 100 private variants. Thus, frequency alone is not a sufficient filter, at least until a significantly higher number of samples has been sequenced.

In a next step the remaining variants are filtered for the assumed mode of inheritance, e.g. dominant or recessive. The efficiency of this step depends on the mode of inheritance and the availability of families: for instance in recessive disorders, filtering for homozygous or compound heterozygous variants effecitvely reduces the number of candidate variants. For dominantly inherited disorders, families with multiple affected members are required for filtering by mode of inheritance.

Variants that fulfill both frequency and mode of inheritance requirements can then be looked up in public databases such as HGMD or ClinVar. Unfortunately, these databases contain a considerable amount of false positive entries. If no known disease causing variants could be identified, novel variants in known disease associated genes are investigated. These variants can be either disease causing or benign, additional evidence for causality is required.

Novel disease associated genes and variants can be identified by formal calculations of statistical significance. Possible methods depend on the mode of inheritance. However, in many projects and diseases a formal calculation of significance is impossible due to the insufficient number of available samples and also if statistical significant candidates can be identified, functional evidence might still be required.

Several *in silico* conservation and prediction scores are available and can be easily investigated. However, decisions on disease causality should not be based solely on these scores, because of the considerable number of wrong classifications. Also functional evidence, such as the expression of a gene in a tissue of interest, from public databases can be used as additional hint on causality, but should be critically examined.

If a convincing candidate variant could be identified, specific functional experiments are often required to clarify if the variant is responsible for the phenotype. The type of experiment depends on the type of variant, e.g. splice site disrupting or loss-of-function, and the type of protein the affected gene encodes, e.g. a membrane channel protein.

3.4. Variant Calling in RNA-Seq Data - RNA editing

RNA editing is a form of posttranscriptional modification. The most important and most common form of RNA editing is *Adenosine-to-Inosine (A-to-I)* editing [91]. Inosine is treated as guanine in splicing and translation, as well as in RNA sequencing.

In humans A-to-I editing is mediated by *adenosine deaminase acting on RNA (ADAR1-ADAR3)* enzymes. ADAR3 has not shown enzymatic activity in experiments, although it shares the functional domains with ADAR1 and ADAR2 and is well conserved[11]. The editing takes place only in double stranded RNA (dsRNA). This might also be the reason

why the vast majority of RNA editing sites are located within repetitive regions, especially *Alus*, in non-coding parts of the RNA, because for repetitive sequences it is more likely that their counterpart lies in close proximity and so a double stranded structure can be formed.

RNA editing is an important mechanism to maintain the physiological function of a cell, especially in the brain[69]. A-to-I editing has been shown to play a role in many diseases, but especially in neurological and psychiatric disorders[79][119]. For instance the glutamate receptor subunit *GRIA2* exon 11 Q/R site is usually edited in the human brain and mediates the Ca^{2+} permeability of glutamate receptors [29]. Down-regulation of editing at this site is believed to correspond to several neuronal diseases such as Amyotrophic Lateral Sclerosis or Epilepsy [79]. A-to-I editing is regulated as a response to various factors, e.g. stages in the cell cycle, and is tissue specific. For example the site in GRIA2 mentioned above is edited at a rate of 100% in kidney tissue and only 33% in adrenal tissue[70].

In principle, identifying RNA editing sites can be done by calling variants from RNA-Seq data that can not be found in corresponding DNA data. However, variant calling in RNA-Seq data is more difficult and error prone than variant calling in DNA data, mainly due to errors from sequencing library preparation and difficulties in accurate read alignment[69] (see also Chapter 1.3.1). Over the last couple of years, several strategies to handle these problems have been published[99][103][101]. These methods mainly focus on stringent filtering of putative variants.

Here, RNA editing is assessed in the context of a large RNA sequencing project of the GEUVADIS consortium[63][125]. In this project mRNA and microRNA from lymphoblastoid cell lines of 462 individuals from the 1000 Genomes Project[128] have been sequenced in seven laboratories, including ours. The main purpose was to assess transcriptome variation within and between populations in the light of genomic variations. In terms of RNA editing, two different aspects have been assessed:

1. Since the proportion of edited bases that are present in RNA-Seq reads at putative RNA editing sites can be counted, RNA editing sites can also be viewed as quantitative traits. To assess if these quantitative traits are regulated by genomic loci, association tests were performed (Chapter 3.4.1).

2. A-to-I editing can create (AA to AI = AG) or destroy (AG to IG = GG) canonical splice sites. Hence, the influence of RNA editing on differential splicing was investigated (Chapter 3.4.2).

Similar to variant calling in exome data (Chapter 3.2), accurate identification of RNA editing sites depends largely on the quality of the raw data and read alignment. There has been a large emphasis on quality control[125] and read mapping[82] in this project. Variant calling has then been performed for all samples together with SAMtools mpileup[68] at 42,039 known editing sites from the DARNED database[52] for the association study (Chapter 3.4.1) and genome wide with GATK UnifiedGenotyper[85][24] for assessment of effects on splice sites (Chapter 3.4.2). To reduce the number of false positive RNA editing events a set of very stringent filters have been applied:

1. a minimum median coverage of 10 at all called sites has been required

2. at least 10 samples had to have a non-reference “genotype” at each site

3. all variants called with SAMtools had to pass the SAMtools varFilter script, whereas the variants called with GATK had to survive variant quality score recalibration.

4. the variant quality at all sites had to be above 100

5. Furthermore, to ensure that the observed variants are true RNA editing events and not due to genetic variants, two things were required:
   a) there should not be a corresponding variant in the 1000 Genomes Phase 1 data set
   b) all variants had to be located within the set of accessible regions defined by the 1000 Genomes project to ensure that a variant would be present in the genetic variant data if it was present at the DNA level

Of the 462 samples in this project, 421 have been part of 1000 Genomes Phase 1. Since whole genome sequencing has not been performed for the remaining 41 samples at this point, these samples have been excluded from further analysis.

3.4.1. RNA Editing as a Quantitative Trait - editQTLs

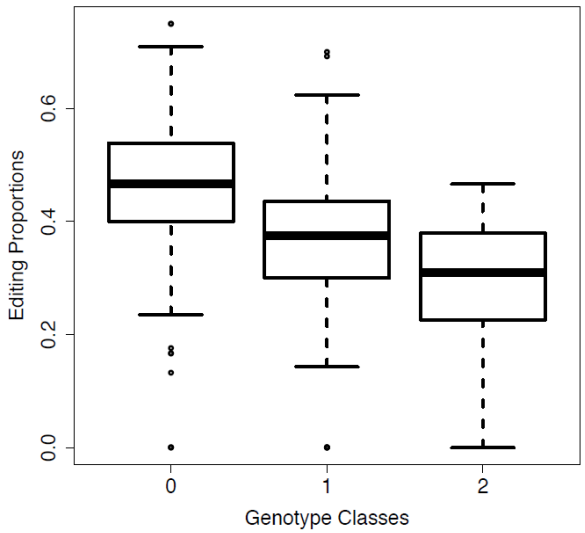


Figure 3.39.: Example for an RNA editing site that is associated with a genetic variant (editQTL). On the x-axis the genotype of the genetic variant is shown. Genotype Class 0 stands for homozygous reference, 1 for heterozygous and 2 for homozygous alternative. The y-axis shows the proportion of edited bases.

Altogether, SAMtools mpileup called non-reference variants (i.e. at least one sample had a non-reference "genotype") at 24,680 of the 42,039 sites from DARNED. After filtering, only 100 edited sites remained. Eight of these sites showed genetic associations to the proportion of editing (FDR5%). An example is shown in Figures 3.39 and 3.40. The associated genetic loci (editQTLs) are close to the RNA editing site (median 304 bp).

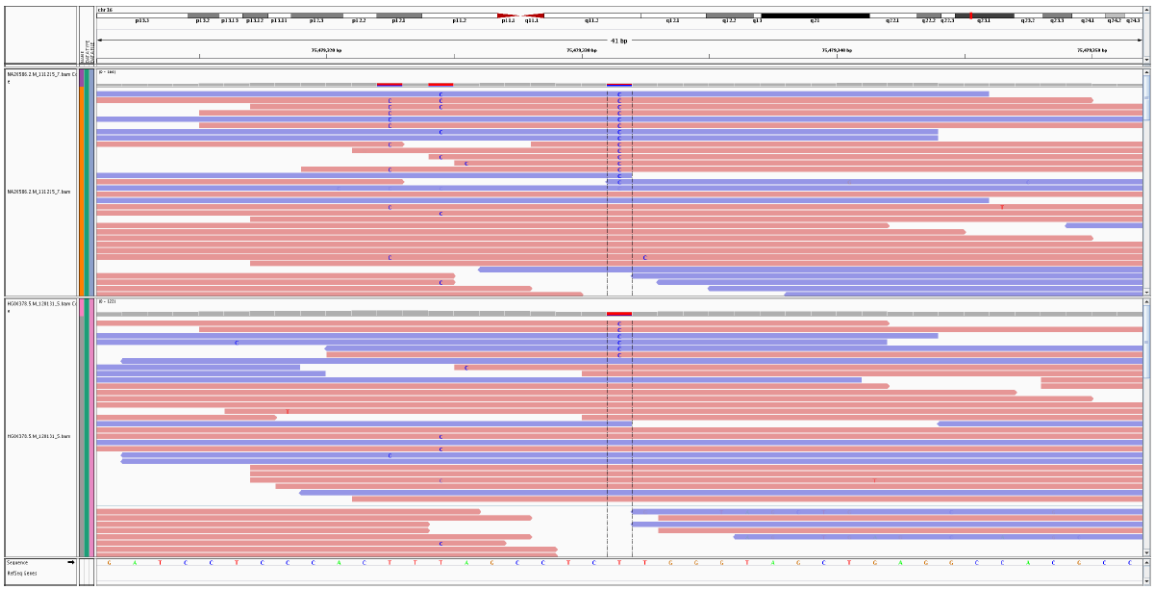
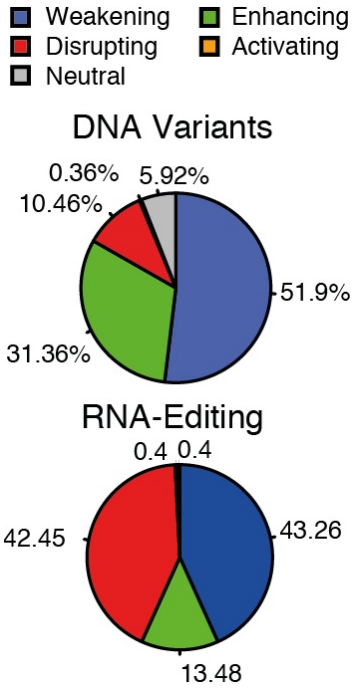


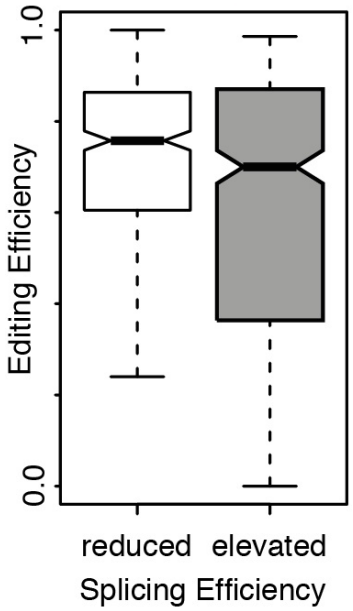
Figure 3.40.: The example RNA editing site from Figure 3.39 in two different samples. The read color is showing the read orientation and the edited site is in the middle. Reads are sorted by allele.

3.4.2. RNA Editing of Splice Sites

After filtering approximately 500 splice sites (0.1% of all splice sites) showed evidence for RNA editing. About 85% of RNA editing sites decrease or disrupt splicing efficiency, compared to about 60% of genomic variants (Figure 3.41a). Interestingly, RNA editing efficiency in splicing disrupting variants is anti-correlated with splicing efficiency (Figure 3.41b). This suggests that RNA editing takes place before splicing.



(a) Effect of RNA editing sites compared to DNA variants on splice sites.



(b) Proportion of edited reads at RNA editing sites vs. splicing efficiency. RNA editing efficiency in splicing disrupting variants is anti-correlated with splicing efficiency.

Figure 3.41.: Effect of RNA editing on splicing

Part IV.

Discussion

4. Discussion

4.1. Data Quality

The last two parts of this thesis showed the development of methods and guidelines for the detection of disease causing variants and their usage in diverse genetic disorders. It could be shown that the quality of NGS data is crucial for the detection of variants with sufficient quality. The read depth is especially important. The data presented here led to the decision to generate between 8 and 12 Gb of data for each sequenced exome, leading to an average of 95.6% of targeted bases covered more than 20 times. This amount of sequence is higher than in other published exome studies. For example, de Ligt *et al.*[23] sequenced 100 intellectual disability trios with an average of 5.4 Gb of sequence per sample, leading to an average of 80% of targeted bases covered more than 20 times. They identified 79 *de novo* mutations. In a similar study conducted during this PhD project[104], 51 intellectual disability trios have been sequenced with an average of 10.2 Gb using the same version of exome enrichment kit. 90% of the targeted bases were covered more than 20 times. 87 *de novo* mutations could be identified and confirmed in this dataset. At least a part of the substantially higher amount of detected variants might be due to the higher coverage. However, with decreasing sequencing costs, the amount of sequence generated per exome sample becomes less important compared to the costs for the exome enrichment kit. At some point whole genome sequencing will be even more cost effective than exome sequencing, which will be beneficial in terms of coverage but raises other issues (see Chapter 5.2).

4.2. Benchmarks for Variant Calling

Benchmarks of variant callers showed that all tested callers (SAMtools mpileup, GATK UnifiedGenotyper and GATK HaplotypeCaller) showed comparable results in terms of sensitivity and specificity of SNVs. This has also been shown by Liu *et al.*[75]. However, they did not investigate the performance of GATK HaplotypeCaller and recommend to use GATK UnifiedGenotyper due to its better performance in multi sample calling. O'Rawe *et al.*[93] did investigate also GATK HaplotypeCaller, but they focused on the concordance between different callers rather than on the performance of individual callers. Discrepancies between callers are mainly due to low quality variants in problematic regions that are called differently between the assessed callers. O'Rawe *et al.* recommend using multiple callers to maximize sensitivity. In the benchmarks discussed in Chapter 3.2, GATK HaplotypeCaller performed better in terms of indel calling, probably due to the local *de novo* assembly it performs to call variants. Accordingly, it is recommended to use GATK HaplotypeCaller for future analysis, also because it is still under active development. There was no evidence that multi sample calling improves the sensitivity of calling singletons.

However, it performs better at calling more common variants in low coverage regions[75]. Multi sample calling also has the advantage that it outputs a genotype for every called sample at each site where at least one sample has a variant, if the number of reads at the position is sufficient. At singleton positions in single sample calls, one does not know if there is no variant in the other samples or if the number of reads is insufficient. Thus, multi sample calling should be preferred over single sample calling, if possible. However, until recently, multi sample calling was too complex for large numbers of samples due to the high number of large BAM files that had to be considered. One had to partition the set of samples into subsets and merge the results, or use so-called reduced BAM files. In these files, multiple reads were collapsed at positions where they did not differ from the reference genome. This reduced computational costs of multi sample calling, but was still time consuming. Additionally, for every new sample the whole process of multi sample calling had to be repeated, an issue referred to as "N+1 problem". Recently, GATK changed its recommended multi sample calling procedure: it is now possible to produce so-called *gVCF* files using GATK HaplotypeCaller. These files contain genotype likelihoods also at homozygous reference sites. To save disk space, adjacent reference sites with similar likelihoods can be collapsed into blocks. It is recommended to produce gVCF files separately for each sample and then call genotypes over all gVCF files using the GATK GenotypeGVCFs module. This tool is fast enough to be run every time new samples have to be included.

Systematic benchmarking of SVs and CNVs proved to be difficult. Large deletions, i.e. $\geq$100 kbp, in 11 samples that were previously detected using array-CGH could also be found in exome data using ExomeDepth (Figure 2.8 shows one of these deletions). However, intermediate sized variants, i.e. $\geq$ 50 bp and $\leq$ 100 kbp, can not be assessed systematically, due to the lack of gold standard datasets for variants of this size. For instance, ExomeDepth[102] was tested on a dataset by Conrad *et al.* [19] who calculated that they were able to detect only approximately 40% of all CNVs. Pindel[137] was tested on a dataset containing only indels with a size of 1 to 16 bp. A gold standard dataset of intermediate sized variants will be essential for benchmarking of SV and CNV discovering algorithms, especially with increasing numbers of whole genome datasets (see Chapter 5.2) which will improve the ability of calling such variants.

4.3. Identifying Disease Causing Variants

The guidelines to identify disease causing variants discussed in Chapter 3.3 try to cover as many cases as possible. However, if this filtering strategy can be applied successfully depends to a great extent on the study design. The number and type of sequenced samples must be appropriate for the mode of inheritance, incidence and genetic architecture of the investigated disease.

For instance, for familial cases it can be sufficient to sequence only two distantly related, affected family members to identify a disease causing variant (e.g. in familial Parkinson's Disease[139]).

Another example are sporadic cases of diseases caused by *de novo* mutations. Filtering variants for frequency and a *de novo* mode of inheritance effectively reduces the number of putative disease causing variants to around 1 variant per sample (see Chapter 3.3.2).

If the identified *de novo* variants of a sample include a known disease causing variant, or a novel variant in a disease associated gene with additional evidence, sequencing of a single parent-child trio is sufficient. This is the desired case in a diagnostic setting (see Chapter 5.3). If no known variant could be identified, a sufficient number of trios has to be sequenced in order to identify statistically significant disease associated genes carrying multiple *de novo* mutations. Which number of trios is sufficient depends on the genetic architecture of a disease. For instance, intellectual disability is a heterogenous disease that can be caused by mutations in many different genes. Thus, the chance that two unrelated samples carry disease causing mutations in the same gene is low and therefore a higher number of samples has to be sequenced in order to identify genes carrying multiple *de novo* mutations. The number of *de novo* mutations per gene that have to be found in order to reach genome wide significance depends on the number of investigated samples, the mutation rate of the gene and also the type of mutations. MacArthur *et al.*[80] describe an example from sequencing 945 autistic children and their parents in four different studies[110][96][89][43]: four independent *de novo* mutations have been found in the gene *TTN*. This is the largest protein coding gene in the human genome. By considering its size, mutation rate and coverage, two *de novo* mutations were expected by chance in these 945 trios, which is not significantly different than the four identified mutation. Thus, the four *de novo* mutations in *TTN* are not significant for this dataset. This example illustrates the need for well-defined statistical methods for the detection of significant disease causing variants and disease associated genes as described in Chapter 3.3.4. However, in the future more complete and robust models that are easier to use will be required.

4.4. Variant Calling in RNA-Seq Data

In 2011, Li *et al.*[72] reported approximately 10,000 differences between DNA and RNA of 27 samples. In addition to the previously known A-to-I (see Chapter 3.4) and C-to-U editing, they also reported all other possible substitutions, for which no mechanism was known. Several groups investigated their data and found that the vast majority of these unknown differences were due to sequencing artifacts, alignment errors and duplicated genomic regions[54][73][100][101]. Li *et al.* replied to these comments by reconsidering and improving some of their analysis[71], but the community remained skeptical.

This example shows the difficulties in the detection of RNA editing sites. The analysis described in Chapter 3.4 has been conducted with these pitfalls in mind. There was a large emphasis on the quality of the raw data and the alignment. Additionally, very stringent filters have been applied to the raw variant calls, because variant calling in RNA-Seq data is error-prone, especially at exon boundaries. Recently, methods for variant calling in RNA-Seq data were implemented in GATK[1]. GATK HaplotypeCaller now provides a RNA-Seq mode that improves the handling of split alignments. This feature is still under active development. Also Variant Score Recalibration is not working at the time of writing. However, development of specialized variant callers for RNA-Seq data will eventually improve the ability to call variants from RNA-Seq data and the detection of RNA editing.

[1] `http://www.broadinstitute.org/gatk/guide/best-practices?bpm=RNAseq` - Last accessed: 29.07.2014

Part V.

Outlook

5. Outlook

5.1. New Developments of Sequencing Technology

The first NGS device from Illumina, the Genome Analyzer I, used in early publications[6] produced approximately 1 Gb per run. Current HiSeq 2500 instruments are capable of producing 1,000 times as much sequence (1 Tb) for about the same price. Illumina recently announced the HiSeq X Ten, consisting of ten single instruments that are not sold separately[1]. Every single instrument can produce between 1.6 and 1.8 Tb per run, dropping the price for sequencing a human genome at 30x coverage below $1,000. Illumina sells these instruments for "population scale sequencing projects" and only supports the sequencing of whole genome samples, i.e. no targeted sequencing is supported.

Also other companies improved the output and cost efficiency of their instruments. Life Technologies improved their Ion Torrent Sequencers to produce up to 60 Gb of sequence in four hours. The Ion Torrent instruments also perform SBS, but they work differently than the instruments described in Chapter 1.2.1[109]: DNA fragments are captured in separate microwells and unmodified nucleotides are added, one type at a time. If the currently added nucleotide is appropriate, it is incorporated into the nucleic acid chain, which releases a hydrogen ion. This leads to subtle changes of the pH that can be measured. An advantage of this technology over technologies using fluorescent dyes, such as Illumina SBS, is, that it uses relatively cheap semiconductor chips instead of expensive optics. Thus, the costs of an instrument are relatively low. However, the output and sequencing costs are insufficient for high throughput sequencing.

5.1.1. Third Generation Sequencing

Third Generation Sequencing (TGS) methods are techniques that allow the sequencing of single molecules in real time, i.e. no stationary amplification is required and sequencing is continuous and not divided into cycles[112]. Pacific Biosciences introduced their *Single Molecule, Real-Time (SMRT)* technology in 2009[27]. At the time of writing, they offer the only available TGS instrument. Sequencing is performed on a so-called *SMRT cell* containing holes with tens of nanometers in diameter, the *Zero-Mode Waveguides (ZMWs)*(Figure 5.1). DNA polymerases are immobilized at the bottom of these ZMWs. Differently labeled dNTPs are floating above the ZMWs and diffuse into them. The polymerase detects the appropriate dNTP and incorporates it into the DNA strand. During incorporation, the fluorescent label is cleaved off. To measure the incorporation of each dNTP, the SMRT cell is illuminated from the bottom by laser light with a wavelength of approximately 600 nm. Since the wavelength is an order of magnitude longer than the diameter of the ZMWs, only

[1] `http://res.illumina.com/documents/products/datasheets/datasheet-hiseq-x-ten.pdf` - Last Accessed: 30.07.2014

the bottom 30 nm of each ZMW are illuminated. Only signals from the labels of dNTPs that get incorporated into the DNA strand stay long enough in this illuminated area to be detected. Current SMRT cells consist of 150,000 ZMWs. However, both the DNA polymerases and the single stranded templates are delivered to the ZMWs through diffusion. Thus, only approximately 50,000 ZMWs are actually used in each run. The current average read length is approximately 8,500 bp with a maximum read length of >30,000 bp, resulting in an output of up to 500 Mb per three hours run. Due to the low throughput and relatively high sequencing costs, SMRT sequencing is currently mainly used in *de novo* sequencing in microbiology, because the sequenced genomes are relatively small, but *de novo* assembly benefits from the longer reads. SMRT sequencing has an error rate of 5%. Most errors are indels due to missed incorporations of dNTPs or incorporations that take longer than expected. However, these errors occur randomly and can therefore be corrected if the read depth is sufficient.

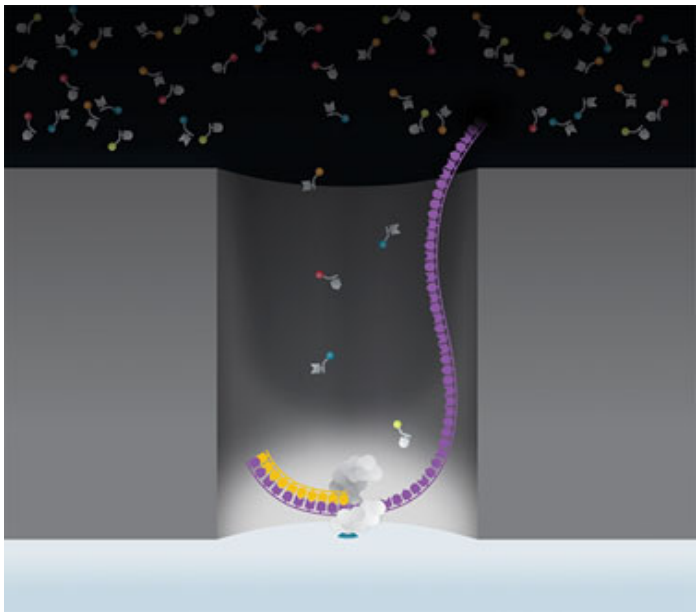


Figure 5.1.: Scheme of a single Zero Mode Waveguide (ZMW) of a Pacific Biosciences SMRT cell. A DNA polymerase is immobilized at the bottom and can be observed in real time while incorporating labeled dNTPs into the DNA strand. Picture adapted from Pacific Biosciences.

Oxford Nanopore is currently developing a TGS technology using so-called *nanopores*, i.e. holes in a membrane with a diameter only slightly larger than a DNA strand, to sequence single DNA molecules[112][18][122]. By applying a current across a synthetic polymer membrane, the current flows through an engineered nanopore protein that pierces the membrane (Figure 5.2). A single stranded DNA (ssDNA) molecule is sequenced while passing through the nanopore by measuring the current flowing through the nanopore. The four different bases can be distinguished by characteristic disruptions of the current.

Oxford Nanopore shipped the first batch of their MinION instrument, a sequencing instrument in the size of a USB flash drive with 512 nanopores, to test users in early 2014. At the time of writing no official dataset has been released, so the performance of these

devices in terms of read length, yield and error rate remains to be seen.

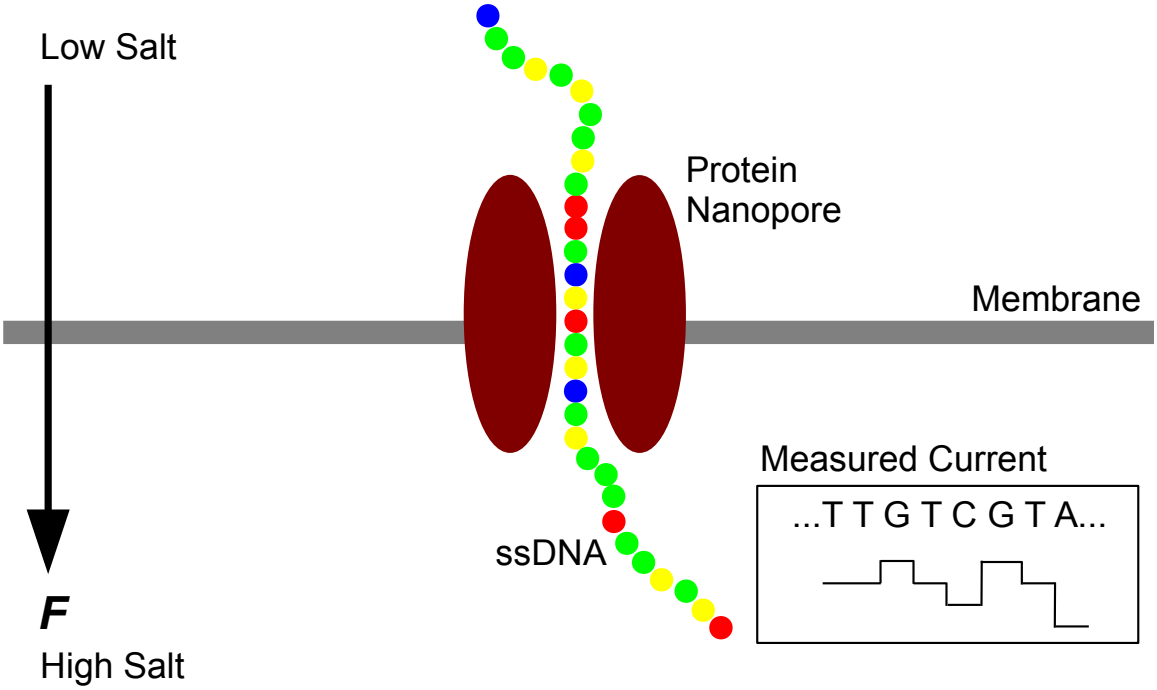


Figure 5.2.: Scheme of the nanopore sequencing technology currently developed by Oxford Nanopore. A voltage applied across a synthetic polymer membrane leads to a current flowing through the nanopore protein that pierces the membrane. A single stranded DNA (ssDNA) molecule is sequenced while passing through the nanopore by measuring the current. The four different bases can be distinguished by characteristic disruptions of the current (see graph in the box).

With the development of TGS technologies new requirements on data analysis arise. The biggest advantage of these technologies is the significantly longer read length compared to NGS. Longer reads could make complete *de novo* assembly from larger mammalian genomes possible as well. However, increased error rates, especially if these errors are indels, will require algorithms with more sophisticated error models.

5.2. Implications of Whole Genome Sequencing

The advances of NGS and TGS technologies discussed in the last chapter will further reduce sequencing costs. With lower sequencing costs, most researchers and clinicians (see Chapter 5.3) would perform whole genome sequencing (WGS) rather than exome sequencing or other forms of targeted resequencing. Compared to exome sequencing, whole genome sequencing has several advantages but also introduces a number of novel problems.

As discussed in Chapter 3.1.1 the coverage distribution of WGS is much more even than the coverage distribution of exome sequencing, especially if the libraries have been pre-

pared with a PCR free kit. This leads to an almost complete coverage of coding regions, i.e. >99% of RefSeq coding regions are covered more than 20 times for samples with an average read depth of >30x. The even coverage also simplifies the detection of SVs and CNVs. Larger CNVs can even be seen in raw data (Figure 5.3) and can be detected by simple sliding window read depth approaches without the need for control samples. An example for a program utilizing such an approach is CNVnator[1]. Also insert size and split read approaches, as implemented for example by breakdancer and Pindel, benefit from WGS. To detect a SV using these approaches, both breakpoints of the SV must be covered. For exome sequencing, this means that both breakpoints must be located in targeted exons which is often not the case.

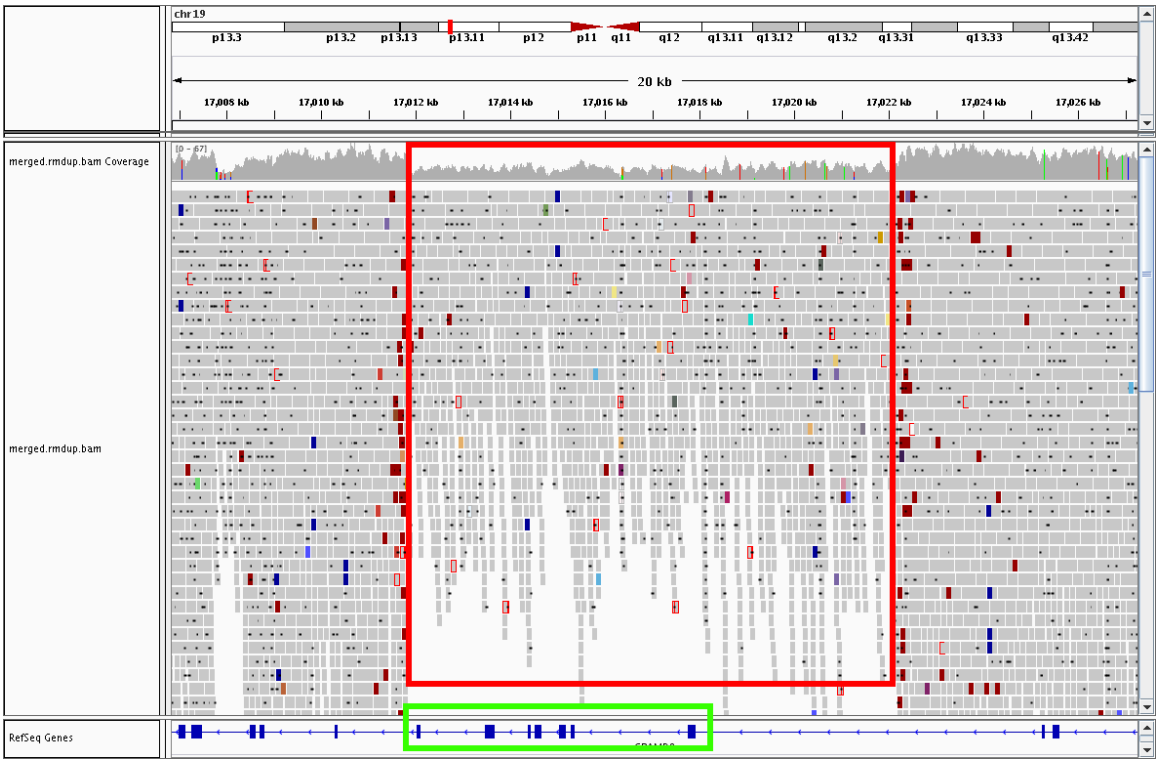


Figure 5.3.: Example of a 10 kbp deletion (red rectangle) in a WGS sample. Due to the even coverage distribution of WGS samples prepared with PCR free kits, SVs/CNVs can even be seen in the raw data. This deletion has been also detected in exome data from the same sample, but it was significantly smaller (green rectangle) because the actual break points are located in introns.

One problem of the rising number of WGS experiments is the amount of raw data created. Using Illumina SBS, every sequenced gigabase requires approximately one gigabyte of disk space to store the compressed raw data. Thus, sequencing a human genome at an average coverage of 30x requires approximately 100 Gb of disk space just for the raw data. Depending on the used analysis pipeline, two to three times as much disk space is required, at least while the pipeline is running.

The pipeline described in this thesis stores approximately 70,000 (23,000 coding + 47,000

surrounding) variants per exome sample in a relational database. This database can then be queried in order to identify putative disease causing variants (see Chapter 3.3). However, doing the same for WGS samples is problematic, because every WGS sample has approximately 3,000,000 variants. Other technologies, such as distributed or in-memory databases, might be required in order to enable querying these large datasets in reasonable time.

Chapter 3.3 showed that the interpretation of variants is often difficult, even if only the approximately 11,000 non-synonymous coding variants of exome samples are considered. Interpretation of the 2,980,000 additional non-coding variants presented by a WGS sample is much more difficult. A first step to make sense of these variants is to focus on other functional elements in the genome. Functional elements can be identified using three approaches[47]:

1. **Genetic approach** - Genetic approaches identify functional elements by investigating changes of a phenotype due to sequence variants. These variants can be identified in individuals with a certain phenotype or disease as described in Chapter 3.3 or they can be artificially created, e.g. in gene knock-out experiments in mice. Genetic approaches have been used primarily to assign functions to protein coding genes. However, they can be also used to identify the function of other genomic regions, such as regulatory elements. Genetic approaches are often labor-intensive, thus their throughput is limited.

2. **Evolutionary approach** - Evolutionary approaches use the sequence conservation between different species as an indication for functional elements. It is assumed that if a site is functionally important, it is under strong purifying selection and therefore higher conserved than sites that are less functionally important. However, evolutionary approaches provide no information on the type of function that is carried out by conserved sites. Additionally, which functional elements are conserved depends on the investigated species. For instance, if human sequences are compared to other mammalian sequences, only functional elements important in mammals are conserved and functional elements specific to humans are not conserved. Conservation scores are investigated in Chapter 3.3.5.

3. **Biochemical approach** - Biochemical approaches identify putative functional elements by their biochemical properties, such as DNAse accessability or the binding of specific transcription factors. The ENCODE project assessed genome wide biochemical properties in 147 different cell types[7]. This data is publicly available and can, for instance, be used to annotate variants in elements such as promoters and enhancers. However, ENCODE assigns at least one biochemical function in at least one cell type to 80.4% of the human genome. Thus, the data must be carefully filtered and interpreted in order to produce useful annotations.

In addition to the annotation of non-coding areas of the human genome, public databases providing information on already known variants are required for the analysis of WGS data[44]. Specifically, two types of information are required: (i) Frequency information from WGS data for all observed variants in order to filter WGS datasets for variants with frequencies appropriate for the investigated phenotype. This kind of information will help

to drastically reduce the number of variants to investigate. Currently, the 1000 Genome Project dataset provides frequency information for whole genome variants, but data from many more samples will be required in the future. This will require a public database where researchers and clinicians can submit their data in an anonymized manner in order to satisfy legal and ethical requirements on data security. (ii) In detail information on the influence, i.e. causal or benign, of variants on a certain phenotype. As described in Chapter 3.3.3, HGMD and ClinVar provide this kind of information. However, especially ClinVar is still in an early stage and requires more data to be a useful resource for researchers and clinicians. Information from more samples will also help to reduce the number of false positive entries.

Also statistical methods, as described in Chapter 3.3.4, can be used to identify causal variants from WGS data. Most current methods for the analysis of rare variants rely on the definition of genomic regions that should be tested as a unit. In addition to genes, other functional elements can be used for this purpose or the genome can be divided into artificial windows. However, novel methods might be required to define units for testing.

5.3. Next-Generation Sequencing in Clinical Diagnosis

Due to the decreasing costs, NGS is increasingly being used as a diagnostic tool in human genetics [2]. NGS is more cost effective than traditional Sanger sequencing, especially if many genes must be investigated or if the genes of interest are large. Nowadays NGS can also be used for the rapid diagnosis of newborns with a likely genetic disorder. However, using NGS in clinical diagnosis introduces some problems.

Three types of NGS experiments can be used in diagnostics[105]: targeted sequencing of gene panels, exome sequencing and WGS. All three approaches have advantages and disadvantages. Sequencing of gene panels has the advantage that capturing can be designed to ensure very high coverage of the genes of interest. Also the amount of required sequence is relatively low and the amount of data that has to be analyzed is small. The biggest disadvantage of this approach is that only a set of predefined genes is analyzed. Thus, the data can not be used to identify putative novel disease associated genes and if novel disease associated genes are identified in the future, new panels must be designed to include them. Also the design and production of such panels is relatively expensive and only economically reasonable if large numbers of samples are analyzed. Exome sequencing overcomes these shortcomings by attempting to capture all coding genes. However, as has been shown in this thesis, coverage of targeted regions is incomplete in exome sequencing and some genes or exons of interest might not be targeted at all. These limitations are overcome by whole genome sequencing, but this technique also has caveats, as discussed in the last chapter. The technical limitations of NGS, such as the inability of spanning larger repeats due to limited read length, apply to all three experiment types.

An example for NGS in clinical diagnosis has been shown in Chapter 3.3.3: exome sequencing has been performed in six samples with suspected familial breast cancer in order to identify variants in a defined set of known disease associated genes. There are some recommendations on what to report to the patients or physicians in charge in such cases[107][2]. Variants in disease associated genes with a known pathogenic effect are often considered as disease causing and reported. Variants in these genes without a known

effect can be reported as candidate disease causing variants but their causality is uncertain and might require additional research (Chapter 3.3.5). If no variant has been identified in disease associated genes, limitations of the approach, e.g. insufficiently covered regions, should be reported.

A question that is often discussed in the context of NGS in clinical diagnosis is if and which *incidental or secondary findings* should be reported[2][33]. Incidental findings are putative disease causing variants or strong risk factors for diseases other than the initially investigated disease. Often patients must sign an informed consent form in which they state if they want to be informed of incidental findings. However, this depends on the respective laboratory that offers the NGS diagnosis. Some institutions specifically only analyze genes that are associated with the investigated disease, i.e. they do not look for secondary findings. Others report only variants for diseases where preventative measures and/or treatments are available, as suggested by the American College of Medical Genetics and Genomics (ACMG), which recently published a list of 56 such genes[33].

As has been shown in Chapter 3.3.3, and by others[136], NGS can be used in diagnostics. However, guidelines for data quality and analysis will be required in the future to allow NGS to become a standard clinical test. Also specifically trained clinicians and genetic counselors are required to communicate results as well as benefits and risks of NGS to the patients.

Bibliography

[1] Alexej Abyzov, Alexander E Urban, Michael Snyder, and Mark Gerstein. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research*, 21(6):974–84, June 2011.

[2] ACMG Board of Directors. Points to consider in the clinical application of genomic sequencing. *Genetics in medicine : official journal of the American College of Medical Genetics*, 14(8):759–61, August 2012.

[3] Ivan a Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–9, April 2010.

[4] Michael J. Bamshad, Sarah B. Ng, Abigail W. Bigham, Holly K. Tabor, Mary J. Emond, Deborah a. Nickerson, and Jay Shendure. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745–755, September 2011.

[5] Callum J Bell, Darrell L Dinwiddie, Neil a Miller, Shannon L Hateley, Elena E Ganusova, Joann Mudge, Ray J Langley, Lu Zhang, Clarence C Lee, Faye D Schilkey, Vrunda Sheth, Jimmy E Woodward, Heather E Peckham, Gary P Schroth, Ryan W Kim, and Stephen F Kingsmore. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Science translational medicine*, 3(65):65ra4, January 2011.

[6] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, Jonathan M Boutell, Jason Bryant, Richard J Carter, R Keira Cheetham, Anthony J Cox, Darren J Ellis, Michael R Flatbush, Niall a Gormley, Sean J Humphray, Leslie J Irving, Mirian S Karbelashvili, Scott M Kirk, Heng Li, Xiaohai Liu, Klaus S Maisinger, Lisa J Murray, Bojan Obradovic, Tobias Ost, Michael L Parkinson, Mark R Pratt, Isabelle M J Rasolonjatovo, Mark T Reed, Roberto Rigatti, Chiara Rodighiero, Mark T Ross, Andrea Sabot, Subramanian V Sankar, Aylwyn Scally, Gary P Schroth, Mark E Smith, Vincent P Smith, Anastassia Spiridou, Peta E Torrance, Svilen S Tzonev, Eric H Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D Alam, Carole Anastasi, Ify C Aniebo, David M D Bailey, Iain R Bancarz, Saibal Banerjee, Selena G Barbour, Primo a Baybayan, Vincent a Benoit, Kevin F Benson, Claire Bevis, Phillip J Black, Asha Boodhun, Joe S Brennan, John a Bridgham, Rob C Brown, Andrew a Brown, Dale H Buermann, Abass a Bundu, James C Burrows, Nigel P Carter, Nestor Castillo, Maria Chiara E Catenazzi, Simon Chang, R Neil Cooley, Natasha R Crake, Olubunmi O Dada, Konstantinos D Diakoumakos, Belen Dominguez-Fernandez, David J Earnshaw, Ugonna C Egbujor, David W Elmore,

Sergey S Etchin, Mark R Ewan, Milan Fedurco, Louise J Fraser, Karin V Fuentes Fajardo, W Scott Furey, David George, Kimberley J Gietzen, Colin P Goddard, George S Golda, Philip a Granieri, David E Green, David L Gustafson, Nancy F Hansen, Kevin Harnish, Christian D Haudenschild, Narinder I Heyer, Matthew M Hims, Johnny T Ho, Adrian M Horgan, Katya Hoschler, Steve Hurwitz, Denis V Ivanov, Maria Q Johnson, Terena James, T a Huw Jones, Gyoung-Dong Kang, Tzvetana H Kerelska, Alan D Kersey, Irina Khrebtukova, Alex P Kindwall, Zoya Kingsbury, Paula I Kokko-Gonzales, Anil Kumar, Marc a Laurent, Cynthia T Lawley, Sarah E Lee, Xavier Lee, Arnold K Liao, Jennifer a Loch, Mitch Lok, Shujun Luo, Radhika M Mammen, John W Martin, Patrick G McCauley, Paul McNitt, Parul Mehta, Keith W Moon, Joe W Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M Novo, Michael J O'Neill, Mark a Osborne, Andrew Osnowski, Omead Ostadan, Lambros L Paraschos, Lea Pickering, Andrew C Pike, Alger C Pike, D Chris Pinkard, Daniel P Pliskin, Joe Podhasky, Victor J Quijano, Come Raczy, Vicki H Rae, Stephen R Rawlings, Ana Chiva Rodriguez, Phyllida M Roe, John Rogers, Maria C Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K Roth, Natalie J Rourke, Silke T Ruediger, Eli Rusman, Raquel M Sanches-Kuiper, Martin R Schenker, Josefina M Seoane, Richard J Shaw, Mitch K Shiver, Steven W Short, Ning L Sizto, Johannes P Sluis, Melanie a Smith, Jean Ernest Sohna Sohna, Eric J Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L Tregidgo, Gerardo Turcatti, Stephanie Vandevondele, Yuli Verhovsky, Selene M Virk, Suzanne Wakelin, Gregory C Walcott, Jingwen Wang, Graham J Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C Mullikin, Matthew E Hurles, Nick J McCooke, John S West, Frank L Oaks, Peter L Lundberg, David Klenerman, Richard Durbin, and Anthony J Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–9, November 2008.

[7] Bradley E Bernstein, Ewan Birney, Ian Dunham, Eric D Green, Chris Gunter, and Michael Snyder. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[8] Felix Beuschlein, Sheerazed Boulkroun, Andrea Osswald, Thomas Wieland, Hang N Nielsen, Urs D Lichtenauer, David Penton, Vivien R Schack, Laurence Amar, Evelyn Fischer, Anett Walther, Philipp Tauber, Thomas Schwarzmayr, Susanne Diener, Elisabeth Graf, Bruno Allolio, Benoit Samson-couterie, Arndt Benecke, Marcus Quinkler, Francesco Fallo, Pierre-francois Plouin, Franco Mantero, Thomas Meitinger, Paolo Mulatero, Xavier Jeunemaitre, Richard Warth, Bente Vilsen, Maria-christina Zennaro, Tim M Strom, and Martin Reincke. Somatic mutations in ATP1A1 and ATP2B3 lead to aldosterone-producing adenomas and secondary hypertension. *Nature Genetics*, (Advanded Online Publication 17.02.2013), 2013.

[9] Judith a Blake, Carol J Bult, Janan T Eppig, James a Kadin, and Joel E Richardson. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic acids research*, 42(Database issue):D810–7, January 2014.

[10] Keith R Bradnam, Joseph N Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, Inanç Birol, Sébastien Boisvert, Jarrod a Chapman, Guillaume Chapuis, Rayan

Chikhi, Hamidreza Chitsaz, Wen-Chi Chou, Jacques Corbeil, Cristian Del Fabbro, T Roderick Docking, Richard Durbin, Dent Earl, Scott Emrich, Pavel Fedotov, Nuno a Fonseca, Ganeshkumar Ganapathy, Richard a Gibbs, Sante Gnerre, Elénie Godzaridis, Steve Goldstein, Matthias Haimel, Giles Hall, David Haussler, Joseph B Hiatt, Isaac Y Ho, Jason Howard, Martin Hunt, Shaun D Jackman, David B Jaffe, Erich D Jarvis, Huaiyang Jiang, Sergey Kazakov, Paul J Kersey, Jacob O Kitzman, James R Knight, Sergey Koren, Tak-Wah Lam, Dominique Lavenier, François Laviolette, Yingrui Li, Zhenyu Li, Binghang Liu, Yue Liu, Ruibang Luo, Iain Maccallum, Matthew D Macmanes, Nicolas Maillet, Sergey Melnikov, Delphine Naquin, Zemin Ning, Thomas D Otto, Benedict Paten, Octávio S Paulo, Adam M Phillippy, Francisco Pina-Martins, Michael Place, Dariusz Przybylski, Xiang Qin, Carson Qu, Filipe J Ribeiro, Stephen Richards, Daniel S Rokhsar, J Graham Ruby, Simone Scalabrin, Michael C Schatz, David C Schwartz, Alexey Sergushichev, Ted Sharpe, Timothy I Shaw, Jay Shendure, Yujian Shi, Jared T Simpson, Henry Song, Fedor Tsarev, Francesco Vezzi, Riccardo Vicedomini, Bruno M Vieira, Jun Wang, Kim C Worley, Shuangye Yin, Siu-Ming Yiu, Jianying Yuan, Guojie Zhang, Hao Zhang, Shiguo Zhou, and Ian F Korf. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2(1):10, January 2013.

[11] C X Chen, D S Cho, Q Wang, F Lai, K C Carter, and K Nishikura. A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. *RNA (New York, N.Y.)*, 6(5):755–67, May 2000.

[12] Ken Chen, John W Wallis, Michael D McLellan, David E Larson, Joelle M Kalicki, Craig S Pohl, Sean D McGrath, Michael C Wendl, Qunyuan Zhang, Devin P Locke, Xiaoqi Shi, Robert S Fulton, Timothy J Ley, Richard K Wilson, Li Ding, and Elaine R Mardis. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6(9):677–81, September 2009.

[13] Kevin Chen and Nikolaus Rajewsky. The evolution of gene regulation by transcription factors and microRNAs. *Nature reviews. Genetics*, 8(2):93–103, February 2007.

[14] Sung Chun and JC Fay. Identification of deleterious mutations within three human genomes. *Genome research*, pages 1553–1561, 2009.

[15] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3):213–9, March 2013.

[16] P Cingolani, A Platts, M Coon, T Nguyen, L Wang, SJ Land, X Lu, and DM Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118;. *Fly*, 6(2):80–92, 2012.

[17] Michael J Clark, Rui Chen, Hugo Y K Lam, Konrad J Karczewski, Rong Chen, Ghia Euskirchen, Atul J Butte, and Michael Snyder. Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology*, advance on, September 2011.

[18] James Clarke, HC Wu, Lakmal Jayasinghe, and Alpesh Patel. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature . . .*, 4(April), 2009.

[19] Donald F Conrad, Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, Yujun Zhang, Jan Aerts, T Daniel Andrews, Chris Barnes, Peter Campbell, Tomas Fitzgerald, Min Hu, Chun Hwa Ihm, Kati Kristiansson, Daniel G Macarthur, Jeffrey R Macdonald, Ifejinelo Onyiah, Andy Wing Chun Pang, Sam Robson, Kathy Stirrups, Armand Valsesia, Klaudia Walter, John Wei, Chris Tyler-Smith, Nigel P Carter, Charles Lee, Stephen W Scherer, and Matthew E Hurles. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–12, April 2010.

[20] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis a Albers, Eric Banks, Mark a DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, 27(15):2156–8, August 2011.

[21] K Danhauser*, SW Sauer*, TB Haack*, T Wieland*, C Staufner, E Graf, J Zschocke, T M Strom, T Traub, J G Okun, T Meitinger, G F Hoffmann, H Prokisch, and S Kolker. DHTKD1 mutations cause 2-aminoadipic and 2-oxoadipic aciduria. *Am. J. Hum. Genet.*, 91(6):1082–1087, 2012.

[22] Eugene V Davydov, David L Goode, Marina Sirota, Gregory M Cooper, Arend Sidow, and Serafim Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology*, 6(12):e1001025, January 2010.

[23] Joep de Ligt, Marjolein H. Willemsen, Bregje W.M. van Bon, Tjitske Kleefstra, Helger G. Yntema, Thessa Kroes, Anneke T. Vulto-van Silfhout, David a. Koolen, Petra de Vries, Christian Gilissen, Marisol del Rosario, Alexander Hoischen, Hans Scheffer, Bert B.a. de Vries, Han G. Brunner, Joris a. Veltman, and Lisenka E.L.M. Vissers. Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *New England Journal of Medicine*, page 121003140044006, October 2012.

[24] Mark a DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony a Philippakis, Guillermo del Angel, Manuel a Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–8, May 2011.

[25] S Dusi, L Valletta, T B Haack, Y Tsuchiya, P Venco, S Pasqualato, P Goffrini, M Tigano, N Demchenko, T Wieland, T Schwarzmayr, T M Strom, F Invernizzi, B Garavaglia, A Gregory, L Sanford, J Hamada, C Bettencourt, H Houlden, L Chiapparini, G Zorzi, M A Kurian, N Nardocci, H Prokisch, S Hayflick, I Gout, and V Tiranti. Exome sequence reveals mutations in CoA synthase as a cause of neurodegeneration with brain iron accumulation. *Am. J. Hum. Genet.*, 94(1):11–22, January 2014.

[26] Sebastian H Eck. *Identification of genetic variation using Next-Generation Sequencing*. Phd thesis, Technische Universität München, 2014.

[27] J Eid, A Fehr, J Gray, K Luong, J Lyle, and G Otto. Real-time DNA sequencing from single polymerase molecules. *Science*, 472(January):431–55, January 2009.

[28] Pär G Engström, Tamara Steijger, Botond Sipos, Gregory R Grant, André Kahles, Gunnar Rätsch, Nick Goldman, Tim J Hubbard, Jennifer Harrow, Roderic Guigó, and Paul Bertone. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature methods*, 10(12):1185–91, December 2013.

[29] Sanaz Farajollahi and Stefan Maas. Molecular diversity through RNA editing: a balancing act. *Trends in genetics : TIG*, 26(5):221–30, May 2010.

[30] X Gai, D Ghezzi, M A Johnson, C A Biagosch, H E Shamseldin, T B Haack, A Reyes, M Tsukikawa, C A Sheldon, S Srinivasan, M Gorza, L S Kremer, T Wieland, T M Strom, E Polyak, E Place, M Consugar, J Ostrovsky, S Vidoni, A J Robinson, L J Wong, N Sondheimer, M A Salih, E Al-Jishi, C P Raab, C Bean, F Furlan, R Parini, C Lamperti, J A Mayr, V Konstantopoulou, M Huemer, E A Pierce, T Meitinger, P Freisinger, W Sperl, H Prokisch, F S Alkuraya, M J Falk, and M Zeviani. Mutations in FBXL4, encoding a mitochondrial protein, cause early-onset mitochondrial encephalomyopathy. *Am. J. Hum. Genet.*, 93(3):482–495, September 2013.

[31] Manuel Garber, Mitchell Guttman, Michele Clamp, Michael C Zody, Nir Friedman, and Xiaohui Xie. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics (Oxford, England)*, 25(12):i54–62, June 2009.

[32] RA Gibbs, JW Belmont, P Hardenbol, and TD Willis. The international HapMap project. *Nature*, 63 Suppl 1:29–34, December 2003.

[33] Robert C Green, Jonathan S Berg, Wayne W Grody, Sarah S Kalia, Bruce R Korf, Christa L Martin, Amy L McGuire, Robert L Nussbaum, Julianne M O'Daniel, Kelly E Ormond, Heidi L Rehm, Michael S Watson, Marc S Williams, and Leslie G Biesecker. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in medicine : official journal of the American College of Medical Genetics*, 15(7):565–74, July 2013.

[34] T B Haack, M Gorza, K Danhauser, J A Mayr, B Haberberger, T Wieland, L Kremer, V Strecker, E Graf, Y Memari, U Ahting, R Kopajtich, S B Wortmann, R J Rodenburg, U Kotzaeridou, G F Hoffmann, W Sperl, I Wittig, E Wilichowski, G Schottmann, M Schuelke, B Plecko, U Stephani, T M Strom, T Meitinger, H Prokisch, and P Freisinger. Phenotypic spectrum of eleven patients and five novel MTFMT mutations identified by exome sequencing and candidate gene screening. *Mol. Genet. Metab.*, 111(3):342–352, 2014.

[35] T B Haack, B Haberberger, E M Frisch, T Wieland, A Iuso, M Gorza, V Strecker, E Graf, J A Mayr, U Herberg, J B Hennermann, T Klopstock, K A Kuhn, U Ahting, W Sperl, E Wilichowski, G F Hoffmann, M Tesarova, H Hansikova, J Zeman, B Plecko, M Zeviani, I Wittig, T M Strom, M Schuelke, P Freisinger, T Meitinger, and

H Prokisch. Molecular diagnosis in mitochondrial complex I deficiency using exome sequencing. *J. Med. Genet.*, 49(4):277–283, April 2012.

[36] T B Haack, P Hogarth, M C Kruer, A Gregory, T Wieland, T Schwarzmayr, E Graf, L Sanford, E Meyer, E Kara, S M Cuno, S I Harik, V H Dandu, N Nardocci, G Zorzi, T Dunaway, M Tarnopolsky, S Skinner, S Frucht, E Hanspal, C Schrander-Stumpel, D Heron, C Mignot, B Garavaglia, K Bhatia, J Hardy, T M Strom, N Boddaert, H H Houlden, M A Kurian, T Meitinger, H Prokisch, and S J Hayflick. Exome sequencing reveals de novo WDR45 mutations causing a phenotypically distinct, X-linked dominant form of NBIA. *Am. J. Hum. Genet.*, 91(6):1144–1149, 2012.

[37] T B Haack, R Kopajtich, P Freisinger, T Wieland, J Rorbach, T J Nicholls, E Baruffini, A Walther, K Danhauser, F A Zimmermann, R A Husain, J Schum, H Mundy, I Ferrero, T M Strom, T Meitinger, R W Taylor, M Minczuk, J A Mayr, and H Prokisch. ELAC2 mutations cause a mitochondrial RNA processing defect associated with hypertrophic cardiomyopathy. *Am. J. Hum. Genet.*, 93(2):211–223, August 2013.

[38] T B Haack, C Makowski, Y Yao, E Graf, M Hempel, T Wieland, U Tauer, U Ahting, J A Mayr, P Freisinger, H Yoshimatsu, K Inui, T M Strom, T Meitinger, A Yonezawa, and H Prokisch. Impaired riboflavin transport due to missense mutations in SLC52A2 causes Brown-Vialetto-Van Laere syndrome. *J. Inherit. Metab. Dis.*, 35(6):943–948, November 2012.

[39] Tobias B Haack, Katharina Danhauser, Birgit Haberberger, Jonathan Hoser, Valentina Strecker, Detlef Boehm, Graziella Uziel, Eleonora Lamantea, Federica Invernizzi, Joanna Poulton, Boris Rolinski, Arcangela Iuso, Saskia Biskup, Thorsten Schmidt, Hans-Werner Mewes, Ilka Wittig, Thomas Meitinger, Massimo Zeviani, and Holger Prokisch. Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency. *Nature genetics*, 42(12):1131–4, December 2010.

[40] Xin He, Stephan J Sanders, Li Liu, Silvia De Rubeis, Elaine T Lim, James S Sutcliffe, Gerard D Schellenberg, Richard a Gibbs, Mark J Daly, Joseph D Buxbaum, Matthew W State, Bernie Devlin, and Kathryn Roeder. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS genetics*, 9(8):e1003671, January 2013.

[41] Xuesong Hu, Jianying Yuan, Yujian Shi, Jianliang Lu, Binghang Liu, Zhenyu Li, Yanxiang Chen, Desheng Mu, Hao Zhang, Nan Li, Zhen Yue, Fan Bai, Heng Li, and Wei Fan. pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics (Oxford, England)*, 28(11):1533–5, June 2012.

[42] Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford, England)*, 28(4):593–4, February 2012.

[43] Ivan Iossifov, Michael Ronemus, Dan Levy, Zihua Wang, Inessa Hakker, Julie Rosenbaum, Boris Yamrom, Yoon-Ha Lee, Giuseppe Narzisi, Anthony Leotta, Jude Kendall, Ewa Grabowska, Beicong Ma, Steven Marks, Linda Rodgers, Asya Stepansky, Jennifer Troge, Peter Andrews, Mitchell Bekritsky, Kith Pradhan, Elena Ghiban,

Melissa Kramer, Jennifer Parla, Ryan Demeter, Lucinda L Fulton, Robert S Fulton, Vincent J Magrini, Kenny Ye, Jennifer C Darnell, Robert B Darnell, Elaine R Mardis, Richard K Wilson, Michael C Schatz, W Richard McCombie, and Michael Wigler. De novo gene disruptions in children on the autistic spectrum. *Neuron*, 74(2):285–99, April 2012.

[44] Jennifer J Johnston and Leslie G Biesecker. Databases of genomic variation and phenotypes: existing resources and future needs. *Human molecular genetics*, August 2013.

[45] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, January 2000.

[46] Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research*, 42(Database issue):D199–205, January 2014.

[47] Manolis Kellis, Barbara Wold, Michael P Snyder, Bradley E Bernstein, Anshul Kundaje, Georgi K Marinov, Lucas D Ward, Ewan Birney, Gregory E Crawford, Job Dekker, Ian Dunham, Laura L Elnitski, Peggy J Farnham, Elise a Feingold, Mark Gerstein, Morgan C Giddings, David M Gilbert, Thomas R Gingeras, Eric D Green, Roderic Guigo, Tim Hubbard, Jim Kent, Jason D Lieb, Richard M Myers, Michael J Pazin, Bing Ren, John a Stamatoyannopoulos, Zhiping Weng, Kevin P White, and Ross C Hardison. Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 111(17):6131–8, April 2014.

[48] W. J. Kent. BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12(4):656–664, March 2002.

[49] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, a. M. Zahler, and a. D. Haussler. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, May 2002.

[50] Min-Sik Kim, Sneha M. Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S. Manda, Raghothama Chaerkady, Anil K. Madugundu, Dhanashree S. Kelkar, Ruth Isserlin, Shobhit Jain, Joji K. Thomas, Babylakshmi Muthusamy, Pamela Leal-Rojas, Praveen Kumar, Nandini a. Sahasrabuddhe, Lavanya Balakrishnan, Jayshree Advani, Bijesh George, Santosh Renuse, Lakshmi Dhevi N. Selvan, Arun H. Patil, Vishalakshi Nanjappa, Aneesha Radhakrishnan, Samarjeet Prasad, Tejaswini Subbannayya, Rajesh Raju, Manish Kumar, Sreelakshmi K. Sreenivasamurthy, Arivusudar Marimuthu, Gajanan J. Sathe, Sandip Chavan, Keshava K. Datta, Yashwanth Subbannayya, Apeksha Sahu, Soujanya D. Yelamanchi, Savita Jayaram, Pavithra Rajagopalan, Jyoti Sharma, Krishna R. Murthy, Nazia Syed, Renu Goel, Aafaque a. Khan, Sartaj Ahmad, Gourav Dey, Keshav Mudgal, Aditi Chatterjee, Tai-Chung Huang, Jun Zhong, Xinyan Wu, Patrick G. Shaw, Donald Freed, Muhammad S. Zahari, Kanchan K. Mukherjee, Subramanian Shankar, Anita Mahadevan, Henry Lam, Christopher J. Mitchell, Susarla Krishna Shankar, Parthasarathy Satishchandra,

John T. Schroeder, Ravi Sirdeshmukh, Anirban Maitra, Steven D. Leach, Charles G. Drake, Marc K. Halushka, T. S. Keshava Prasad, Ralph H. Hruban, Candace L. Kerr, Gary D. Bader, Christine a. Iacobuzio-Donahue, Harsha Gowda, and Akhilesh Pandey. A draft map of the human proteome. *Nature*, 509(7502):575–581, May 2014.

[51] Sangwoo Kim, Kyowon Jeong, and Vineet Bafna. Wessim: a whole-exome sequencing simulator based on in silico exome capture. *Bioinformatics (Oxford, England)*, 29(8):1076–7, April 2013.

[52] Anmol Kiran and Pavel V Baranov. DARNED: a DAtabase of RNa EDiting in humans. *Bioinformatics (Oxford, England)*, 26(14):1772–6, July 2010.

[53] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–5, March 2014.

[54] Claudia L Kleinman and Jacek Majewski. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science (New York, N.Y.)*, 335(6074):1302; author reply 1302, March 2012.

[55] Gavin C K W Koh, Pablo Porras, Bruno Aranda, Henning Hermjakob, and Sandra E Orchard. Analyzing protein-protein interaction networks. *Journal of proteome research*, 11(4):2014–31, April 2012.

[56] Augustine Kong, Michael L. Frigge, Gisli Masson, Soren Besenbacher, Patrick Sulem, Gisli Magnusson, Sigurjon a. Gudjonsson, Asgeir Sigurdsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, Wendy S. W. Wong, Gunnar Sigurdsson, G. Bragi Walters, Stacy Steinberg, Hannes Helgason, Gudmar Thorleifsson, Daniel F. Gudbjartsson, Agnar Helgason, Olafur Th. Magnusson, Unnur Thorsteinsdottir, and Kari Stefansson. Rate of de novo mutations and the importance of fatherâs age to disease risk. *Nature*, 488(7412):471–475, August 2012.

[57] C Kornblum, T J Nicholls, T B Haack, S Scholer, V Peeva, K Danhauser, K Hallmann, G Zsurka, J Rorbach, A Iuso, T Wieland, M Sciacco, D Ronchi, G P Comi, M Moggio, C M Quinzii, S DiMauro, S E Calvo, V K Mootha, T Klopstock, T M Strom, T Meitinger, M Minczuk, W S Kunz, and H Prokisch. Loss-of-function mutations in MGME1 impair mtDNA replication and cause multisystemic mitochondrial disease. *Nat. Genet.*, 45(2):214–219, February 2013.

[58] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols*, 4(7):1073–81, January 2009.

[59] Peter W Laird. Principles and challenges of genomewide DNA methylation analysis. *Nature reviews. Genetics*, 11(3):191–203, March 2010.

[60] Melissa J Landrum, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, 42(Database issue):D980–5, January 2014.

[61] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9, April 2012.

[62] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25, 2009.

[63] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter a C 't Hoen, Jean Monlong, Manuel a Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G MacArthur, Monkol Lek, Esther Lizano, Henk P J Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B Montgomery, Peter Donnelly, Mark I McCarthy, Paul Flicek, Tim M Strom, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Angel Carracedo, Stylianos E Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G Gut, Xavier Estivill, and Emmanouil T Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–11, September 2013.

[64] Seunggeung Lee, Gonçalo R Abecasis, Michael Boehnke, and Xihong Lin. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *American journal of human genetics*, 95(1):5–23, July 2014.

[65] H Leonard and X Wen. The epidemiology of mental retardation: challenges and opportunities in the new millennium. *Mental retardation and developmental ...*, 2002.

[66] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*, 00(00):1–3, 2013.

[67] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60, July 2009.

[68] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9, August 2009.

[69] Jin Billy Li and George M Church. Deciphering the functions and regulation of brain-enriched A-to-I RNA editing. *Nature neuroscience*, 16(11):1518–22, November 2013.

[70] Jin Billy Li, Erez Y Levanon, Jung-Ki Yoon, John Aach, Bin Xie, Emily Leproust, Kun Zhang, Yuan Gao, and George M Church. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science (New York, N.Y.)*, 324(5931):1210–3, May 2009.

[71] M. Li, I. X. Wang, and V. G. Cheung. Response to Comments on "Widespread RNA and DNA Sequence Differences in the Human Transcriptome". *Science*, 335(6074):1302–1302, March 2012.

[72] Mingyao Li, Isabel X Wang, Yun Li, Alan Bruzel, Allison L Richards, Jonathan M Toung, and Vivian G Cheung. Widespread RNA and DNA sequence differences in the human transcriptome. *Science (New York, N.Y.)*, 333(6038):53–8, July 2011.

[73] Wei Lin, Robert Piskol, Meng How Tan, and Jin Billy Li. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science (New York, N.Y.)*, 335(6074):1302; author reply 1302, March 2012.

[74] Kerstin Lindblad-Toh, Manuel Garber, Or Zuk, Michael F Lin, Brian J Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, Evan Mauceli, Lucas D Ward, Craig B Lowe, Alisha K Holloway, Michele Clamp, Sante Gnerre, Jessica Alföldi, Kathryn Beal, Jean Chang, Hiram Clawson, James Cuff, Federica Di Palma, Stephen Fitzgerald, Paul Flicek, Mitchell Guttman, Melissa J Hubisz, David B Jaffe, Irwin Jungreis, W James Kent, Dennis Kostka, Marcia Lara, Andre L Martins, Tim Massingham, Ida Moltke, Brian J Raney, Matthew D Rasmussen, Jim Robinson, Alexander Stark, Albert J Vilella, Jiayu Wen, Xiaohui Xie, Michael C Zody, Jen Baldwin, Toby Bloom, Chee Whye Chin, Dave Heiman, Robert Nicol, Chad Nusbaum, Sarah Young, Jane Wilkinson, Kim C Worley, Christie L Kovar, Donna M Muzny, Richard a Gibbs, Andrew Cree, Huyen H Dihn, Gerald Fowler, Shalili Jhangiani, Vandita Joshi, Sandra Lee, Lora R Lewis, Lynne V Nazareth, Geoffrey Okwuonu, Jireh Santibanez, Wesley C Warren, Elaine R Mardis, George M Weinstock, Richard K Wilson, Kim Delehaunty, David Dooling, Catrina Fronik, Lucinda Fulton, Bob Fulton, Tina Graves, Patrick Minx, Erica Sodergren, Ewan Birney, Elliott H Margulies, Javier Herrero, Eric D Green, David Haussler, Adam Siepel, Nick Goldman, Katherine S Pollard, Jakob S Pedersen, Eric S Lander, and Manolis Kellis. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–82, October 2011.

[75] Xiangtao Liu, Shizhong Han, Zuoheng Wang, Joel Gelernter, and Bao-Zhu Yang. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS ONE*, 8(9):e75619, September 2013.

[76] Xiaoming Liu, Xueqiu Jian, and Eric Boerwinkle. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Human mutation*, 32(8):894–9, August 2011.

[77] Xiaoming Liu, Xueqiu Jian, and Eric Boerwinkle. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Human mutation*, 34(9):E2393–402, September 2013.

[78] Ruibang Luo, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Liu, Jingbo Tang, Gengxiong Wu, Hao Zhang, Yujian Shi, Yong Liu, Chang Yu, Bo Wang, Yao Lu, Changlei Han, David W Cheung, Siu-Ming Yiu, Shaoliang Peng, Zhu Xiaoqian, Guangming Liu, Xiangke Liao, Yingrui Li, Huanming Yang, Jian Wang, Tak-Wah Lam, and Jun Wang. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1):18, January 2012.

[79] Stefan Maas, Yukio Kawahara, KM Tamburro, and Kazuko Nishikura. A-to-I RNA editing and human disease. *RNA biology*, 3(1):1, 2006.

[80] D. G. MacArthur, T. a. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis, D. R. Adams, R. B. Altman, S. E. Antonarakis, E. a. Ashley, J. C. Barrett, L. G. Biesecker, D. F. Conrad, G. M. Cooper, N. J. Cox, M. J. Daly, M. B. Gerstein, D. B. Goldstein, J. N. Hirschhorn, S. M. Leal, L. a. Pennacchio, J. a. Stamatoyannopoulos, S. R. Sunyaev, D. Valle, B. F. Voight, W. Winckler, and C. Gunter. Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497):469–476, April 2014.

[81] M Macdonald. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72(6):971–983, March 1993.

[82] Santiago Marco-Sola, Michael Sammeth, Roderic Guigó, and Paolo Ribeca. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods*, October 2012.

[83] J A Mayr, T B Haack, E Graf, F A Zimmermann, T Wieland, B Haberberger, A Superti-Furga, J Kirschner, B Steinmann, M R Baumgartner, I Moroni, E Lamantea, M Zeviani, R J Rodenburg, J Smeitink, T M Strom, T Meitinger, W Sperl, and H Prokisch. Lack of the mitochondrial protein acylglycerol kinase causes Sengers syndrome. *Am. J. Hum. Genet.*, 90(2):314–320, February 2012.

[84] Kerensa E McElroy, Fabio Luciani, and Torsten Thomas. GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC genomics*, 13(1):74, January 2012.

[85] A. H. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. Depristo. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, July 2010.

[86] KJ McKernan and HE Peckham. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome . . .* , pages 1527–1541, 2009.

[87] Michael L Metzker. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1):31–46, 2010.

[88] AP Morris, BF Voight, and TM Teslovich. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *. . . genetics*, 44(9):981–990, 2012.

[89] Benjamin M. Neale, Yan Kou, Li Liu, Avi Maâayan, Kaitlin E. Samocha, Aniko Sabo, Chiao-Feng Lin, Christine Stevens, Li-San Wang, Vladimir Makarov, Paz Polak, Seungtai Yoon, Jared Maguire, Emily L. Crawford, Nicholas G. Campbell, Evan T. Geller, Otto Valladares, Chad Schafer, Han Liu, Tuo Zhao, Guiqing Cai, Jayon Lihm, Ruth Dannenfelser, Omar Jabado, Zuleyma Peralta, Uma Nagaswamy,

Donna Muzny, Jeffrey G. Reid, Irene Newsham, Yuanqing Wu, Lora Lewis, Yi Han, Benjamin F. Voight, Elaine Lim, Elizabeth Rossin, Andrew Kirby, Jason Flannick, Menachem Fromer, Khalid Shakir, Tim Fennell, Kiran Garimella, Eric Banks, Ryan Poplin, Stacey Gabriel, Mark DePristo, Jack R. Wimbish, Braden E. Boone, Shawn E. Levy, Catalina Betancur, Shamil Sunyaev, Eric Boerwinkle, Joseph D. Buxbaum, Edwin H. Cook, Bernie Devlin, Richard a. Gibbs, Kathryn Roeder, Gerard D. Schellenberg, James S. Sutcliffe, and Mark J. Daly. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, pages 5–9, April 2012.

[90] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics*, 12(6):443–51, June 2011.

[91] Kazuko Nishikura. Functions and Regulation of RNA Editing by ADAR Deaminases. *Annual review of biochemistry*, December 2010.

[92] J Ockenga, M Stuhrmann, M Ballmann, N Teich, V Keim, T Dörk, and M P Manns. Mutations of the cystic fibrosis gene, but not cationic trypsinogen gene, are associated with recurrent or chronic idiopathic pancreatitis. *The American journal of gastroenterology*, 95(8):2061–7, August 2000.

[93] Jason O'Rawe, Tao Jiang, Guangqing Sun, Yiyang Wu, Wei Wang, Jingchu Hu, Paul Bodily, Lifeng Tian, Hakon Hakonarson, W Evan Johnson, Zhi Wei, Kai Wang, and Gholson J Lyon. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome medicine*, 5(3):28, March 2013.

[94] Brian P O'Sullivan and Steven D Freedman. Cystic fibrosis. *Lancet*, 373(9678):1891–904, May 2009.

[95] J Oyston. Online Mendelian Inheritance in Man. *Anesthesiology*, 89(3):811–812, 1998.

[96] Brian J. OâRoak, Laura Vives, Santhosh Girirajan, Emre Karakoc, Niklas Krumm, Bradley P. Coe, Roie Levy, Arthur Ko, Choli Lee, Joshua D. Smith, Emily H. Turner, Ian B. Stanaway, Benjamin Vernot, Maika Malig, Carl Baker, Beau Reilly, Joshua M. Akey, Elhanan Borenstein, Mark J. Rieder, Deborah a. Nickerson, Raphael Bernier, Jay Shendure, and Evan E. Eichler. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, pages 1–7, April 2012.

[97] Philipp Pagel, Stefan Kovac, Matthias Oesterheld, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Pekka Mark, Volker Stümpflen, Hans-Werner Mewes, Andreas Ruepp, and Dmitrij Frishman. The MIPS mammalian protein-protein interaction database. *Bioinformatics (Oxford, England)*, 21(6):832–4, March 2005.

[98] Peter J Park. ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10):669–80, October 2009.

[99] Ernesto Picardi, Angela Gallo, Federica Galeano, Sara Tomaselli, and Graziano Pesole. A Novel Computational Strategy to Identify A-to-I RNA Editing Sites by RNA-Seq Data: De Novo Detection in Human Spinal Cord Tissue. *PloS one*, 7(9):e44184, January 2012.

[100] Joseph K Pickrell, Yoav Gilad, and Jonathan K Pritchard. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science (New York, N.Y.)*, 335(6074):1302; author reply 1302, March 2012.

[101] Robert Piskol, Zhiyu Peng, Jun Wang, and Jin Billy Li. Lack of evidence for existence of noncanonical RNA editing. *Nature Biotechnology*, 31(1):19–20, January 2013.

[102] Vincent Plagnol, James Curtis, Michael Epstein, Kin Y Mok, Emma Stebbings, Sofia Grigoriadou, Nicholas W Wood, Sophie Hambleton, Siobhan O Burns, Adrian J Thrasher, Dinakantha Kumararatne, Rainer Doffinger, and Sergey Nejentsev. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics (Oxford, England)*, 28(21):2747–54, November 2012.

[103] Gokul Ramaswami, Wei Lin, Robert Piskol, Meng How Tan, Carrie Davis, and Jin Billy Li. Accurate identification of human Alu and non-Alu RNA editing sites. *Nature Methods*, (April), April 2012.

[104] Anita Rauch, Dagmar Wieczorek, Elisabeth Graf, Thomas Wieland, Sabine Endele, Thomas Schwarzmayr, Beate Albrecht, Deborah Bartholdi, Jasmin Beygo, Nataliya Di Donato, Andreas Dufke, Kirsten Cremer, Maja Hempel, Denise Horn, Juliane Hoyer, Pascal Joset, Albrecht Röpke, Ute Moog, Angelika Riess, Christian T Thiel, Andreas Tzschach, Antje Wiesener, Eva Wohlleber, Christiane Zweier, Arif B Ekici, Alexander M Zink, Andreas Rump, Christa Meisinger, Harald Grallert, Heinrich Sticht, Annette Schenck, Hartmut Engels, Gudrun Rappold, Evelin Schröck, Peter Wieacker, Olaf Riess, Thomas Meitinger, André Reis, and Tim M Strom. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet*, 380(9854):1674–82, November 2012.

[105] Heidi L Rehm, Sherri J Bale, Pinar Bayrak-Toydemir, Jonathan S Berg, Kerry K Brown, Joshua L Deignan, Michael J Friez, Birgit H Funke, Madhuri R Hegde, and Elaine Lyon. ACMG clinical laboratory standards for next-generation sequencing. *Genetics in medicine : official journal of the American College of Medical Genetics*, 15(9):733–47, September 2013.

[106] Boris Reva, Yevgeniy Antipin, and Chris Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, 39(17):e118, September 2011.

[107] C Sue Richards, Sherri Bale, Daniel B Bellissimo, Soma Das, Wayne W Grody, Madhuri R Hegde, Elaine Lyon, and Brian E Ward. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genetics in medicine : official journal of the American College of Medical Genetics*, 10(4):294–300, April 2008.

[108] JT Robinson and H Thorvaldsdóttir. Integrative genomics viewer. *Nature ...*, 29(1):24–26, 2011.

[109] Nicole Rusk. Torrents of sequence. *Nature Methods*, 8(1):44–44, December 2010.

[110] Stephan J. Sanders, Michael T. Murtha, Abha R. Gupta, John D. Murdoch, Melanie J. Raubeson, a. Jeremy Willsey, a. Gulhan Ercan-Sencicek, Nicholas M. DiLullo, Neelroop N. Parikshak, Jason L. Stein, Michael F. Walker, Gordon T. Ober, Nicole a. Teran, Youeun Song, Paul El-Fishawy, Ryan C. Murtha, Murim Choi, John D. Overton, Robert D. Bjornson, Nicholas J. Carriero, Kyle a. Meyer, Kaya Bilguvar, Shrikant M. Mane, Nenad Šestan, Richard P. Lifton, Murat Günel, Kathryn Roeder, Daniel H. Geschwind, Bernie Devlin, and Matthew W. State. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, pages 1–6, April 2012.

[111] F Sanger and S Nicklen. DNA sequencing with chain-terminating. 74(12):5463–5467, 1977.

[112] Eric E Schadt, Steve Turner, and Andrew Kasarskis. A Window into Third Generation Sequencing. *Human molecular genetics*, 19(R2):R227–240, September 2010.

[113] Jana Marie Schwarz, Christian Rödelsperger, Markus Schuelke, and Dominik Seelow. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature methods*, 7(8):575–6, August 2010.

[114] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature biotechnology*, 26(10):1135–45, October 2008.

[115] 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 research*, 29(1):308–11, January 2001.

[116] Hashem a Shihab, Julian Gough, David N Cooper, Peter D Stenson, Gary L a Barker, Keith J Edwards, Ian N M Day, and Tom R Gaunt. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human mutation*, 34(1):57–65, January 2013.

[117] Adam Siepel, KS Pollard, and David Haussler. New methods for detecting lineage-specific selection. *Research in Computational Molecular ...*, 2006.

[118] Jared T Simpson, Kim Wong, Shaun D Jackman, Jacqueline E Schein, Steven J M Jones, and Inanç Birol. ABySS: a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–23, June 2009.

[119] W Slotkin and K Nishikura. Adenosine-to-inosine RNA editing and human disease. *Genome Med*, 5(11):105, November 2013.

[120] ML Stallings-Mann. Alternative splicing of exon 3 of the human growth hormone receptor is the result of an unusual genetic polymorphism. *Proceedings of the ...*, 93(October):12394–12399, 1996.

[121] Peter D Stenson, Edward V Ball, Matthew Mort, Andrew D Phillips, Jacqueline a Shiel, Nick S T Thomas, Shaun Abeysinghe, Michael Krawczak, and David N Cooper. Human Gene Mutation Database (HGMD): 2003 update. *Human mutation*, 21(6):577–81, June 2003.

[122] David Stoddart, Andrew J Heron, Ellina Mikhailova, Giovanni Maglia, and Hagan Bayley. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences of the United States of America*, 106(19):7702–7, May 2009.

[123] B Strandvik, E Björck, M Fallström, E Gronowitz, J Thountzouris, a Lindblad, D Markiewicz, J Wahlström, L C Tsui, and J Zielenski. Spectrum of mutations in the CFTR gene of patients with classical and atypical forms of cystic fibrosis from southwestern Sweden: identification of 12 novel mutations. *Genetic testing*, 5(3):235–42, January 2001.

[124] Andrew I Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith a Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, Michael P Cooke, John R Walker, and John B Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–7, April 2004.

[125] P A 't Hoen, M R Friedlander, J Almlof, M Sammeth, I Pulyakhina, S Y Anvar, J F Laros, H P Buermans, O Karlberg, M Brannvall, J T den Dunnen, G J van Ommen, I G Gut, R Guigo, X Estivill, A C Syvanen, E T Dermitzakis, T Lappalainen, S E Antonarakis, A Brazma, P Flicek, S Schreiber, P Rosenstiel, T Meitinger, T M Strom, H Lehrach, R Sudbrak, A Carracedo, M van Iterson, J Monlong, E Lizano, G Bertier, P G Ferreira, P Ribeca, T Griebel, S Beltran, M Gut, K Kahlem, T Giger, H Ongen, I Padioleau, H Kilpinen, M Gonzalez-Porta, N Kurbatova, A Tikhonov, L Greger, M Barann, D Esser, R Hasler, T Wieland, T Schwarzmayr, M Sultan, and V Amstislavskiy. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.*, 31(11):1015–1022, November 2013.

[126] J W Taanman. The mitochondrial genome: structure, transcription, translation and replication. *Biochimica et biophysica acta*, 1410(2):103–23, February 1999.

[127] Jacob a Tennessen, Abigail W Bigham, Timothy D O'Connor, Wenqing Fu, Eimear E Kenny, Simon Gravel, Sean McGee, Ron Do, Xiaoming Liu, Goo Jun, Hyun Min Kang, Daniel Jordan, Suzanne M Leal, Stacey Gabriel, Mark J Rieder, Goncalo Abecasis, David Altshuler, Deborah a Nickerson, Eric Boerwinkle, Shamil Sunyaev, Carlos D Bustamante, Michael J Bamshad, and Joshua M Akey. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science (New York, N.Y.)*, 337(6090):64–9, July 2012.

[128] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(V):56–65, 2012.

[129] Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–92, March 2013.

[130] Francis O Walker. Huntingtonâs disease. *Lancet*, 369:218–28, 2007.

[131] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164, September 2010.

[132] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 2008.

[133] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*, 42(Database issue):D1001–6, January 2014.

[134] David a Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhijani, G Thomas Roth, Xavier Gomes, Karrie Tartaro, Faheem Niazi, Cynthia L Turcotte, Gerard P Irzyk, James R Lupski, Craig Chinault, Xing-zhi Song, Yue Liu, Ye Yuan, Lynne Nazareth, Xiang Qin, Donna M Muzny, Marcel Margulies, George M Weinstock, Richard a Gibbs, and Jonathan M Rothberg. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–6, April 2008.

[135] Dagmar Wieczorek, William G Newman, Thomas Wieland, Tea Berulava, Maria Kaffe, Daniela Falkenstein, Christian Beetz, Elisabeth Graf, Thomas Schwarzmayr, Sofia Douzgou, Jill Clayton-Smith, Sarah B Daly, Simon G Williams, Sanjeev S Bhaskar, Jill E Urquhart, Beverley Anderson, James O'Sullivan, Odile Boute, Jasmin Gundlach, Johanna Christina Czeschik, Anthonie J van Essen, Filiz Hazan, Sarah Park, Anne Hing, Alma Kuechler, Dietmar R Lohmann, Kerstin U Ludwig, Elisabeth Mangold, Laura Steenpaß, Michael Zeschnigk, Johannes R Lemke, Charles Marques Lourenco, Ute Hehr, Eva-Christina Prott, Melanie Waldenberger, Anne C Böhmer, Bernhard Horsthemke, Raymond T O'Keefe, Thomas Meitinger, John Burn, Hermann-Josef Lüdecke, and Tim M Strom. Compound Heterozygosity of Low-Frequency Promoter Deletions and Rare Loss-of-Function Mutations in TXNL4A Causes Burn-McKeown Syndrome. *American journal of human genetics*, pages 698–707, November 2014.

[136] Yaping Yang, Donna M. Muzny, Jeffrey G. Reid, Matthew N. Bainbridge, Alecia Willis, Patricia a. Ward, Alicia Braxton, Joke Beuten, Fan Xia, Zhiyv Niu, Matthew Hardison, Richard Person, Mir Reza Bekheirnia, Magalie S. Leduc, Amelia Kirby, Peter Pham, Jennifer Scull, Min Wang, Yan Ding, Sharon E. Plon, James R. Lupski, Arthur L. Beaudet, Richard a. Gibbs, and Christine M. Eng. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *New England Journal of Medicine*, page 131002140031007, October 2013.

[137] Kai Ye, Marcel H Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, 25(21):2865–71, November 2009.

[138] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5):821–9, May 2008.

[139] Alexander Zimprich, Anna Benet-Pagès, Walter Struhal, Elisabeth Graf, Sebastian H Eck, Marc N Offman, Dietrich Haubenberger, Sabine Spielberger, Eva C Schulte, Peter Lichtner, Shaila C Rossle, Norman Klopp, Elisabeth Wolf, Klaus Seppi, Walter Pirker, Stefan Presslauer, Brit Mollenhauer, Regina Katzenschlager, Thomas Foki, Christoph Hotzy, Eva Reinthaler, Ashot Harutyunyan, Robert Kralovics, Annette Peters, Fritz Zimprich, Thomas Brücke, Werner Poewe, Eduard Auff, Claudia Trenkwalder, Burkhard Rost, Gerhard Ransmayr, Juliane Winkelmann, Thomas Meitinger, and Tim M Strom. A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease. *American journal of human genetics*, 89(1):168–75, July 2011.

[140] Justin M Zook, Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide, and Marc Salit. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature biotechnology*, 32(3), February 2014.

Thomas Wieland

Lebenslauf

Persönliche Daten

✍ Addresse	Loferer Str. 17, 81671 München
Telefon	+49 152 56190626
✉ E-mail	thomas.wieland@gmx.at
Nationalität	Österreich
Geburtsdatum	12. Juni 1986
Geburtsort	Scheibbs (Österreich)
Familienstand	ledig

Ausbildung

2010– **Dr. rer. nat.**, *Helmholtz Zentrum München*, München.
- Thema: Next-Generation Sequencing Data Analysis
- Bioinformatik

2009–2010 **Master of Science**, *King's College*, London.
- Bioinformatics
- Abschluss mit Auszeichnung im September 2010

2006–2009 **Bachelor of Science**, *UMIT–Private Universität für Gesundheitswissenschaften, Medizinische Informatik und Technik*, Hall/Tirol.
- Biomedizinische Informatik
- Abschluss im Juli 2009, GPA: 88/100

2000–2005 **Matura**, *HTBLA St. Pölten* .
- Abteilung für Elektronische Datenverarbeitung und Organisation
- Abschluss mit Auszeichnung im Juni 2005

Berufliche Laufbahn

Sep 2008 **Nibelungenheim (Pflegeheim)**, Ybbs, Österreich.
- Webseite: http://www.lph-ybbs.at/
- IT Administrator

Sep 2007 **Krammer Clinic Consulting GmbH**, Scheibbs, Österreich.
- Webseite: http://www.kcc.at/
- Programmierer

Jan–Jun 2006 **Präsenzdienst**, *Pionierbatallion 3*, Melk, Österreich.

Jul–Dez 2005 **Anton Haubenberger GmbH**, *Bäckerei*, Petzenkirchen, Österreich.
- Webseite: http://www.haubis.at
- MS SQL Server- und Netzwerkadministrator

2002–2010 **Rudolf Haubenberger GmbH**, *Kanalreinigung und -inspektion*, Kemmelbach, Österreich.
- Webseite: http://www.haubenberger.com
- IT Administrator

Vorträge

Mai 2014 **Munich Multicentric Variant Database**, *Illumina User Group Meeting 2014*, Heidelberg, Deutschland.

Jun 2013 **Somatic mutations in *ATP1A1* and *ATP2B3* lead to aldosterone-producing adenomas and secondary hypertension**, *European Human Genetics Conference 2013*, Paris, Frankreich.

Apr 2013 **Bioinformatic Aspects of Next Generation Sequencing**, 3^{rd} *Workshop of Genetic Epidemiology*, Grainau, Deutschland.

Sep 2012 **Functional effect of rare variants and mutation rate estimates derived from exome sequencing**, *Genome Informatics 2012*, Cambridge, Großbritannien.

Jun 2012 **Identification of de novo variants in 51 sporadic patients with unspecific severe intellectual disability (ID) and 20 controls by exome sequencing**, *European Human Genetics Conference 2012*, Nürnberg, Deutschland.

Apr 2012 **Analysis of Next Generation Sequencing Data**, 2^{nd} *Workshop of Genetic Epidemiology*, Grainau, Deutschland.

Fähigkeiten und Interessen

Spezielle Interessen	next generation sequencing, big data, high performance computing
Computer Fähigkeiten	Linux/Unix, Microsoft Windows, MS Office, LaTeX
Programmier-sprachen	Perl, Java, C, C++, VBA, PL/SQL, SQL, R
Datenbanken	Oracle, MS SQL Server, MySQL

Sprachen

Deutsch	**Muttersprache**
Englisch	**fließend**

Persönliche Interessen

Vereine	- Mitglied der biotechnologischen Studenteninitiative (btS) e.V. von 2010 bis 2014. Organisation verschiedener Veranstaltungen für Lifescience Studenten, unter anderem die Berufsinformationsmesse ScieCon
Sport	- Schifahren - Fitness
Verschiedenes	- Jagd - Lesen - Snooker

Literatur

[1] M Auer-Grumbach, H Bode, T R Pieber, M Schabhuttl, D Fischer, R Seidl, E Graf, **T Wieland**, R Schuh, G Vacariu, F Grill, V Timmerman, T M Strom, and T Hornemann. Mutations at Ser331 in the HSN type I gene SPTLC1 are associated with a distinct syndromic phenotype. *Eur J Med Genet*, 56(5):266–269, 2013.

[2] C Beetz, T R Pieber, N Hertel, M Schabhuttl, C Fischer, S Trajanoski, E Graf, S Keiner, I Kurth, **T Wieland**, R E Varga, V Timmerman, M M Reilly, T M Strom, and M Auer-Grumbach. Exome sequencing identifies a REEP1 mutation involved in distal hereditary motor neuropathy type V. *Am. J. Hum. Genet.*, 91(1):139–145, July 2012.

[3] F Beleggia, Y Li, J Fan, N H Elcio?lu, E Toker, **T Wieland**, I H Maumenee, N A Akarsu, T Meitinger, T Strom, R Lang, and B Wollnik. CRIM1 haploinsufficiency causes defects in eye development in human and mouse. *Hum. Mol. Genet.*, January 2015.

[4] F Beuschlein, S Boulkroun, A Osswald, **T Wieland**, H N Nielsen, U D Lichtenauer, D Penton, V R Schack, L Amar, E Fischer, A Walther, P Tauber, T Schwarzmayr, S Diener, E Graf, B Allolio, B Samson-Couterie, A Benecke, M Quinkler, F Fallo, P F Plouin, F Mantero, T Meitinger, P Mulatero, X Jeunemaitre, R Warth, B Vilsen, M C Zennaro, T M Strom, and M Reincke. Somatic mutations in ATP1A1 and ATP2B3 lead to aldosterone-producing adenomas and secondary hypertension. *Nat. Genet.*, 45(4):440–444, April 2013.

[5] F Beuschlein, M Fassnacht, G Assie, D Calebiro, C A Stratakis, A Osswald, C L Ronchi, **T Wieland**, S Sbiera, F R Faucz, K Schaak, A Schmittfull, T Schwarzmayr, O Barreau, D Vezzosi, M Rizk-Rabin, U Zabel, E Szarek, P Salpea, A Forlino, A Vetro, O Zuffardi, C Kisker, S Diener, T Meitinger, M J Lohse, M Reincke, J Bertherat, T M Strom, and B Allolio. Constitutive activation of PKA catalytic subunit in adrenal Cushing's syndrome. *N. Engl. J. Med.*, 370(11):1019–1028, 2014.

[6] N C Bramswig, H J Ludecke, Y Alanay, B Albrecht, A Barthelmie, K Boduroglu, D Braunholz, A Caliebe, K H Chrzanowska, J C Czeschik, S Endele, E Graf, E Guillen-Navarro, P O Kiper, V Lopez-Gonzalez, I Parenti, J Pozojevic, G E Utine, **T Wieland**, F J Kaiser, B Wollnik, T M Strom, and D Wieczorek. Exome sequencing unravels unexpected differential diagnoses in individuals with the tentative diagnosis of Coffin-Siris and Nicolaides-Baraitser syndromes. *Hum. Genet.*, February 2015.

[7] K Burk, F J Kaiser, S Tennstedt, L Schols, F R Kreuz, **T Wieland**, T M Strom, T Buttner, R Hollstein, D Braunholz, J Plaschke, G Gillessen-Kaesbach, and C Zuhlke. A novel missense mutation in CACNA1A evaluated by in silico protein modeling is associated with non-episodic spinocerebellar ataxia with slow progression. *Eur J Med Genet*, January 2014.

[8] L Crotti, C N Johnson, E Graf, G M De Ferrari, B F Cuneo, M Ovadia, J Papagiannis, M D Feldkamp, S G Rathi, J D Kunic, M Pedrazzini, **T Wieland**, P Lichtner, B M Beckmann, T Clark, C Shaffer, D W Benson, S Kaab, T Meitinger, T M Strom, W J Chazin, P J Schwartz, and A L George. Calmodulin mutations associated with recurrent cardiac arrest in infants. *Circulation*, 127(9):1009–1017, 2013.

[9] K Danhauser*, SW Sauer*, TB Haack*, **T Wieland***, C Staufner, E Graf, J Zschocke, T M Strom, T Traub, J G Okun, T Meitinger, G F Hoffmann, H Prokisch, and S Kolker. DHTKD1 mutations cause 2-aminoadipic and 2-oxoadipic aciduria. *Am. J. Hum. Genet.*, 91(6):1082–1087, 2012.

[10] S Dusi, L Valletta, T B Haack, Y Tsuchiya, P Venco, S Pasqualato, P Goffrini, M Tigano, N Demchenko, **T Wieland**, T Schwarzmayr, T M Strom, F Invernizzi, B Garavaglia, A Gregory, L Sanford, J Hamada, C Bettencourt, H Houlden, L Chiapparini, G Zorzi, M A Kurian, N Nardocci, H Prokisch, S Hayflick, I Gout, and V Tiranti. Exome sequence reveals mutations in CoA synthase as a cause of neurodegeneration with brain iron accumulation. *Am. J. Hum. Genet.*, 94(1):11–22, January 2014.

[11] D Fischer, M Schabhuttl, **T Wieland**, R Windhager, T M Strom, and M Auer-Grumbach. A novel missense mutation confirms ATL3 as a gene for hereditary sensory neuropathy type 1. *Brain*, 137(Pt 7):e286, July 2014.

[12] A Freischmidt*, **T Wieland***, B Richter*, W Ruf*, V Schaeffer, K Muller, N Marroquin, F Nordin, A Hubers, P Weydt, S Pinto, R Press, S Millecamps, N Molko, E Bernard, C Desnuelle, M H Soriani, J Dorst, E Graf, U Nordstrom, M S Feiler, S Putz, T M Boeckers, T Meyer, A S Winkler, J Winkelman, M de Carvalho, D R Thal, M Otto, T Brannstrom, A E Volk, P Kursula, K M Danzer, P Lichtner, I Dikic, T Meitinger, A C Ludolph, T M Strom, P M Andersen, and J H Weishaupt. Haploinsufficiency of TBK1 causes familial ALS and fronto-temporal dementia. *Nat. Neurosci.*, 2015.

[13] X Gai, D Ghezzi, M A Johnson, C A Biagosch, H E Shamseldin, T B Haack, A Reyes, M Tsukikawa, C A Sheldon, S Srinivasan, M Gorza, L S Kremer, **T Wieland**, T M Strom, E Polyak, E Place, M Consugar, J Ostrovsky, S Vidoni, A J Robinson, L J Wong, N Sondheimer, M A Salih, E Al-Jishi, C P Raab, C Bean, F Furlan, R Parini, C Lamperti, J A Mayr, V Konstantopoulou, M Huemer, E A Pierce, T Meitinger, P Freisinger, W Sperl, H Prokisch, F S Alkuraya, M J Falk, and M Zeviani. Mutations in FBXL4, encoding a mitochondrial protein, cause early-onset mitochondrial encephalomyopathy. *Am. J. Hum. Genet.*, 93(3):482–495, September 2013.

[14] L Greger, J Su, J Rung, P G Ferreira, T Lappalainen, E T Dermitzakis, A Brazma, M Sammeth, M R Friedlander, P A Hoen, J Monlong, M A Rivas, M Gonzalez-Porta, N Kurbatova, T Griebel, M Barann, **T Wieland**, M van Iterson, J Almlof, P Ribeca, I Pulyakhina, D Esser, T Giger, A Tikhonov, M Sultan, G Bertier, D G MacArthur, M Lek, E Lizano, H P Buermans, I Padioleau, T Schwarzmayr, O Karlberg, H Ongen, S Beltran, M Gut, K Kahlem, V Amstislavskiy, M Pirinen, S B Montgomery, P Donnelly, M I McCarthy, P Flicek, T M Strom, H Lehrach, S Schreiber, R Sudbrak, A Carracedo, S E Antonarakis, R Hasler, A C Syvanen, G J van Ommen, T Meitinger, P Rosenstiel, R Guigo, I G Gut, and X Estivill. Tandem RNA chimeras contribute to transcriptome diversity in human population and are associated with intronic genetic variants. *PLoS ONE*, 9(8):e104567, 2014.

[15] T B Haack, M Gorza, K Danhauser, J A Mayr, B Haberberger, **T Wieland**, L Kremer, V Strecker, E Graf, Y Memari, U Ahting, R Kopajtich, S B Wortmann, R J Rodenburg, U Kotzaeridou, G F Hoffmann, W Sperl, I Wittig, E Wilichowski, G Schottmann, M Schuelke, B Plecko, U Stephani, T M Strom, T Meitinger,

H Prokisch, and P Freisinger. Phenotypic spectrum of eleven patients and five novel MTFMT mutations identified by exome sequencing and candidate gene screening. *Mol. Genet. Metab.*, 111(3):342–352, 2014.

[16] T B Haack, P Hogarth, M C Kruer, A Gregory, **T Wieland**, T Schwarzmayr, E Graf, L Sanford, E Meyer, E Kara, S M Cuno, S I Harik, V H Dandu, N Nardocci, G Zorzi, T Dunaway, M Tarnopolsky, S Skinner, S Frucht, E Hanspal, C Schrander-Stumpel, D Heron, C Mignot, B Garavaglia, K Bhatia, J Hardy, T M Strom, N Boddaert, H H Houlden, M A Kurian, T Meitinger, H Prokisch, and S J Hayflick. Exome sequencing reveals de novo WDR45 mutations causing a phenotypically distinct, X-linked dominant form of NBIA. *Am. J. Hum. Genet.*, 91(6):1144–1149, 2012.

[17] T B Haack, R Kopajtich, P Freisinger, **T Wieland**, J Rorbach, T J Nicholls, E Baruffini, A Walther, K Danhauser, F A Zimmermann, R A Husain, J Schum, H Mundy, I Ferrero, T M Strom, T Meitinger, R W Taylor, M Minczuk, J A Mayr, and H Prokisch. ELAC2 mutations cause a mitochondrial RNA processing defect associated with hypertrophic cardiomyopathy. *Am. J. Hum. Genet.*, 93(2):211–223, August 2013.

[18] T B Haack, C Makowski, Y Yao, E Graf, M Hempel, **T Wieland**, U Tauer, U Ahting, J A Mayr, P Freisinger, H Yoshimatsu, K Inui, T M Strom, T Meitinger, A Yonezawa, and H Prokisch. Impaired riboflavin transport due to missense mutations in SLC52A2 causes Brown-Vialetto-Van Laere syndrome. *J. Inherit. Metab. Dis.*, 35(6):943–948, November 2012.

[19] S Jansen, B Aigner, H Pausch, M Wysocki, S Eck, A Benet-Pages, E Graf, **T Wieland**, T M Strom, T Meitinger, and R Fries. Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. *BMC Genomics*, 14:446, 2013.

[20] R Kopajtich, T J Nicholls, J Rorbach, M D Metodiev, P Freisinger, H Mandel, A Vanlander, D Ghezzi, R Carrozzo, R W Taylor, K Marquard, K Murayama, **T Wieland**, T Schwarzmayr, J A Mayr, S F Pearce, C A Powell, A Saada, A Ohtake, F Invernizzi, E Lamantea, E W Sommerville, A Pyle, P F Chinnery, E Crushell, Y Okazaki, M Kohda, Y Kishita, Y Tokuzawa, Z Assouline, M Rio, F Feillet, B de Camaret, D Chretien, A Munnich, B Menten, T Sante, J Smet, L Regal, A Lorber, A Khoury, M Zeviani, T M Strom, T Meitinger, E S Bertini, R Van Coster, T Klopstock, A Rotig, T B Haack, M Minczuk, and H Prokisch. Mutations in GTPBP3 Cause a Mitochondrial Translation Defect Associated with Hypertrophic Cardiomyopathy, Lactic Acidosis, and Encephalopathy. *Am. J. Hum. Genet.*, 95(6):708–720, 2014.

[21] C Kornblum, T J Nicholls, T B Haack, S Scholer, V Peeva, K Danhauser, K Hallmann, G Zsurka, J Rorbach, A Iuso, **T Wieland**, M Sciacco, D Ronchi, G P Comi, M Moggio, C M Quinzii, S DiMauro, S E Calvo, V K Mootha, T Klopstock, T M Strom, T Meitinger, M Minczuk, W S Kunz, and H Prokisch. Loss-of-function mutations in MGME1 impair mtDNA replication and cause multisystemic mitochondrial disease. *Nat. Genet.*, 45(2):214–219, February 2013.

[22] A Kuechler, M H Willemsen, B Albrecht, C A Bacino, D W Bartholomew, H van Bokhoven, M J van den Boogaard, N Bramswig, C Buttner, K Cremer, J C Czeschik,

H Engels, K van Gassen, E Graf, M van Haelst, W He, J S Hogue, M Kempers, D Koolen, G Monroe, S de Munnik, M Pastore, A Reis, M S Reuter, D H Tegay, J Veltman, G Visser, P van Hasselt, E E Smeets, L Vissers, **T Wieland**, W Wissink, H Yntema, A M Zink, T M Strom, H J Ludecke, T Kleefstra, and D Wieczorek. De novo mutations in beta-catenin (CTNNB1) appear to be a frequent cause of intellectual disability: expanding the mutational and clinical spectrum. *Hum. Genet.*, 134(1):97–109, January 2015.

[23] A Kuechler, A M Zink, **T Wieland**, H J Ludecke, K Cremer, L Salviati, P Magini, K Najafi, C Zweier, J C Czeschik, S Aretz, S Endele, F Tamburrino, C Pinato, M Clementi, J Gundlach, C Maylahn, L Mazzanti, E Wohlleber, T Schwarzmayr, R Kariminejad, A Schlessinger, D Wieczorek, T M Strom, G Novarino, and H Engels. Loss-of-function variants of SETD5 cause intellectual disability and the core phenotype of microdeletion 3p25.3 syndrome. *Eur. J. Hum. Genet.*, August 2014.

[24] T Lappalainen, M Sammeth, M R Friedlander, P A 't Hoen, J Monlong, M A Rivas, M Gonzalez-Porta, N Kurbatova, T Griebel, P G Ferreira, M Barann, **T Wieland**, L Greger, M van Iterson, J Almlof, P Ribeca, I Pulyakhina, D Esser, T Giger, A Tikhonov, M Sultan, G Bertier, D G MacArthur, M Lek, E Lizano, H P Buermans, I Padioleau, T Schwarzmayr, O Karlberg, H Ongen, H Kilpinen, S Beltran, M Gut, K Kahlem, V Amstislavskiy, O Stegle, M Pirinen, S B Montgomery, P Donnelly, M I McCarthy, P Flicek, T M Strom, H Lehrach, S Schreiber, R Sudbrak, A Carracedo, S E Antonarakis, R Hasler, A C Syvanen, G J van Ommen, A Brazma, T Meitinger, P Rosenstiel, R Guigo, I G Gut, X Estivill, E T Dermitzakis, E Dermitzakis, S Antonarakis, A Palotie, J F Deleuze, H Lerach, I Gut, U Gyllensten, H Brunner, J Veltman, P A Hoen, A Cambon-Thomsen, J Mangion, D Bentley, and A Hamosh. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, September 2013.

[25] U D Lichtenauer, G Di Dalmazi, E P Slater, **T Wieland**, A Kuebart, A Schmittfull, T Schwarzmayr, S Diener, D Wiese, W E Thasler, M Reincke, T Meitinger, M Schott, M Fassnacht, D K Bartsch, T M Strom, and F Beuschlein. Frequency and clinical correlates of somatic Ying Yang 1 mutations in sporadic insulinomas. *J. Clin. Endocrinol. Metab.*, page jc20151100, 2015.

[26] J A Mayr, T B Haack, E Graf, F A Zimmermann, **T Wieland**, B Haberberger, A Superti-Furga, J Kirschner, B Steinmann, M R Baumgartner, I Moroni, E Lamantea, M Zeviani, R J Rodenburg, J Smeitink, T M Strom, T Meitinger, W Sperl, and H Prokisch. Lack of the mitochondrial protein acylglycerol kinase causes Sengers syndrome. *Am. J. Hum. Genet.*, 90(2):314–320, February 2012.

[27] A Rauch*, D Wieczorek*, E Graf*, **T Wieland***, S Endele, T Schwarzmayr, B Albrecht, D Bartholdi, J Beygo, N Di Donato, A Dufke, K Cremer, M Hempel, D Horn, J Hoyer, P Joset, A Ropke, U Moog, A Riess, C T Thiel, A Tzschach, A Wiesener, E Wohlleber, C Zweier, A B Ekici, A M Zink, A Rump, C Meisinger, H Grallert, H Sticht, A Schenck, H Engels, G Rappold, E Schrock, P Wieacker, O Riess, T Meitinger, A Reis, and T M Strom. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet*, 380(9854):1674–1682, November 2012.

[28] M Reincke, S Sbiera, A Hayakawa, M Theodoropoulou, A Osswald, F Beuschlein, T Meitinger, E Mizuno-Yamasaki, K Kawaguchi, Y Saeki, K Tanaka, **T Wieland**,

E Graf, W Saeger, C L Ronchi, B Allolio, M Buchfelder, T M Strom, M Fassnacht, and M Komada. Mutations in the deubiquitinase gene USP8 cause Cushing's disease. *Nat. Genet.*, 47(1):31–38, January 2015.

[29] M Schabhuttl, **T Wieland**, J Senderek, J Baets, V Timmerman, P De Jonghe, M M Reilly, K Stieglbauer, E Laich, R Windhager, W Erwa, S Trajanoski, T M Strom, and M Auer-Grumbach. Whole-exome sequencing in patients with inherited neuropathies: outcome and challenges. *J. Neurol.*, 2014.

[30] P A 't Hoen, M R Friedlander, J Almlof, M Sammeth, I Pulyakhina, S Y Anvar, J F Laros, H P Buermans, O Karlberg, M Brannvall, J T den Dunnen, G J van Ommen, I G Gut, R Guigo, X Estivill, A C Syvanen, E T Dermitzakis, T Lappalainen, S E Antonarakis, A Brazma, P Flicek, S Schreiber, P Rosenstiel, T Meitinger, T M Strom, H Lehrach, R Sudbrak, and A Carracedo. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.*, 31(11):1015–1022, November 2013.

[31] D Wieczorek, W G Newman, **T Wieland**, T Berulava, M Kaffe, D Falkenstein, C Beetz, E Graf, T Schwarzmayr, S Douzgou, J Clayton-Smith, S B Daly, S G Williams, S S Bhaskar, J E Urquhart, B Anderson, J O'Sullivan, O Boute, J Gundlach, J C Czeschik, A J van Essen, F Hazan, S Park, A Hing, A Kuechler, D R Lohmann, K U Ludwig, E Mangold, L Steenpass, M Zeschnigk, J R Lemke, C M Lourenco, U Hehr, E C Prott, M Waldenberger, A C Bohmer, B Horsthemke, R T O'Keefe, T Meitinger, J Burn, H J Ludecke, and T M Strom. Compound Heterozygosity of Low-Frequency Promoter Deletions and Rare Loss-of-Function Mutations in TXNL4A Causes Burn-McKeown Syndrome. *Am. J. Hum. Genet.*, 95(6):698–707, 2014.

[32] J Winkelmann, L Lin, B Schormair, B R Kornum, J Faraco, G Plazzi, A Melberg, F Cornelio, A E Urban, F Pizza, F Poli, F Grubert, **T Wieland**, E Graf, J Hallmayer, T M Strom, and E Mignot. Mutations in DNMT1 cause autosomal dominant cerebellar ataxia, deafness and narcolepsy. *Hum. Mol. Genet.*, 21(10):2205–2210, 2012.

[33] M Zech, F Castrop, B Schormair, A Jochim, **T Wieland**, N Gross, P Lichtner, A Peters, C Gieger, T Meitinger, T M Strom, K Oexle, B Haslinger, and J Winkelmann. DYT16 revisited: exome sequencing identifies PRKRA mutations in a European dystonia family. *Mov. Disord.*, 29(12):1504–1510, 2014.

***** Authors contributed equally

München, 1. Oktober 2015

Anlage 5

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die bei der promotionsführenden Einrichtung bzw. Fakultät

Informatik

der TUM zur Promotionsprüfung vorgelegte Arbeit mit dem Titel:

Next-Generation Sequencing Data Analysis

in Bioinformatik

(Lehrstuhl bzw. Fachgebiet oder Klinik)

unter der Anleitung und Betreuung durch

Prof. Dr. Burkhard Rost

ohne sonstige Hilfe erstellt und bei der Abfassung nur die gemäß § 6 Abs. 6 und 7 Satz 2 angegebenen Hilfsmittel benutzt habe.

(X) Ich habe keine Organisation eingeschaltet, die gegen Entgelt Betreuerinnen und Betreuer für die Anfertigung von Dissertationen sucht, oder die mir obliegenden Pflichten hinsichtlich der Prüfungsleistungen für mich ganz oder teilweise erledigt.

(X) Ich habe die Dissertation in dieser oder ähnlicher Form in keinem anderen Prüfungsverfahren als Prüfungsleistung vorgelegt.

( ) Die vollständige Dissertation wurde in .....

veröffentlicht. Die promotionsführende Einrichtung.....
hat der Vorveröffentlichung zugestimmt.

( ) Ich habe den angestrebten Doktorgrad **noch nicht** erworben und bin **nicht** in einem früheren Promotionsverfahren für den angestrebten Doktorgrad endgültig gescheitert.

( ) Ich habe bereits am .....

bei der Fakultät für .....

der Hochschule .....

unter Vorlage einer Dissertation mit dem Thema .....

.....
die Zulassung zur Promotion beantragt mit dem Ergebnis:

.....

Die öffentlich zugängliche Promotionsordnung der TUM ist mir bekannt, insbesondere habe ich die Bedeutung von § 28 (Nichtigkeit der Promotion) und § 29 (Entzug des Doktorgrades) zur Kenntnis genommen. Ich bin mir der Konsequenzen einer falschen Eidesstattlichen Erklärung bewusst.

Mit der Aufnahme meiner personenbezogenen Daten in die Alumni-Datei bei der TUM bin ich

(X) einverstanden
( ) nicht einverstanden

München, den ....
Unterschrift