Freie Universität Berlin

Fachbereich Mathematik und Informatik der Freien Universität Berlin

Aspects of Quality Control for Next Generation Sequencing Data in Medical Genetics

Verena Heinrich

Eingereicht am: 16.08.2016

Disputation am: 22.11.2016

Dissertation zur Erlangung des akademischen Grades Doktor rerum naturalium (Dr. rer. nat.), eingereicht beim Fachbereich Mathematik und Informatik der Freien Universität Berlin

Erstgutachter: Prof. Dr. Martin Vingron

Zweitgutachter: Prof. Dr. Peter N. Robinson

I hereby declare that I have created this work completely on my own and used no other sources or tools than the ones listed, and that I have marked any citations accordingly. Furthermore I have not submitted this work in any earlier doctoral procedure.

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe. Desweiteren habe ich die vorliegende Arbeit nicht schon einmal in einem früheren Promotionsverfahren eingereicht.

Berlin, 16.08.2016
Verena Heinrich

Preface

Since the discovery of the double-helical structure of the *deoxyribonucleic acid* (*DNA*) by Watson and Crick in 1953 molecular biologists achieved a remarkable understanding of the mechanisms undergoing in the expression of genes. The invention of (*Sanger*) *DNA* sequencing in the 1970s and the decoding of the complete sequence of a humans genome at the end of the last century finally gave geneticists the possibility to get a more detailed look into the genetic code. Ever since then steadily increasing improvement in technology changed the pace of research but also new challenges have arisen regarding data analysis and interpretation.
Since more and more technical advancements based on high throughput sequencing were achieved in the paste decade, a broad range of applicability has been opened, including *de novo* or *whole exome sequencing* (*WES*) at comparable low prices. Further, exome screens have already emerged to an indispensable tool to discover genetic variations that appear in *Mendelian* disorders which are often characterized by a high phenotypic and genetic heterogeneity. The final identification of a pathogenic mutation amongst several thousands of benign variants often relies on filtering techniques that simply reduce the search space and approaches such as segregation or linkage analysis are used for prioritization.
Compared to traditional *Sanger* sequencing the results obtained by *next generation sequencing* (*NGS*) can be affected by a range of artefacts and platform-specific biases that arise during library

preparation and the actual sequencing process which have a great impact on the genotyping quality of the sequenced data. Especially the detection of *heterozygous* variants is associated with a considerably high error rate compared to the identification of *homozygous* mutations, even at comparable levels of sequence coverage. Consequently, it still remains a major challenge to make a trustworthy distinction between falsely assigned *heterozygous* variants and true *de novo* mutations although several methods for *quality control* (*QC*) were introduced during the last years.

Further, most of the current approaches include the simultaneous analysis of several unrelated individuals as for example in association tests or family members which is applied as standard procedure in the detection of *de novo* candidates. However, as these approaches rely on a correct assignment for sequence samples a robust method to check for the relatedness between them should be a constant part of *quality control* downstream of further analysis steps.

During my *Ph.D.* I concentrated on three different levels of *QC* in *WES* experiments, which also make up the main components of this work.

First I modelled the amplification of sequence fragments during library preparation of *WES* experiments as a stochastic *(Bienayme-) Galton-Watson* (*BGW*) branching process. The resulting variance of the distribution of *allele frequencies* (*AFs*) at *heterozygous* positions can be used to draw conclusions about how to reduce the stochastic fluctuations which originally arise from the amplification step and serve as an indicator of the quality of an *WES* experiment.

Further, to indicate the exome-wide accuray of a *WES* sample I developed a method that estimates the similarity between a sample and a *Reference* set of good quality sequenced by the *1000 genomes project* (*1KGP*) based on a metric that emphasizes rare variants.

Finally, a *likelihood ratio* (*LR*) based approach provides a robust

technique to infer relatedness from *WES* family data and can be used to clarify sample *mix-ups* that would otherwise corrupt further analysis strategies.
The results of all three approaches are summarized in the following publications:

1. V. Heinrich, J. Stange, T. Dickhaus, P. Imkeller, U. Krüger, S. Bauer, S. Mundlos, P.N. Robinson, J. Hecht, and P.M. Krawitz. The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. *Nucleic Acids Res.*, 40(6):2426–31, 2012. doi: https://dx.doi.org/10.1093%2Fnar%2Fgkr1073

2. V. Heinrich, T. Kamphans, J. Stange, D. Parkhomchuk, J. Hecht, T. Dickhaus, P.N. Robinson, and P.M. Krawitz. Estimating exome genotyping accuracy by comparing to data from large scale sequencing projects. *Genome Medicine*, 5(7):69, 2013. doi: https://dx.doi.org/10.1186%2Fgm473

3. V. Heinrich, T. Kamphans, S. Mundlos, P.N. Robinson, and P.M. Krawitz. A likelihood ratio-based method to predict exact pedigrees for complex families from next-generation sequencing data. *Bioinformatics*, 33(1):72–8, 2016. doi: https://doi.org/10.1093/bioinformatics/btw550

A part of this work [115] was also adapted for another approach which was developed by my colleague Zhu Na. She evaluated different matching strategies for the selection of suitable controls in *rare variant association studies* (*RVAS*) and I contributed to this study by the computation of distance matrices based on different metrics:

4. N. Zhu, V. Heinrich, T. Dickhaus, J. Hecht, P.N. Robinson, S. Mundlos, T. Kamphans, and P.M. Krawitz. Genome analysis Strategies to improve the performance of rare variant

association studies by optimizing the selection of controls. *Bioinformatics*, 31(22):3577–83, 2015

Further, during my research time I also contributed to other projects which are not part of this thesis:

5. Alexander Hruscha, Peter Krawitz, Alexandra Rechenberg, V. Heinrich, Jochen Hecht, Christian Haass, and Bettina Schmid. Efficient CRISPR/Cas9 genome editing with low off-target effects in zebrafish. *Development*, 140(5):4982–4987, 2013

6. T. Kamphans, P. Sabri, N. Zhu, V. Heinrich, S Mundlos, P.N. Robinson, D. Parkhomchuk, and P.M. Krawitz. Filtering for Compound Heterozygous Sequence Variants in Non-Consanguineous Pedigrees. *PLoS One*, 8(8):1–6, 2013

7. Peter M. Krawitz, Daniela Schiska, Ulrike Krger, Sandra Appelt, V. Heinrich, Dmitri Parkhomchuk, Bernd Timmermann, Jose M. Millan, Peter N. Robinson, Stefan Mundlos, Jochen Hecht, and Manfred Gross. Screening for single nucleotide variants, small indels and exon deletions with a next-generation sequencing based gene panel approach for usher syndrome. *Molecular Genetics & Genomic Medicine*, 2(5):393–401, 2014

8. N. Ehmke, A. Caliebe, R. Koenig, S.G. Kant, Z. Stark, D. Wieczorek, G. Gillessen-kaesbach, K. Hoff, A. Knaus, N. Zhu, V. Heinrich, C. Huber, I. Harabula, M. Spielmann, D. Horn, H. Manzke, and S. Mundlos. Homozygous and Compound-Heterozygous Mutations in TGDS Cause Catel-Manzke Syndrome. *Am J Hum Genet.*, 95(6):763–70, 2014

9. D. Emmerich, T. Zemojtel, J. Hecht, P. Krawitz, M. Spielmann, J. Kühnisch, K. Kobus, M. Osswald, V. Heinrich, P. Berlien, U. Müller, V. Mautner, K. Wimmer, P.N. Robinson, M. Vingron, S. Tinschert, S. Mundlos, and M. Kolanczyk. Somatic neurofibromatosis type 1 (NF1) inactivation events

in cutaneous neurofibromas of a single NF1 patient. *Eur J Hum Genet.*, 23(6):870–3, 2015

10. H. Lademann, B. Gerber, D.M. Olbertz, E. Darvin, L. Stauf, K. Ueberholz, V. Heinrich, J. Lademann, and V. Briese. Non-Invasive Spectroscopic Determination of the Antioxidative Status of Gravidae and Neonates. *Skin Pharmacol Physiol.*, 28(4):189–95, 2015

11. D. Lupianez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J.M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S.A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, and S. Mundlos. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, 161(5):1012–25, 2015

12. M. Franke, D.M. Ibrahim, G. Andrey, W. Schwarzer, V. Heinrich, R. Schöpflin, K. Kraft, R. Kempfer, I. Jercovic, W.L. Chan, M. Spielmann, B. Timmermann, I. Kurth, P. Cambiaso, O. Zuffardi, G. Houge, L. Lambie, F. Brancati, A. Pombo, M. Vingron, F. Spitz, and S. Mundlos. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538(7624):265–9, 2016

Contents

List of Figures

List of Tables

List of Algorithms

1 | Introduction

1.1 Biological Background

1.1.1 Next Generation Sequencing

DNA Sequencing

The general process of *DNA* sequencing can be describes as the determination of the order of the four nucleotides or bases *adenine* (*A*), *guanine* (*G*), *thymine* (*T*) and *cytosine* (*C*) within a *DNA* molecule. Over the last years several technologies for sequencing have evolved and became indispensable for basic biological research and in applied fields such as medical diagnostics.

From Sanger Sequencing to Next Generation Sequencing

Up until recently most *DNA* sequencing was performed by applying the chain termination method developed by Frederick Sanger, which is still considered as the gold standard regarding sequencing accuracy [104]. In this approach chemically altered *nucleotides* cause the copying process of a *DNA* strand to stop each time it is incorporated into the growing chain. A repetition of this method with all four bases provides the exact information of all breakpoints and can be joint to build a complete *DNA* sequence.

In the last two decades limitations regarding costs and throughput led to a fundamental shift away from the application of classical *Sanger* sequencing for genome analysis which is now often replaced by so-called *next generation sequencing* (*NGS*) techniques which represent the next phase in the evolution of *DNA* sequencing technologies, also referred to as 2^{nd} generation sequencing methods [20]. To date, *NGS* techniques represent the next phase in the evolution of DNA sequencing technology at dramatically reduced cost compared to traditional *Sanger* sequencing.

Miniaturazation and massive parallelization yielded new sequencing platforms that are now referred to as *NGS* methods, which are able to generate large amounts of high-throughput data in a considerably less expensive way compared to the *Sanger* method [56]. This enabled the resequencing of human genomes which was first realized by Craig Venter in 2001 [120] and initialized a new field of research, where the discovering and understanding of variations in the *DNA* and their influence on health and diseases became one of the major tasks in clinical diagnostics. Additionally, the targeting of specific subsets of the human genome, for example the protein-coding region, the exome, is used to lower sequencing costs and to increase the sequence coverage of regions of interest. *whole exome sequencing* (*WES*) is a common approach in genetic diagnostic as it is especially effective to identify diseases associated with rare *Mendelian* variants. This diseases are often related to very rare variants which are present in a small set of individuals and most likely occur in the protein coding sequence of a genome.

Polymerase Chain Reaction

The capturing of a targeted subset requires an amplification step during library preparation that is often realized by *polymerase chain reaction* (*PCR*). The *PCR* is a technology in molecular biology which is used to generate thousands to millions of copies of a small initial number of *DNA* segments. During the reaction repeated cycles of *DNA* replication is performed in a tube that contains free nucleotides, DNA replication enzymes, short nucleotide sequences (*primers*), and template *DNA* molecules. Introduced in 1986 by Mullis et al. [87] *PCR* has quickly developed into a common technique in research labs for a variety of applications and is now among others a crucial part in the amplification process in *next generation sequencing*.

Usually the experimental setup of a *PCR* reaction consist of $25-40$ [53] repeats of temperature changes, referred to as cycles. Each cycle consists of usually three steps of changing temperatures. Initially, the double-stranded *DNA* template is denaturated at high temperatures, resulting in single strands . In a second step, the temperature is decreased and primers anneal to the now single stranded *DNA* template. The polymerase binds to the primer-template hybrid and begins to synthesize new strands of *DNA*.

At the end of the first cycle, each double stranded *DNA* molecule consist of a new and an old *DNA* strand. The *PCR* then continues with additional cycles that repeat the aforementioned steps where the newly synthesized DNA segments serve as templates for the following cycles (illustrated in Figure 1.1).

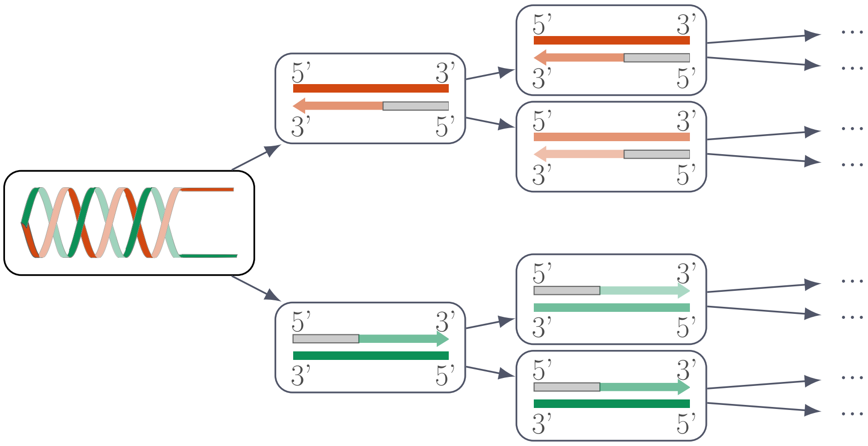


Figure 1.1: *Illustration of a polymerase chain reaction.* PCR consists of of a series of usually 25 – 40 *cycles*, in which the double stranded DNA is denatured, primers (gray bars) are annealed and a new strand is synthesized.

Different Sequencing Technologies

The challenge of getting fast, cheap and accurate genome information has inspired the development of novel sequencing technologies and the diversity of NGS platforms is growing rapidly. The major commercial 2^{nd} generation sequencing technologies which came into existence after the success of the *human genome project* (*HGP*) include *Roche/454* [1], *Illumina* [6] and *SOLiD* [14]. Each of these platforms comprises a number of methods including template preparation, sequencing and imaging followed by data analysis and comprehensive overviews are for instance given by Mardis et al. [80] and Liu et al. [78]).

3^{rd} Generation Sequencing

Beyond the 2^{nd} generation sequencing technologies, so called 3^{rd} generation instruments have already arisen in the last few years such as *Helicos Heliscope* [5], *Pacific Biosciences SMRT* [12] and *Oxford Nanopore* [11]. The major difference between the technologies of both generations is that the initial step of DNA amplification

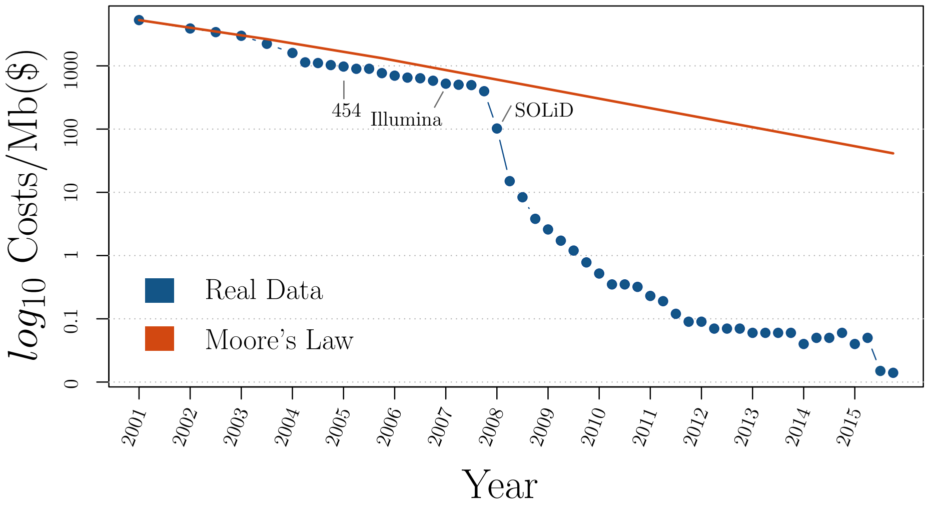


Figure 1.2: *Costs per mb of DNA sequenced in the last* 15 *years.*] The *NHGRI* (data source: *http://www.genome.gov/sequencingcosts/*) has tracked the costs associated with *DNA* sequencing performed at different sequencing centers in the last 15 years. The plot shows the massive decrease of sequencing costs over the last decade, which is mainly influenced by the increase of commercial vendors since 2005. This finding is also supported by observing the hypothetical data process defined by *Moore's Law* (red line), which is a commonly used rule to measure the growth rate of digital technology. It can be clearly seen that the rate of the reduction of the sequencing costs per *Mb* passes this rule starting from approximately 2008.

became unnecessary in the latest technologies [20].

Moore's Law

The rate of growth in *DNA* sequencing generation and the associated decrease of costs has already passed the expected approximation, commonly described by *Moore's Law* paradigm [25], which is usually used to estimate the expansion rate of digital technology such as the growth of hardware speed (see Figure 1.2). *Moore's Law* describes a driving force of technological growths and states that the rate of growths doubles every 2 years, which can be formularized as a function K, dependent on a time parameter t:

$$K(t) = K(t_0) * 2^t, \tag{1.1}$$

whereas $K(t)$ describes the technical entity, for example speed or cost, at time point t.

By different methods, each 2^{nd} generation technology implies the amplification of single short sequence *reads* of a fragment library by *PCR* followed by a sequencing reactions on the amplified *DNA* fragments. Since most of the data shown in this thesis are based on the *Illumina* *NGS* sequencing technology, I will describe the main steps of this method in the following.

Library Preparation and Cluster Generation

The fragment libraries are generated by annealing platform-specific adapters to blunt-ended fragments that were prepared by random fragmentation of the *DNA* as illustrated in Figure 1.3 **a) - c**). The resulting adapter-ligated fragments are then amplified via *PCR* and gel purified. For the following generation of clusters, the fragment library is loaded on a flow cell, where each single stranded *DNA* fragment is annealed to another surface-bound complementary oligonucleotide (Figure 1.3 **d)-f)**).

Sequencing-By-Synthesis

The Illumina system utilizes a sequencing-by-synthesis approach in which all four bases are fluorescently labeled. 3-OH blocked nucleotides are added simultaneously to the flow cell channels, along with *DNA* polymerase, for incorporation into the oligo-primed cluster strands, which are then extended by one nucleotide. After an imaging step, the $3'$ blocking group is chemically removed to prepare each strand for the next incorporation by *DNA* polymerase, which continues for a specific number of cycles which is usually in the range of $50 - 300$ [6].

Coverage Biases in NGS Platforms

Next generation sequencing determines the order of nucleotides within each *DNA* fragment and allows the detection of genetic variations. However, a limitation in *NGS* techniques is the high frequency with which alleles are wrongly weighted due to artefacts introduced in the sequencing process. Sequence dependent deviations in quality as well as several types of biases that can occur during sample preparation lead to a partly strong deviation in coverage from the expected uniformly distributed sequencing *reads* across

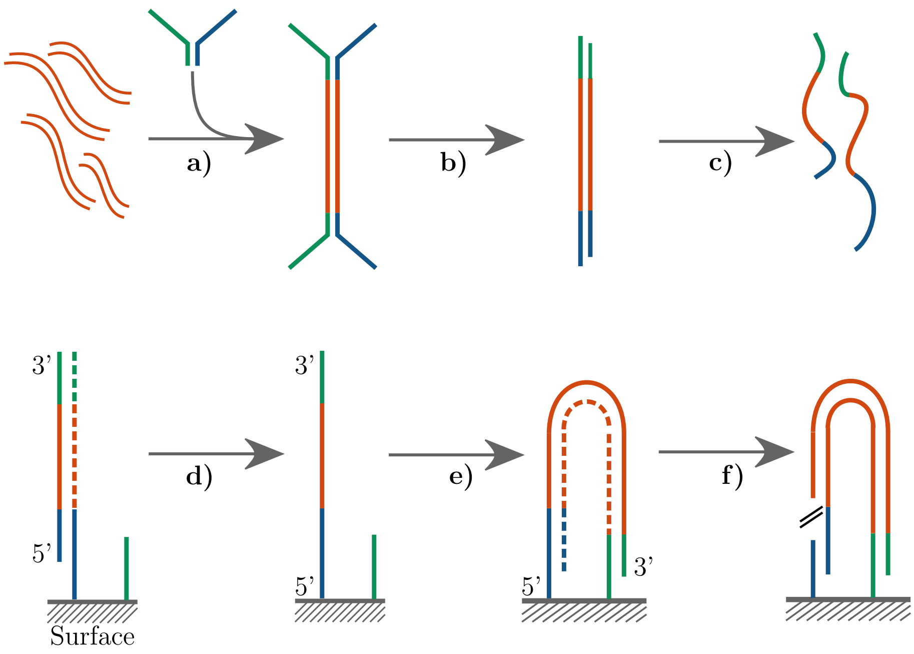


Figure 1.3: *Preparation of sequencing samples.*(adapted from [22]) Genomic *DNA* is randomly divided into small fragments of a few hundred base pairs. The double stranded fragments are joined to a pair of oligonucleotides in a forked adaptor configuration (**a)**). The obtained ligands are then copied by *PCR* using two oligonucleotide primers, resulting in blunt-ended double-strands with a different adaptor sequence on each end (**b)**), which are then denaturated into single strand fragments (**c)**). Primed from the $3'$-end of complementary surface-bound oligonucleotides, a new fragment is copied from the original strand by polymerase-directed single base extension. The original strand, that served as a template is then removed by denaturation, whereby the complementary strand is left as template for the next sequencing reaction (**d)**). The adaptor sequence at the $3'$-end of each copied strand is annealed to another surface-bound complementary oligonucleotide, forming a bridge and generating a new synthesis template for a second strand (**e)**). The resulting double strand is then denatured, generating a new single-stranded template for the next bridge amplification (**f)**). The cycle of annealing, extension and denaturation is repeated several times, resulting in clusters of surface-bounded *DNA* strands of approximately one million copies of the original single *DNA* molecule that initiated the amplification process.

Table 1.1: *Comparison of technical specifications of different sequencing platforms.*

	Illumina	SOLiD	454
Max. *read* Length (*bp*)	2×150 [6]	2×50 [14]	400 [1]
# Reads/Run ($\times 10^6$)	$3.4 - 3000$	$840 - 1410$	$0.1 - 1$
Output/Run (*Gb*) [50]	$1.2 - 600$	$71.4 - 155.1$	$0.5 - 9$
Run Time [50]	$26h - 14d$	$8 - 12d$	$10 - 20h$
Error Frequency (%) [46]	0.1	$2 * 0.01$	1
Most Frequent Error Type [46]	Single Nucleotide Substitution	A↔T bias	*INDELs*
Min. Costs/*Mb* ($) [50]	0.04	0.07	7

the genome [46]. Moreover, sequencing and imaging biases show specific pattern for the respective *NGS* platforms. For instance, *Illumina* and *ABI Solid* sequencing platforms tend to introduce wrong nucleotides in regions with high or low *GC* content which leads to an uneven coverage or even entire gaps in the sequenced *read* distribution [102, 98, 40, 56]. Error profiles which are especially specific for *Illumina* sequencing platforms were described in several publications [102, 88, 40] involving inverted repeats, palindromic sequences and *GC* sequences. Each of the aforementioned patterns can influence the base elongation process during sequencing and leads to an uneven coverage profile. Hence, low quality *reads* are not randomly distributed, but are rather localized at specific mapping regions and should be treated with caution in further analysis steps.

Error Frequencies in Sequencing Platforms

Current *NGS* technologies have considerably higher error rates than traditional *Sanger* sequencing, which is the reason why *Sanger* is still considered as gold-standard regarding sequencing accuracy which is about 99.999%. As summarized in Table 1.1, different *NGS* platforms are prone to various types of sequencing errors [46, 105] and different strategies of downstream analysis are required. The majority of errors that occur in *454* sequencing platforms are related to *insertions or deletions* (*INDELs*), whereby *reads* generated with *SOLiD* have a higher *A*↔*C* bias and the *Illumina* sequencing technology tends to introduce single nucleotide substitutions. During the sequencing-by-synthesis approach utilized in the *Illumina* sequencing technology, fluorescently labelled nucleotides are added for the polymerization, which are then illuminated by a red laser for *A* and *C* and a green laser for *G* and *T*. A strong correlation of the complementary bases that cause similar emission spectre of the fluorophores and limitations in signal separation lead then to false assignments of nucleotides in approximately 1 out of 1000 *base pair*s (*bp*s) [105].

1.1.2 The Human Reference Genome

A fundamental first step in bioinformatics processing of high throughput sequencing data is to compare the sequenced *reads* to a suitable high quality reference assembly which is a haploid representation of an organism's genome, also referred to as *reference genome* (*Ref*) in the following. Such an assembly is represented as a series of characters over the nucleotide alphabet, referred to as *contigs*, partitioned into a set of *scaffolds* which is an ordered and oriented set of continuous and gap-free *contigs*. In case of an uncertain *DNA* sequence between *contigs*, wild-card characters are inserted to fill the gaps which are typically encoded as *'N'*.

The accuracy of a reference genome is crucial for the analysis and annotation of the sequenced data which is an even more complicated challenge considering the high population diversity in the human genome. All reference genomes are based on one of the standard assemblies, initially produced by the *human genome project* (*HGP*). Nowadays all assemblies are produced by the *international human genome sequencing consortium* (*IHGSC*) and all released versions differ in their length and quality as summarized in Table 1.2. Fur-

International Human Genome Sequencing Consortium thermore, several minor variants were released by the *national center for biotechnology information* (*NCBI*) that additionally integrated the mitochondrial sequence into the assembly. The *IHGSC* constructed a map of the whole human genome uitlizing a *clone-based* approach, generating an overlapping series of clones that cover the entire genome sequence whereas each clone represents a haplotype from one single donor [84].

Table 1.2: *Human reference genome build development.* Listed are major human reference assemblies as defined by the *GRC* and their main global statistics. The associated alternative descriptions of the different builds, released by the *UCSC*, are displayed in brackets. The table is based on extracted data from the *NCBI* website [9].
In summary, the *GRCh38* assembly has the best global statistics compared to earlier releases: It is interrupted by less gaps and has a higher scaffold N50, which is a measure of continuity (50% of the bases reside in a scaffold of this length or greater).

Description	Year	Total Length (*bp*)	Total Gaps (*bp*)	Scaffold N50
GRCh38	2015	3,231,297,122	161,368,151	67,794,873
GRCh37 (hg19)	2013	3,234,834,689	243,140,514	46,395,641
NCBI36 (hg18)	2006	3,104,054,490	222,405,369	38,509,590

A sequenced assembly is considered as highly accurate, with a high per-base accuracy of 99.99%, which translates to one expected error every 10000 bases in a sequence of about 3 billion nucleotides. The newest assembly version comprises a total consensus sequence length of ~ 3.23 billion base pairs (Table 1.2) and accounts for genomic regions of excess diversity, such as repetitive sequences in the centromer region, by including additional scaffold sequences and alternate loci, which was managed by introducing gaps in the earlier version *GRCh37*. Nevertheless, most variant calling algorithms expect a haploid assembly model and thus penalize *reads* that have more than one single location in the genome [29].

Population Bias in the Human Genome

The generation of such an assembly is mostly based on the *DNA* of several anonymous donors and thus represents a mixture of *haplotypes* rather than the genome of a single individual. The version of the human genome assembly produced by *IHGSC*, *GRCh37*, which is used in this work, is also based on sequences from a combination of about a few dozen individuals, although most of the *haplotypes* are derived from donors from Buffalo, New York, [41] and thus introduces a population bias that neglects the global wide diversity of the human genome. In the future, a graph-based alternative illustration of a human assembly would represent the population-based human genome in a more intuitive way rather than a linear sequence [29]. However an adaption of this model to existing analysis pipelines will still take substantial efforts and for now other detours have to be considered to account for population biases in the human genome.

1.1.3 Genetic Variability in the Human Genome

In diploid organisms each somatic cell contains two copies of a chromosome that is not involved in sex determination. The different forms of such a copy at each genomic locus are referred to as *alleles* whereas one has been inherited from the mother and one from the father respectively. Within a population many different alleles can exist for one locus and their frequency is used to analyse the genetic variation within a population. In one individual a genomic position is described as *homozygous* for an allele, if the two copies are the same and *heterozygous* otherwise. This constitution of such a position is then referred to as *genotype* (*GT*).

Hardy-Weinberg Equilibrium

Both the *allele frequency* (*AF*) and the genotype frequency within a population are central to population genetics as they provide insight into evolutionary changes over time. Hardy and Weinberg formulated a general law for *bi-allelic* loci, referred to as *Hardy-*

Weinberg equilibrium (*HWE*), that states that both allele and genotype frequency will remain in equilibrium between generations if no other evolutionary influences take place such as mutations or selection [110]. A deviation of this ideal condition can be an indicator of selection or a genotyping error [15]. To give an example, the concept of the *HWE* can be applied to compare the observed and expected level of *heterozygosity* which can be used as an indicator of the level of consanguinity within a population.

In consideration of the overall genetic diversity the term *human genome* must be assessed as an artificial construction since there are not two humans in the world with the exact same *DNA* sequence. A comparison between any two randomly chosen individuals would lead into roughly 1 difference in 1000 genomic positions [65]. But also closely related individuals differ from each other due to novel mutations that did not occur in the parents and *structural variation*s (*SV*s) that appear during development. Any form of mutation can either arise in the *germ line* which can be passed to the next generation or in the *somatic cell line* which is then referred to as *de novo* mutation.

Structural Variations

The term *SV*s comprises a number of different genomic rearrangements which span at least $1kb$ of the genome [48]. In particular this includes inversions, translocations or genomic imbalances such as duplications, deletions or insertions which are commonly referred to as *copy number variation*s (*CNV*s) (Figure 1.4 **a) - e)**). Additionally, medium-sized *INDELs*, ranging approximately $1 - 10000$ nucleotides [85], will play an essential role in personalized medicine as many of these map to functionally important sites within human genes [86].

Single Nucleotide Variant, SNV

The most frequent changes in the human genome are single *base pair* (*bp*) mutations which occur in about 4 billion positions per individual. If at least one *allele* differs from the reference genome at a specific location, this single base-pair mutation is called *single nucleotide variant* (*SNV*) which can be further characterized,

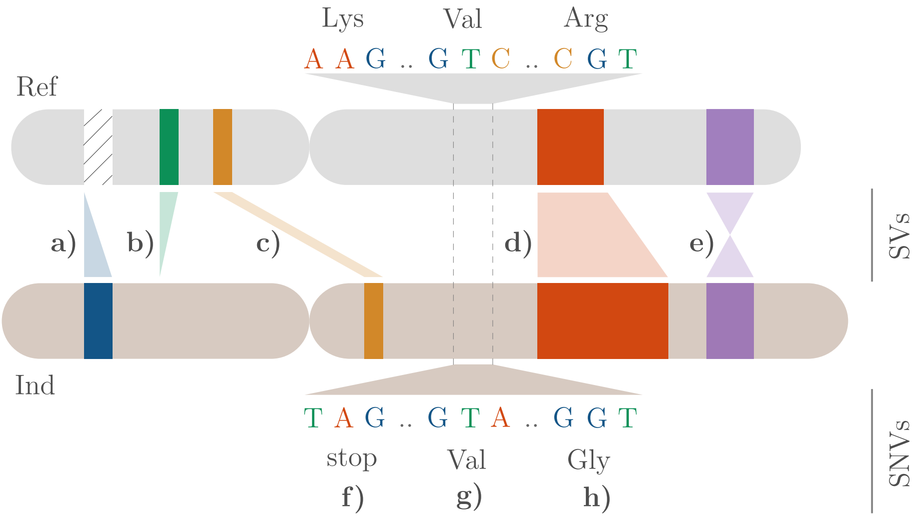


Figure 1.4: *Types of structural variations and single nucleotide variants.* Depicted are two chromosomes whereas one represents the *reference genome* (*Ref*) and the second originates from a single individual (*Ind*). All possible *structural variation*s (*SV*s) are described in **a) - e)** and the examples shown in **f) - h)** visualize different types of *single nucleotide variant*s (*SNV*s).
a) an insertion differs in one or more additional nucleotides from the reference genome. The contrary effect is seen in deletions (**b)**), where a part of the reference is not present in the sample's genome. **c)** a translocation describes the transfer from a chromosomal part to another location in the genome. Insertions, deletions and duplications (**d)**) are all *SV*s where the total amount of the *DNA* differs from the *DNA* in the *Ref* and are therefore referred to as *copy number variation*s (*CNV*s). **e)** an inversion describes an structural rearrangement in which a part of the reference genome is reversed. There are several types of *SNV*s, depending on the effect on the transcriptional level. Either the mutation leads to a *stop-codon* (**f)**), the mutation has no effect at all and is therefore called *synonymous* or *silent* (**g)**) or it results in a *non-synonymous* substitution, where a completely different amino acid is translated (**h)**).

depending on the effect on the translational level (Figure 1.4 **f) - h)**). A *synonymous* (or *silent*) substitution has no effect on the translated *amino acid* (*AA*), since several triplets of nucleotides encode for the same chemical construction, whereas *non-synonymous* mutations lead to a completely different *AA* or even a stop-codon (*nonsense* mutation).

Single Nucleotide Polymorphism, SNP

A variation from the reference genome that occurs commonly within a population in at least 1 out of 100 individuals is referred to as *single nucleotide polymorphism* (*SNP*) [26] and explains about 90% of the heterogeneity within the human genome [30]. Within populations the *AF* for *SNP*s can differ, meaning that a *SNP allele* may be common in a geographical or ethical group but is rare in another.

The Database for Common Variants, dbSNP

A public archive for genetic variation, the *single nucleotide polymorphism database* (*dbSNP*) [124], established by the *NCBI* in collaboration with the *NHGRI* lists a range of molecular variations that contains in addition to *SNP*s also other molecular variations such as for example *INDELs*.

Several hundred thousand *SNP*s with large differences in allele frequencies are observed when comparing individuals from populations with different backgrounds as shown by Durbin et al. [42]. As a consequence difficulties may arise when analysing cohorts with different population substructures to identify mutations that are associated with diseases, especially as rare variants appear to account much more for the genetic diversity compared to common variants [28].

1.1.4 The 1000 Genomes Project

The *1000 genomes project* was launched in 2008 and established the biggest map of human genetic variation that is publicly available until today [31, 32, 33, 34]. Three *pilot* projects have been designed and completed until 2010 which contain variants with a genotype frequency of 1% or greater [31], using a combination of high and low-coverage *whole genome sequencing* (*WGS*) and *WES*. Genomes of two *Trio*s were sequenced at high coverage $20 - 60$ times in the first *pilot* project, using different sequencing technologies [8]. The genomes generated in the 2^{nd} *pilot* study complemented the first phase with data from 179 people but were sequenced at low coverage for economic reasons.

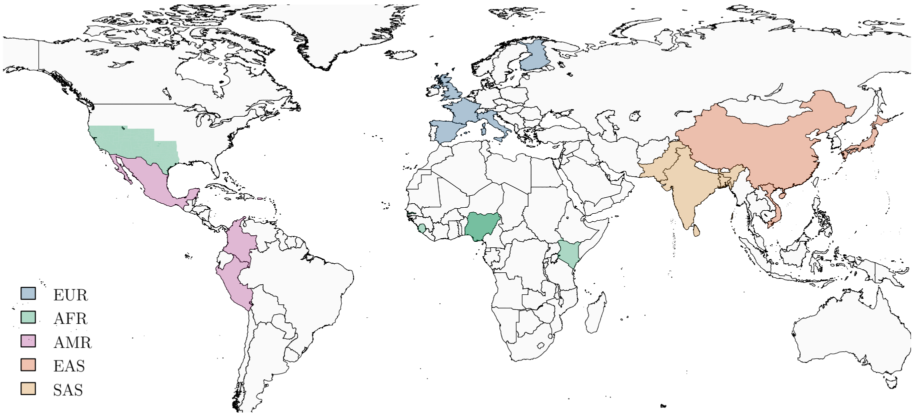


Figure 1.5: *A map of different populations split into several geographic groups (adapted from [7]).* A detailed description of every population captured in the *1KGP* is given in Table 1.3.

Altogether exons of 1000 genes in about 700 individuals were sequenced in the third *pilot* project that concentrated on the coding regions. At the end of 2012 1092 individuals were sequenced with a combination of low-coverage whole-genome and exome sequencing, comprising data from individuals from 14 different ethnic background populations [32] (*phase 1*). Up to now 2535 individuals from 26 different geographical groups were completed [33, 34] (*final phase*). An overview of all background populations is illustrated in Figure 1.5 and a detailed description is given in Table 1.3.

Table 1.3: *Different populations used in the 1KGP.* (adapted from [7]) All *population*s (*POP*s) can be further clustered into super populations: *AFR* (african), *AMR* (mixed american), *EAS* (east asian), *EUR* (european) and *SAS* (south asian).

POP	*POP* Description	Super *POP*	#
CHB	Han Chinese in Bejing, China	EAS	103
JPT	Japanese in Tokyo, Japan	EAS	104
CHS	Southern Han Chinese	EAS	108
CDX	Chinese Dai in Xishuangbanna, China	EAS	99
KHV	Kinh in Ho Chi Minh City, Vietnam	EAS	101
CEU	Utah Residents (CEPH) with Northern and Western Ancestry	EUR	99
TSI	Toscani in Italia	EUR	108
FIN	Finnish in Finland	EUR	99
GBR	British in England and Scotland	EUR	92
IBS	Iberian Population in Spain	EUR	107
YRI	Yoruba in Ibadan, Nigeria	AFR	109
LWK	Luhya in Webuye, Kenya	AFR	101
GWD	Gambian in Western Divisions in the Gambia	AFR	113
MSL	Mende in Sierra Leone	AFR	85
ESN	Esan in Nigeria	AFR	99
ASW	Americans of African Ancestry in SW USA	AFR	66
ACB	African Caribbeans in Barbados	AFR	96
MXL	Mexican Ancestry from Los Angeles USA	AMR	67
PUR	Puerto Ricans from Puerto Rico	AMR	105
CLM	Colombians from Medellin, Colombia	AMR	94
PEL	Peruvians from Lima, Peru	AMR	86
GIH	Gujarati Indian from Houston, Texas	SAS	106
PJL	Punjabi from Lahore, Pakistan	SAS	96
BEB	Bengali from Bangladesh	SAS	86
STU	Sri Lankan Tamil from the UK	SAS	103
ITU	Indian Telugu from the UK	SAS	103
ALL	All Individuals	*1KG*	2535

1.1.5 A Broad Outlook of Next Generation Sequencing in Human Genetics

Up until today, 2^{nd} generation sequencing technologies had a great impact in every field of molecular research und diagnostics mainly due to immense cost reduction and an increased throughput compared to Sanger sequencing and paved the way for studying the molecular mechanisms of human diseases.

Whole Genome vs. Whole Exome Sequencing

WES had a great impact in the characterization of hundreds of novel disease-associated genes within the past five years [109] and is particularly effective in the study of rare *Mendelian* diseases. Each sequencing run identifies approximately 20000 variants in coding regions of which 90% can be found in publicly available data bases [90]. Nethertheless, assuming unlimited resources and time, *whole genome sequencing* (*WGS*) provides clear benefits compared to *WES* as it allows *SNV*s, *INDELs*, *SV*s and *CNV*s to be interrogated in both the ~ 1% part of the genome that encodes for protein sequences and the remaining ~ 99% of the non-coding genome. This results in ~ 5 million reported variants of which about 144000 variants are not listed in any database [90]. Additionally the overall coverage of a *WGS* sample is uniformly distributed and thus superior to *WES* whose captured probes can result in regions with little or low sequence coverage. This is due to the difficulties to design suitable capture baits in regions of the genome with low sequence complexity (such as *GC*-rich regions) resulting in off-target capture effects. On the contrary, an up front enrichment step isn't required during library preparation in *WGS* which reduces the potential of such biases. Furthermore, *WES* capture probes often tend to preferentially enrich reference alleles at *heterozygous* sites producing false negative *SNV* calls.

Another advantage of *whole genome sequencing* is the ability to take advantage of longer sequencing reads, which is restricted in *WES* since the majority of human exons are restricted to a maximum of 200 *bp*s. Longer sequencing *reads* simplify the identification of large *CNV*s, genomic rearrangements and other *SV*s.

However, *WES* still benefits from one unbeatable argument which is the advantage to quickly sequence an individual at a low price level compared to *WGS*. Reduced costs make it feasible to increase the number of samples to be sequenced, enabling large population based comparisons which are not yet feasible with *WGS*. But with the launch of *Illumina*'s *HiSeq X* platform [6] the possibility of sequencing genomes at a substantially reduced price level (1000$ per genome) is already within reach.

On the practical side, the amount of both raw as well as processed data for *WES* experiments is far smaller compared to *WGS* [90]. New challenges regarding the analysis of sequencing *reads* in the dimensions of *WGS* experiments arise and not only existing algorithms have to be adapted (e.g. for *read* assembly), but also different error rates and quality measurements have to be taken into account.

1.2 Bioinformatics Processing of Next Generation Sequencing Data

1.2.1 Methods for Sequence Alignment

The outcome of each *NGS* platform is a bunch of multiple short *DNA* fragments, or *reads*, in the magnitude of *giga base pair* (*Gb*)s. Depending on the fragment library that is used during sample preparation either *single-end* or *paired-end reads* are produced whereas the latter requires sequencing of both opposite ends of one *DNA* fragment. Once sequencing *reads* are obtained, the first essential step is to align them against a known reference sequence (see Section 1.1.2).

Different Alignment Approaches

Several applications that align short-read sequences independently to a reference genome have been developed over the last years, for example *MAQ* [74], *NovoAlign* [10], *BWA* [71, 72, 70], *Bowtie* [68] and *Bowtie2* [67], and comprehensive overviews which compare the different approaches exist [73, 103, 57]. Although many alignment tools are available, they vary a lot in runtime and accuracy which affect the identification of *SV*s and *SNV*s. However, none of them outperforms the others in all metrics as already pointed out by Hatem et al. [57]. The performance rather depends on many different factors and each tool reveals its own strengths and weaknesses.

Build Index

The majority of existing alignment algorithms construct auxiliary data structures for the reference genome, referred to as build *index*, which is used to search for the corresponding genomic position for each *read*. *Index* based approaches are commonly used to efficiently search a big amount of text that is too large to store it all in a computers main memory. The need for not only compressing the textual sequence itself but also the *index* led to the development of many algorithms which address this problem. Varying in runtime, sensitivity and memory usage, two main *index* types can be distinguished that are either based on hash tables as done by

MAQ [74] and *NovoAlign* [10] or suffix/prefix tries as applied by *Bowtie* [68], *Bowtie* 2 [67] and *BWA* [71, 72, 70]. Algorithms based on suffix/prefix tries store multiple identical sub-strings of a reference just once because these copies collapse on a single path in the trie [73] which is an advantage over *indices* based on hash tables. Consequently trie based approaches are much faster and require less memory space compared to hash table based algorithms, but perform with a lower sensitivity.

Strategies to Identify Inexact Matches

To identify inexact matches, all algorithms apply different approaches where a certain number of mismatches is allowed. *MAQ* uses a split strategy, *NovoAlign* adopts an alignment scoring system based on the *Needleman-Wunsch algorithm* (*NWA*) and *Bowtie*, *Bowtie2* and *BWA* apply a backtracking strategy based on *Burrows-Wheeler transform* (*BWT*). *BWA* implements an upper bound for the number of mismatches in a *read* which makes the algorithm more efficient compared to other similar methods [71].

Accuracy and Quality Scores

The performance of sequence alignments, as well as downstream analysis of *SV*, can be increased by using *paired-end* instead of *single-end reads* in terms of sensitivity as well as specificity as it provides additional position information (shown by Shrestha et al. [108]). However, not all available alignment tools provide a suitable statistical adaption of *paired-end* data. Hence, for instance *Bowtie*, *Bowtie2* and *NovoAlign* require preliminary information about the mean and standard deviation of the genomic distance between two *read* mates whereas *BWA* estimates the fragment size distribution from uniquely mapped pairs.

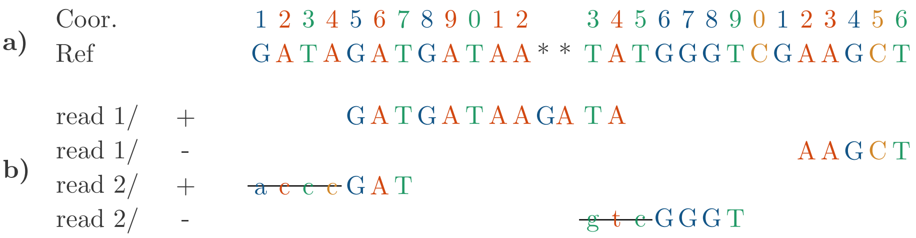
Integrating per-base quality scores can additionally help to increase the mapping accuracy as they lower the penalty for an error prone mismatch [73]. The sequencing-by-synthesis process is captured in a series of fluorescence images, as described in Section1.1.1, and all sequencing platforms record a measurement of uncertainty for the fluorescence of each base that is represented by a *phred* quality score:

$$Q_{\text{Base Call}} = -10log_{10}\mathbb{P}(\text{base calling error}) \qquad (1.2)$$

A *phred* score of $Q_{\text{Base Call}} = 20$ would therefore correspond to a 1% error rate in base calling.

Mapping Quality Score

An overall *mapping quality score* (*MAPQ*) for each sequencing *read* is additionally used to improve downstream analysis. The mapping quality is a *phred*-scaled posterior probability of the accuracy that a *read* is mapped correctly which is an indicator of the uniqueness of the aligned *read* but also incorporates per-base qualities and the number of mismatches within a *read*. Quality scores



c)

```
@SQ SN:Ref LN:26
read 1  136  Ref  5   30  8M2I2M   =  22   22  GATGATAAGA TA  *
read 1  83   Ref  22  30  5M       =  5   -22  AAGCT          *
read 2  0    Ref  4   30  4H3M     *  0    0   GAT            NM:i:1
read 2  16   Ref  16  30  3H4M     *  0    0   GGGT           NM:i:0
```

Figure 1.6: *The SAM format.* **a)** A reference genome/assembly (*Ref*) is a representative example of a species set of nucleic assids, that does not necessarily represent the genome of any single individual. Deviations from the reference assembly are then classified as single nucleotide variants (SNV), copy-number variations (*) or inversions. **b)** Short paired-end sequence *reads* are mapped against a reference assembly (*Ref*). *Read1* represents a *read* pair whereas *read2* constitues a chimeric *read* that consists of two *read* parts that map to different locations on the reference genome. **c)** The corresponding *SAM* format to the alignment shown in b). The header (starting with '@') is always *prior* to the alignment. Each line consists of at least 11 mandatory fields: query name, bitwise flag, reference name, leftmost position, overall mapping quality, CIGAR string, reference name of mate *read*, leftmost position of mate *read*, observed template length, segment sequence and per-base quality (not shown) .

reported by *BWA*, *NovoAlign* and *Bowtie2* are calculated in different ways which makes a direct comparison difficult. For instance, the maximum *MAPQ* score generated by *Bowtie2* is 42 whereas the maximum value reported by *BWA* is 37.

Sequence Alignment/Map format

All available alignment tools produce output in the *sequence alignment/map* (*SAM*) format, a tab-delimited text file consisting of a header section, starting with *'@'*, and an alignment section that lists all sequence *read*, its mapping position and additional information such as the mapping quality score. A reduced example of an alignment in *SAM* format is given in Figure 1.6.

1.2.2 Variant Calling and Allele Frequencies at Heterozygous Positions

A crucial step in the analysis of *NGS* data is the identification of *SNV*s and other genetic variation from *NGS* experiments (as described in Section 1.1.3) which depends on a number of different factors, including per-base and mapping quality scores, *read* length and the depth of coverage. The choice of the applied strategy depends also on the type of study design as different properties have to be taken into account. In this work the focus lies on the identification of germline mutations which is a main part in the identification of rare *Mendelian* diseases. A common representation of all identified or *called* variants within a *NGS* sample is the *variant calling format* (*VCF*) [36] that is described in Figure 1.8.

Allelic Inballance in NGS data

The key challenge in variant calling approaches is to distinguish between systematic noise, arising for example from specific platform biases (see Table 1.1), *PCR* artefacts or local misalignments and actual variations, especially at *heterozygous* gene loci. Other sources of an *allelic imbalance* (*AI*) of both *alleles* in diploid organisms can be associated with the epigenetic inactivation of one of the two gene copies, genetic variation in regulatory regions [122] or tumor development.

Several approaches to identify variants in a sequenced sample exists which are either based on a varity of heuristic thresholds or statistical models.

Heuristic Approach

Assuming the result of an high throughput experiment would be error free and well covered with a reasonable number of *reads* per locus, a strict threshold for the *AF* would be sufficient to distinguish between *homozygous* and *heterozygous* variants. For instance, a simple approach to detect *heterozygous* genomic loci was described by Bell, et. al [21] as a sensitive method for *SNV* calling at a coverage of at least 20 *reads* per locus. Hence a genomic position is classified as a *heterozygous* variant if at least two different *alleles* were detected and the *reference allele frequency* (*RAF*) is between 14% and 86%. At *bi-allelic* loci the *RAF* can be defined by the ratio of the number of *alleles* equal to the reference and the total number of *reads* at one position.

Deviations in the Distribution of Allele Frequencies

However, when performing analyses on real *NGS* data this robust but uncompromising decision rule is not very popular as it cannot properly account for noise and biases in the input data that can be introduced either during the sequencing procedure or by the alignment process. With larger deviations in between the number of reference and alternating *alleles*, frequency heuristics cannot always separate true *heterozygous* variants from noise as exemplarily shown in Figure 1.7. The *RAF*s of 42757 *allele*-mixed positions of a human exome sample which are well covered with at least 20 *reads* are plotted and the thresholds defined by Bell et al. [21] are highlighted (Figure 1.7 **a)**). Figure 1.7 **b)** shows a position which has a *RAF* ≤ 0.14 and is therefore classified as *homozygous*. Relying on this approach, this potentially disease causing variant would have been completely ignored although the *AI* might be originating from previously introduced artefacts or biases.

Another effect can be observed as the detected mean *RAF* is shifted from the expected value of 0.5 to slightly higher values. This shift can be explained by two well known biases which occur during in vitro as well as in silico processes. The *SureSelect* exon enrichment workflow [51] applies 120*bp* antisense oligonucleotides, or *baits*, which are designed for the haploid reference sequence of the latest Human Genome Build to select targets out of a set of randomly sheared, adaptor ligated and *PCR*-amplified total human *DNA*. Hybridisation of fragments containing common variants may be weaker as compared to hybrids without mismatches which leads to a slight advantage for the reference *allele* to be enriched. A sec-

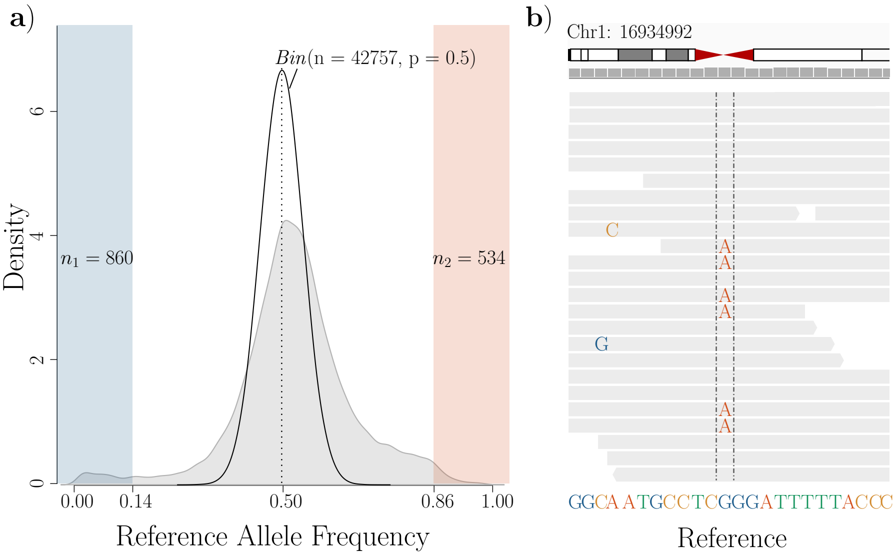
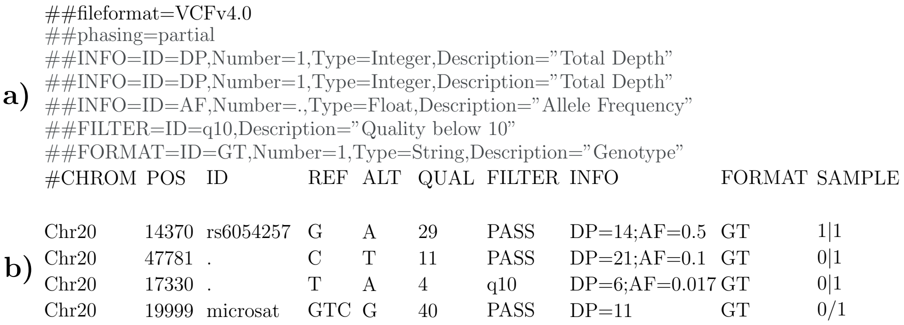


Figure 1.7: *Mis-ballance of allele frequencies at heterozygous positions.* **a)** The *reference allele frequency* (*RAF*) at 42757 *Allele*-mixed loci of a human exome (*HG00119*), that are at least covered with 20 *reads* (gray distribution). Compared to the expected binomial distribution (solid line) the detected mean *RAF* of 0.5 is slightly shifted towards higher values, which can be explained either by a bias introduced in the *PCR* amplification step or by mapping artefacts. **b)** Displayed is an enlarged view of a position that is covered with a total number of 93 *reads* (not all *reads* are shown), of which 11 *reads* are representing an alternating *allele* (*A*). Using the simple ratio-based approach as proposed by Bell et al. [21] a *RAF* of ~ 0.88 would classify this position as *homozygous* for the reference *allele* *G* and would not be considered in further analysis steps. On the contrary, a probabilistic model (such as *GATK*) identifies this locus as a *heterozygous* variant.

ond bias may be found in the alignment process, where all short sequence *reads* are mapped to the haploid reference sequence. A combination of short *read* lengths, low base quality, caused by non-reference variants in the fragments, and a low sequence complexity may result in mapping errors and consequently leads to a reduced mapping ratio of non-reference *allele* fragments [37, 101]. Though these biases lead to a slight deviation of the detected mean reference *AF*, they do not influence the variance of the *AF* distribution.

Probabilistic Methods

Over the last years, several probabilistic methods have been developed to identify *SNV*s from *NGS* experiments by generating robust estimates of the probabilities of each of the possible *GT*s and a comprehensive overview is given for example by [89]. To date,



```
    ##fileformat=VCFv4.0
    ##phasing=partial
    ##INFO=ID=DP,Number=1,Type=Integer,Description="Total Depth"
    ##INFO=ID=DP,Number=1,Type=Integer,Description="Total Depth"
a)  ##INFO=ID=AF,Number=.,Type=Float,Description="Allele Frequency"
    ##FILTER=ID=q10,Description="Quality below 10"
    ##FORMAT=ID=GT,Number=1,Type=String,Description="Genotype"
    #CHROM POS   ID         REF ALT QUAL FILTER INFO              FORMAT SAMPLE

    Chr20  14370 rs6054257  G   A   29   PASS   DP=14;AF=0.5      GT     1|1
b)  Chr20  47781 .          C   T   11   PASS   DP=21;AF=0.1      GT     0|1
    Chr20  17330 .          T   A   4    q10    DP=6;AF=0.017     GT     0|1
    Chr20  19999 microsat   GTC G   40   PASS   DP=11             GT     0/1
```

Figure 1.8: *The VCF format.* (adapted from [7]) The *variant calling format* is a tab-delimited text format, that consist of meta-information, including the header line (**a)**) and data lines for each position in the genome (**b)**). The header line starts with *'#'* and lists eight fixed, mandatory fields for the chromosome (*CHROM*), the genomic position (*POS*), the *dbSNP* identifier (*ID*), the reference base (*REF*), a comma separated list of alternating *Alleles* (*ALT*), the phred-scaled quality score for the variance call (*QUAL*), a column that gives information if the position has passed all filtering criteria (*FILTER*) and additional information such as the total *read* depth, *DP*, of a position or the *allele frequency* (*AF*) (*INFO*). These information is followed by a *FORMAT* column header and an arbitrary number of *SAMPLE* identifiers. The *genotype* (*GT*) for each sample is always encoded with a 0 for the reference *allele* and a 1 for the alternating *allele*. Phased *GT*s are characterised with a separating '|' and with '/' otherwise.

commonly used probabilistic methods for variant calling apply algorithms based on *Bayes'* Theorem, named after Thomas Bayes (1702 – 1761) (e.g *SAMtools* [69] or the *Unified Genotyper* of *The Genome Analysis Toolkit* (*GATK*) [83, 38, 117]). This allows to compute the conditional probability for a *GT* dependent on the available data D which are the aligned *reads*:

$$\mathbb{P}(GT|D) = \frac{\mathbb{P}(GT)\mathbb{P}(D|GT)}{\sum_i \mathbb{P}(GT_i)\mathbb{P}(D|GT_i)}, \tag{1.3}$$

whereas $\mathbb{P}(GT)$ defines the *prior* probability of seeing this *GT* and $\mathbb{P}(D|GT)$ gives the likelihood of GT. The notation GT_i refers to the ith out of 10 possible diploid *GT*s at each position, $GT \in \{AA, AC,$ ..., GT, $TT\}$. Tools that utilize this approach differ, dependent of the applied models for the calculation of the *prior* probabilities and the likelihood of the *GT*. The variant calling approach implemented in the software package *GATK* (*Unified Genotyper*) utilizes an algorithm that relies on a likelihood function that is based on the decomposed *haplotypes* H_1 and H_2 for every *read* j [3] [83]:

$$\mathbb{P}(D|GT) = \prod_j \left(\frac{\mathbb{P}(D_j|H_1)}{2} + \frac{\mathbb{P}(D_j|H_2)}{2} \right), \tag{1.4}$$

assuming $GT = H_1H_2$.
Finally, the likelihood function $\mathbb{P}(D_j|H)$ uses the pileup of bases b and associated reversed *phred* quality scores ε at a given locus with $\mathbb{P}(D_j|H) = \mathbb{P}(D_j|b)$ and

$$\mathbb{P}(D_j|b) = \begin{cases} 1 - \varepsilon_j & D_j = b \\ \frac{\varepsilon_j}{3} & \text{otherwise} \end{cases} \tag{1.5}$$

The *prior* probability of observing a *GT* is usually calculated assuming an underlying *binomial* distribution of the *read* count data, as also shown in Figure 1.7 **a)**.

Another variant caller of the *GATK* package, the *HaplotypeCaller*, is considered as the *state-of-art* variant calling algorithm as it performs a local *de novo* assembly of *haplotypes* around each genomic site and identifies variants with high accuracy, especially on *INDELs*. However, in a direct comparison the *Genotype Caller* outperforms the *Haplotype Caller* in terms of run time.

1.3 Quality Control and Filtering Techniques

1.3.1 Quality Measurements in NGS Data

Once sequencing *Reads* from *NGS* experiments are obtained and aligned against a suitable reference genome and variants are detected, several filtering criteria can be applied, on the one hand to shrink the search space for variant detection and on the other hand to reduce sources of platform and sequencing specific errors and biases.

Sequencing Depth

The average sequencing depth or *coverage* is often used to denote the breadth of sequence *coverage* of a target region which is defined as the percentage of the region that is sequenced a given number of times. A commonly used target region which is applied in diverse *WES* studies is the consensus coding sequence as defined by the *collaborative consensus coding sequence* (*CCDS*) project [95]. In practice, a good quality exome sample should yield an average sequencing *coverage* of 50× and achieve a 90% breadth of *coverage* of the target region at a minimum depth of 10 *reads* [80]. This is also illustrated in Figure 1.9, where the *coverage* distribution over the *CCDS* target region of 123 *Illumina* *WES* experiments is shown. In general sequencing of more *reads* at higher depths improves the confidence level in downstream variant calls. However platform and sequence specific biases, such as a *GC* bias that is introduced during *DNA* amplification, yield varying sequencing *coverages* at different genomic areas or even result in regions with no *coverage* at all. The distribution of the mean *GC* content against the mean *coverage* of 57 exome samples sequenced by the *1000 genomes project* (*1KGP*) is shown in Figure 1.10.

Per-Base Coverage

The term *coverage* can also refer to the number of *reads* that align to a particular position in the genome which correlates strongly with the degree of confidence for variant discovery that is higher

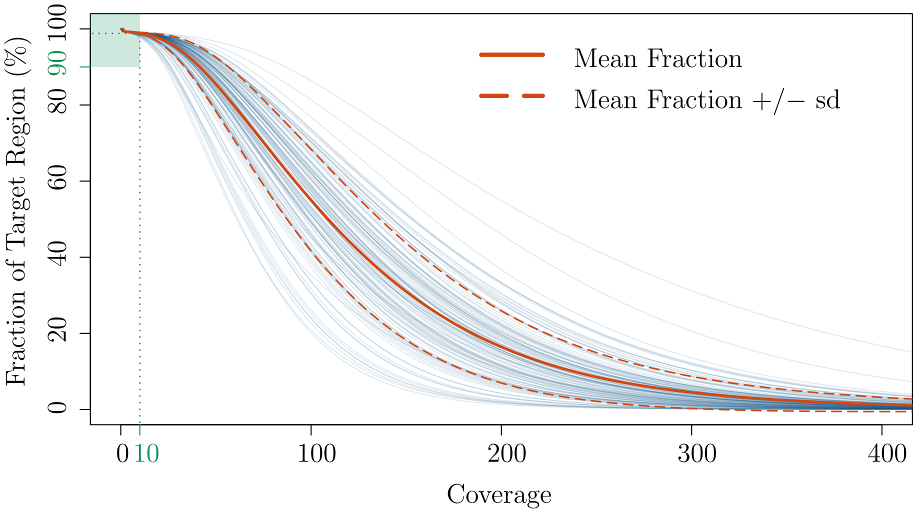


Figure 1.9: *Coverage of target region.* Exomes of 123 samples were sequenced on a HiSeq 1500 platform [6] and exonic regions were captured using the SureSelect Human All Exon *V*5 kit from *Agilent* [2]. Duplicated sequencing *reads* were removed from each alignment using *SAMtools* [13] and a threshold of a minimum $MAPQ \geq 1$ was applied. The figure visualizes the fraction of the captured target region that is covered with a minimum number of *reads* (blue lines). For instance taking the mean fraction of all samples (solid red line) 0.98% of the target region is covered by a minimum of 10 *reads* (gray dotted line). A good quality sample is considered if at least 90% of all positions in the target region were covered by at least 10 *reads* (green box).

for well covered positions. This is why a minimum per-base *coverage* threshold is often used as a first quality criterion to filter for high confidence variant calls.

Genotype Quality

In most *NGS* analysis pipelines a *phred*-scaled quality score is provided for each identified *GT* which is a measure of confidence for the call (*QUAL*, see Figure 1.8):

$$Q_{\text{Genotype Call}} = -10log_{10}\mathbb{P}(\text{wrong genotype}) \tag{1.6}$$

Algorithms that assume a certain distribution of the sequenced *reads* at *heterozygous* positions may introduce biases for locations with *AFs* deviating from the expected value of 0.5 and the *phred*-

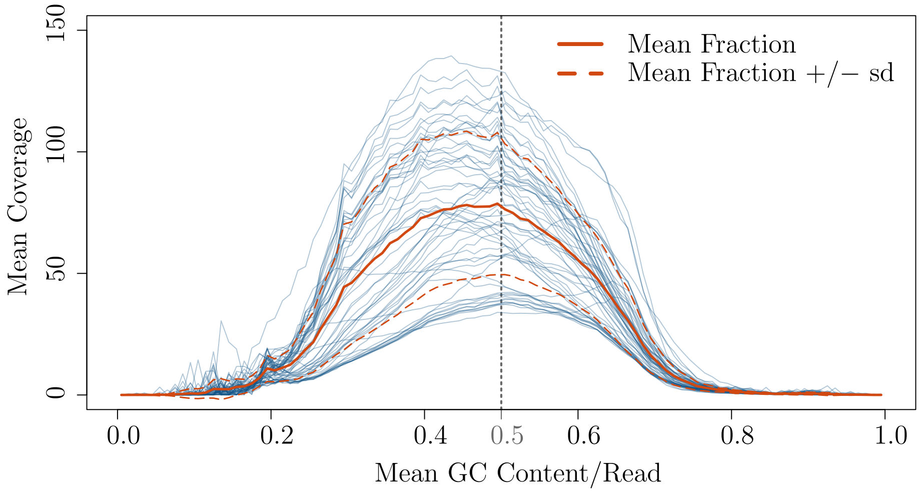


Figure 1.10: *GC content against coverage.* *WES* sequence alignments from 57 individuals produced by the *1KGP* were used to calculate the mean GC content and sequence *coverage* in a window of 400 *bp*s in for the *CCDS* target region. A strong deviation of the distribution from the expected mean of 0.5 can be explained by the low *coverage* in GC-poor or GC-rich regions.

scaled quality score would decrease the more the ratio of *reads* supporting the alternate *allele* deviates from the expected mean. However, this quality score not only depends on the raw data but also on the mapping algorithms and probability models that were used for variant calling. Processing the same raw data by different bioinformatic pipelines may result in varying distributions of quality scores suggesting different genotyping error profiles for the same exome sample. Even variant calling approaches that are based on similar *Bayesian* methods do not yield the same GT probabilities due to different *priors* and methods of quality score recalibration cannot completely adjust for that effect (see Table 1.4) [115].

The variance of reference AFs as shown in Figure 1.11 may also be an indicator for quality in *NGS* experiments as it is associated with the error rates in *heterozygous* variant detection. As exemplarily shown in Figure 1.7 variants with an AF that is strongly deviating from the expected mean of 0.5 may be missed by variant calling

Table 1.4: *Quality measurements in VCF files.* (adapted from Heinrich et al. [115]) Short sequence *reads* of test Sample #1 and #2 and a sample from the *1KG* reference set, *NA06986*, were down-sampled to comparable mean per-base *coverages* over the target region. The mean *GT* quality scores for all three samples drop with a decreasing *coverage*, indicating an increasing false positive error rate. It can also be observed that different *priors* in the genotyping models used by *SAM*tools and *GATK* result in different mean *GT* quality scores for the same alignments (**a**)). Quality score recalibration with *GATK VariantRecalibrator* performed on *SAM*tools- and *GATK*-called variants adds an adjusted quality score, *VQSLOD* (log odds ratio of being a true variant versus being false under a trained *Gaussian* mixture model) which diverge greatly between the two variant calling tools (**b**)). All variants that pass the *VQSLOD* score cutoffs identified by the *VariantRecalibrator* are then defined as a set of highly confidential calls. The percentage of this set shows an irregular behaviour with respect to the mean per-base *coverage* of the different samples (**c**)).

	Sample #1			**Sample #2**				**NA06986**			
Coverage	30	50	65	30	65	100	142	30	65	100	313
a) Mean Genotype Quality											
*SAM*tools	83.4	88.0	89.7	84.7	91.8	93.4	94.4	77.2	87.3	90.4	94.8
GATK	81.6	87.7	89.8	81.7	91.2	93.4	94.7	73.1	85.1	89.4	94.7
b) Mean *VQSLOD* Score											
*SAM*tools	14.12	13.69	14.46	14.58	11.32	14.46	14.89	12.69	1.94	15.68	1.86
GATK	4.21	3.33	3.09	3.33	-0.34	-0.05	0.51	2.15	2.02	1.94	1.59
c) % High Confidential Calls											
*SAM*tools	93.8	94.0	93.9	92.3	91.8	92.1	96.5	91.4	85.2	91.2	88.0
GATK	90.9	91.0	90.9	82.7	81.7	84.4	86.2	85.7	83.3	81.4	76.9

algorithms.

Additional Quality Metrics

Other quality metrics can be applied to the entire set of variant calls in a *VCF* file such as the percentage of *SNV*s that are already listed in databases such as *dbSNP* [107], the ratio of *homozygous* and *heterozygous* variant calls and the ratio of *transition*s (*Ti*s) and *transvertion*s (*Tv*s). There are two types of *DNA* substitutions

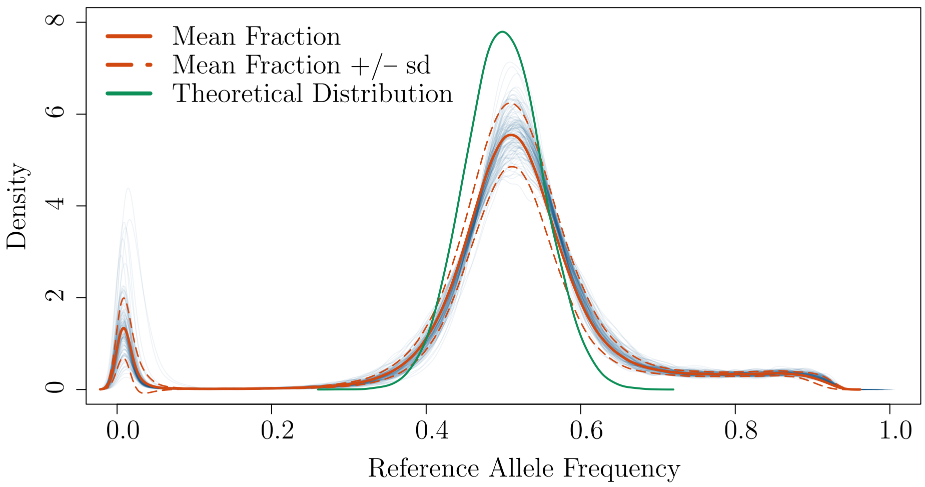


Figure 1.11: *Allele frequencies at heterozygous positions.* The reference *allele frequencies* (*AFs*) of 123 exome samples were calculated for *heterozygous* loci with at least one alternating *allele*. By comparing the average distribution over all samples (red solid line) and the assumed *binomial* distribution, one recognizes that the detected mean reference *AF* differs from the expected value of 0.5 which is due to well known biases introduced during in vitro as well as in silicio processes.

in *SNV*s: *transvertion*s are interchanges of purine for pyrimidine bases, *transition*s describe substitutions within purines or pyrimidines. For the *CCDS* region, the *Ti*/*Tv* ratio should be close to 1 : 3 and the proportion of variants that are not listed in *dbSNP*, also referred to as *singletons*, should be below 10% [18, 115]. However, the *Ti*/*Tv* ratio is influenced by the target region and the correlating amount of non-coding variants, whereas the number of novel variants may also correlate with the background population. For example, higher ratios of novel variants may be observed if the sequenced sample is of a population that is poorly represented in the variant databases [115].

Table 1.5: *Additional quality parameters in VCF files.* Shown are the mean (μ) and the standard deviation (σ) of different quality measurements obtained from 123 *VCF* files based on *WES*, that were restricted to the exome target region defined by *CCDS*: the total number of *SNV*s in the target region, the percentage of *SNV*s that are present in *dbSNP* (version 138), the ratio between *heterozygous* and *homozygous* *GT*s (*het/hom*) and the ratio between *transition*s and *transvertion*s (*Ti*/*Tv*).

	# of *SNV*s	% in *dbSNP*	*het/hom* Ratio	*Ti/Tv* Ratio
μ	24646	0.98	1.70	0.88
$[-/+\sigma]$	[24195 – 25098]	[0.98 – 0.98]	[1.59 – 1.81]	[0.84 – 0.93]

1.3.2 Strategies to Filter for Diseases in NGS Studies

Annotation of Variants

With *WES* experiments yielding about 20000 to 24000 protein altering variants, depending on the background population [19], the interpretation of all *SNV*s remains challenging in terms of complexity and time expenditure. The separation of disease-related variants which are associated with *Mendelian* or complex traits from the background of non-pathogenic polymorphisms is still the key issue. Several approaches have been developed to simplify the search space by applying annotations for variants, including *allele frequencies* (*AFs*), the functional impact on gene expression, pathway informations and predictions for the expected pathogenicity (e.g. *MutationTaster* [106], *ANNOVAR* [123], *VariantDB* [118], *GeneTalk* [60]). However, several tools applied for annotations such as *ANNOVAR* or *MutationTaster* may rather be used for prioritization as the assessment of human expertise may not be replaced yet [60].

Filter for Inheritance

Mendelian traits are passed down by *recessive* or *dominant alleles* of only one gene and depending on the inheritance model and available sequenced relatives, different filtering strategies can be applied. The additional sequening of family members is common practice when studying *Mendelian* disorders and filtering strategies and *linkage* analysis are usually applied. Genetic *linkage* analysis is a powerful approach to identify two loci in close proximity which are likely to be inherited together via the calculation of *logarithm of the odds* (*LOD*) scores.

When analysing *trio*s (parents and one offspring) variants which are present in the patient that could not have been inherited from the parents can be easily filtered (Non-*Mendelian* trait). These would most likely be an indication for either sequencing artefacts [92] or *de novo* mutations, especially whenever the disorder is highly heterogeneous and *de novo* mutations are the most promising candidates [119]. To give another example, when analysing sequencing samples from patients with rare recessive diseases the most likely underlying inheritance model is *compound heterozygosity*, at least in non-consanguineous families [61]. Additionally non-related individuals can help to filter for non-pathogenic variants. However, these filtering approaches rely on the presence of correct pedigree information and have to be additionally adapted when analysing highly consanguineous families. Another popular approach for detecting variants at genomic loci that are associated especially with complex traits which are hard to resolve via the analysis of just one single pedigree is referred to as *genome wide association studies* (*GWAS*). This technique is based on the hypothesis of strong associations between common *SNP*s and common diseases such as heart diseases or psychiatric disorders. While common variants typically have modest effect sizes, rare variants, especially those in coding regions, can have larger effect sizes with greater potential to influence disease. While population-based variant callers such as *GATK* have improved the accuracy of *GT*s for low frequency variants, they perform poorly when identifying singletons and doubletons [26]. Therefore rare variants have a high *heterozy-*

Association Studies

gote to *homozygote* error rate. Additionally *GWAS* for common variants usually require a large amount of case and control samples to guarantee sufficient power and possible population substructures or varying data quality within the samples may influence the disease gene discovery [127].

2 | Thesis Organization

This work is devoted to various aspects of quality measurements in *next generation sequencing* (*NGS*) and especially in *whole exome sequencing* (*WES*) studies. Three strategies for *quality control* (*QC*) will be presented which focus on different parts of an *WES* experiment.

During the introduction (Chapter 1) various levels in the analysis of exome experiments were highlighted and existing tools and strategies were introduced. The following chapters are focusing on remaining uncertainties in quality assessment of *WES* experiments. To be more specific, in the third chapter (Chapter 3) I will study the distribution of the variance of *allele frequencies* (*AFs*) at *heterozygous* genomic loci as measured in *NGS* data sets as this is a crucial pre-requisite for variant calling algorithms based on *Bayesian* models such as *SAM tools* or *GATK* as explained in Section 1.2.2. For this I will define the amplification of sequence fragments during library preparation as described in Section 1.1.1 as a two-type *(Bienayme-) Galton-Watson* (*BGW*) branching process. I will then analyze the effect of different parameter values analytically as well as on simulations and finally compare the results with real *NGS* data sets.

In the fourth chapter (Chapter 4) I will demonstrate the importance of different background populations that have to be taken into account during the analysis of sequencing samples. I will introduce a new similarity metric, weighted with population based *allele frequencies* (*AFs*) that can be used to access the overall genotyping accuracy of an exome. For this I will make use of a high quality reference set comprising 2535 genotyped exome samples which were sequenced by the *1000 genomes project* (*1KGP*) as introduced in Section 1.1.4.

Section 1.3.2 highlighted different strategies for efficient filtering and prioritization of possible disease causing variants, which are depending on correct pedigree information. In Chapter 5 I will systematically describe *likelihood ratio* (*LR*) based approaches to

reconstruct entire pedigrees which also take rare marker loci into account. I will give a description for the calculation of *logarithm of the odds*s (*LOD*s) scores for different genotype combinations based on five predefined relatedness hypotheses and study the effects of the number of variants and the degree of consanguinity on the precision values.
Finally I summarize and discuss the major results from the separated chapters and give an outlook for future research projects (Chapter 5).

Additionally I would like to give some general information about the formal structure of this thesis. Each of the three main chapters (3, 4 and 5) is constructed with a similar layout. First, each chapter starts with a short overview (*Overview of this Chapter*) and introduces some basic knowledge that is related to the specific topic of the chapter (*Introduction to [...]*). The following sections describe the actual work I have done during my *Ph.D.*. Finally, each chapter includes a section where the obtained results are validated (*Experimental Validation*) and a short summary (*Summary of this chapter*).

3 | Distribution of Short Read Fragments in Next Generation Sequencing Experiments

3.1 Overview of this Chapter

This chapter focuses on the distribution of *allele frequencies* (*AFs*) at *heterozygous* genomic loci as measured in *whole exome sequencing* (*WES*) data sets. A lot of variant detection tools rely on the assumption that the *allele frequencies* (*AFs*) at *heterozygous* positions follow a *binomial* distribution. But as exemplarly shown in Figure 1.7 the variance of the *reference allele frequency* (*RAF*) can be much broader compared to an expected *binomial* distribution, which can lead to a misclassification of variants. This can be explained by a bias introduced in the *polymerase chain reaction* (*PCR*) amplification step during library preparation before sequencing as described in Figure 1.3. In the following sections the amplification of sequence fragments is described as a stochastic process, depending on the number of *cycles*, the initial amount of fragments and the amplification efficiency. Hence I will first introduce the concept of stochastic processes, *Markov* chains and *(Bienayme-) Galton-Watson* (*BGW*) branching processes in particular. These stochastic concepts are then applied to the fragment amplification step which results in a single function of the variance of the fragment distribution. The effects of these parameters are analysed on simulated data as well as on real life *WES* samples.

Note that the results of Section 3.4.2 were obtained in collaboration with Prof. Dr. Peter Imkeller, Jun.-Prof. Dr. Thorsten Dickhaus and Jens Stange from the Department of Mathematics at the Humboldt-University in Berlin whom I would like to acknowledge at this point.

3.2 Introduction to (Bienayme-) Galton-Watson (BGW) Branching Processes

3.2.1 Stochastic Processes

Stochastic Process X_k

In probability theory a stochastic or random process is a collection $\{X_k : k \in K\}$ of random variables on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ whereby the sample set Ω is defined as the set of all possible outcomes, the σ-algebra $\mathscr{F}$ is a collection of subsets of possible outcomes and the probability measure function $\mathbb{P}$ assigns probabilities to each subset of Ω [23]. Although K can be quite arbitrary, in most cases, K is a set of integers and k is interpreted as time.
To every time point k corresponds a random variable, for instance $X_k : \Omega \to \mathbb{R}$, which means that to every outcome $\omega \mapsto X_k(\omega)$, $\omega \in \Omega$, corresponds a realization of the stochastic process which is a function defined on the index set K and values in $\mathbb{R}$.

Distribution Function $F_{\hat{k}}(x)$

A stochastic process can be characterized by its finite distribution function for any fixed $\hat{k}$:

$$F_{\hat{k}}(x) = \mathbb{P}(X_{\hat{k}} \leq x), \forall x \in \mathbb{R}. \tag{3.1}$$

Considering a more generalized form with several quantities $X_{k_1}, X_{k_2}, ..., X_{k_n},\ n \in \mathbb{N}$, the corresponding joint $n-$dimensional distribution function is denoted as following:

$$F_{\vec{k}}(\vec{x}) = F_{k_1,...,k_n}(x_1, ..., x_n) = \mathbb{P}(X_{k_1} \leq x_1, ..., X_{k_n} \leq x_n). \tag{3.2}$$

Consistency Conditions

Each family of finite dimensional distribution functions as defined in Equation 3.2 satisfy the following consistency conditions [62]

1. *(Symmetry)* The $n-$ dimensional distribution function is symmetric in all pairs (x_j, k_j), so that $F_{\vec{k}}(\vec{n})$ remains invariant for every permutation of $(j_1, j_2, ..., j_n)$:

$$F_{k_{j_1},...,k_{j_n}}(x_{j_1},...,x_{j_n}) = F_{k_1,...,k_n}(x_1,...,x_n) \qquad (3.3)$$

2. *(Consistency)* Knowing the $n-$dimensional distribution implies knowing all lower order distributions:

$$\lim_{x_n \to \infty} F_{k_1,...,k_n}(x_1,...,x_n) = F_{k_1,...,k_{n-1}}(x_1,...,x_{n-1}) \qquad (3.4)$$

Classification of Stochastic Processes

Stochastic processes can be classified by imposing suitable restrictions on their n-dimensional distribution functions:

1. *(Stationary) process* A stochastic process is stationary if its finite dimensional distributions are invariant under arbitrary translation of the parameter $k \in K$.

2. *(Gaussian) process* For *gaussian* or *normal* processes the joint distribution functions are multivariate *normal*.

3. *(Markov) process* Given a random variable X_k, the value of $(X_s)_{s \neq k}$ does not depend on the values of $(X_u)_{u \neq k,s}$ for $u < k < s$.

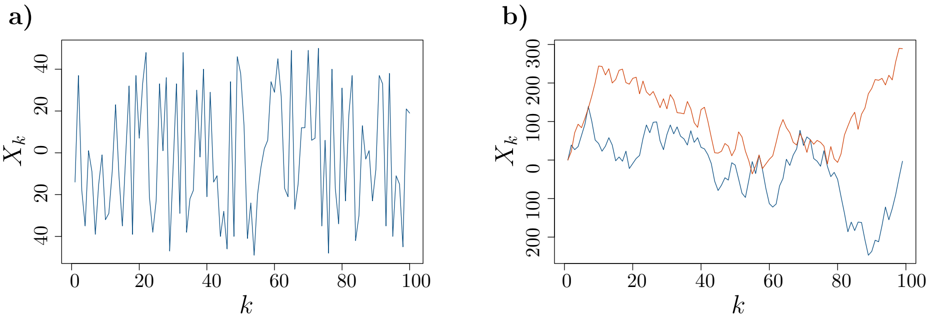


Figure 3.1: *Examples of simple stochastic processes.* **a)** A realization of the *i.i.d noise* process which consists of a series of uncorrelated random variables $\{X_k : k \in \mathbb{Z}\}$. **b)** Two sample paths of the *random walk* R_n with $\{X_k : k \in \mathbb{N}\}$ and $R_n = X_1 + X_2 + \cdots + X_n$.

Examples of Stochastic Processes

One fundamental stochastic process from which many other stationary processes are derived is the so-called *i.i.d noise* process, which is defined as a series of uncorrelated random variables $\{X_k : k \in \mathbb{Z}\}$, each with a zero mean and the same finite variance as illustrated in Figure 3.1 **a)**. Another classic example of a stochastic process is the *random walk*, R_n, which can be defined as a sequence of accumulated independent random numbers $R_n = X_1 + X_2 + \cdots + X_n$ with $\{X_k : k \in \mathbb{N}\}$, usually starting with $X_{k_0} = 0$ [23] as shown in Figure 3.1 **b)**.

Markov Processes

A special case of stochastic processes is the *Markov* process, named after Andrei Markov, which is described as a random process with the so-called *Markov* property which is that the conditional probability distribution of future states of the process depends only on most recent present state and not on the past states (also referred to as *memoryless*):

$$\mathbb{P}(X_{k+1} = s_{k+1} | X_0 = s_0, X_1 = s_1, ..., X_k = s_k) = \mathbb{P}(X_{k+1} = s_{k+1} | X_k = s_k), \quad (3.5)$$

with $s_0, ..., s_k, s_{k+1} \in S$ whereas S defines the state space and $\mathbb{P}(X_0 = s_0, X_1 = s_1, ..., X_k = s_k) \neq 0$.

A distinction is made between a *Markov* process in discrete and continuous time depending on the nature of S and the ordered index set K.

Markov Chains

If both variables are based on discrete time values, $k \in \mathbb{N}_0$ the stochastic process is referred to as *Markov* chain with the following transition probability function:

$$p_{lm}(k) = \mathbb{P}(X_{k+1} = m | X_k = l), \quad (3.6)$$

with $m, l \in S$ where $l, m = 1, ..., n$ and S is a denumerable index set.

One special class of *Markov* chains which is often adapted to population dynamics is the *(Bienayme-) Galton-Watson* (*BGW*) branching process and will be described in the following section.

3.2.2 (Bienayme-) Galton-Watson Branching Processes

Consider a not further specified evolution population of particles with non-overlapping generations. Each individual of the population has a fixed lifetime of one time unit $k \in \mathbb{N}_0$ and reproduces a random number of descendants according to the same distribution before death, independent of the number of ancestors. Starting with X_{k_0} particles at time point $k = 0$ each of it splits independently of the others into a random number of offspring, namely the first generation, according to the transition probability function $p_{lm}(k)$ as defined in Equation 3.6. The process continues and the number of descendants produced by a single particle at any given time k is independent of the history of the process and other existing particles. In the following most definitions were taken from Athreya and Ney [17] unless stated otherwise.

Transition Function

The transition probability function given a reproduction distribution $(p_m)_{m\geq 0}$ with $(p_m)_{m\geq 0} \geq 0$ can be defined as follows:

$$p_{lm}(k) = \mathbb{P}(X_{k+1} = m | X_k = l) = \begin{cases} p_m^{*(l)} & , \quad l \geq 1, m \geq 0 \\ \delta_{0m} & , \quad l = 0, m \geq 0 \end{cases} , \qquad (3.7)$$

whereas $p_m^{*(l)}$ is the $l-$fold convolution of $(p_m)_{m\geq 0}$ and δ_{0m} defines the *Kronecker* symbol:

$$\delta_{0m} = \begin{cases} 1 & , l = m \\ 0 & , \text{else} \end{cases} . \qquad (3.8)$$

Let $\{X_k, k \in \mathbb{N}_0\}$ be a branching process with initially X_0 particles and X_k particles in the k^{th} generation. Further denote the probability generating function $F_k(s)$:

$$F_k(s) = \mathbb{E}[s^{X_k}] = \sum_{m=0}^{\infty} p_m s^m, \forall |s| \leq 1 \tag{3.9}$$

with the reproduction distribution $(p_m)_{m\geq 0}$ and the following notations:

$$F_0(s) = s \tag{3.10}$$

$$F_1(s) = F(s) \tag{3.11}$$

$$F_{k+1}(s) = F[F_k(s)] \tag{3.12}$$

Multi-Type (Bienayme-) Galton-Watson Processes

By repeated application of the branching process X_k, additive property of the BGW process can be used which can be understood as a set of l independent copies of the branching process. To be more specific the multi-type BGW branching process $\{\vec{X}_k : k \in \mathbb{N}_0\} = (X_k(1), X_k(2), ..., X_k(d))$ is denoted as a set of $d \in \mathbb{N}$ independent processes at time k. To define the particle production of the d-type branching process d generating functions $\vec{F}_k(s) = \big(F_k^{(1)}(s), ..., F_k^{(d)}(s)\big)$ are needed. The i^{th} generating function, $\big\{F_k^{(i)}(s)\big\}_{i\in 1,...,d}$, will determine the distribution of the number of offspring of various types to be produced by a type i particle.

Probability Generating Function

As motivated in Equation 3.9 one can define the probability generating function for the multi-type BGW process $\{\vec{X}(k); k \in \mathbb{N}_0\}$ as done by Yakovlev and Yanev ([126]):

$$\begin{aligned}\vec{F}_k(s) &= \mathbb{E}\big[s^{\vec{X}_k}|X_0(1)\big] \\ &= \mathbb{E}\big\{s_1^{X_k(1)} s_2^{X_k(2)} \cdots s_d^{X_k(d)}|X_0(1) = 1\big\}\end{aligned} \tag{3.13}$$

with $s = (s_1, s_2, ..., s_d)$ and $|s_i| \leq 1$, $i = 1, ..., d$.

In other words, $\vec{F}_k(s)$ denotes the number of particles in the k^{th} generation with initially 1 particle. Unlike one-dimensional BGW branching processes multi-type processes allows to study a number of distinguishable particles with different probabilistic behaviour. As the main focus in the theory of branching processes is on probabilistic characteristics of the multi-type process $\vec{X}_k$ and its asymptotic behaviour, I will concentrate on the behaviour of $\vec{X}_k$ when the initial number of ancestors is large and the time point k of observation is fixed. This aspect will become especially important when focusing on relative frequencies of different particle types as described in the next section.

3.3 Asymptotic Behaviour of the Relative Frequencies

The analysis of relative fractions of a set of distinguishable types of particles is useful when studying for example proliferation or differentiation of cells. However the main focus in previous studies was rather on the number of particles but not on their relative frequencies. Yakovlev and Yanev [126] suggested an approach to study the asymptotic behaviour of the fractions of different particle types T_i, which is denoted as $\Delta_i(k; N)$, as the initial number of ancestors, denoted as N tends to infinity ($N \to \infty$) assuming a fixed time point k.
Consider a multi-type BGW process $\vec{X}_k$ with $X_k^{(i)}$ defined as the total number of particles of a finite number of different particle types T_i and $U(k) = \sum_{i=1}^{d} X_k^{(i)}$ denotes the total number of particles at time point $k \in \mathbb{N}_0$.
All moments of a branching process can be obtained from the probability generating function $\vec{F}_k(s)$ as motivated in Equation 3.13 and expressed in terms of the derivates of $\vec{F}_k(s)$ evaluated at $\vec{s} = (1, 1, ..., 1)$.

Mean $m_i(k)$

Therefore the mean of the process $X_k^{(i)}$ can be determined as follows:

$$m_i(k) = \mathbb{E}[X_k^{(i)} | X_0^{(1)} = 1] = \frac{\delta}{\delta_i} \vec{F}_k(s) \mid_{\vec{s}=(1,1,...,1)} . \tag{3.14}$$

The normalized mean $q_i(k)$ will be important for the long-time behaviour of multi-type BGW processes:

$$q_i(k) = \frac{m_i(k)}{\sum_{j=1}^{d} m_j(k)}. \tag{3.15}$$

Variance $\sigma_i^2(k)$

As stated by Georgii [54] the variance of the process can be obtained as follows:

$$\sigma_i^2(k) = Var[X_k^{(i)}|X_0^{(1)} = 1] = \frac{\delta^2}{\delta_i^2}\vec{F}_k(s) \mid_{s=(1,1,\ldots,1)} + m_i(k) - (m_i(k))^2 \tag{3.16}$$

Covariance $||C_{ij}(k)||$

With the assumption of a finite covariance matrix $C_{ij}(k)$, it follows:

$$C_{ij}(k) = ||Cov[X_k^{(i)}, X_k^{(j)}]||_{i \neq j} = \frac{\delta^2}{\delta_i \delta_j}\vec{F}_k(s) \mid_{s=(1,1,\ldots,1)} - m_i(k)m_j(k) \tag{3.17}$$

Relative Frequency $\Delta_i(k)$

Let $\Delta_i(k)$ be defined of the relative frequency of the non-extinction set $\{U(k) > 0\}$, the total number of particles at time point k is denoted as follows:

$$\Delta_i(k) = \frac{X_i(k)}{U(k)} \tag{3.18}$$

with the propoerty $\sum_{i=1}^{d} \Delta_i(k) = 1$.

Number of Initial Particles N

The development of the process $X_k^{(k)}$ strongly depends on the number of initital particles $X_0^{(i)} = N$ and thus the following notation holds (as motivated by Yakovlev et al. [126]):

$$\vec{X}_k(N) = (X_k^{(1)}(N), X_k^{(2)}(N), .., X_k^{(d)}(N)) \tag{3.19}$$

with the property

$$X_k^{(i)}(N) = \sum_{n=1}^{N} X_k^{(i)}(n) \tag{3.20}$$

under the assumption that all particles develop independently with $\left\{X_k^{(i)}(n)\right\}$ defined as $i.i.d$ copies of the branching process $\{X_k^{(i)}(n)\}_{n=1}^{N}$. Further the total number of particles at time point k is denoted as:

$$U(k; N) = \sum_{i=1}^{d} X_k^{(i)}(N) = \sum_{n=1}^{N} U^{(n)}(k) > 0 \tag{3.21}$$

with $U^{(n)}(k) = \sum_{i=1}^{d} X_k^{(i)}(n)$ and $\mathbb{E}[U(k;N)] = \sum_{i=1}^{d} m_i(k)$.
Hence, with Equation 3.18, 3.19 and 3.21 it follows:

$$\Delta_i(k;N) = \frac{\sum_{n=1}^{N} X_k^{(i)}(n)}{\sum_{n=1}^{N} U^{(n)}(k)}. \tag{3.22}$$

The Strong Law of Large Numbers

The *law of large numbers* (*LLN*) will serve as a basis for the following steps. In probability theory the *LLN* describes the result of performing the same experiment for a large number of trials. Let $X_1, ..., X_n$ be a sequence of *i.i.d* distributed random variables with finite expected value $\mu = \mathbb{E}[X_1] = \mathbb{E}[X_2] = ... = \mathbb{E}[X_n] < \infty$. The strong *LLN* by Cantelli and Kolmogoroff states that the sample average converges *a.s.* to the expected value μ as $n \to \infty$.
Hence, as stated by Georgii [54]:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \overset{a.s.}{\to} 0. \tag{3.23}$$

which is equivalent to

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{a.s.}{\to} \mu. \tag{3.24}$$

For each particle type i the normalized mean $q_i(k)$ as defined in Equation 3.15, can be interpreted as the probability for a randomly chosen particle i at time point k to be of type T_i. Further, considering the fractions $\Delta_i(k;N)$ as defined in Equation 3.22 , $\Delta_i(k;N)$ can be seen as a strongly consistent estimator for $q_i(k)$. With a strong *LLN* one can obtain that $\Delta_i(k;N)$ converges to $q_i(k)$ as $N \to \infty$:

$$\Delta_i(k;N) \overset{a.s.}{\to} q_i(k). \tag{3.25}$$

As stated by Yakovlev and Yanev [126], the difference $\Delta_i(k;N) - q_i(k)$ can be obtained using Equation 3.22:

$$
\begin{aligned}
\Delta_i(k;N) - q_i(k) &= \frac{1}{\sum\limits_{n=1}^{N} U^{(n)}(k)} \left\{ \sum_{n=1}^{N} X_k^{(i)}(n) \right\} - q_i(k) \\
&\overset{+0}{=} \frac{1}{\sum\limits_{n=1}^{N} U^{(n)}(k)} \left\{ \sum_{n=1}^{N} \Big(X_k^{(i)}(n) + m_i(k) - m_i(k) \Big) \right\} - q_i(k) \\
&= \frac{1}{\sum\limits_{n=1}^{N} U^{(n)}(k)} \left\{ \sum_{n=1}^{N} \Big(X_k^{(i)}(n) - m_i(k) \Big) + \sum_{n=1}^{N} m_i(k) \right\} - q_i(k) \\
&\overset{Equ.3.21}{=} \frac{1}{U(k;N)} \left\{ \sum_{n=1}^{N} \Big(X_k^{(i)}(n) - m_i(k) \Big) + \sum_{n=1}^{N} m_i(k) \right\} - q_i(k) \\
&= \frac{1}{U(k;N)} \left\{ \sum_{n=1}^{N} \Big(X_k^{(i)}(n) - m_i(k) \Big) + \sum_{n=1}^{N} m_i(k) - q_i(k) U(k;N) \right\} \\
&\overset{Equ.3.21}{=} \frac{1}{U(k;N)} \left\{ \sum_{n=1}^{N} \Big(X_k^{(i)}(n) - m_i(k) \Big) + \sum_{n=1}^{N} m_i(k) - q_i(k) \sum_{i=1}^{d} X_k^{(i)}(N) \right\} \\
&\overset{Equ.3.20}{=} \frac{1}{U(k;N)} \left\{ \sum_{n=1}^{N} \Big(X_k^{(i)}(n) - m_i(k) \Big) + \sum_{n=1}^{N} m_i(k) - q_i(k) \sum_{n=1}^{N} \sum_{i=1}^{d} X_k^{(i)}(n) \right\} \\
&\overset{\cdot \frac{q_i(k)}{q_i(k)}}{=} \frac{1}{U(k;N)} \left\{ \sum_{n=1}^{N} \Big(X_k^{(i)}(n) - m_i(k) \Big) - q_i(k) \Big(\sum_{n=1}^{N} \Big[\sum_{i=1}^{d} X_k^{(i)}(n) \Big] - \frac{m_i(k)}{q_i(k)} \Big) \right\} \\
&\overset{Equ.3.15}{=} \frac{1}{U(k;N)} \left\{ \sum_{n=1}^{N} \Big(X_k^{(i)}(n) - m_i(k) \Big) - q_i(k) \Big(\sum_{n=1}^{N} \sum_{i=1}^{d} X_k^{(i)}(n) - \sum_{j=1}^{d} m_j(k) \Big) \right\} \\
&= \frac{1}{U(k;N)} \left\{ \sum_{n=1}^{N} \Big(X_k^{(i)}(n) - m_i(k) \Big) \right. \\
&\qquad \left. - q_i(k) \sum_{n=1}^{N} \sum_{j \neq i}^{d} \Big(X_k^{(j)}(n) - m_j(k) + X_k^{(i)}(n) - m_i(k) \Big) \right\} \\
&= \frac{\sqrt{N}}{U(k;N)} \left\{ \sigma_i(k)[1 - q_i(k)] V_i(k;N) - q_i(k) \sum_{j \neq i}^{d} \sigma_i(k) V_j(k;N) \right\}
\end{aligned}
$$

with

$$V_i(k;N) = \frac{\sum_{n=1}^{N} \left(X_k^{(i)}(n) - m_i(k) \right)}{\sigma_i(k)\sqrt{N}}$$

and $\mathbb{E}[V_i(k;n)] = 0$ and $Var[V_i(k;N)] = 1$.

With the Central Limit Theorem by Lindeberg and Levy [23] the random variable $V_i(k;N)$ converges in distribution to a normally distributed variable:

$$V_i(k;N) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1). \tag{3.26}$$

A vector of *i.i.d* random variables is said to be multivariate normally distributed if every linear combination of its components has an univariate normal distribution [54]. By this the following holds:

$$(V_1(k;N), ..., V_d(k;N)) \xrightarrow{\mathcal{D}} (X_k^{(1)}(N), ..., X_k^{(d)}(N)) \tag{3.27}$$

whereas the random variables $\vec{X}_k(N)$ have a joint normal distribution with $\mathbb{E}[X_k^{(i)}] = 0$ and $Var[X_k^{(i)}] = 1$.

With the notations:

$$W_i(k;N) = \sum_{i=1}^{d} m_i(k)\sqrt(N)[\Delta_i(k;N) - q_i(k)] \tag{3.28}$$

and the linear combination of multivariate normal random variables:

$$Y_i(k) = \sigma_i(k)[1 - q_i(k)]V_i(k;N) - q_i(k)\sum_{j \neq i}^{d} \sigma_i(k)V_j(k;N) \tag{3.29}$$

the following convergence holds in distribution with $N \to \infty$:

$$W_i(k;N) \xrightarrow{\mathcal{D}} Y_i(k) \tag{3.30}$$

These observations are used in the next chapter, where the obtained limiting result of multi-type BGW branching processes is of advantage in the study of the evolution of two different allele types A_1 and A_2 in an amplification process with an initial large number of fragments.

3.4 Sequence Fragments after Amplification

3.4.1 Fragment Amplification as a BGW Branching Process

As introduced in Section 1.1.1 the enrichment of *DNA* fragments during sample preparation includes an amplification step of a certain number of cycles of *PCR* which is also illustrated in Figure 1.1. The assumption of *binomially* distributed *allele frequencies* (*AFs*) at *heterozygous* loci is commonly used as prior knowledge in many currently used variant calling programs (e.g. *SAM*tools [69] or *GATK* [83, 38, 117], Section 1.2.2). However, as shown in Figure 1.11, the distribution of *AFs* of *WES* experiments deviates strongly from the expected mean of 0.5 which can partly be explained by a bias occuring in the enrichment step before sequencing (Section 1.2.2). Additionally, read mapping algorithms tend to favor the reference allele. Both systematic biases are rather systematic which lead to a shifted mean of the *allele frequency* (*AF*) distribution but leads the variance unaffected, which cannot be observed in real data (Figure 1.7 and Figure 1.11).

Fragment Amplification as an inhomogenous Markov Chain

In the following the crucial step of amplification of sequence fragments during library preparation is modelled as a stochastic process or, to be more precise as an inhomogeneous *Markov* chain, with transition probabilities $p_{lm}(k)$ which are dependent of the parameter k, $k \in K$ (see Section 3.2.1).

Polya's Urn Model

To illustrate this model, one can think of a tube that initially contains a set of fragments with different *alleles* such as depicted in Figure 3.2. The amplification process can be seen as a *Polya's Urn* model whereas balls are drawn from an urn and thrown back together with additional balls from the same color. In the context of fragment amplification, performing a finite number of *PCR* cycles

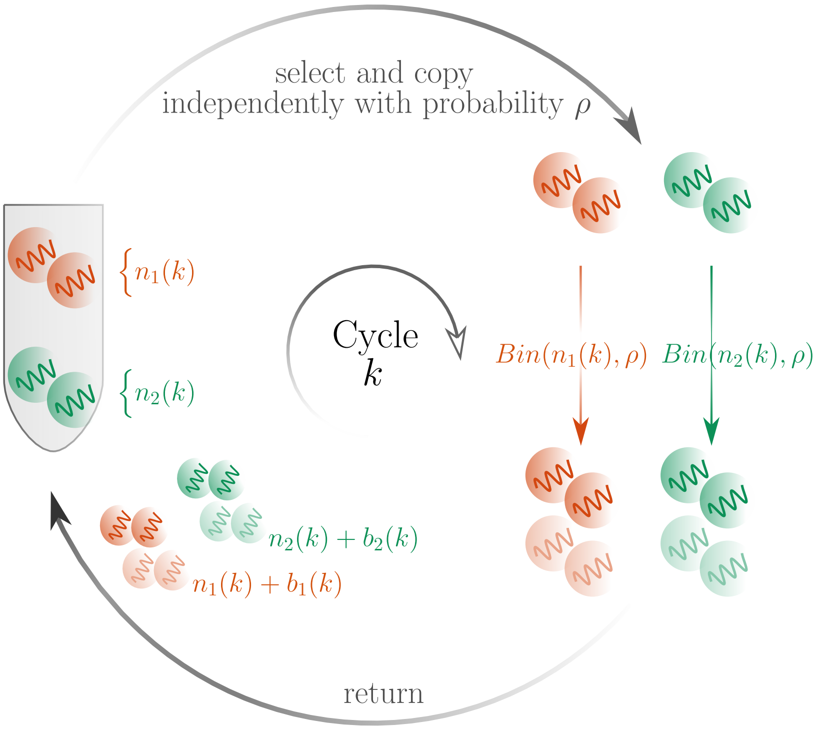


Figure 3.2: *The amplification of heterozygous alleles before sequencing.* (Adapted from [116]) Consider a tube that initially contains a set of different *alleles*. In each *PCR* step k a certain number of *allele* type A_i, $n_i(k)$, is drawn from this tube and replaced by $n_i(k) + b_i(k)$ whereas $b_i(k)$ is the result of a *binomially* distributed random variable depending on the cycle- and *allele* independent variable ρ that gives gives the probability that a fragment is copied.

$K_{\geq 0}$ on each *allele* $A_{i_{i \in \mathbb{N}}}$ is the same as performing a *Polya's Urn* experiment. Note that the process can be applied to all sequence variants but will just be studied on *bi-allelic* (*autosomal*) loci in the following. Hence, considering only fragments which contain one variable base of a *SNV* two classes of fragments can be distinguished containing either *allele* A_1 or *allele* A_2 with $n_i(k)_{i=\{1,2\}}$ being the number of *alleles* of type i at cycle step k.
The following assumptions are made:

1. The fragmentation should ideally be random and unbiased

which means that the extensions into both directions from the variable position is uniform and only limited by the fragment size.

2. The initial number of *alleles* $n_1(k=0)$ and $n_2(k=0)$ are in the same order of magnitude with the assumption that the *DNA* originates from many cells containing the *DNA* of the same diploid genome with a total initial number of *alleles* $N = \sum_{i=1}^{2} n_i(0)$.

Efficiency of the PCR reaction ρ

The distribution of *allele frequencies* (*AFs*) after amplification depends partly on the cycle independent efficiency of the *PCR* reaction represented as parameter ρ_i for each *allele* and partly on the probability that an *allele* is amplified. Additionally $\rho = \rho_1 = \rho_2$ if considering an *allele* independent amplification process as done in the following.

Prior to sequencing sequencing adaptor oligomers are ligated to the fragments and a *PCR* is applied for K cycle steps. For a successful amplification adaptors must be attached to both ends of the fragment and the attachment of the polymerase to the adaptor can be seen as a prerequisite.

The success of this event only depends on the total number of polymerase molecules which remains stable for k and N and in this work it is preconditioned that a constant number of sequence fragments will always be bound by polymerase in every cycle.

The probability that a randomly chosen *allele* A_i is copied in each *PCR* cycle k can be described as the transition probability of a *Markov* chain. Due to the different amount of fragments in each cycle k the process is further characterized as *inhomogeneous*.

With the *Markov* condition:

$$\mathbb{P}\Big[\Big((n_1(k), n_2(k)\Big)\Big|\Big(n_1(k-1), n_2(k-1)\Big), ..., \Big(n_1(0), n_2(0)\Big)\Big] \\ = \mathbb{P}\Big[\Big((n_1(k), n_2(k)\Big)\Big|\Big(n_1(k-1), n_2(k-1)\Big)\Big]$$

the probabilities of the three possible transitions of a randomly chosen *allele* A_i in cycle step k can be denoted as

$$\begin{aligned}
\mathbb{P}\big(n_1(k), n_2(k)\big) \to (n_1(k)+1, n_2(k)) &= \frac{n_1(k)}{n_1(k)+n_2(k)} \cdot \rho \\
\mathbb{P}\big(n_1(k), n_2(k)\big) \to (n_1(k), n_2(k)+1) &= \frac{n_2(k)}{n_1(k)+n_2(k)} \cdot \rho \\
\mathbb{P}\big(n_1(k), n_2(k)\big) \to (n_1(k), n_2(k)) &= 1-\rho,
\end{aligned}$$

whereby ρ is the probability that a fragment is copied. The ratio $\big[n_i(k)/\sum_i n_i(k)\big]_{i\in\{1,2\}}$ describes the proportion of *allele* A_i after the k^{th} amplification step and can be interpreted as the expected AF measured by sequencing multiple read fragments in this pool. With this definitions the system can be reduced to:

$$\Big(n_1(k+1), n_2(k+1)\Big) = \Big(n_1(k)+b_1(k), n_2(k)+b_2(k)\Big), \quad (3.31)$$

whereas $b_i(k)_{i\in\{1,2\}}$ are realizations of *binomially* distributed random variables $B(n_i(k),\rho)_{i\in\{1,2\}}$.

Fragment Amplification as a BGW Branching Process

The distribution of the fragments of two *allele* types A_1 and A_2 at the $k+1^{th}$ cycle step can be described as a cycle-dependent BGW branching process.

Therefore, let $\xi^{(i)}_{n.k}$ denote two independent triangular arrays of stochastically independent random *bernoulli* distributed variables for *allele* types $A_i \in \{1,2\}$:

$$\mathbb{P}(\xi^{(i)}_{n,k}=2) = \rho_i = 1-\mathbb{P}(\xi^{(i)}_{n,k}=1), \quad (3.32)$$

whereas n ranges from 1 to $n_i(k-1)$ and the parameter $\rho_i, i\in\{1,2\}$, reflects the probability of a successful amplification of one fragment. This leads to a two-type BGW branching process $(X^1_k, X^2_k)_{k\geq 0}$ with an initial number of $n_i(0)$ fragments. The following recursive definition holds:

$$X_k^{(i)} = \xi_{1,k}^{(i)} + \cdots + \xi_{(X_{k-1}^{(i)}),\, k}^{(i)} = \sum_{n=1}^{X_{k-1}^{(i)}} \xi_{n,k}^{(i)} \tag{3.33}$$

The focus in this work will be on the proportion of *alleles* of type A_i after k *PCR* cycles as defined in Equation 3.22. For two *allele* types A_1 and A_2, this can be described as

Proportion of alleles A_i after k Cycles, $\Delta_k^{(i)}$

$$\Delta_k^{(i)} = \frac{X_k^{(i)}}{X_k^{(1)} + X_k^{(2)}} \tag{3.34}$$

To study the asymptotic behaviour of $\Delta_k^{(i)}$ the *central limit theorem* (*CLT*) as proven by Yakovlev and Yanev [126] will be utilized for $n_1(0) \approx n_2(0) \approx N$ and $N \to \infty$:

$$(m_1(k) + m_2(k))\sqrt{N}(\Delta_k^{(i)} - q_i(k)) \xrightarrow{\mathcal{D}} Z^{(i)} \tag{3.35}$$

with

$$\begin{aligned} m_i(k) :=& \mathbb{E}[X_k(i) \mid X_0(i) = 1] \textit{ and} \\ q_i(k)) =& \frac{m_i(k)}{m_1(k) + m_2(k)}, i \in \{1, 2\}. \end{aligned}$$

It should be noted, that the limiting variable $Z^{(i)}$ is normally distributed with mean 0 and variance $a_i(k)^2$:

$$a_i(k)^2 = \sigma_1^2(k)(1 - q_i(k))^2 + \sigma_2^2(k)(q_i(k))^2, \tag{3.36}$$

whereby

$$\sigma_i^2(k) = Var[X_k^{(i)} \mid X_0^{(i)} = 1]_{i \in \{1,2\}}.$$

The entire process is determined by its probability generating function as introduced in Section 3.2.2 and will be used in Section 3.4.2

to derive the first two moments of the two-type *BGW* branching process $(X_k^{(1)}, X_k^{(2)})_{k\geq 1}$.

3.4.2 Variance of Δ_k^i

Derivation of the PGF for a Two-Type BGW Branching Process

Assuming independence of positions and individuals the marginal *PGF* of the two branching processes $X_k^{(1)}$ and $X_k^{(2)}$ can be considered for one *allele* type A_1 which is denoted by $\vec{F}_k(s)$ as defined in Equation 3.13.
Using Equation 3.33:

$$\begin{aligned}
\vec{F}_k(s) &= \mathbb{E}\Big[s^{X_k^{(1)}}\Big|n_1(0)\Big] \\
&= \mathbb{E}\Big[\mathbb{E}[s^{X_k^{(1)}}|X_{k-1}^{(1)}]\Big|X_0^{(1)} = n_1(0)\Big] \\
&= \mathbb{E}\Big[\mathbb{E}[s^{\sum_{n=1}^{X_{k-1}^{(1)}} \xi_{n,k}^{(1)}}|X_{k-1}^{(1)}]\Big|X_0^{(1)} = n_1(0)\Big]
\end{aligned}$$

By Equation 3.32 it follows for any $|s| \leq 1$ that:

$$\begin{aligned}
\mathbb{E}\Big[s^{\xi_{n,k}^{(1)}}\Big] &= \mathbb{P}[\xi_{n,k}^{(1)} = 1] \cdot s^{\Big[\xi_{n,k}^{(1)} = 1\Big]} + \mathbb{P}[\xi_{n,k}^{(1)} = 2] \cdot s^{\Big[\xi_{n,k}^{(1)} = 2\Big]} \\
&= (1 - \rho_1)s + \rho_1 s^2
\end{aligned}$$

and assuming stochastic independence of the ξ's leads to:

$$\vec{F}_k(s) = \mathbb{E}\Big[\Big((1-\rho_1)s + \rho_1 s^2\Big)^{X_k^{(1)}}\Big|n_1(0)\Big] \tag{3.37}$$

In the following $\big((1-\rho_1)s + \rho_1 s^2\big)$ will be substituted by $\varphi(s)$ and with this the *PGF* can be formulated as:

$$\vec{F}_k(s) = \mathbb{E}\Big[\varphi(s)^{X_{k-1}^{(1)}}\Big|X_0^{(1)} = n_1(0)\Big] \tag{3.38}$$

Composition of Functions

For the following the concept of *function composition* will be applied which is a point-wise application of one function $f_1(x)$ to the

result of another function $f_2(x)$ with $f_1(x) \circ f_2(x) = f_1(f_2(x))$.
With the notation $\varphi^1 = \varphi$ and the *composition* $\varphi^k = \varphi \circ \varphi^{k-1}$ for $k \geq 2$ it follows:

$$\begin{aligned}\vec{F}_k(s) &= \mathbb{E}\Big[\varphi \circ \varphi(s)^{X_{k-2}^{(1)}} \Big| X_0^{(1)} = n_1(0)\Big]\\ &= \mathbb{E}\Big[\varphi \circ \cdots \circ \varphi(s)^{X_0^{(1)}} \Big| X_0^{(1)} = n_1(0)\Big]\\ &= \Big(\varphi^k(s)\Big)^{n_1(0)}\end{aligned}$$

With Equation 3.14 and 3.16 one can formulate the following limiting behaviours:

$$m_1(k) = \lim_{s \to 1} \frac{\delta}{\delta s} \vec{F}_k(s) \tag{3.39}$$

$$\sigma_1^2(k) = \lim_{s \to 1} \frac{\delta s}{\delta s^2} \vec{F}_k(s) + m_1(k) - (m_1(k))^2. \tag{3.40}$$

For the special case $n_1(0) \equiv 1$ and $\varphi^k(s)$ is considered as a smooth function one can calculate the first and second derivative of $\varphi^k(s)$ with respect to s in 1.
Note that $\varphi(1) = ((1-\rho_1) \cdot 1 + \rho_1 \cdot 1^2) = (1 - \rho + \rho) = 1$ and $\varphi^k(1) = 1$ for all $k \geq 1$. Moreover, the first derivative of φ with respect to s is given by:

$$\varphi^{'}(s) = \frac{\delta}{\delta s}\varphi(s) = (1-\rho_1) + 2\rho_1 s, \tag{3.41}$$

which leads to $\varphi^{'}(1) = 1 + \rho_1$.
The second derivative is then given by:

$$\varphi^{''}(s) = \frac{\delta s}{\delta s^2}\varphi(s) = 2\rho_1. \tag{3.42}$$

By application of the *chain rule*, which can be generalized as $(f_1 \circ f_2)^{'} = (f_1^{'} \circ f_2)f_2^{'}$, one can obtain the following:

$$\begin{aligned} \frac{\delta}{\delta s}\varphi^k(s) &= \frac{\delta}{\delta s}\Big(\varphi(\varphi^{k-1}(s))\Big) \\ &= \varphi^{'}(\varphi^{k-1}(s)) \cdot (\varphi^{k-1}(s))^{'} \end{aligned}$$

and with iterating it follows:

$$\begin{aligned} \frac{\delta}{\delta s}\varphi^k(s) \mid_{s=1} &= (\varphi^k(1))^{'} \\ &\overset{Equ.3.41}{=} (1+\rho_1)^k \\ &\overset{Equ.3.39}{=} m_1(k). \end{aligned} \tag{3.43}$$

For the calculation of the second derivative $\frac{\delta^2}{\delta s^2}\varphi^k(s) \mid_{s=1}$, the *generalized chain rule* by Faà di Bruno [39] can be applied, which can be formulated as follows:

$$\frac{\delta^2}{\delta s^2}(f_1 \circ f_2)(x) = f_1^{''}(f_2(x))(f_2^{'})^2 + f_1^{'}(f_2(x))f_2^{''}(x). \tag{3.44}$$

And with $\varphi^k = \varphi \circ \varphi^{k-1}$ this leads to:

$$\frac{\delta^2}{\delta s^2}\varphi^k(s) = \varphi^{''}(\varphi(s)^{k-1})(\varphi^{'}(s)^{k-1})^2 + \varphi^{'}(\varphi(s)^{k-1})(\varphi^{''}(s)^{k-1})$$

Iterating leads to:

$$\begin{aligned}
\frac{\delta^2}{\delta s^2}\varphi^k(s)\mid_{s=1} &\overset{\varphi(1)^k=1}{=} \varphi^{''}(1)\varphi^{'}(1)^{2(k-1)} + \varphi^{'}(1)(\varphi^{''}(1)^{k-1}) \\
&= \varphi^{''}(1)\varphi^{'}(1)^{2(k-1)} + \varphi^{'}(1)\Big(\varphi^{''}(1)\varphi^{'}(1)^{2(k-2)} + \varphi^{'}(1)\varphi^{''}(1)^{k-2}\Big) \\
&= \varphi^{''}(1)\varphi^{'}(1)^{2(k-1)} + \varphi^{''}(1)\varphi^{'}(1)\varphi^{'}(1)^{2(k-2)} + \varphi^{'}(1)^2\varphi^{''}(1)^{k-2} \\
&= [...] \\
&= \sum_{l=0}^{k-1}\varphi^{''}(1)\varphi^{'}(1)^{l}\varphi^{'}(1)^{2(k-l-1)} \\
&= \varphi^{''}(1)\varphi^{'}(1)^{2k-2}\sum_{l=0}^{k-1}\varphi^{'}(1)^{-l} \\
&\overset{Equ.3.41}{=} \varphi^{''}(1)(1+\rho_1)^{2k-2}\sum_{l=0}^{k-1}(1+\rho_1)^{-l} \\
&\overset{Equ.3.42}{=} 2\rho_1(1+\rho_1)^{2k-2}\sum_{l=0}^{k-1}(1+\rho_1)^{-l}
\end{aligned}$$

By utilizing the sum formula of the first k terms of the *geometric series* $\sum_{l=0}^{k-1}(1+p_1)^{-l}$ this leads to:

$$\begin{aligned}
\frac{\delta^2}{\delta s^2}\varphi^k(s)\mid_{s=1} &= 2\rho_1(1+\rho_1)^{2k-2}\left(\frac{1-(\frac{1}{1+\rho_1})^{-k}}{1-(\frac{1}{1+\rho_1})}\right) \\
&= 2\rho_1(1+\rho_1)^{2k-2}\left(\frac{1-(1+\rho_1)^k}{\frac{\rho_1}{1+\rho_1}}\right) \\
&= 2(1+\rho_1)^{2k-1}-(1+\rho)^{k-1}
\end{aligned}$$

And with Equation 3.40 it follows:

$$\sigma_1^2(k) = (1+\rho_1)^{2k}\Big(2(1+\rho_1)^{-1}-2(1+\rho_1)^{-k-1}+(1+\rho_1)^{-k}-1\Big) \quad (3.45)$$

With Equation 3.36, the variance of the asymptotic normal distribution $\Delta_k^{(i)}$ is given by:

$$Var(\Delta_k^{(i))} = \frac{a_1(k)^2}{N(m_1(k)+m_2(k))^2}, \tag{3.46}$$

whereas for $i = 1$:

$$\begin{aligned}
a_1(k)^2 &= \sigma_1^2(k)(1-q_1(k))^2 + \sigma_2^2(k)(q_1(k))^2 \\
&\overset{Equ.3.15}{=} \sigma_1^2(k)\Big(1 - \frac{m_1(k)}{m_1(k)+m_2(k)}\Big)^2 + \sigma_2^2(k)\Big(\frac{m_1(k)}{m_1(k)+m_2(k)}\Big)^2 \\
&\overset{Equ.3.43}{=} \sigma_1^2(k)\Big(1 - \frac{(1+\rho_2)^k}{(1+\rho_1)^k + (1+\rho_2)^k}\Big)^2 \\
&\qquad + \sigma_2^2(k)\Big(\frac{(1+\rho_1)^k}{(1+\rho_1)^k + (1+\rho_2)^k}\Big)^2
\end{aligned}$$

Assuming the special case $\rho_1 = \rho_2 = \rho$, we get that $m_1(k) = m_2(k)$, $q_1(k) = q_2(k) = \frac{1}{2}$ and $\sigma_1^2(k) = \sigma_2^2(k) = \sigma^2(k)$, $\forall k \geq 1$. Furthermore, the asymptotic normal distributions $\Delta_k^{(1)}$ and $\Delta_k^{(2)}$ coincide if $n_1(0) \approx n_2(0)$.

With these pre-assumptions and Equation 3.45, the a asymptotic variance of $\Delta_k^{(1)}$ and $\Delta_k^{(2)}$ respectively is given by considering $a(k)^2$ for $\rho_1 = \rho_2 = \rho$:

$$\begin{aligned}
a(k)^2 &= \frac{1}{2}\sigma^2(k) \\
&= \frac{1}{2}(1+\rho)^{2k}\Big(2(1+\rho)^{-1} - 2(1+\rho)^{-k-1} + (1+\rho)^{-k} - 1\Big).
\end{aligned}$$

Utilization of Equation 3.46 finally gives the variance of the two-type BGW branching process:

$$
\begin{aligned}
&Var(\Delta_k^{(i)})|_{i\in\{1,2\}} = \\
&= \frac{\frac{1}{2}(1+\rho)^{2k}\Big(2(1+\rho)^{-1} - 2(1+\rho)^{-k-1} + (1+\rho)^{-k} - 1\Big)}{N(2(1+\rho)^k)^2} \\
&= \frac{\frac{1}{2}(1+\rho)^{2k}\Big(2(1+\rho)^{-1} - 2(1+\rho)^{-k-1} + (1+\rho)^{-k} - 1\Big)}{8N(1+\rho)^{2k}} \\
&= \frac{\Big(2(1+\rho)^{-1} - 2(1+\rho)^{-k-1} + (1+\rho)^{-k} - 1\Big)}{8N}
\end{aligned}
\tag{3.47}
$$

3.5 Experimental Validation

3.5.1 Experimental Whole Exome Sequencing Data

Sample Collection and Sequencing

Altogether, 17 anonymized donors where used for *WES*, obtaining either human blood or tissue samples. Additionally, 9 independent samples of the same individual were collected and further processed independently which will be referred to as technical replicates in the following subsections.

Genomic *DNA* was enriched for the target region of all human *CCDS* exons, with *Agilent*'s *SureSelect Human All Exon Kit* [2] for each sample and subsequently sequenced on an *Illumina Genome Analyzer II* [6] with 100 *bp single end reads.*

Number of PCR Cycle Steps k

The standard protocol including an amplification step of 18 *PCR* cycles was applied to all samples with the exception of one exome which was run with 36 cycles of *PCR* to further analyze the effect of the cycle number on the *allele frequency* (*AF*) distribution.

Additionally 35 cycles of *PCR* were run for all samples in a Cluster Generation step that follows the *PCR* step in the standard protocol (as described in Section 1.1.1) to increase the fluorescent signal of a fragment on the sequencing flow cell.

Alignment and Variant Calling

The raw sequencing data (≈ 5 *Gb* per sample) was mapped to the haploid reference genome *GRCh37* using *NovoAlign* [10] that yielded a mean *per-base coverage* of the exome target region (*CCDS*) of 50×.

3.5.2 Independency of Positions and Individuals

Dependency Between Positions

The dependency between genomic loci was tested by comparing the distribution of all *heterozygous AFs* in a pooled set of 17 *WES* data sets (Section 3.5.1) to a smaller randomly chosen subset of these positions (see Figure 3.3). A test of independence by χ^2 statistic did not show significant differences ($p = 0.234$) which suggest an association between both distributions.

Dependency Between Individuals

Dependence between individuals was tested by comparing the differences of *heterozygous allele* distributions between different individuals and technical replicates of the same individual. Differences in frequency distributions between individuals is statistically not significant and fluctuations in these distributions are comparable to those observed in technical replicates of the same individual.

3.5.3 Simulation of Allele Frequencies After Amplification

For an experimental validation of the analytically derived variance of *allele frequencies* (*AFs*) after amplification, $Var(\Delta_k^{(i)})$ as given in Equation 3.47 the variance was simulated for different parameter settings of the amplification efficiency ρ, the initial number of *alleles* N and for the cycle number k (see Algorithm 1).
Each amplification step was simulated for a *per-base coverage* of 20× for $10,000$ *heterozygous* positions which is in the expected order of magnitude for *heterozygous* variant calls in a human *WES* sample.

Simulations for Fixed N and k

The simulations are well approximated by the analytical result for initial numbers of fragments $N \geq 5$. In case of fixed parameters $k = 18$ and $N \in \{1, 3, 5, 10\}$ the function of $Var(\Delta_k^{(i)})$ reaches its maximum at an amplification efficiency of $\rho \sim 0.2$ and decreases towards $\rho = 1$ which corresponds to a perfect amplification (Figure 3.4). Overall as N increases the simulated as well as the derived variance (Equation 3.47) is constantly shrinking and approaches a fixed level. This can be explained solely by the variance introduced by the measurement process of sequencing.

Adaptation of $Var(\Delta_k^{(i)})$

By adapting the analytically derived variance as well as the simulated results towards an additional contribution of variance which

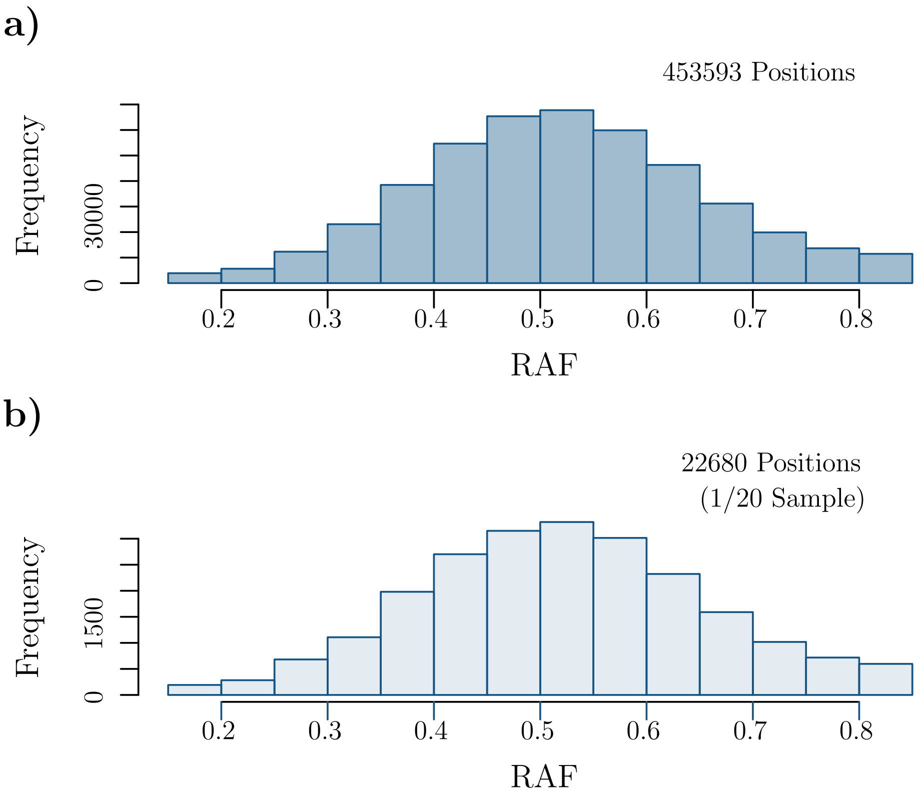


Figure 3.3: *Allele frequency distributions at heterozygous sites are position- and individual independent.* (Adapted from [116]) **a)** Frequency distribution of the reference *allele* for all positions and all individuals pooled. **b)** Frequency distribution for a random set comprising 5% of all positions in a randomly chosen individual. Comparison of the two distribution shown in **a)** and **b)** with *Pearsons* χ^2 test yields a p-value of $p = 0.234$, which corresponds to the distributions having a not further specified association and are not independent.

is introduced during sequencing values comparable to real *WES* experiments can be achieved (see Algorithm 2).

Further statements about the limiting behaviour of the variance can be made by an alternating fixating of two out of the three depending parameters N, k and ρ (Figure 3.5).

Constant variance level for $k \geq 15$ and $N \geq 10$

The variance increases with a growing number of *PCR* cycles k, for fixed values of N and ρ, and approaches a constant level for $k = 15$ which leads to the argumentation, that an increase in the number of cycles during library preparation ($k \geq 15$) as well as amplification of the cluster generation step that succeeds the library preparation will only contribute marginally to the total variance.

Overall a generalized conclusion can be drawn: assuming one *allele* type A_i is preferred to be sequenced it is easier for this *allele* to gain predominance in the pool of *alleles* that is sequenced if first the initial set of alleles is small ($N \leq 10$), the amplification efficiency is low ($\rho \leq 0.2$) and enough *PCR* cycles are run ($k \geq 15$).

Algorithm 1: Simulation of $Var(\Delta_k^{(i)})$.

```
Result: Simulation of the variance of AFs after the
        amplification process: $Var(\Delta_k^{(i)}(N,\rho))_{simulated}$
Data:
$\rho$: fixed amplification efficiency;
$N$: fixed number of initial fragments;
$k$: fixed number of cycles;
Ratio_of_Alleles[‖SNVs‖]: ratio after amplification;
foreach $b$ in number of SNVs do
    Intital number of alleles $A_i, i \in \{1,2\}$;
    $n_i(0) = N$;
    foreach $\hat{k}$ in $k$ do
        The number of copies of allele $A_i$ in Cycle $\hat{k}$ is drawn
        from a binomial distribution;
        $Number\_of\_Copies_{A_i} = Bin(n_i(\hat{k}-1), p)$;
        Add number of copies of allele $A_i$ to $n_i(\hat{k}-1)$;
        $n_i(\hat{k}) = n_i(\hat{k}-1) + Number\_of\_Copies_{A_i}$;
    end
    Ratio of allele $A_1$ after amplification;
    Ratio_of_Alleles[$b$] $= n_1(k) / \sum_{i=1}^{2} n_i(k)$
end
$Var(\Delta_k^{(i)}(N,\rho))_{simulated} = Var($Ratio_of_Alleles$)$;
```

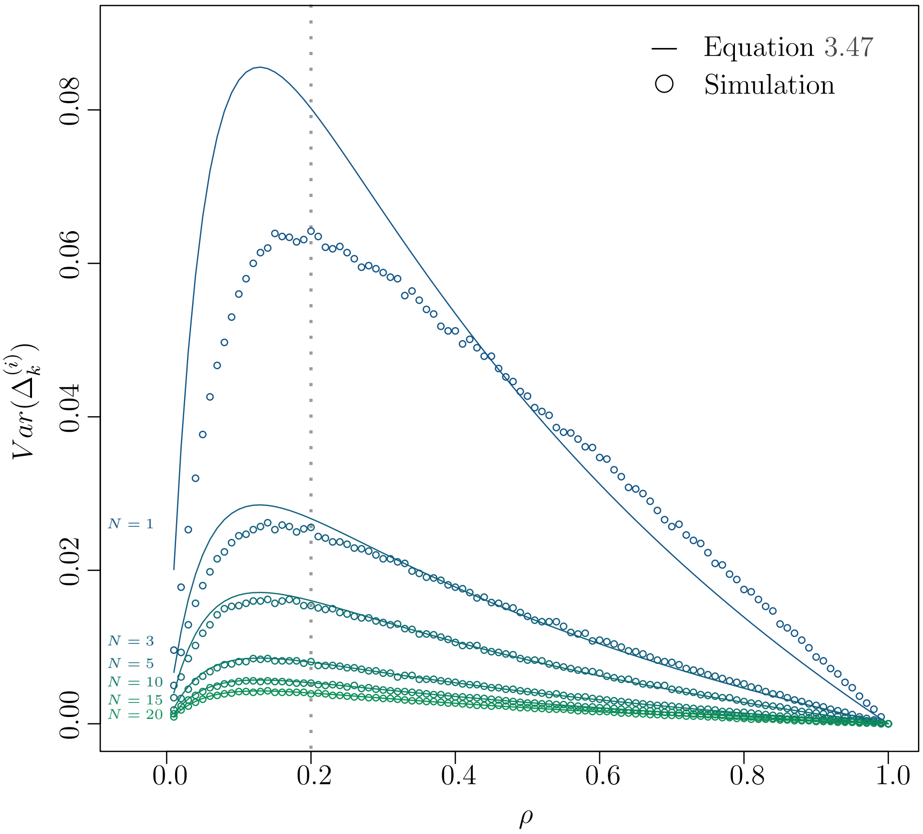


Figure 3.4: *Simulation of* $Var(\Delta_k^{(i)})$. (Modified from [116]) The variance of of the *allele frequency* (AF) after amplification $Var(\Delta_k^{(i)})$ was sampled from simulations which is described in Algorithm 1 for fixed values of the cycle number $k = 18$ and various values of the initial number of fragments $N \in \{1, 3, 5, 10, 15, 20\}$ (circles). Additionally the variance is analytically derived from Equation 3.47 (solid lines).
The function reaches its maximum around an amplification efficiency of $\rho = 0.2$ (gray dotted line) and decreases for ρ towards 1 which correspond to a perfect amplification.
For very small initial numbers of N the simulated results do not fit perfectly to the analytically derived curve but are in agreement starting from an initial number of *alleles* of $N \geq 5$.

Algorithm 2: Adaptation of $Var(\Delta_k^{(i)})$

```
Result: $Var(\Delta_k^{(i)})$ adapted by an additional variance
        introduced during sequencing: $Var(\Delta_k^{(i)}(N))_{\text{adapt}}$
Data:
$c$: per-base coverage;
$Var\left(\Delta_k^{(i)}(N)\right)$: Variance After Amplification (Equ. 3.47);
foreach $b$ in number of SNVs do
    The AF after amplification is a random number drawn from a
    normal distribution:;
    $\text{AF\_After\_Amplification} \in \mathcal{N}(0.5, \sqrt{Var\left(\Delta_k^{(i)}(N)\right)});$
    The AF after sequencing is a random number drawn from a
    binomial distribution:;
    $\text{AF\_After\_Sequencing}[b] \in Bin(c, \text{AF\_After\_Amplification});$
end
$Var(\Delta_k^{(i)}(N))_{\text{adapt}} = Var(\text{AF\_After\_Sequencing}[b]);$
```

Allele Frequency Distribution of Real WES samples

The distribution of *AFs* at *heterozygous* genomic positions was also analysed in 17 real human *WES* data sets that were generated following the standard protocol (see section 3.5.1), including $k = 18$ cycles of *PCR* to verify the findings obtained in the previous simulations. Furthermore, one *WES* sample was sequenced twice with the solely difference that $k = 36$ cycles of *PCR* steps were applied in order to experimentally check if the variance remains stable above a cycle number of $k \geq 15$, as observed in Figure 3.5 *a)*.

Down-Sampling of Coverage

For a fair comparison to the analytical derived results each position in all samples was randomly down-sampled to a *per-base coverage* of 20 *reads*. A position was defined as *heterozygous* by applying a simple *percent rule call* , $0.14 < RAF < 0.86$ as proposed by *Bell et al.* [21] and as already described in Section 1.2.2.

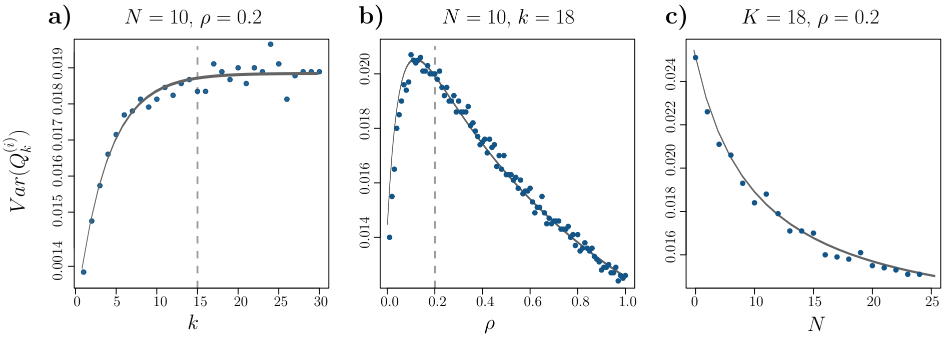


Figure 3.5: *The fragment amplification as a stochastic branching process.* (Adapted from [116]) The variance of the *allele frequency* (*AF*) was sampled from simulations as well as analytically derived (solid line) for different initial values for the cycle number k, the amplification efficiency ρ and the number of starting alleles N. Every simulation of the measurement process of sequencing was done for a *read coverage* of 20×. The variance sampled from 10 000 simulated *heterozygous* *SNV*s (blue circles) is well approximated by the analytical results of Equation 3.47 (solid gray lines).
a) The variance approaches a fixed level for a cycle number $k > 15$ and fixed values of ρ and N (gray dashed line). **b)** It reaches its maximum for an amplification efficiency around $\rho = 0.2$ (gray dashed line). **c)** For an increasing number of starting *alleles* N before amplification the variance approximates a fixed level which is explained solely by the variance introduced by the measurement process of sequencing. This could already be observed in Figure 3.4 where the variance started to approach a fixed level at around $N = 10$.

Simulation of Allele Frequencies

Random numbers in the dimension of the measured *AFs* were drawn from a *binomial* distribution whereby the number of trials is equal to the number of initial fragments $n_1(0) = n_2(0) = N$ and the success parameter is equal to the mean of the empirically measured *AFs* (see Figure 3.6 **a)**).

By comparing the two distribution one can observe that the variance at *heterozygous* positions which was obtained by the measured *RAF* distribution is 0.017 and thus much larger than the variance of 0.012 which is expected by hypothetical sequencing before amplification, derived by a *binomial* distribution. That leads to the

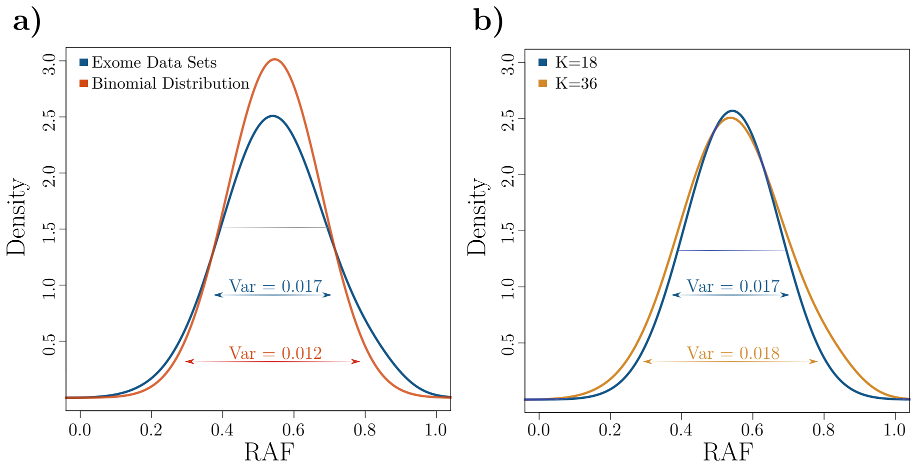


Figure 3.6: *Measured variance in WES data sets.* (Adapted from [116]) **a)** The variance of the distribution of *heterozygous allele frequencies* (*AFs*) measured in *WES* data sets at a *per-base coverage* of 20× (blue) is significantly larger compared to the theoretical distribution expected before amplification (red). **b)** An exome of the same individual was sequenced following $k = 18$ and $k = 36$ cycles of amplification and the variance of the *AFs* only slightly increases after the additional 18 cycles. This is in agreement with the finding observed in Figure 3.5 **a)** where $Var(\Delta_k^{(i)}(N))$ approaches a fixed level at around $k \geq 15$ with fixed parameters ρ and N.

conclusion that the sequence fragments in a short *read* alignment, on which variant calling is performed, are not properly represented by a random sample of the distribution of initial fragments but the effect of the amplification process on this distribution has to be taken into account.

Comparison of different Cycle Numbers k

A single *WES* sample was sequenced using 18 and 36 rounds of *PCR* cycles and the same analysis was applied. As already observed in the analytical and simulated results in Figure 3.5 **a)** no significant increase in the variance of the derived *AFs* could be detected (see figure 3.6) **b)**.

Amplification Efficiency $\rho \in [0.3, 0.5]$

In all simulations a constant amplification efficiency ρ is assumed over all *PCR* cycles k which is a reasonable simplification of the process, taking the relatively low number of usually ~ 18 *PCR* cycles into account, which are used in *NGS* library preparation protocols (see Section 1.1.1). An amplification efficiency of $\rho \in [0.3, 0.5]$ yields a variance of the *AFs* that is close to the variance observed in real *WES* data sets (compare Figure 3.5 **b)** and Figure 3.2 **a)**). A value of $5ng$ of initial fragmented *DNA* ($k = 0$) and $5 - 10\mu g$ of *DNA* after $k = 18$ cylces of *PCR* was measured, which corresponds to an amplification by a factor of $(1 - 2 \times 10^3)$ which is in agreement of an amplification efficiency of $\rho \in [0.3, 0.5]$.

3.5.4 Influence on Error Rates in heterozygous Variant Detection

Strong Deviations in Allele Frequencies

As already exemplary illustrated in Section 1.2.2 *heterozygous* variants with *AFs* that are strongly deviating from the expected mean of 0.5 correlate to a high variance in the distribution of *AFs*. These may be misclassified by existing variant calling algorithms which assume an underlying *binomial* distribution (see Figure 1.7).

Another example is given in Figure 3.7 **a)** where a common *heterozygous SNP* position is highlighted that is present in 5 technical replicates. In 4 out of 5 replicates the position is classified as *heterozygous* but due to low frequency in the fifth sample (v) this position was misclassified as homozygous.

Based on these observations the hypothesis can be formulated that a certain rate of *true positive* (*TP*) *heterozygous* variants will not be detected by commonly used variant calling algorithm due to the high variance in *AFs* after amplification. To proof this the influence of the variance of *AFs* on error rates in the detection of *heterozygous* variant calls will be analysed in more detail in the following.

Assuming a comparable quality for all *Reads* each variant call is based on a random sample that is drawn from the set consisting of *alleles* A_1 and A_2 after amplification which is of size $n_1(k) + n_2(k)$. Consequently the sequencing depth at a *SNV* location is equal to the size of the random sample on which the call is based on.

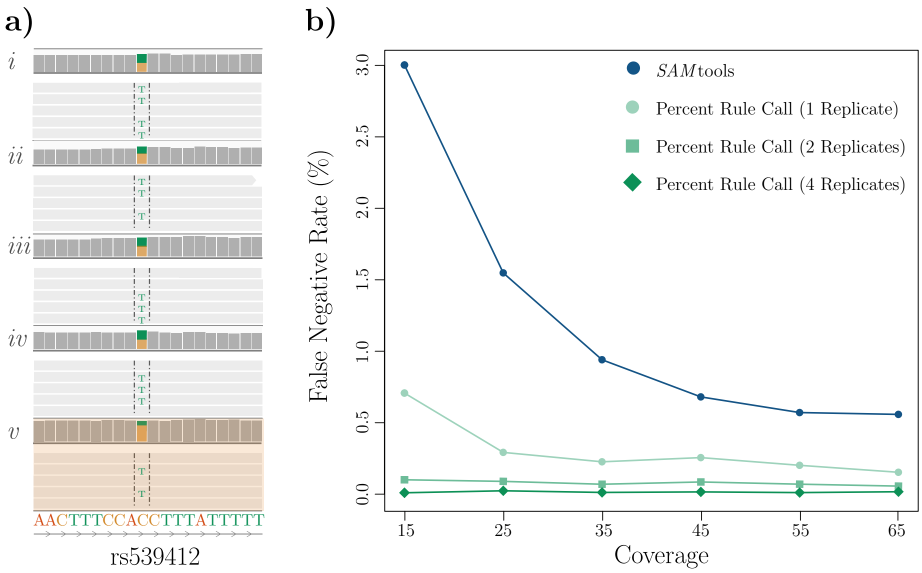


Figure 3.7: *Influence of variance in measured allele frequency on variant calling.* (Adapted from [116]) **a)** The *GT* at a common *SNP* position *rs539412* could be identified as *heterozygous* variant in 4 technical replicates (*i* - *iv*) but could not been detected in the fifth replicate due to low *AF* (*v*). **b)** To measure the *false negative rate* (*FNR*) for *SNP* positions that were classified as *heterozygous* by *SAM*tools or by the *PRC* (14% <*RAF*≤ 86%), a *TP* call was defined, when at least six out of nine replicates where called *heterozygous* with the chosen calling algorithm. Variant calling was performed on randomly drawn sets of alleles to achieve comparable *per-base coverages* of 15×, 25×, 35×, 45×, 55× and 65×. Overall the *FNR* decreases with an increasing sequencing depth. The classification of a *heterozygous* variant based on a frequency interval (*percent rule call* (*PRC*), green circles) is more sensitive than a calling algorithm that uses a *binomial* prior distribution (*SAM*tools, blue circles). At low total sequencing depth the *FNR* can be markedly reduced by creating pools of two (green squares) and four (green rhombus) replicates for which the *PRC* was applied.

Definition of TP heterozygous Variant Calls

A *TP heterozygous* position was defined by using nine technical replicates of a *WES* sample and the following conditions:

1.) The variant is listed in *dbSNP*.

2.) The variant is covered by at least 15 *reads*.

3.**a)** The variant is called in at least 6 out of 9 technical replicates by the *percent rule call* ($0.14 < RAF < 0.86$) [21] or

3.**b)** The variant is called in at least 6 out of 9 technical replicates by *SAM*tools [13].

FNR and FPR for heterozygous Variant Calls

Using this *TP* set as a *gold standard*, the *FNR* as well as the *false positive rate* (*FPR*) could be computed for different categories of *coverage* for *heterozygous* calls on each *WES* data set as listed in Table 3.1. Overall the *FNR* decreases with an increasing sequencing depth and over the whole exome a *FNR* between $1.0 - 3.0\%$ and a *FPR* in the range of $0.1 - 2.9\%$ was measured depending on the *coverage* category and the calling approach.

Table 3.1: *FNR and FPR for variant calling depending on different coverages.* (Adapted from [116]) The *GT* calling approach proposed by *Bell et al.* [21] (*PRC*) classifies a genomic position as *heterozygous* variant if $0.14 < RAF \leq 0.86$ whereas *SAM*tools utilizes a prior *binomial* distribution. Depending on the *per-base* depth all error rates decrease, whereby the frequency based method (*PRC*) has lower values of *FNR* as *SAM*tools for all categories of *coverage*. Additionally the *FNR* as well as the *FPR* approaches a stable level for sequencing depth around ≥ 35 independent of the calling algorithm. For *per-base coverages* of ≥ 35 the values for *FNR* as well as *FPR* remain relatively stable independent of the calling algorithm.

Coverage	15-24	25-34	35-44	45-54	55-64	65-74	75-84	85-94	95-104
FNR *SAM*tools	3.0	1.5	0.9	0.6	0.6	0.6	0.7	0.5	0.6
FPR *SAM*tools	2.6	0.8	0.3	0.1	0.1	0.1	0.1	0.1	0.1
FNR *PRC*	0.7	0.3	0.2	0.3	0.2	0.2	0.2	0.2	0.2
FPR *PRC*	2.9	1.0	0.4	0.2	0.2	0.2	0.2	0.2	0.3

For the condition 3.**a)** pools of 2 and 4 technical replicates were considered and the mean of all *FNR*s for all replicates in a pool was calculated (Figure 3.7 **b)**). At low sequencing depths the error rate could be markedly reduced by considering pools of technical replicates instead of just one sample. This observation indicates that, once a sufficient sequencing depth is achieved an additional reduction of the total error rates can only be reached by technical replication.

Pools of Replicates

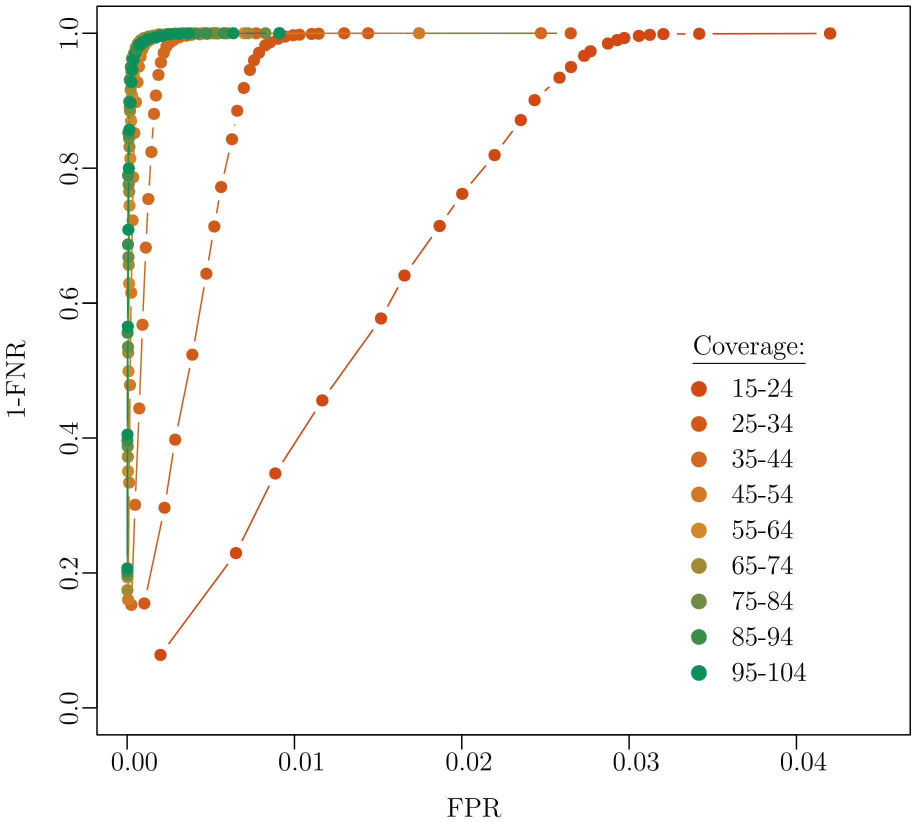


Figure 3.8: *ROC curves for heterozygous genotype calling.* (Adapted from [116]) A genomic position was defined as *heterozygous* if $| 0.5 - RAF | < c$, with cut-off thresholds $c \in [0.02, 0.04, ..., 0.48]$. The area under the *ROC* curve (*AUC*) increases considerably for the first three categories of *coverage*. However, for a *per-base coverage* $\geq 35\times$ no significant change can be observed in *AUC* and for | 0.5 - *RAF* |< 0.36, the *FPR*s are comparable to the error rates of *SAM*tools

Usually one would expect about $10\,000 - 15\,000$ *heterozygous* variants in an *WES* sample and the obtained *FNR*s indicate that ~ 100 *heterozygous* variants will be missed just due to stochastic fluctuations of the *AFs* after amplification.

Sensitivity vs. Specificity

These observations lead to the conclusion that a variant calling approach that is simple based on a *heterozygous* *AF* interval (*PRC*, 14% $<$*RAF*$\leq$ 86%) has a higher sensitivity at a comparable specificity that an algorithm that uses a *prior* distribution that is not perfectly suited for the *allele frequencies* (*AFs*). This can also be observed in the *ROC* analysis displayed in Figure 3.8 for different categories of *coverage* depending on different cut-offs $c \in [0.02, 0.04, ..., 0.48]$ for *heterozygous* variant calling: $\mid 0.5 - RAF \mid < c$.

The *AUC* increases markedly for the first three categories of *coverage* whereas for a *per-base coverage* ≥ 35 no difference can be seen. However, the *FPR* are comparable to the error rates of *SAM*tools as shown in Table 3.1 for $\mid 0.5 - RAF \mid < 0.36$.

3.6 Summary of this Chapter

In this chapter the distribution of *allele frequencies* (*AFs*) at *heterozygous* genomic positions was studied and it's implication on variant calling was analysed.

Examples as given in Figure 1.7 indicate that the total variance of *AFs* at *heterozygous* genomic loci is strongly influenced by an amplification process during library preparation in *next generation sequencing* protocols (as introduced in Section 1.1.1). Furthermore a solid knowledge of the distribution of *AFs* is essential as many variant calling approaches such as *SAM*tools or *GATK* rely on this *prior* information.

Therefore the fragment amplification of sequence fragments was described as a two-type *(Bienayme-) Galton-Watson* branching process with discrete time k steps which are interpreted as *cycles* in a *polymerase chain reaction*. The variance of this distribution $Var(\Delta_k^{(i)})$ could be accurately described by Equation 3.47 for two *allele* types $a_i, i \in \{1,2\}$, and validations based on simulations and real *whole exome sequencing* data correspond well with the analytical derived findings.

The final equation is depending on different parameters namely the amplification efficiencies ρ_i, the number of *PCR cycles* k and the amount of initial fragments before amplification $n_i(k = 0)$ whereby all simulations were done assuming the simplification that $\rho_1 = \rho_2 = \rho$ and $n_1(0) + n_2(0) = N$.

It could be shown that for typical values of the amplification efficiency $\rho \in [0.3, 0.5]$ and for a sequencing depth of 20× the analytical derived variance is considerably higher than the variance of the corresponding *binomial* distribution that is usually used as *prior* probability distribution in variant calling algorithm based on *Bayesian* methods.

The correlation of high variances in *AFs* at *heterozygous* positions and error rates in variant detection could be demonstrated: the higher the variance the higher the *false negative rate*. In the analysis of single *WES* samples the *FNR* approaches a fixed level at a *per-base coverage* of about ≥ 35 *reads*. However, pooling of technical replicates of *whole exome sequencing* samples is an efficient approach to further decrease the *FNR* and yields even better results than just sequencing more *reads* from the same library.

Overall conclusions can be drawn to minimize stochastic fluctuations arising from the amplification step during library preparation in *WES* experiments. Increasing both N and ρ and simultaneously decreasing the number of *PCR cycles* k in 2^{nd} generation sequencing protocols will reduce the overall variance in the distribution of *AFs* at *heterozygous* loci.

4 | Exome Genotyping Accuracy in Next Generation Sequencing Data

4.1 Overview of this Chapter

This chapter will address the importance to measure an overall quality to a whole set of *GT* data obtained from a single *WES* experiment, which was already discussed in Section 1.3.1.

Various quality control measurements can be applied to an entire variant call set and recommendations for the sequencing depth and *coverage* over the target region of an exome exist [80]. However, these parameters are valuable indicators of quality, but they do not directly indicate the accuracy of a sequenced exome. For instance, the *Ti*-*Tv* ratio is strongly influenced by the target region and the number of novel variants within an exome also depends on the background population. Further, the *phred-like* quality score that is provided for each called variant, is based on a certain likelihood model for genotypes and depends not just on the raw data but also on the applied alignment and calling algorithm. Applying different algorithms to the same sample yields different quality scores.

The following chapter will show a new approach that calculates the distance of an exome to a *reference set* of high quality and uses this as an indicator for the genotyping accuracy.

A short introduction will provide an overview about similarities, distance metrics and methods to reduce the dimensionality of a high dimensional dataset. The exome-wide *GT* accuracy will then be estimated from the distance to a good quality reference set that is given by the 2535 individuals which are sequenced by the *1000 genomes project*, as introduced in Section 1.1.4. Finally, the overall accuracy will be estimated by a *standardized dissimilarity score* (*SDS*) that is based on simulated error rates.

Different exomes of varying quality will be analysed, depending on the *GC* content, sequencing *coverage*, sequencing platform and the size of the target region.

4.2 Introduction to Distance Metrics

4.2.1 Distance Metrics and Similarity Measures

Measures of distance or similarity are key features in classification or clustering of data points [91] and the choice of a suitable distance metric for a given data set is crucial.

Distance Metric Conditions

A distance metric can be defined as a (non-negative) real-valued function $d(x_1, x_2)$ on the *cartesian* product $X \times X$ that assigns a distance between any two pairs x_1 and x_2 whereby $x_1, x_2 \in X$. Further, the following conditions must be satisfied for every $(x_1, x_2, x_3) \in X$ [75] [27].:

I. *(Non-Negativity)* $d(x_1, x_2) \geq 0$

II. *(Identity Axiom)* $d(x_1, x_2) = 0$ iff $x_1 \equiv x_2$

III. *(Symmetry Axiom)* $d(x_1, x_2) = d(x_2, x_1)$

IV. *(Triangle Inequality)* $d(x_1, x_2) + d(x_2, x_3) \geq d(x_1, x_3)$

Difference between Distances and Similarities

Although the concept of distance metrics is well defined it seems that no consistent formal definition exists for similarity measures as stated by Chen *et al.* [27]. Generally, a similarity between two objects cannot be transferred into a metric or euclidean space, but can be calculated by using a distance measure. However, a commonly used definition for similarity measure is that it has to satisfy conditions $I - III$, but must not necessarily fulfil the *triangle inequality* which makes a similarity measure less stringent than a distance metric.

By that definition, a similarity measure $s(x_i, x_j)$ is not a metric but coefficients that express similarity in the range $[0, 1]$ can be transformed into a (distance) metric by using for instance the following equation [52]:

$$d(x_i, x_j) = \sqrt{1 - s(x_i, x_j)}$$

But more generally any monotonically decreasing transformation can be applied to transform similarity measures into a distance metric.

Different Distance Functions

Consider each instance x_i as a vector of N measures for attributes with $\vec{x_i} = (x_i^{(1)}, .., x_i^{(N)})$. For non-nominal scales the distance function $d(\vec{x_1}, \vec{x_2})$ can be generalized as

$$d(\vec{x_1}, \vec{x_2}) = \Big(\sum_{n=1}^{N} |x_1^{(n)} - x_2^{(n)}|^{\lambda} \Big)^{\frac{1}{\lambda}}, \tag{4.1}$$

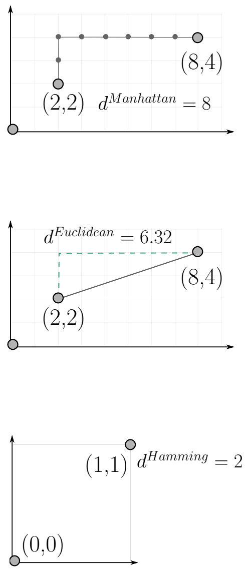
Minkowski Metric

which is also referred to as *Minkowski* Metric for $\lambda \geq 1$. Depending on the value λ, different distance functions can be derived, such as the *Euclidean* distance ($d^{\text{Euclidean}}$, $\lambda = 2$), which uses the *Pythagorean* formula, or the *Manhattan* distance ($d^{\text{Manhattan}}$, $\lambda = 1$) (as illustrated in Box 4.2.1). The *Hamming* distance is another special case of the family of *Minkowski* metrics, that operates over a binary alphabet.

$$d^{Manhattan}(\vec{x_1}, \vec{x_2}) = \sum_{n=1}^{N} \mid x_1^{(n)} - x_2^{(n)} \mid$$

$$d^{Euclidean}(\vec{x_1}, \vec{x_2}) = \sqrt{\sum_{n=1}^{N} \Big(x_1^{(n)} - x_2^{(n)} \Big)^2}$$

$$d^{Hamming}(\vec{x_1}, \vec{x_2}) = \sum_{n=1}^{N} \mid x_1^{(n)} - x_2^{(n)} \mid_{\vec{x_1},\vec{x_2} \in \{0,1\}}$$



Box 4.2.1: Different distance functions for two instances $\vec{x_1}$ and $\vec{x_2}$, derived from Equation 4.1.

Given two vectors $\vec{x_1}$ and $\vec{x_2}$ the *Hamming* distance d^{Hamming} is then the number of entries in which they differ which is equivalent to Equation 4.1 for $\lambda = 1$ and $(\vec{x_1}, \vec{x_2}) \in \{0,1\}^N$.

Distance Matrix D

A summary of the pairwise distances between any object in X can then be collected in a distance matrix D. The properties of such a matrix are directly related to the properties of a (distance) metric:

$$D(X) = \begin{bmatrix} x_{11} = 0 & x_{12} = x_{21} & [...] & x_{1i} = x_{i1} \\ x_{21} = x_{12} & x_{22} = 0 & [...] & x_{2i} = x_{i2} \\ [...] & [...] & [...] & [...] \\ x_{i1} = x_{1i} & x_{i2} = x_{2i} & [...] & x_{II} = 0 \end{bmatrix} \tag{4.2}$$

Usually such high-dimensional data is difficult to interpret and techniques to reduce the complexity of D and visualize the data in a lower dimensional space are applied to provide a manageable representation of the dataset.

4.2.2 Embedding of the Distance Matrix D

Dimension Reduction Techniques

Data reduction techniques create an n-dimensional embedding in $\mathbb{R}^n$, $n < N$ which displays the relative positions of a number of objects $x_i \in X$ given the distances between them in an $I \times I$ matrix D. Different approaches exist, among them *principal component analysis* (*PCA*), *classical* (metric) *multi dimensional scaling* (*MDS*) and *non-metric MDS*. All techniques assume different properties of the input distance matrix D (Equation 4.2).

However, the primary outcome of each technique is a spatial n-dimensional configuration where each object x_i is represented as a point $\varphi(x_i)$ in a new embedded coordinate system, arranged in a way such that distances between each pair $\Big(\varphi(x_i), \varphi(x_j)\Big)_{i \neq j}$ correspond to their similarities.

Principal Component Analysis

The principle of *PCA* can be understood as an orthogonal linear transformation of D. By this means, the coordinates represented in the distance matrix D are transformed to a new coordinate sys-

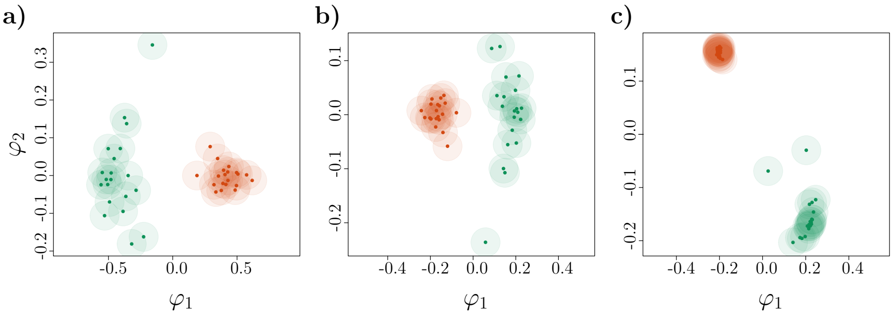


Figure 4.1: *Different visualizations of data reduction techniques.* Simulated (non euclidean) distances were embedded into the 2-dimensional space with *PCA* (**a**)), *metric MDS* (**b**)) and *non-metric MDS* (**c**)). Whereas the results of *PCA* and *non-metric MDS* are similar, *non-metric MDS* leads to a better clustering of the data.

tem, where the n biggest variances of D, also referred to as *principal components*, are represented by the new axes. To do so, the *eigenvectors* ν of the centred *covariance* matrix of D are calculated and ordered by the corresponding *eigenvalues* λ. The new $n \times N$-dimensional map D_n is then the result of the multiplication of D with a matrix containing the first n *eigenvectors* with the biggest *eigenvalues*, $\Gamma = (\vec{\nu_1}, \vec{\nu_2}, ..., \vec{\nu_n})$: $D_n = D\Gamma$

Metric MDS

The approach of *metric MDS* [64] assumes metric properties in the input matrix D and preserves the intervals and ratios between the new embedded coordinates $\varphi(x_i)$ and $\varphi(x_j)$ such that *Euclidean* distances between them approximate the given distances $d(x_i, x_j)$ or short d_{ij}:

$$\parallel \varphi(x_i) - \varphi(x_j) \parallel \approx D_{ij} \tag{4.3}$$

This can be archived by minimizing the goodness of fit, also referred to as *STRESS* function:

$$STRESS_{\text{metric}}(x_1, ...x_n, \varphi) \sim \Big(\sum_{i \neq j}^{n} (d_{ij} - ||\varphi(x_i) - \varphi(x_j)||)^2\Big)^{\frac{1}{2}} \tag{4.4}$$

The *classical MDS* which was first introduced by Torgerson [113] replaces this cost function by a related function, called *STRAIN*:

$$STRAIN(x_1, ...x_n, \varphi) \sim \left(\sum_{i \neq j}^{n} (b_{ij} - < \varphi(x_i), \varphi(x_j) >)^2 \right)^{\frac{1}{2}} \tag{4.5}$$

Here the error of scalar products is minimized and the new coordinate matrix D_n is derived by *eigenvalue* decomposition of $B = D_n D_n^{'}$. B can then be computed from D by using *double centering*. The *classical MDS* yields the same results as *PCA*, if *Euclidean* distances were used to obtain D [35] and thus *PCA* can be seen as a sub-form of the *classical MDS*. In contrast to other ordination methods, *non-metric MDS* makes few assumption about the nature of the data, as for instance *PCA* requires a linear relationship between the objects. The most common *STRESS* function was introduced by Kruskal [64]:

Non-Metric MDS

$$STRESS_{\text{non-metric}}(x_1, ...x_n, \varphi) = \left(\frac{\sum_{i \neq j}^{n} (d_{ij} - ||\varphi(x_i) - \varphi(x_j)||)^2}{\sum_{i \neq j}^{n} ||\varphi(x_i) - \varphi(x_j)||^2} \right)^{\frac{1}{2}} \tag{4.6}$$

The embedded coordinates $\varphi(x_i)$ and $\varphi(x_j)$ are regressed against the original distances d_{ij} and the configuration is improved by moving the positions in ordination space by a small amount in the direction of the steepest descent, where the $STRESS_{\text{non-metric}}$ function changes most rapidly (as illustrated in Figure 4.2).

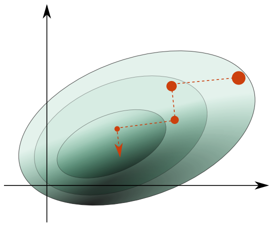


Figure 4.2: *Method of steepest descent.*

Results for different data reduction techniques for simulated (non euclidean) distances are shown in Figure 4.1. Since *non-metric MDS* yields the best clustering result and does not make any assumptions about the distribution of the underlying high-dimensional data, this data reduction technique will be used in this work.

4.3 Generation and Processing of Exome Data

Genomic *DNA* of 241 european individuals was enriched for the *CCDS* target region with *Agilent*s SureSelect Human All Exon Kit [2] and sequenced on an *Illumina* HiSeq 2000 [6]. This *in-house* dataset consists mainly of individuals of European and Arabian origin, but some also have an African or Asian population background. In the following this dataset will be referred to as *BER* (*samples obtained from Berlin*).

All short sequence reads were mapped to the human *reference* genome *GRCh37* using *novoalign* [10] and *BWA* [70] and variants were detected using *SAM*tools [70] and *GATK* [83, 38].

Additionally variant calls of 2535 individuals with different ethnic backgrounds were obtained from the *1000 genomes project* (*1KGP*) [7, 33, 34] and are referred to as *reference* set in the following sections. An overview of the *1KG* dataset was already given in section 1.1.3 and a more detailed description of the 26 background populations is shown in Table 1.3.

All variants were restricted to *bi-allelic* *SNV* positions and to the exome target *consensus* region of size K (~ 29*Mb*) as defined by the *1KGP*. Variant calls that are classified as technical artifacts by the *1KGP* were ignored in the analysis.

Further, to test the predicted accuracies exome variants of the *exome sequencing project* (*ESP*)[4] were used.

4.4 An Error Sensitive Genotype-Weighted Metric

4.4.1 Computation of the Weighted Distance Function

Weighted Distance d_{ij}^{W}

The similarity or distance $d_{ij} \equiv d(x_i, x_j)$ between any two individuals x_i and x_j can be calculated by a weighted *indicator* function, $I_{ij} \cdot W_{ij}^{(k)}$:

$$I_{ij}^{(k)} = I\Big(x_i^{(k)}, x_j^{(k)}\Big) = \begin{cases} 1, & \text{if } x_i^{(k)} \equiv x_j^{(k)} \\ 0, & \text{if } x_i^{(k)} \neq x_j^{(k)} \end{cases}, \tag{4.7}$$

whereby $1 - I_{ij}$ is equivalent to $d^{Hamming}$ as described in Section 4.2.1. That means, if two individuals x_i and x_j differ at a genomic position k, this will contribute to the function with the weight $W_{ij}^{(k)}$:

$$W_{ij}^{(k)} = \frac{2}{f(x_i^{(k)}) + f(x_j^{(k)})}. \tag{4.8}$$

Pre-defined Genotype Frequencies $f(x_i^{(k)})$

The genotype frequencies $f(x_i^{(k)})$ are pre-defined by the *reference* set, taking from the *1KGP*, for all *bi-allelic* positions.

To give an example, consider a genotype for individual x_i at a given position k: $x_i^{(k=CHR\ 6:\ 79595096)} = CC$. If the same genotype CC occurs in the *1KG* dataset 2534 times and one individual has genotype AC, the frequency at this position would be $f(x_i^{(k)}) = 2534 \setminus 2535 = 0.9996055$ (see Table 4.1).

For the sake of completeness, *SNV* positions which are not present in the *reference* set, but in another sample, are initialized with frequency $\dfrac{1}{(N+1)}$, whereby N is the total number of individuals in the *reference* set.

Table 4.1: *Example genotype counts.*

Position	AA	AC	AG	AT	CC	CG	CT	GG	GT	TT	$\sum$
CHR 6: 79595096	0	1	0	0	2534	0	0	0	0	0	2535

Thus, the weighted distance, or dissimilarity between two individuals x_i and x_j is given by d_{ij}^W:

$$d_{ij}^W = 1 - \frac{1}{C_{ij}} \sum_{k=1}^{K} I_{ij}^{(k)} \cdot W_{ij}^{(k)} \tag{4.9}$$

where

$$C_{ij} = \sum_{k=1}^{K} W_{ij}^{(k)}$$

is utilized as a normalizing constant and K is the size of the *consensus* target region.

As each position is weighted with the inverse of it's genotype frequency rare variants that appear less often in a population than common variants contribute more to the overall distance d_{ij}^W. Further, a disagreement at a position with low variability in the *reference* set has a stronger impact than a disagreement at highly variable positions. This also translates to positions with falsely called genotypes, which are induced either during sequencing or during the bioinformatics workflow. A metric that induces a topology that is sensitive to rare variants is in an analogous manner sensitive for genotype errors.

Hamming Distance d_{ij}^H

Additionally, other related metrics have been used to compare different effects on the data. The *Hamming* distance was already described in Section 4.2.1 and is equivalent to an unweighted distance metric d_{ij}^W with $W_{ij}^{(k)} = 1$. In this context this metric will be referred to as d_{ij}^H:

$$d_{ij}^H = 1 - \frac{1}{K} \sum_{k=1}^{K} I_{ij}^{(k)}. \tag{4.10}$$

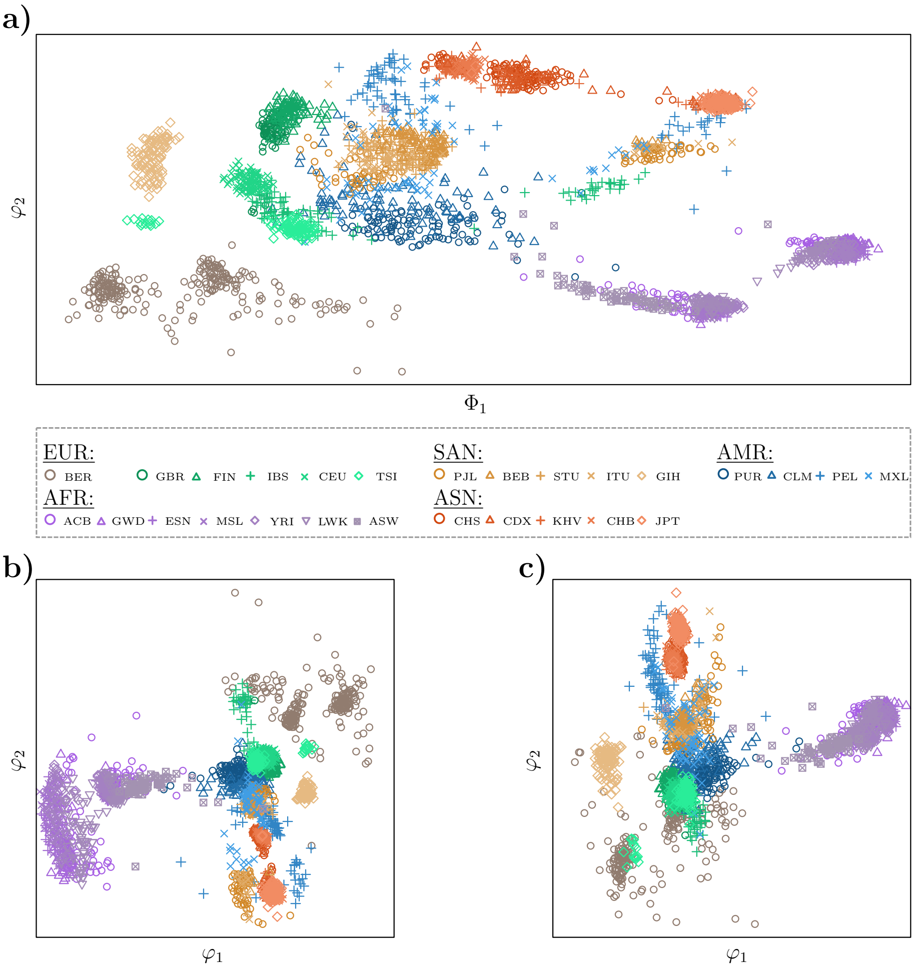


Figure 4.3: *Comparison of different similarity metrics.* Distances between 2535 individuals with 26 different population backgrounds and 241 *in-house* individuals (*BER*) were calculated using different metrics and visualized with *non-metric multi dimensional scaling*. All populations can be clustered by their 5 super-populations: European (*EUR*), South Asian (*SAS*), East Asian (*EAS*), Mixed American (*AMR*) and African (*AFR*) (see also Table 1.3). The weighted metric d_{ij}^W (**a)**) separates more clearly between the sub-populations than the *Hamming* distance d_{ij}^H (**b)**) and the entropy based approach d_{ij}^E (**c)**).

To observe a counterpart to d_{ij}^W a metric that puts much emphasis on common variants has also been studied. In this context *Shannon*'s entropy was used to calculate the similarity d_{ij}^E between two samples x_i and x_j where the entropy of all 10 genotype frequencies f_a given by the *1KG* data (see Table 4.1) was added up for each position k:

Entropy Based Distance Metric d_{ij}^E

$$d_{ij}^E = 1 - \frac{1}{K} \sum_{k=1}^{K} \sum_{a=1}^{10} f_a^{(k)} \cdot log(f_a^{(k)}). \tag{4.11}$$

The influence of the different metrics is displayed in Figure 4.3, where all pairwise distances are visualized with *non-metric MDS* (see Section 4.2.2). All three distance metrics can successfully cluster the different super-populations, but the weighted approach d_{ij}^W (Figure 4.3 **a)**) distinguishes better between sub-populations.

To access the genotype quality of a single individual x_i, that is not included in the *reference* set the distances d_{ij}^W between x_i and all *1KG* individuals x_j of the best suiting background population is calculated. Visualization of the data with *non-metric MDS* gives a first impression of the overall genotype accuracy of x_i assuming perfect data quality in the *1KG* dataset. The distance of samples with comparable quality is closer than the distance between samples with strongly deviating qualities. This can already been observed in Figure 4.3 when comparing individuals from *BER* and other European sub-populations

Different Background Populations

4.4.2 Computation of a Standardized Dissimilarity Score (SDS)

To assign a comparable value to the dissimilarities which are collected in the distance matrix D and visualized via *non-metric MDS* a *SDS* is calculated for every sample that is compared to the *reference* set.

To do so the distances between all *1KG* individuals of each background population are calculated and the average of the mean

$m_{\text{reference}}$ and *inter qartile range* (*IQR*) $IQR_{\text{reference}}$ of all columns of the distance matrix is computed and stored. For each sample that does not belong to the *reference* set the median m_{x_i} of the distances d_{ij}^W to all *1KG* samples with an appropriate background population is computed and normalized by the pre-calculated median and *IQR* of the *reference* set:

$$SDS_{x_i} = \frac{m_{x_i} - m_{\text{reference}}}{IQR_{\text{reference}}}. \tag{4.12}$$

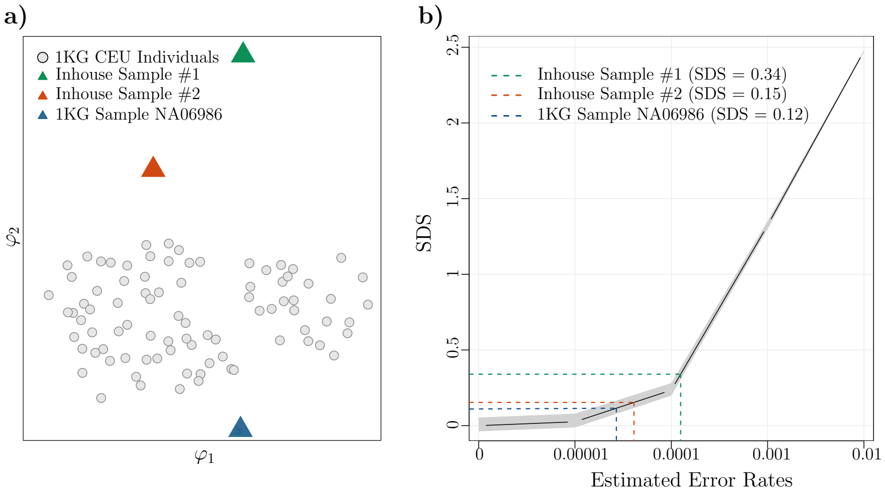


Figure 4.4: *Standardized dissimilarity score.* (Adapted from [115]) **a)** The dissimilarities of three samples of different qualities were embedded with *non-metric MDS* in the two-dimensional plane. The *reference* set, consisting of the best suited background population for all three samples (*European* descent, *CEU*) builds a homogeneous cluster (gray circles). The larger distance of *in-house* sample #1 (green triangle) to the *reference* set indicates a lower genotyping accuracy than *in-house* sample #2 (green triangle) and the *1KG* sample *NA06986* (blue triangle).
b) The *SDS reference* curve (black line) with it's 5% and 95% quantiles (gray) is calculated based on simulated error rates to the *reference* set. The estimated error rate of *in-house* sample #1 (> 0.0001) is considerably higher than the error rates of *in-house* sample #2 and the *1KG* sample *NA06986*.

Simulated Genotyping Accuracies

The *SDS* score was computed as described for different simulated genotyping accuracies. For this a 100% accuracy is assumed in the *reference* set and genotyping errors were introduced at random positions. Most of these positions have a low variability in the *reference* set and the contribution of genotyping errors could be approximated by adding twice a *binomially* distributed random variable $Bin(N,p)$ to the normalizing constant C_{ij} where p is the specific genotyping error ($p \in \{0.00001, 0.0001, 0.001, 0.01\}$) and $N = 2.8 \cdot 10^7$ is the total size of the *consensus* target region.

By calculating the *SDS* score for all simulated error rates an exome sample can be directly linked to its estimated genotyping accuracy. To give an example, Figure 4.4 shows two *in-house* samples of different qualities and one *1KG* sample (*NA06986*) which was processed with the same analysis pipeline.

Exomes with Different Genotyping Accuracies

All three samples were compared to the best suited background population within the *1KG* dataset (*CEU*) and the result is visualized with *non-metric MDS* (Figure 4.4 **a)**). The distance of *in-house* sample #2 and *1KG* sample *NA06986* to the *reference* set is similar while *in-house* sample #1 is clearly separated from the *reference* cluster. This result can also be seen in the *SDS* curve displayed in Figure 4.12 **b)** where the *SDS* of *in-house* sample #2 is corresponding to higher error rates than the other two samples.

4.5 Experimental Validation

4.5.1 Analysis of Exomes With Different Genotyping Accuracies

The distance from an individual to a high quality *reference* set increases when the genotyping quality of the sample decreases. Figure 4.4 **a)** displays the dissimilarities of two *in-house* samples with strongly deviating distances to the *reference* set.

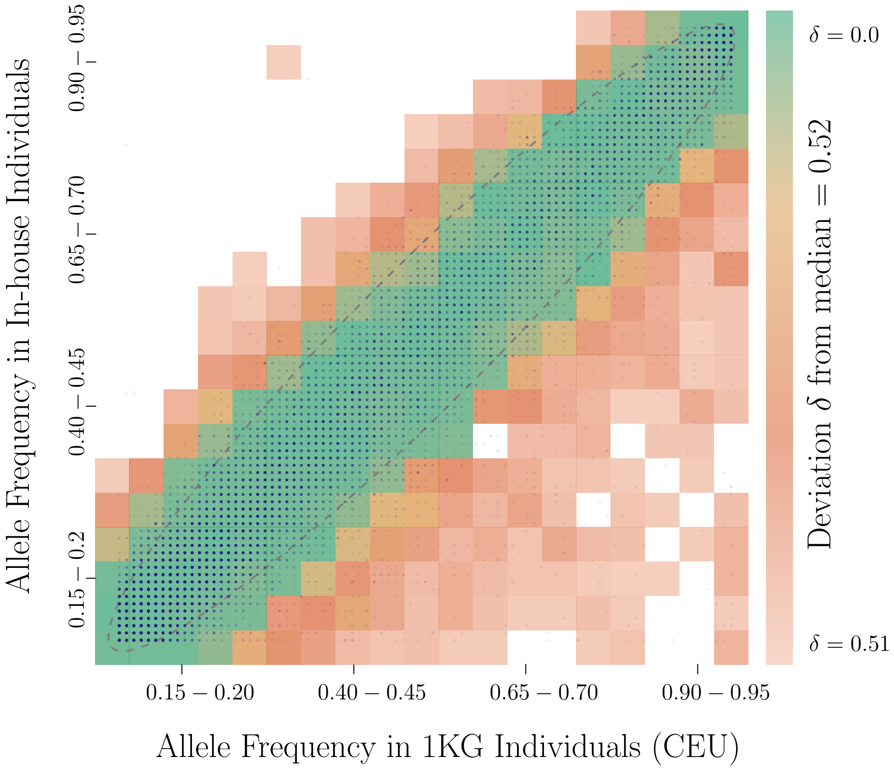


Figure 4.5: *Allele frequencies in 1KG and in-house samples.* (Adapted from [115]) Genotype *allele frequencies* (*AFs*) of 85 samples from the *1KGP* and from 85 *in-house* samples of the same ethnicity (*CEU*) were computed. The gray ellipse displays twice the standard deviation assuming a *binomial* model for the *AF* p. Variants with a strongly deviating *AF* as seen in the right lower quadrant, were called with a lower probability in the *in-house* data and are characterized by a *GC* content that deviates from the expected median of 0.52.

Although the basic quality parameters as the *Ti*-*Tv* ratio (both ~ 3.2), the percentage of *SNP*s as listed in *dbSNP* (both $\sim 97\%$) and the mean variant genotyping quality as listed in the *VCF* files (89.7 in sample #2, 94.8 in sample #1) are comparable these two samples result in different estimated accuracies.

Differences in Quality between In-house and 1KG Datasets

One hypothesis states that variants that are often detected in exomes of the *1KGP* but not in the two *in-house* samples, might point to a subset that requires high data quality to be properly detected. To investigate this *allele frequency* based on 85 *CEU* individuals from the *1KG* set and from 85 *in-house* samples were computed and compared, which is shown in Figure 4.5.

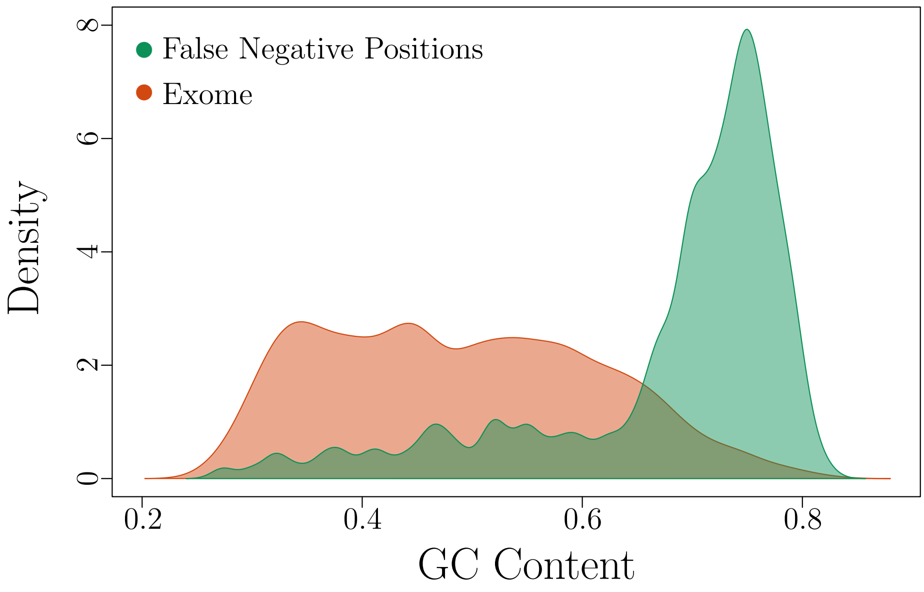


Figure 4.6: *GC content at false negative positions.* (Adapted from [115]) The *GC* content of 100 *bp*s flanking each variant that was present in at least half of the samples from the *1KGP* but in just one *in-house* sample was calculated as classified as *false negative positions* (green distribution). Additionally the *GC* content of randomly drawn *SNV*s was computed (red distribution).
This comparison shows that variants in exomes with a large distance to the high quality *reference* set are overrepresented in exome regions with a high *GC* content.

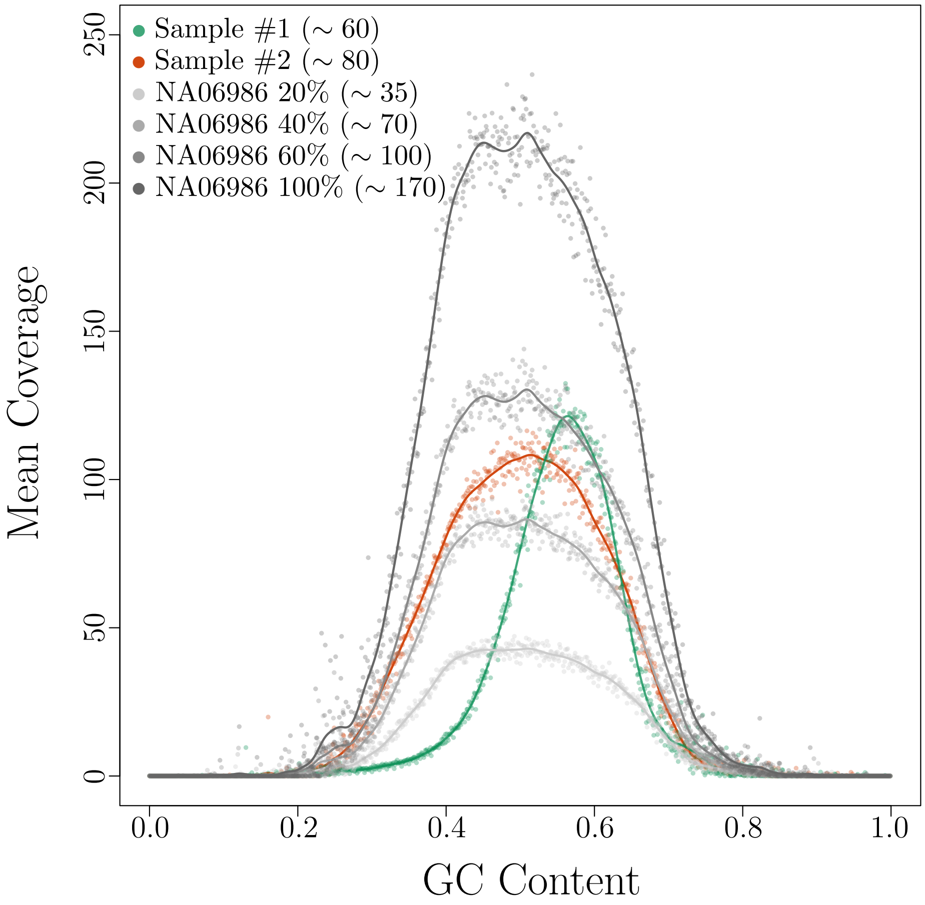


Figure 4.7: *Coverage against GC content.* (Adapted from [115]) The mean *per-base coverage* of an exemplary chosen *1KG* exome with *European* descent (*NA06986*) was downsampled, yielding 20%, 40% and 60% of the original *coverage*, ∼ 170 *reads* per *bp* (gray distributions). Sample #2 has a particularly low *coverage* in *GC*-rich regions (green distribution) in comparison to *in-house* sample #1 (red distribution) and sample *NA06986*.

In case of *technical replicates* one would expect that all *AF* value pairs would lie close to the diagonal. For two sample sets of the same size that are drawn from the same background population a certain variance in the measured *AF* is expected. However, for a finite sample size one would assume that about 95% of the frequency value pairs would fall inside the displayed ellipse that is based on a *bernoulli* distribution. Instead there are considerably more outliers in the right lower corner than expected by chance.

Influence of the GC-Content

To further analyse similarities within the unequal *AF* pairs shown in Figure 4.5 the *GC* content of 100 *bp*s flanking each variant that was present in at least half of the samples from the *1KGP* but in at most one *in-house* sample was computed. The resulting distribution strongly deviates from that of randomly drawn variants as displayed in Figure 4.6.

On this basis it can be concluded that *GC*-rich sequences lead to an increase in *false negative rate*s (*FNR*s). Thus the difference in quality between the two *in-house* samples can partly be explained by a low sequence read *coverage* of regions with an extreme *GC*-content in sample #2 as shown in Figure 4.7.

4.5.2 The Influence of Coverage and Error Rates on d_{ij}^{W}

To investigate the influence of low sequence *coverage* and an increased error rate on the distance metric d_{ij}^{W}, the *per-base coverage* of *1KG* sample *NA06986* was iteratively reduced and in addition detection artefacts were simulated and randomly distributed over the *consensus* target region.
The resulting samples were compared to the *1KG* *reference* set (*CEU*) and visualized with *non-metric* *MDS* (Figure 4.8). The embedded points follow a trajectory that departs from the reference cluster with growing error rate. A similar behaviour can be observed for the samples with a decreasing *coverage*.

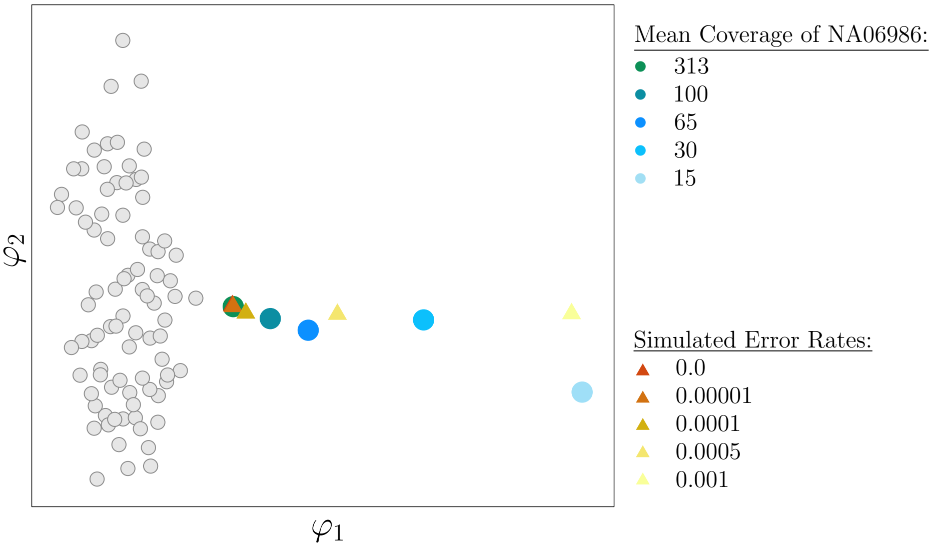


Figure 4.8: *Influence of decreasing coverage and increasing error rate.* (Adapted from [115]) A reduction of the *per-base coverage* for the *1KG* sample *NA06986* (blue circles) and simulated genotyping error rates (red triangles) reduces the similarity to the *reference set* (gray circles, *CEU*).

4.5.3 Comparison of WES Data from Different NGS Studies and Target Sizes

Varying SDS Scores in Different NGS Studies

In addition to the *in-house* samples and the exomes from the *1KGP* also simulated individuals based on genotype frequencies from the *exome sequencing project* (*ESP*) [4] and 100 exomes that were already studied by de Ligt *et al.* [76] were analysed. In comparison the mean *SDS* for the the simulated *ESP* exomes is comparable to the *in-house* data and lower than the *SDS* of the *de Ligt* exomes (4.9 **a)**) which could partly be explained by out-dated error-prone genotyping algorithms.

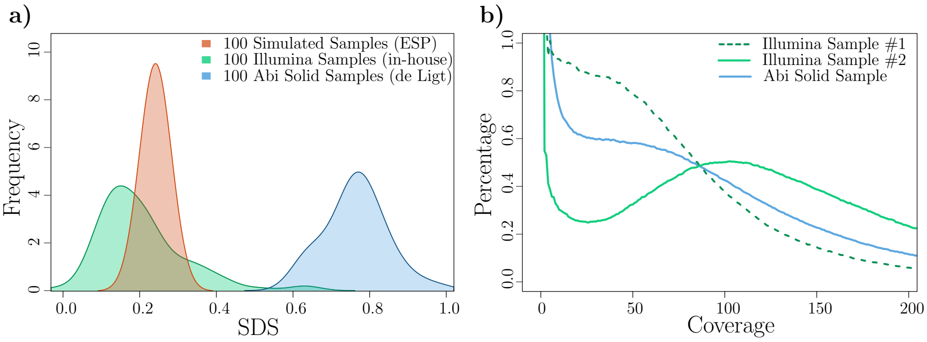


Figure 4.9: *Comparison of different platforms.* (Adapted from [115]) Altogther 100 *in-house* samples were sequenced on the *Illumina* platform [6] with a mean *coverage* above 60×.
a) The obtained mean *SDS* for the *Illumina* samples (green distribution) is comparable to value of 100 exomes that were simulated based on genotype frequencies from the *exome sequencing project* (red distribution) [4], although the variance is much smaller due to missing *haplotype* information in the *ESP* data that leads to a higher self-similarity in the simulated data. In comparison the mean *SDS* of 100 exomes that were sequenced on an *Abi Solid* [14] platform is considerably higher. This can be explained by a lower and less uniform sequence distribution over the target region (see **b)**) and due to less accurate variant calls. **b)** The *coverage* distribution of two *Illumina* samples (green) and one *Abi Solid* sample (blue).

Low Sequencing Coverage

An additional reason might be the low mean sequencing *coverage* which is depicted in Figure 4.9 **b)** for an exemplarly chosen exome, which was sequenced on an *Abi Solid* platform [14].

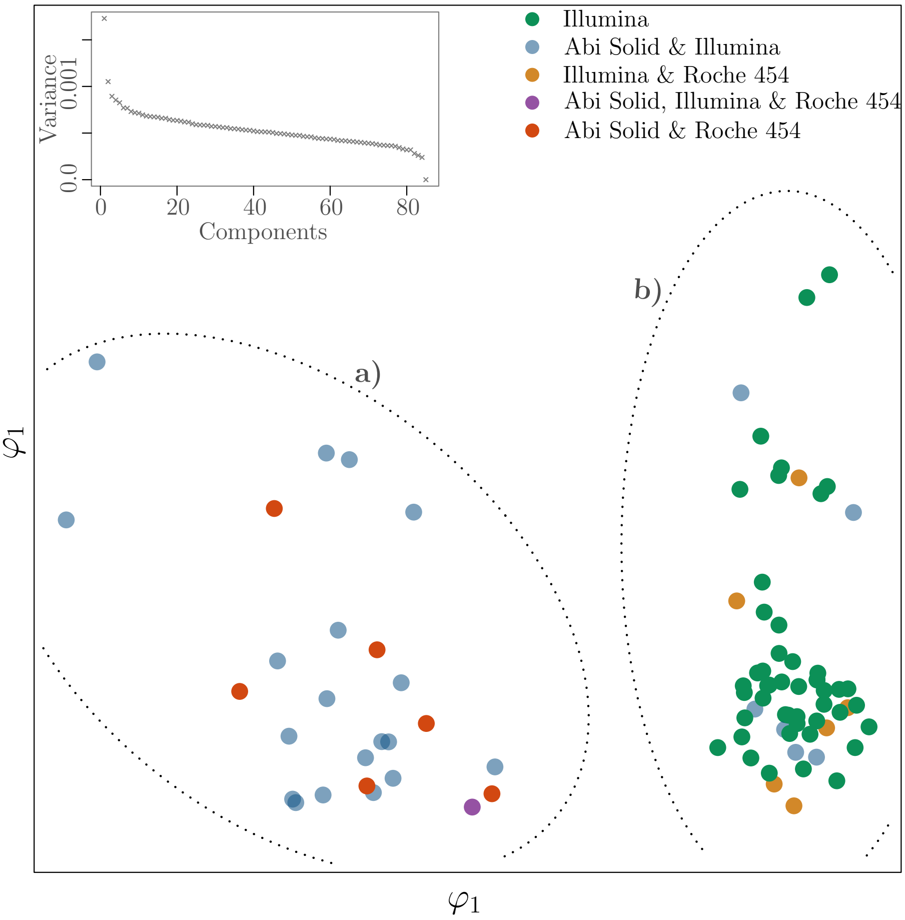


Figure 4.10: *PCA reveals platform specificity.* (Adapted from [115]) In both *PCA* (shown) and *metric MDS* (not shown) a clear substructure is visible in the *reference set* (*CEU*) that is specific to the sequencing platform and also explained by the two biggest principal components, which show the biggest variation within the data (upper left cluster). However, the effect of the sequencing platform for predicting the genotyping accuracy is small and the estimated error rates are comparable even if the *reference* set is restricted to a specific sequencing platform (not shown).

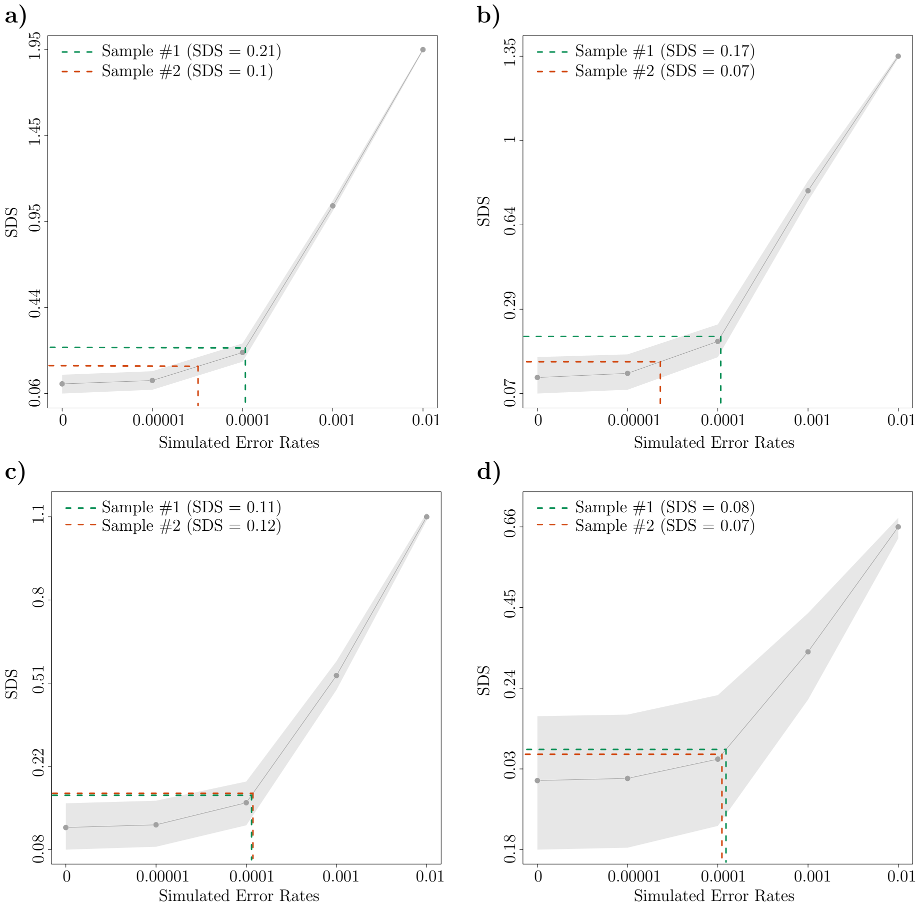


Figure 4.11: *Different target sizes.* (Adapted from [115]) The distances of two *in-house* samples and their corresponding *SDS* were computed for different target regions that differ in size. Originally, this calculation was done for the *consensus* target region (∼ 29 *Mb*) which is displayed in Figure 4.4.
a) The *HPO* panel [100] comprises all exons of genes that are associated with phenotypic features (∼ 5.8*Mb*). **b)** The *Kingsmore* panel [21] includes 548 genes of known inherited diseases (∼ 1.2*Mb*). **c)** Chromosome 22 of the *CCDS* target region [95] comprises a size of about 600*kb*. **d)** The *GPI* panel contains all genes involved in the *GPI*-anchor synthesis (∼ 45*kb*).
The larger the target region the larger is the estimated precision of the estimated error rates. On the other hand, with a shrinking size of the target region the confidence intervals (5% and 95% quantiles, gray area) of the reference curve for the *standardized dissimilarity score* expand. Further, gene panels below 1 *Mb* are too small to clearly distinguish the different genotyping qualities between both samples.

PCA Reveals Platform-Dependent Clustering

To further investigate the influence of the sequencing platform on the results, the *reference* set (*CEU*) was restricted to samples that were sequenced with the same technology. Interestingly visualization of the weighted distances d_{ij}^{W} with *PCA* as well as with *metric MDS* showed a clustering of the different platforms (see Figure 4.10) and the *in-house* samples were closest to the *Illumina* cluster (not shown). However, the effect of the sequencing platform on the accuracy prediction (*SDS*) is just marginal (not shown).

Although the *standardized dissimilarity score* seems to be robust and independent of the underlying sequencing platform, the approach requires a minimum amount of positions to achieve high precision as shown in Figure 4.11.

4.6 Similarity Metrics in Rare Variant Association Studies

Recently another application for a similarity-based approach was shown by Zhu et al. [127] which emphasizes the importance of the right choice of a suitable metric.

In the analysis of cohorts with extremely rare disorders the sample size is usually small which is simultaneously associated with varying ethnicities and heterogeneous data qualities. The usual approach in *common variant association studies* (*CVAS*), such as *GWAS*, is to correct for population stratification in the dataset. However, Mathieson and McVean [82] have already demonstrated that this approach is not suitable for *RVAS* as they cannot account for the sub-structure that is introduced by rare variants. As existing techniques fail to correct for confounding effects, Zhu et al. analysed the effects of different similarity-based matching strategies for case and control group setups.

All three metrics introduced in Section 4.4.1 were applied to different cohorts and found that the best performance is achieved when applying a metric that puts a stronger weight on sharing rare *alleles* (d_{ij}^{W}), especially if the data quality is heterogeneous.

4.7 Summary of this Chapter

This chapter described an approach to estimate the overall exome genotyping quality and different datasets were applied to evaluate the results.

Although various quality control measurements as introduced in Section 1.3.1 exist they do not directly indicate the accuracy of an entire exome. It was shown that the genotyping quality of an exome can be estimated by calculating the distance to a high quality *reference set* with a suitable background population, as the individuals sequenced by the *1000 genomes project* using a metric that is sensitive for rare variants and genotyping errors, d_{ij}^{W}. The computed *standardized dissimilarity score* is platform independent and suitable for a comprehensive quality control in exome samples.

Further, *non-metric multi dimensional scaling* was the best technique to reduce the high dimensionality of the distance matrix D whereas *principal component analysis* and *non-metric MDS* revealed a sub-clustering of the *reference* set based on different sequencing techniques.

Exomes with poor data quality based on low *coverage* in GC-rich regions could be identified although common quality control measurements such as the *Ti*-*Tv* ratio, the percentage of *SNP*s in databases such as *dbSNP* and the mean *phred*-like genotype quality score were comparable to other exome samples of good quality.

Comparisons of target regions of different sizes showed that a minimum size of $\sim$ 10 *Mb* that corresponds to about 10000 variant positions is necessary to yield reliable results.

Even in high quality data sets, such as the *1000 genomes project*, up to 2% of the variants cannot be validated, not even if the exome is re-sequenced by an alternative sequencing platform. In a target region that comprises $\sim$ 30*Mb*, as in the *consensus* target region, one would detect an average of 15000 variants depending on the background population. From this, approximately 300 variants are falsely detected which corresponds to an error rate of about 10^{-5}. Thus every estimated *SDS* that corresponds to a similar error rate would indicate a sample of high genotyping quality.

Finally another parameter for the quality assessment of exomes can be applied, namely the variance of *allele frequencies* (*AFs*) at *heterozygous* variant calls which was introduced in Chapter 3. The lower the variance the lower is the expected error rate from amplification artefacts. This effect could also be observed in the discussed two *in-house* samples of deviating qualities. The poor-quality sample with a high estimated error rate yields a variance of 0.009 which is far below the average value for the tested *in-house* samples (~ 0.017).

5 | Prediction of Family Structures in Next Generation Sequencing Data

5.1 Overview of this Chapter

The identification of disease causing mutations is often based on the analysis of several family members, especially if segregation or linkage analysis is used to prioritize suitable candidate variants which are often *de novo* mutations associated with highly heterogeneous disorders [119]. However, these approaches rely on correct pedigree information and a method to test for underlying relationships between individuals is crucial. In particular sample *mix-ups* can lead to erroneous conclusions when filtering for potentially pathogenic variants.

The following chapter focuses on different methods to infer relationship classes for related *whole exome sequencing* (*WES*) individuals which are sequenced by the *1000 genomes project* (*1KGP*) as well as *in-house*, including several degrees of relationship. Therefore genetic identity coefficients such as *identical-by-state* (*IBS*) and *identical-by-descent* (*IBD*) are introduced and a likelihood based approach to reconstruct entire pedigrees is systematically analysed. Pairwise *logarithm of the odds* (*LOD*) scores for predefined hypotheses for different classes of relatedness are calculated and evaluated depending on the number of positions and the degree of consanguinity.

5.2 Introduction to Approaches to Analyse Family Structures

5.2.1 Genetic Identity Coefficients

Identical-By-Descent & Identical-By-State

During *meiosis*, crossover recombination occur between homologous chromosomes. Thereby two individuals may receive identical *allele* segments which are inherited from a common ancestor and defined as *identical-by-descent* (*IBD*) as illustrated in Figure 5.1 **a)**. If the two *alleles* which are observed in different individuals are identical but may be derived from different ancestors then they are called *identical-by-state* (*IBS*) [77]. For both definitions a distinction is made weather none, one or two *alleles* are shared by two individuals (also referred to as *dyad*) as displayed in Figure 5.1 **b)**. The expected proportion of the genome that shares common *alleles* between a *dyad* can be formulated as a function of their genetic relatedness and several methods that are based on this approach already exist [55, 111].

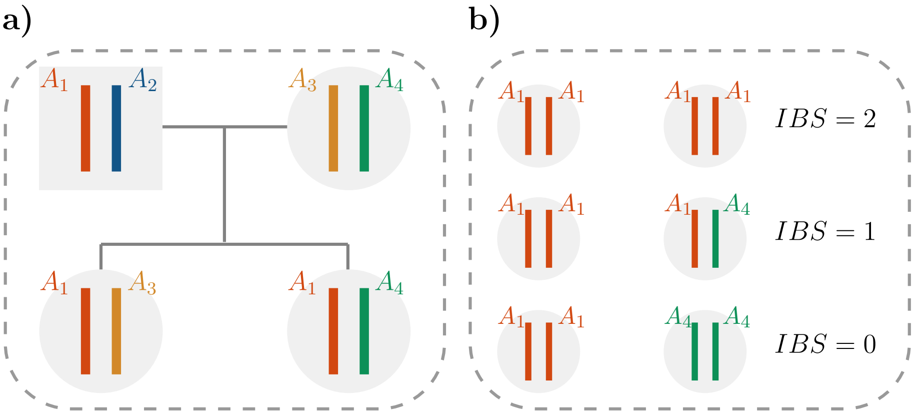


Figure 5.1: *Identical-By-Descent and Identical-By-State.* **a)** Two *alleles* are *identical-by-descent* (*IBD*) if they are copies of the same ancestral *allele* as exemplarily shown for *allele* A_1. **b)** A pair of diploid individuals can be *identical-by-state* (*IBS*) (or *IBD*) for none, one or two *alleles* at a certain locus.

Kinship Coefficient & Coeffient of Relatedness

The probability that two *alleles* are *IBD* is given by the *kinship coefficient* F which is a measure of the likelihood that two randomly drawn *alleles* at the same *autosomal* locus from different individuals are identical. To give an example, if a randomly sampled *allele* from a child is given at a certain locus then there is a probability of 50% that this *allele* originates from the mother or from the father respectively. Due to the independence of these two probabilities the *kinship coefficients* is then given by $F = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. Table 5.1 lists several common relatedness cases and their *kinship coefficients* F as well as the associated *coefficients of relatedness* $R = 2F$ [125].

Table 5.1: *Kinship coefficients and coefficients of relatedness.* Common cases of pair-wise relationships and their associated *kinship coefficients* F and *coefficients of relatedness* R are listed. The categories *grand-parent - grand-offspring*, *half siblings* and *aunt/uncle - niece/nephew* are summarized as 2^{nd} Order relationships whereas *first cousins* and *great grand-parent - great grand-offspring* are combined as 3^{rd} Order relationships.

Relationship	Kinship Coefficient	Coefficient of Relatedness
Identic /Monozygotic Twins	0.5	1
Parent-Child	0.25	0.5
Full Siblings	0.25	0.5
2^{nd} Order	0.125	0.25
3^{rd} Order	0.0625	0.125

Cryptic Relationships

The *kinship coefficient* is also often used to identify *cryptic* relationships or population sub-structures within large population-based studies [111] which may avoid confounders in *case-control* association studies [121, 127] as selection will increase the number of *alleles* that are *IBD* among individuals within a population.

5.2.2 Likelihood Ratios

Likelihood Ratios

In general a *likelihood ratio* (*LR*) is a statistical procedure to compare the *goodness-of-fit* of two models. Mathematically the ratio between the likelihood of Θ taking a value, for example Θ_A, and the likelihood L of Θ under the *null*-hypothesis, for example Θ_0, given the same underlying data D is denoted as given by Marshall [81]:

$$LR(\Theta_0, \Theta_A|D) = \frac{L(D|\Theta_A)}{L(D|\Theta_0)}. \tag{5.1}$$

LOD Score

For practical reasons human geneticists have used the related *logarithm of the odds* (*LOD*) score for hypothesis testing of genetic linkage which is defined as the base 10 logarithm of the *LR* score:

$$LOD = log_{10}\frac{L(D|\Theta_A)}{L(D|\Theta_0)}. \tag{5.2}$$

Broadly speaking the interpretation of a *LOD* score is that the alternative hypothesis Θ_A is 10^{LOD} times more likely than the *null*-hypothesis Θ_0.

Linkage Analysis

This method is often applied in *linkage* analysis which is a powerful approach to detect the chromosomal location of disease genes based on the knowledge that genes residing physically close on a chromosome remain linked during *meiosis* [96].

LOD Score to Infer Relationships

Finally the *LOD* score can also be used to discriminate between different classes of relationship between two individuals x_1 and x_2 (*dyad*) which are formulated as hypotheses Θ_i (e.g. x_1 and x_2 are related) and Θ_j (e.g. x_1 and x_2 are *unrelated*) given the underlying *SNP genotype* (*GT*) data $D = (gt_{x_1}, gt_{x_2})$ for both samples:

$$LOD\big(\Theta_i, \Theta_j \mid D\big) = log_{10}\frac{\mathbb{P}\big(D = (gt_{x_1}, gt_{x_2}) \mid \Theta_i\big)}{\mathbb{P}\big(D = (gt_{x_1}, gt_{x_2}) \mid \Theta_j\big)}. \tag{5.3}$$

Information Content

The majority of approaches to infer relatedness to a pair of individuals which are based on either *IBD* or *LR* make use of a pre-defined marker set, mostly including *SNP*s or *simple sequence repeat*s (*SSR*s), which are characterized by a high *information content* (*IC*) [24, 93]. *SSR*s or *microsatellites* are variable in length, highly informative and in general accurate predictions about the relatedness between two samples can be achieved by analysing just $20 - 30$ *SSR*s. On the contrary, *SNP*s that are in general easier to genotype than *microsatellites* [94] are mostly *bi-allelic* and not as heterogeneous as *SSR*s. Additionally there is a chance that *bi-allelic* markers with a high population frequency are also *IBS* by pure chance.

In the following sections the effectiveness of *LOD* scores to assign relationships to *dyads* by using all *GT*s that were obtained by *WES*, including rare variants, will be studied.

5.3 Definition of Hypotheses and LOD Scores

As introduced in Section 5.2.2 *LOD* scores can be applied to evaluate the *goodness-of-fit* between two pre-defined hypotheses for relatedness given the genotype *SNP* data of two individuals x_1 and x_2. In this work five different hypotheses $\Theta_i, i \in I = \{0, 1, 2, 3, 4\}$, are defined:

Θ_0: x_1 and x_2 are *unrelated*

Θ_1: x_1 and x_2 are technical/biological *replicates* or monozygotic twins

Θ_2: x_1 and x_2 are *full siblings*

Θ_3: x_1 is a parent of sample x_2 (*parent - child* relationship)

Θ_4: x_1 and x_2 have a 2^{nd} *order* relationship
(including *grandparent - grandoffspring*, *half siblings* and *aunt/uncle - niece/nephew*)

Box 5.3.1: *Different hypotheses for relatedness between two samples x_1 and x_2.*

Probability of Genotype Combinations

The probability of the combination of two *genotypes* (*GTs*) at a given locus of a *dyad* can be estimated with the use of population data as already introduced in previous works [81, 16, 112]. For this purpose *allele frequencies* (*AFs*) of 2535 *unrelated* individuals, sequenced by the *1000 genomes project* (*1KGP*) [31, 32, 33, 34] which was already introduced in Section 1.1.4 and summarized in Table 1.3, were extracted and added to the calculations. Depending on the tested hypothesis different numbers of unknown relatives must be taken into account. Examples for two hypotheses, Θ_3 and Θ_4, for a given *dyad* (x_1, x_2) which are both *homozygous* for an

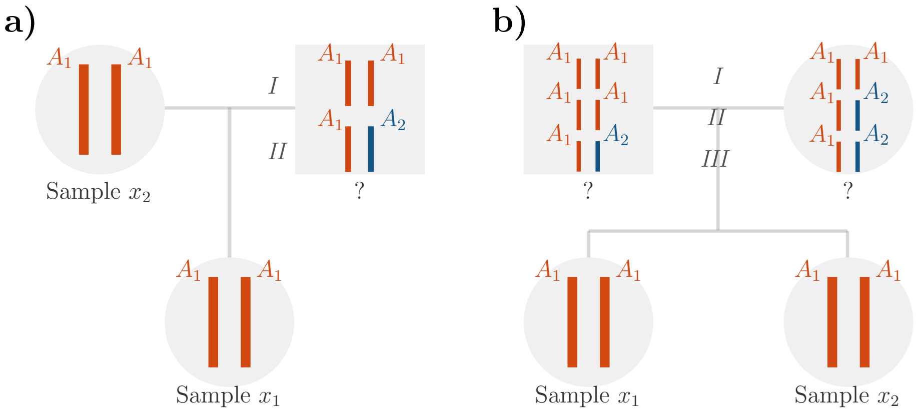


Figure 5.2: *Scheme for genotype combinations.* Displayed are two pedigrees for two possible underlying hypotheses Θ_3 and Θ_2 for samples x_1 and x_2. Imagine that both are *homozygous* for an *allele* A_1.
a) If hypothesis Θ_3 is true (x_1 and x_2 have a *parent - child* relationship) there are two possible GTs for the unknown second *parent*: A_1A_1 (I) or A_1A_2 (II). **b)** If hypothesis Θ_2 is true (x_1 and x_2 are *full siblings*) three GT combinations are possible for the two unknown *parents*: either both are *homozygous* for *allele* A_1 (I), only the second *parent* is *heterozygous* (II) or both unknown *parents* have a *heterozygous* GT A_1A_2 (III).

allele A_1 at any position k is given in Figure 5.2.
A more detailed example of how to calculate the probability for a hypothesis, e.g. Θ_3, given a pair of *genotype*s (GTs), e.g. A_1A_1 and A_1A_1, is described in the following.

Example Calculation for $LR(\Theta_3, \Theta_0)$

If hypothesis Θ_3 is true (samples x_1 and x_2 have a *parent - child* relationship) two GTs are possible for the alleged second *parent* as also depicted in Figure 5.3 **a)**:

- *Case I)* A_1A_1 and
- *Case II)* A_1A_2.

In conformity with the *Hardy-Weinbergs* equilibrium (see Section 1.1.1) the probability that each parent transmits an allele to their offspring is equal to the AF in the population, f_1 for *allele* A_1 and f_2 for *allele* A_2 respectively. Hence, the frequency of a GT A_1A_1 is then defined by $f_1 \cdot f_1 = f_1^2$ and the frequency of A_1A_2 is $2f_1f_2$.

The probability of the GT of the offspring gt_{x_1} (A_1A_1), given the assumed GTs for the two alleged parents can then be calculated by multiplying their frequencies. For $D = (gt_{x_1} = A_1A_1, gt_{x_2} = A_1A_1)$ the probability $\mathbb{P}(D \mid \Theta_3)$ is given by:

$$\begin{aligned}\mathbb{P}(D \mid \Theta_3) &= \overbrace{f_1^2 f_1^2}^{\textit{Case I}} + \overbrace{f_1^2 \frac{1}{2} 2 f_1 f_2}^{\textit{Case II}} \\ &= f_1^4 + f_1^3 f_2.\end{aligned}$$

With the probability that both samples x_1 and x_2 are unrelated formulated by the *null*-hypothesis Θ_0:

$$\mathbb{P}(D \mid \Theta_0) = f_1^2 f_1^2 = f_1^4$$

the *likelihood ratio* $LR(\Theta_3, \Theta_0 \mid D)$ can be calculated as follows:

$$\begin{aligned}LR(\Theta_3, \Theta_0 \mid D) = \frac{\mathbb{P}(D \mid \Theta_3)}{\mathbb{P}(D \mid \Theta_0)} &= \frac{f_1^4 + f_1^3 f_2}{f_1^4} \\ &= \frac{f_1 + f_2}{f_1} \\ &= \frac{1}{f_1},\end{aligned}$$

assuming that $f_1 + f_2 = 1$.
A comprehensive summary of all LRs $LR(\Theta_i, \Theta_0 \mid D = (x_1, x_2))$, $i \in \{1, 2, 3, 4\}$, for all GT combinations of the *dyad* (x_1, x_2) is given in Table 5.2.

Error Constant e

For perfectly genotyped datasets without the occurrence of *De-Novo* mutations some combinations of GTs do not exist and therefore the corresponding probability would simply be zero. To give an example a *parent* cannot be *homozygous* for *allele* A_1 while the child is *homozygous* for another *allele* A_2 at the same position. The incorporation of an additional error constant $e = 0.001$ allows to include these positions into the analysis without upstream filtering accounting for sequencing errors and *de novo* mutations. It should be noted that the choice of e is not trivial as the result can be influenced by deviating values. In this work, e corresponds to the mean genotyping error rate which is observed as upper bound for low quality *WES* samples [115] which is also sufficiently discussed in Chapter 4.

Table 5.2: *Likelihood ratios for genotype combinations.* The variable f_n refers to the *allele* frequency of *allele* A_n and is pre-calculated using *GT* data from the *1KGP* assuming $\sum_n f_n = 1$. Combinations of *GT*s that do not occur for certain relationships (*Mendelian* error) could still be observed due to e.g. erroneous genotyping ($e = 0.001$).

gt_{x_1}	gt_{x_2}	$LR(\Theta_1, \Theta_0)$	$LR(\Theta_2, \Theta_0)$	$LR(\Theta_3, \Theta_0)$	$LR(\Theta_4, \Theta_0)$
A_1A_1	A_1A_1	$\frac{1}{f_1^2}$	$\frac{(f_1+1)^2}{4f_1^2}$	$\frac{1}{f_1}$	$\frac{f_1+1}{2f_1}$
A_1A_1	A_1A_2	$\frac{e}{2f_1^3 f_2}$	$\frac{f_1+1}{4f_1}$	$\frac{1}{2f_1}$	$\frac{2f_1+1}{4f_1}$
A_1A_1	A_2A_2	$f_1^2 f_2^2$	$\frac{1}{4}$	$\frac{e}{f_1^2 f_2^2}$	$\frac{1}{2}$
A_1A_2	A_1A_2	$\frac{1}{2f_1 f_2}$	$\frac{1+f_1 f_2}{4f_1 f_2}$	$\frac{1}{4f_1 f_2}$	$\frac{4f_1 f_2 + f_1 + f_2}{8f_1 f_2}$
A_1A_1	A_2A_3	$\frac{e}{4f_1^2 f_2 f_3}$	$\frac{1}{4}$	$\frac{e}{4f_1^2 f_2 f_3}$	$\frac{1}{2}$
A_1A_2	A_1A_3	$\frac{e}{8f_1^2 f_2 f_3}$	$\frac{2f_1+1}{8f_1}$	$\frac{1}{8f_1}$	$\frac{4f_1+1}{8f_1}$
A_1A_2	A_3A_4	$\frac{e}{8f_1 f_2 f_3 f_4}$	$\frac{1}{4}$	$\frac{e}{8f_1 f_2 f_3 f_4}$	$\frac{1}{2}$

LOD Scores

Finally with the pre-calculated probabilities given in Table 5.2 and Equation 5.3 the *LOD* score for all combinations of *GT*s $gt = (A_nA_n)$, $A_n \in \{A, C, G, T\}$, for each *dyad* (x_1, x_2) can be calculated for all loci $k \in K$ as follows, assuming linkage equilibrium between positions:

$$\begin{aligned} LOD(\Theta_i, \Theta_0 \mid D) &= log_{10} \prod_{k \in K} \frac{\mathbb{P}\Big(D = \big(gt_{x_1}(k), gt_{x_2}(k)\big) \mid \Theta_i\Big)}{\mathbb{P}(D = \big(gt_{x_1}(k), gt_{x_2}(k)\big) \mid \Theta_0)} \\ &= \sum_{k \in K} log_{10} \frac{\mathbb{P}\Big(D = \big(gt_{x_1}(k), gt_{x_2}(k)\big) \mid \Theta_i\Big)}{\mathbb{P}\Big(D = \big(gt_{x_1}(k), gt_{x_2}(k)\big) \mid \Theta_0\Big)} \end{aligned} \tag{5.4}$$

whereby $gt_{x_1}(k)$ refers to the GT of sample x_1 at the kth position. For each *dyad* (x_1, x_2) LOD scores for all hypotheses Θ_i, $i \in \{1, 2, 3, 4\}$, versus the *null*-hypothesis Θ_0 (*unrelated*) are computed and divided by the number of comparisons K.

In the following the notation $LOD(\Theta_i, \Theta_0 \mid D)$ is simplified as $LOD(\Theta_i, \Theta_0)$

5.4 Experimental Validation

5.4.1 Separation Efficiency of LOD Scores

WES Samples

LOD scores were calculated for all *dyads* from 39 *WES* families sequenced by the *1KGP* as well as from 9 *in-house* families including three families with a self-declared consanguineous degree of relationship which are only used in Section 5.4.4. All *in-house* samples were sequenced on an *Illumina HiSeq 2000* [6] and the resulting reads as well as the raw reads from the *1KGP* were mapped to the *GRCh38* reference genome using *BWA-MEM* [70].

Simulation of Technical Replicates

Additionally six *technical replicates* were simulated based on one *1KG* individual (*NA06986*) by randomly reducing the coverage of the original alignment yielding a mean *per-base coverage* of 313, 140, 120, 80, 50 and 30 *reads*.

Multi-sample variant calling was performed using *GATK* [83, 38, 117] and the variant list was restricted to the *consensus* exonic tar-

Table 5.3: *Summary of related individuals.* All related *WES* samples from 39 families sequenced by the *1KGP* as well as from six *in-house* families. Additionally six *technical replicates* of one individual from the *1KGP* (*NA06986*) were simulated, yielding decreasing *coverages* of 313, 140, 120, 80, 50 and 30 *reads* per positions.

Dateset	Relationship Type	Number of *dyads*
1KG	replicates	15
1KG	full siblings	6
1KG	parent - child	16
1KG	2^{nd} order	7
in-house	full siblings	7
in-house	parent - child	24
in-house	2^{nd} order	7
in-house	3^{rd} order	5

get region defined by the *1KGP* [31, 32, 33, 34].
A summary of all families, with the exception of the three consanguineous families, and their underlying relationship types is shown in Table 5.3.

Visualization of LOD Scores

Figure 5.3 displays all *LOD* scores $LOD(\Theta_i, \Theta_0)$, $i \in \{1, 2, 3, 4\}$, for each *dyad* sequenced by the *1KGP* as well as *in-house*. Violin plots visualize the distributions of all unrelated *dyads* for each hypothesis, whereas known relationship types are highlighted. Most *LOD* scores show a segmentation of the *dyads* which corresponds to their *kinship coefficient* (see Table 5.1). To give an example the *LOD* scores $LOD(\Theta_3, \Theta_0)$ cluster the *technical replicates* (red circles) at the right end, *dyads* with a *full sibling* or *parent - child* relationship in the middle area and 2^{nd} *order* related individuals at the left side. The most likely hypothesis maximizes the *LOD* score, $max\{LOD(\Theta_i, \Theta_0)\}_{\geq 0}$ which is emphasized in Figure 5.3 with additional dashed lines.

Higher order Relationship Types

As shown in Table 5.3 five 3^{rd} order relationships are included in the *in-house* dataset which are not explicitly tested with an additional hypothesis as they occur in just one big *in-house* family. In absence of a suited pre-defined hypothesis these *dyads* are marked as *unrelated* but assigned to a 2^{nd} *order* relationship which can be observed in the upper panel of Figure 5.3 ($LOD(\Theta_4, \Theta_0)$) as the violin distribution includes values ≥ 0.

Misclassification due to Low Quality

Misclassification can also occur due to low quality within samples with a high genotyping error rate as shown for three *replicate dyads* with strongly deviating *per-base coverages* in Figure 5.3 (yellow asterix). These *dyads* are rather assigned to a *full sibling* relationship when resolving $max\{LOD(\Theta_i, \Theta_0)\}_{\geq 0}$ as with increasing error rates it becomes more difficult to identify *technical replicates* or *monozygotic twins*.

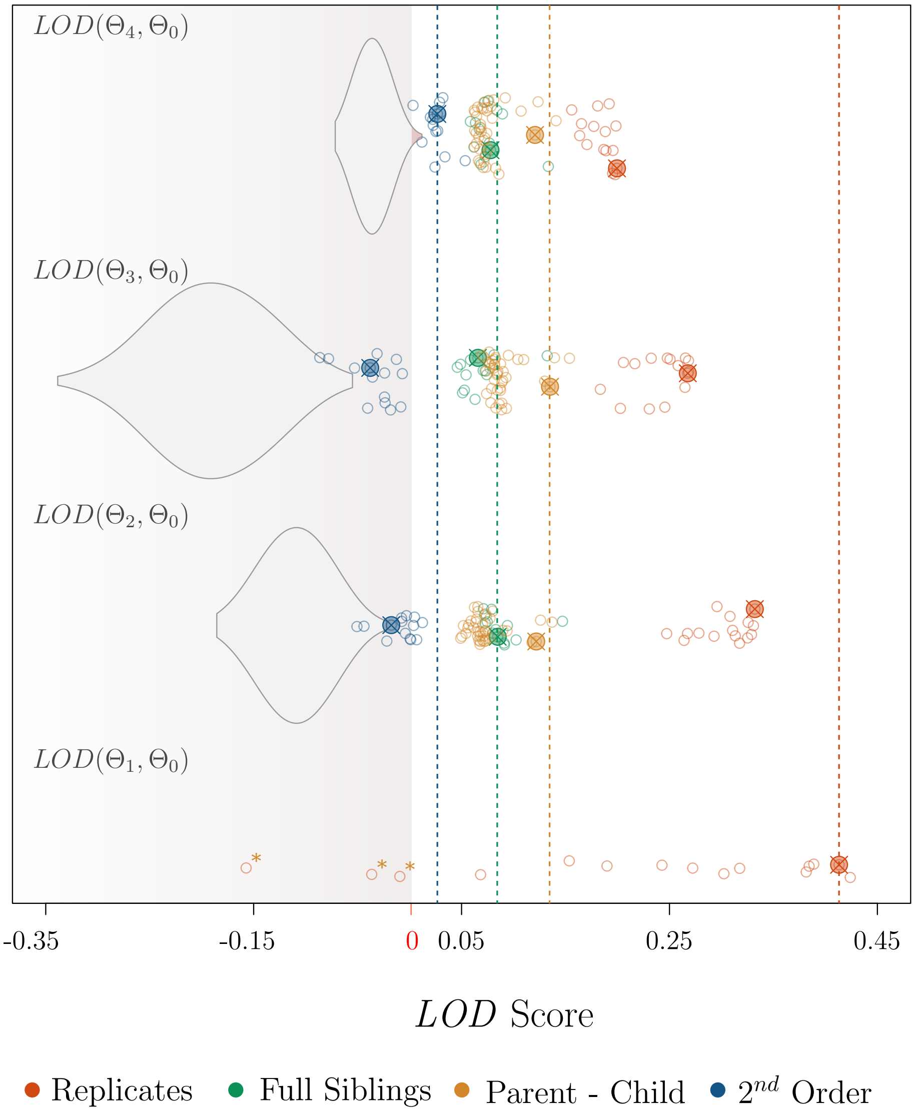


Figure 5.3: *LOD scores based on different hypotheses.* All hypotheses $\Theta_i, i \in \{1, 2, 3, 4\}$ are compared to the *null*-hypothesis Θ_0 (*unrelated*), for each *dyad* included in the *1KG* family dataset as well as in 6 *in-house* families. Individual pairs, that are not related are depicted as gray background distribution whereas true underlying relationship classes are highlighted. Note that most of the *dyads* have small values for $LOD(\Theta_1, \Theta_0)$ and are not displayed in the figure. The variance of the *LOD* scores comparing Θ_1 and Θ_0 is relatively big compared to other hypotheses and therefore most *dyads* have very small *LOD* scores and could not be displayed in the same figure. When comparing the same *dyad* under all hypotheses as exemplarly chosen in the figure (crosses) the most likely relationship class is chosen as $max\{LOD(\Theta_i, \Theta_0)\}_{\geq 0}$ (dashed lines).

One solution could be to adjust the error constant e, corresponding to the mean genotyping quality for each sample which was also discussed in Chapter 4.

5.4.2 Directionality of Pairwise Relationships

All relationship types assigned to each *dyad* by resolving $max\{LOD(\Theta_i, \Theta_0)\}_{\geq 0}$ are not directed which makes it necessary to include at least three individuals of one family into the analysis to reconstruct entire pedigrees. To give an example, if the most likely classification of a *dyad* is a *parent - child* relationship it cannot be distinguished between *parent* and *child*, especially when the sex of each individual is the same (see Figure 5.4 **c)**). On the other hand, additional knowledge about the second *parent* (Figure 5.4 **a)**) or at least one *sibling* (Figure 5.4 **b)**) can be used to clear the assignment of *parent - child* relationships. Further the sex of each individual is determined by the ratio of *heterozygous* SNVs compared to all variants which are located on the X chromosome.

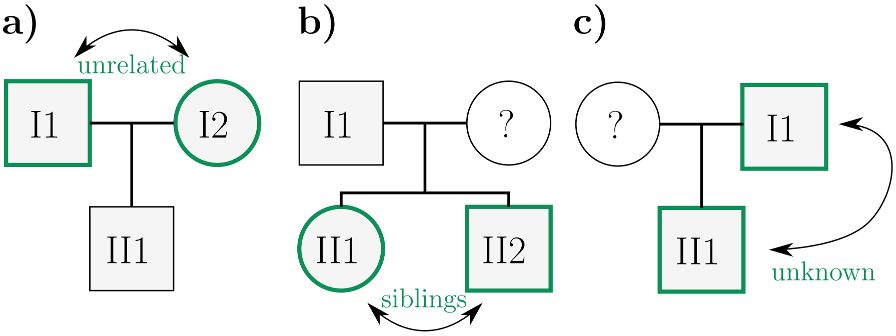


Figure 5.4: *Directionality of parent - child relationships.* At least three related samples are sufficient to classify a *parent - child* relationship with known directionality, either due to the knowledge that both *parents* are unrelated (**a)**) or with the additional information from *siblings* (**b)**). The directionality cannot be resolved with just two sequenced samples, that are involved in the analysis (**c)**). Without additional information, it is impossible to infer whether *I1* or *II2* is the *parent*.

Parents with Different Ethnic Backgrounds

In general without additional knowledge the directionality of a *parent - child* relationship of a single *dyad* cannot be resolved. One special exception is given in the case when both *parents* have strongly deviating ethnic backgrounds, as the offspring will show a combination of two sets of population specific *SNP*s. The *heterozygosity* of the offspring will be higher than of each of the *parent*s as the frequency of many polymorphisms differs, depending on the background population, and the directionality can be resolved without any additional information.

5.4.3 Precision of LOD Scores

The average number of *SNV*s that is expected in the *consensus* exonic target region as defined by the *1KGP* yields about $15000 - 20000$ positions per sample. With these high amount of markers a *precision* of 100% could be achieved for all tested *dyads*.

Effect of Small Target Regions

In order to test the effect of smaller target regions, the number of *SNV*s was randomly reduced for all *1KG* samples and pedigree prediction was performed for all reduced subsets repeatedly for 100 times. The obtained precision remains constant for all *full sibling* and *parent - child dyads* up to 1000 randomly chosen positions but starts to drop with less available markers, as shown in Figure 5.5. At a comparable number of *SNV*s, *dyads* with a 2^{nd} *order* relationship are more difficult to distinguish from other classes and especially unrelated *dyads* are more often falsely assigned to 2^{nd} *order* relationships.

Heterozygosity

The contribution of different *GT* combinations to the *LOD* scores also depend on their *heterozygosity*, h, which is defined as

$$h = 1 - \sum_n f_n^2, \tag{5.5}$$

whereby f_n is the *AF* of the n*th* allele.

The mean *heterozygosity* for *bi-allelic SNV*s in *WES* experiments

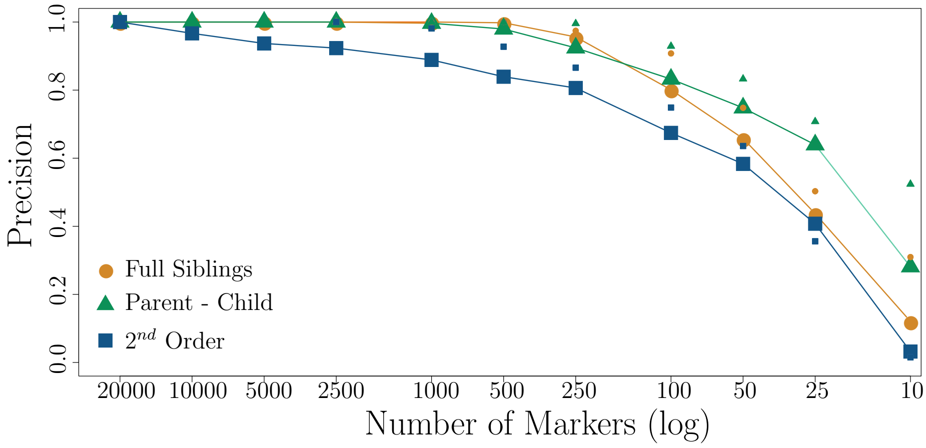


Figure 5.5: *Precision of pedigree prediction with decreasing number of loci.* The number of loci was either randomly reduced (big shapes) or restricted to a subset of highly informative markers (small shapes). In general the positive predicted value or *precision* drops with decreasing number of positions for all three relationship types: *full siblings* (orange circles), *parent-child* (green triangles) and 2^{nd} *order* (blue squares). For *parent-child* as well as for *full sibling* relationships the precision starts to drop when using ≤ 500 randomly chosen markers, whereas 2^{nd} *order* relationships are more difficult to distinguish from unrelated individuals. A higher precision can be achieved when choosing *bi-allelic* positions with a *heterozygosity* above the average value of $h = 0.3$ in *WES* experiments.

is ~ 0.3. When the subset of markers is not chosen randomly but with a high *heterozygosity* above 0.3, higher precision values can be achieved for all relationship types (small shapes in Figure 5.5). This is also in good agreement with approaches in studies, where predominantly markers with a high *IC* are used [45].

5.4.4 Influence of Inbred Structures

In families with a high degree of *consanguinity*, the prediction of exact pedigrees becomes more challenging. The *heterozygosity* decreases in outbred populations and as a result the *AFs* deviates stronger from the expected *null*-model. To test the influence of this known bias on the assignment of relationship types via *LOD*

Simulation of Inbred Structures

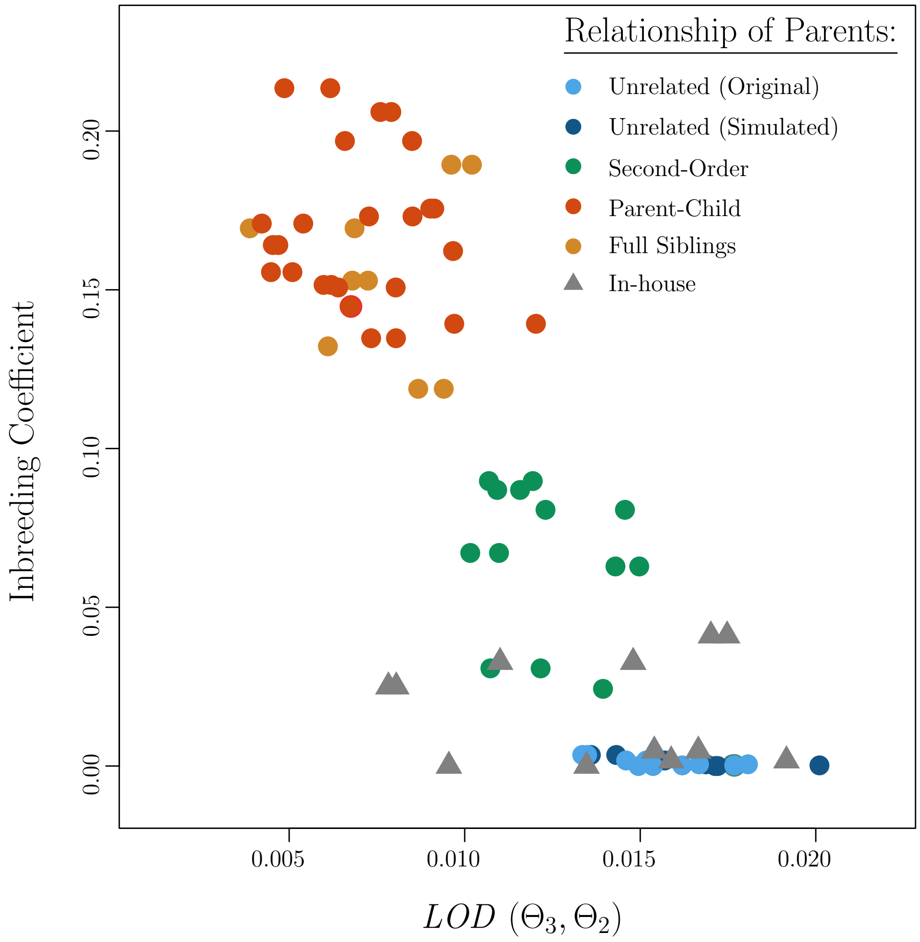


Figure 5.6: *Classification of relationships in highly consanguineous families.* Offspring from related *1KG dyads* with different degrees of relatedness (*full siblings* , *parent - child* and 2^{nd} *order*) were simulated. Additionally the *individual inbreeding coefficient* was calculated for all offsprings and the *LOD* score $LOD(\Theta_3, \Theta_2)$ was computed for all simulated *parent - child* pairs. A correct classification was possible for all simulated *parent - child dyad*s although a strong negative correlation between the degree of inbreeding and $LOD(\Theta_3, \Theta_2)$ exist.

scores not only three *in-house* families with a self-declared consanguineous degree were analysed but additionally inbred structures were simulated using related *1KG dyads*. For this purpose one *allele* from each assumed *parent* was chosen randomly at each locus to create a new *GT* at this position for the simulated offspring.

With this approach different degrees of *consanguinity* could be created, depending on the relationship of the assumed *parents* which can be quantified with the *inbreeding coefficient* described as the proportion of the genome that is *IBD* [125] [49].

Inbreeding Coefficient

As shown in Figure 5.6 high *inbreeding coefficient*s correlate with lower *LOD* scores between relationship models with the same *kinship coefficient*, $LOD(\Theta_3, \Theta_2)$. However, all relationship types could be resolved as *parent* - *child* regardless of the underlying *inbreeding coefficient* of the simulated offsprings as well as of the *consanguin in-house* datasets.

5.5 Summary of this Chapter

This paragraph focused on an approach to reconstruct entire family structures soley based on *whole exome sequencing* (*WES*) sequencing datasets including rare variants. For this purpose *logarithm of the odds* (*LOD*) scores were computed for different models of relatedness utilizing *allele frequencies* (*AFs*) obtained by 2535 non related individuals of the *1000 genomes project* (*1KGP*) to estimate the probability of the combination of two *genotype*s (*GT*s) at any position.
Pedigrees could be derived with high precision for up to 1000 positions for publicly available as well as *in-house* samples with different underlying *coefficients of relatedness.*
The methods reveals some flaws when the *kinship coefficient* between two models becomes close, which is difficult especially in family structures with a high degree of *consanguinity* (as shown in Figure 5.6). Also *dyads* with a higher degree of relationship than 2^{nd} *order* can just be resolved by the analysis of additional family members as the method highly depends on pre-defined hypotheses. Further it was shown that particularly *dyads* with a 2^{nd} *order* relationship benefit from highly polymorphic marker loci as illustrated in Figure 5.5. Especially in paternity testing marker loci with a high *heterozygosity*, such as microsatellites (*SSR*s), were used so far. Positions which are highly polymorphic can be found in the *human leukocyte antigen* (*HLA*) cluster on chromosome 6 that includes 7300 known *alleles* [99]. A comparison between the predictive power of multiple *bi-allelic* marker loci such as most *SNV*s and a single marker as for example the *HLA* locus demonstrates the limited discriminatory power as the *information content* (*IC*), $IC = -\sum_n f_n log f_n$, of 13 unlinked *bi-allelic* markers is comparable to the *IC* of the *HLA* locus with 7300 markers. However it could also be demonstrated that also rare and family specific *SNV*s can be used infer to ancestry.

Another drawback of the approach is, that strongly deviating levels of genotyping quality are not considered, which can lead to falsely classified relationship types as shown in Figure 5.3 for different technical *replicates* with deviating *per-base coverages*. With an decreasing coverage the error rate increases, especially for *heterozygous* variant calls. One solution could be to adapt the error constant e for each sample independently according to the estimated genotyping quality (see Chapter 4).

6 | Discussion

Up to date *next generation sequencing* (*NGS*) has influenced and expedited many biomedical areas. Particularly exome screening that focuses on the protain coding region, has emerged to be an essential tool for the detection of disease causing variants in routine diagnostic.

Still, the detection of pathogenic *single nucleotide variant* (*SNV*) among thousands of potentially disease causing mutations remains challenging and assisting techniques for prioritizing became indispensable. However, all approaches rely on correct sequencing information and the need for methods for *quality control* (*QC*) substantially intensified.

Each sequencing platform is characterized by a specific error profile and stochastic fluctuations may occur in the biological process of library creation. Even in datasets that were sequenced with a high sequencing coverage, up to about 2% of the detected variants couldn't be reproduced if resequenced by a different technique [115]. Further, algorithms for sequence *read* alignment and variant detection cannot compensate for artefacts that were either introduced during the sequencing process or that arise due to error prone regions, such as highly repetitive genomic ranges. Quite the contrary, most variant calling algorithms make strong assumptions about the distribution of the sequencing distribution and deviations affect the sensitivity of *SNV* identification. Especially the error rates in the detection of *heterozygous* *SNV*s is much higher compared to *ho-*

mozygous variants. This is mainly due to a lower sequencing coverage, for exampels in *GC*-rich regions, and the preferential capture of *reference* sequence alleles.

Most of these systematic biases arise during *polymerase chain reaction* (*PCR*) amplification and lead to a relatively heterogeneous profile of the overall sequencing *coverage* over the genome. In contrast, *Whole genome sequencing* (*WGS*) would yield a more uniform distribution of sequencing reads as it does not contain an additional *PCR* and capturing step. Until now, *WES* was considered as the state-of-the-art technique in diagnostic set-ups, as it targets the 1% of the genome that encodes for protein sequences at comparable low costs and less storage efforts compared to *WGS*. However, as the costs to sequence a whole genome has come down significantly this may lead to a fundamental rethinking within the community. Not only is the *coverage* uniformity superior to *WES*, but it would also allow to examine *SNV*s and *copy number variation*s (*CNV*s) within the 99% of the non-coding genomic sequence. However, *WES* still benefits from lower costs which also enables to sequence more samples at comparable prices. This is not only important for large scale population studies but also for linkage analysis and filtering approaches based on family data as for instance the detection of *de-novo* mutations.

Apart from *NGS* or 2^{nd} generation sequencing technologies, 3^{rd} generation instruments have already been introduced in the recent years including *Helicos Heliscope* [5], *Pacific Biosciences SMRT* [12] and *Oxford Nanopore* [11]. One of the main advantages of this new generation is that the initial step of *DNA* amplification becomes unnecessary [20] which theoretically increases the accuracy of a sequencing experiment. On the other hand, these new instruments still suffer from worse accuracy and lower throughput compared to *NGS* technologies as for example reported for *Pacific Biosciences* platforms [97].

To overcome the limitations that are associated with *WES* and 2^{nd}

generation sequencing technologies several measures were recommended to address and improve the quality of an exome sample. Besides suggestions for a minimum sequencing depth, other quality parameters were introduced such as the proportion between *transitions* (*Ti*s) and *transvertions* (*Tv*s) or the percentage of variants that are already listed in databases such as *dbSNP*. However, many of these parameters are influenced by the size of the target region and may also depend on the ethnic background population. These examples illustrate that existing guidelines for *QC* are still ambiguous and methods which help to increase the sensitivity of variant calls would improve the analysis of exome sequencing experiments. This work focused on strategies for *quality control* (*QC*) for experiments based on *WES* whereby three different stages and their specific characteristics of the analysis were examined.

Initially, I described the distribution of *AFs* at *heterozygous* positions as measured in *next generation sequencing* (*NGS*) data sets. The assumption of binomially distributed frequencies that is used as prior distribution in variant call algorithms based on *Bayesian* methods often lead to a misclassification of *heterozygous* variants. I could show that the amplification step during sequencing preparation strongly influences the variance of this distribution and deviates from the widely accepted model. The process of fragment generation during the *polymerase chain reaction* (*PCR*) step could be simulated by a two-type *(Bienayme-) Galton-Watson* (*BGW*) branching process. Conclusions about how to reduce the stochastic fluctuations could be drawn from the analytically derived variance of *heterozygous* *AFs*, such as increasing the efficiency of the adaptor ligation and the number of initially used fragments.

The variance of the *AFs* at *heterozygous* *SNV* positions can further be used as an additional quality parameter to indicate the overall *false negative rate* (*FNR*) of an *WES* experiment. Additionally the simulated distribution could serve as an adapted prior distribution for variant calling algorithms to increase the sensitivity of variant calls in targeted exome screens.

Although the amplification process may become obsolete in the future as *WGS* develop to be more and more feasible, one should keep in mind that simple technical replication of samples substantially decrease the *FNR* of variant calls.

The second topic of this work concentrated on an approach to assess the overall accuracy of an exome sample, independent of the sequencing platform. I developed a distance metric that emphasizes the weight on rare variants and used this to compare *in-house* *WES* samples to a high quality *reference* set with the best suiting background population, which was sequenced by the *1000 genomes project* (*1KGP*). Groups with different simulated accuracies, based on the *1000 Genomes* (*1KG*) exome data, were used to estimate the quality of a single genotyped sample without further knowledge about the sequencing technology or applied bioinformatic pipeline. The derived *standardized dissimilarity score* (*SDS*) and the associated estimated error rate serves as an indicator for the genotyping quality and should yield values comparable to the reference set. In a high quality *WES* sample about $15000 - 20000$ *SNV*s would be detected in a coding region that targets ~ 30*Mb* and one would expect about 300 of them to be false positive. This corresponds to an error rate of ~ 0.00001 and this is also the value one would expect for a good quality *WES* sample.

This approach heavily relies on finding a suiting *reference* background population to which a sample can be compared to. If a *WES* sample is compared to a *reference* set, that is not exactly the best suiting background population, all population specific *single nucleotide polymorphisms* (*SNP*s) that are absent in the *reference* set will be treated as false positive variant calls and influence the result. Similar problems occur when an offspring of two different ethnic backgrounds is analysed. Up to date the *1000 genomes project* (*1KGP*) provides sequencing data for 2535 individuals which are grouped into 26 ethnic sub-populations. In general the *reference* set can be adapted and more high quality exomes can be incorporated, for example if more individuals are sequenced by the *1KGP* or by additional *in-house* data and more fine grained population studies will help to improve the breakdown of genetic diversity within sub-populations in the future.

In the case of an unknown source of the sequenced sample I would suggest to first analyse the population background, regardless of the quality of the sample. This could be done in a similar approach with a distance metric that puts for example more weight on common *SNP*s.

Recently the approach could be adapted to another application by Zhu et al. [127]. In *rare variant association studies*s (*RVAS*s) of cohorts with ultra-rare disorders the size of the available samples is usually very small and approaches such as *GWAS* cannot be applied. In the work of Zhu et al. it was shown that a similarity based matching strategy for the set-up of case-control groups is suitable to correct for possible confounding effects in the underlying data that occur due to varying ethnicities or heterogeneous quality of the data.

In the exploration of high dimensional sequencing data the only strategy to reduce the search space of possible pathogenic variants is often to integrate additional samples into the analysis. These approaches rely on correct pedigree information and can lead to false conclusions in the case of sample *mix-ups*. Therefore, the estimation of relatedness structures between pairs of individuals was the last important topic that was studied in this work.

I developed a tool based on *logarithm of the odds* (*LOD*) scores to test different pre-defined hypotheses of relatedness between two *WES* samples and correct pedigrees could be derived for publicly available *1KG* families as well as for *in-house* data with different *relatedness coefficients*. The most difficult separation occurs between *unrelated* individuals in a highly homogeneous population with consanguineous structures and samples that are related by 2^{nd} *order*.

This is especially challenging when only a few marker loci are available. Usually positions with a high *information content* (*IC*), such as *simple sequence repeat*s (*SSR*s), were used to infer paternity or differentiate between full and half *siblings*. As these markers are not available in *WES* experiments, such an approach is not suitable in this setting without the effort of additional experiments. Up to date, the most polymorphic site that could be obtained via exome sequencing is the *human leukocyte antigen* (*HLA*) cluster on chromosome 6 which is characterized by more than 7300 different alleles [99]. However, a comparison between the predictive power of multiple *bi-allelic* *SNV*s and one single polymorphic marker such as *HLA* shows some limiting disillusionment. The *information content* (*IC*) of 13 *bi-allelic* variants would be the same as the *IC* of the *HLA* locus. However, it was shown in this work that ancestry could also be predicted with the use or rare and family specific variants that are available in exome screens.

It was also illustrated that different levels of qualities within the data influence the *precision* of the pedigree prediction. Until now, just one error constant is used to account for misclassified variants. This could be adapted by combining the estimation of an overall genotyping accuracy for each sample with the prediction of pedigrees.

The estimation of relatedness structures could markedly be improved in combination with *haplotype* reconstruction [58]. However, the derivation of long haplotypes for 2^{nd} generation *WES* experiments that are based on *PCR* amplification is limited but could be realized with *WES*. With an *likelihood ratio* (*LR*) based approach that does not only utilize *GT* frequencies but includes additional *haplotype* information, also difficult relationship types such as 2^{nd} *order* could be resolved with higher precision.

In summary I studied various levels of a exome sequencing experiments and the different possibilities for *quality control* (*QC*). As *whole exome sequencing* (*WES*) is an important part in routine diagnostics, the estimation and improvement of data quality is crucial. For a reliable variant calling, it is not only important to observe the *coverage* of an exome over the target region, as several other biases can influence the sensitivity of an exome screen. To assess a reliable estimation of the quality of an experiment, the overall distribution of the sequence reads has to be taken into account. Further, as several international projects provide good quality data sets by now, these should be taken as a comparable *reference* to estimate the overall accuracy of the whole variance set. Finally, as a simple mislabelling can lead to serious misinterpretation of the data, testing for underlying family structures or cryptic relationships can help to improve and speed up the analysis.

7 | Summary

During the last decade methods based on *NGS* have revolutionized the field of medical genetics. By sequencing the protein coding region via *WES* genetic variations that appear in *Mendelian* disorders can be identified. Further, additional approaches were introduced to reduce the search space for potentially pathogenic mutations, for example by including more family members into the analysis. With a growing rate of technical advances also new challenges have arisen and methods for *QC* are crucial to increase the sensitivity of variant detection. In this work different strategies for *QC* are presented which concentrate on three levels of the analysis of an *WES* experiment.

The distribution of *AFs* at *Heterozygous* positions is associated with the amplification step during library preparation before sequencing. Strong deviations from the expected mean of 0.5 lead to an increased error rate in the detection of genetic variations. It is shown that the variance of this distribution can be modelled with a two-type *BGW* branching process. With this, conclusions can be drawn on how to reduce stochastic fluctuations caused by the amplification step. Additionally the derived variance can be used as an indicator for the error rate of a *WES* sample.

Variant detection is strongly influenced by the ethnic background of an individual as *SNP* frequencies have population specific characteristics. Here the exome wide accuracy is estimated by comparing all variants of an *WES* sample to a good quality *Reference* set with a matching background population, using a distance metric that emphasises weight on rare variants.

Most strategies to filter for potentially pathogenic variants are based on the simultaneous analysis of several family members, for example if filtering for *De-Novo* mutations. However, these techniques strongly rely on correct pedigree information and sample *Mix-Ups* considerably affect the analysis and can lead to false conclusions. In this work relatedness structures between samples are inferred by calculating *LOD* scores based on population *GT* frequencies.

These approaches complement existing *QC* recommendations and help to indicate the accuracy of a *WES* sample.

8 | Zusammenfassung

Durch die Sequenzierung der Protein kodierenden genomischen Region können genetische Variationen identifiziert werden, die Mendelischen Krankheiten zugrunde liegen. Dabei sind Methoden zur Qualitätskontrolle ein essentieller Bestandteil, um die Sensitivitt der Detektion von genetischen Varianten abzuschätzen und zu steigern. In dieser Arbeit werden verschieden Strategien zur Qualitätskontrolle vorgestellt, welche sich auf drei verschiedene Phasen in der Analyse eines Exoms konzentrieren.

Die Verteilung von Allele Frequenzen an heterozygoten Positionen ist mit einem Amplifikationsschritt assoziiert, welcher der Sequenzierung vorrausgeht. Es wurde gezeigt, dass die Varianz dieser Verteilung mit einem Verzweigungsprozess modelliert werden kann. Mithilfe dieser Simulation können Rückschlüsse über die stochastischen Fluktuationen während des Amplifikationsschrittes gezogen werden, womit sich die Fehlerrate eines Experimentes abschätzen lässt.

Die Detektion von Varianten ist stark durch den ethnischen Hintergrund eines Individuums beeinflusst, da SNP Häufigkeiten populationsspezifische Charakteristika aufweisen. Durch den Vergleich aller Varianten eines Exoms mit einem qualitativ guten Referenzset, welches einen ähnlichen Populationshintergrund aufweist, kann die Genauigkeit eines Experimentes abgeschaetzt werden. In diseser Arbeit wurde dafür eine Distanzmetrik verwendet die seltene Varianten stärker gewichtet als Häufige.

Viele Strategien, die angewandt werden um nach möglichen pathogenen Mutationen zu filtern, basieren auf der Analyse mehrerer Familienangehöriger. Allerdings sind diese Ansätze auf korrekte Stammbäume angewie-sen and mögliche Probenverwechslungen behindern die Analyse und führen zu falschen Ergebnissen. In dieser Arbeit wurden Verwandtschaftsbeziehungen mithilfe von Likelihood-Quotienten-Tests ermittelt, welche auf Genotypfrequenzen basieren.

Die vorgestellten Ansätze ergänzen vorhandene Empfehlungen zur Qualitätskontrolle und helfen, die Genauigkeit eines Exom Experimentes zu bestimmen.

9 | Acknowledgments

I would like to express my appreciation to several people for their support and encouragement during my Ph.D. time.

First of all I am grateful to Professor Dr. Stefan Mundlos and Prof. Dr. med. Peter N. Robinson who gave me the opportunity to join the Computational Biology Group at the Institute for Medical Genetics and Human Genetics at the Charité, Berlin. I've had the privilege to work together with wonderful colleagues who never stopped to support me. Thank you for all the productive coffee breaks!

My special thanks go to Dr. med. Peter Krawitz for his never-ending patience, support and continuous advice during the last years.

I would like to express my gratitude to the Deutsche Forschungsgemeinschaft (DFG), the SYBIL EU-Project, the Berlin Institute for Health (BIH) and the German Epigenome Programme (DEEP) for their financial support during my research.

I would also like to thank Prof. Dr. Martin Vingron who is not only supervising my thesis but also gave me the opportunity to continue my research in his Computational Molecular Biology Group at the Max Planck Institute for Molecular Genetics. My sincere thanks go to all members of the group for a very warm welcome and inspiring conversations.

Further, I would like to thank Prof. Dr. Peter Imkeller from the Department of Mathematics at the Humboldt-University in Berlin and Prof. Dr. Thorsten Dickhaus from the Weierstrass Institute for Applied Analysis and Stochastics (WIAS) for stimulating discussions and advice.

I would like to especially thank Prof. Dr. med. Peter N. Robinson, Dr. med. Peter Krawitz, Leon Kuchenbecker, Max Schubach and Sabrina Krakau for reading my thesis and providing suggestions.

I am deeply grateful to my friends for encouraging me and constantly providing moral support. Finally, I am greatly indebted to my parents who have always supported me and without whom I never could have reached this point.

List of Abbreviations

1KG	*1000 Genomes*	
1KGP	*1000 genomes project*	ii
A	*adenine*	2
AA	*amino acid*	13
AI	*allelic imbalance*	22
AF	*allele frequency*	ii
a.s.	*almost surely*	
AUC	*area under the curve*	
bp	*base pair*	9
BER	*samples obtained from Berlin*	
BGW	*(Bienayme-) Galton-Watson*	ii
BWT	*Burrows-Wheeler transform*	20
C	*cytosine*	2
CCDS	*collaborative consensus coding sequence*	28
CEU	*Utah residents (CEPH) with northern and western ancestry*	
CHR	*chromosome*	
CLT	*central limit theorem*	59
CNV	*copy number variation*	12
CVAS	*common variant association studies*	105
dbSNP	*single nucleotide polymorphism database*	14
DNA	*deoxyribonucleic acid*	i

ESP	*exome sequencing project*	89
FNR	*false negative rate*	
FPR	*false positive rate*	76
G	*guanine*	2
GATK	*The Genome Analysis Toolkit*	
Gb	*giga base pair*	19
GT	*genotype*	11
GPI	*glycophos-phatidyl-inositol*	
GRC	*genome reference consortium*	
GRCh37	*genome reference consortium human genome build 37*	
GRCh38	*genome reference consortium human genome build 38*	
GWAS	*genome wide association study*	34
HGP	*human genome project*	4
HLA	*human leukocyte antigen*	128
HPO	*human phenotype ontology*	
HWE	*Hardy-Weinberg equilibrium*	11
IHGSC	*international human genome sequencing consortium*	9
kb	*kilo base (1000bp)*	
i.i.d	*independent and identically distributed*	
IBD	*identical-by-descent*	110
IBS	*identical-by-state*	110
IC	*information content*	114
INDEL	*insertion or deletion*	9
IQR	*inter qartile range*	94
LLN	*law of large numbers*	51
LOD	*logarithm of the odds*	34
LR	*likelihood ratio*	ii
MAPQ	*mapping quality score*	21
Mb	*megabyte*	

MDS	*multi dimensional scaling*	85
µg	*mikrogramm*	
NCBI	*national center for biotechnology information*	10
ng	*nanogramm*	
NGS	*next generation sequencing*	i
NHGRI	*the national human research intsitute*	
NWA	*Needleman-Wunsch algorithm*	20
PCA	*principal component analysis*	85
PCR	*polymerase chain reaction*	3
PGF	*probability generating function*	
POP	*population*	
PRC	*percent rule call*	
QC	*quality control*	ii
RAF	*reference allele frequency*	23
Ref	*reference genome*	9
ROC	*receiving operating characteristics*	
RVAS	*rare variant association studies*	iii
SAM	*sequence alignment/map*	
SDS	*standardized dissimilarity score*	82
SNP	*single nucleotide polymorphism*	14
SNV	*single nucleotide variant*	
SSR	*simple sequence repeat*	114
SV	*structural variation*	12
T	*thymine*	2
Ti	*transition*	31
TP	*true positive*	74
Trio	*a family consisting of mother, father and child*	
Tv	*transvertion*	31
UCSC	*university of California Santa Cruz*	

VCF	*variant calling format*	22
WES	*whole exome sequencing*	i
WGS	*whole genome sequencing*	14

Bibliography

[1] 454. http://www.454.com.

[2] Agilent. https://www.broadinstitute.org/.

[3] Broad institute. https://www.broadinstitute.org/.

[4] Exome sequencing project. http://evs.gs.washington.edu/EVS/.

[5] Helicos heliscope. http://www.helicosbio.com.

[6] Illumina. http://www.illumina.com.

[7] The international genome sample resource. http://www.1000genomes.org.

[8] Mcdonell genome institute. http://genome.wustl.edu.

[9] Ncbi. http://www.ncbi.nlm.nih.gov.

[10] Novocraft. http://www.novocraft.com/.

[11] Oxford nanopores. http://www.nanoporetech.com.

[12] Pacific biosciences. http://www.pacificbiosciences.com.

[13] Samtools. http://samtools.sourceforge.net/.

[14] Solid. http://www.appliedbiosystems.com.

[15] Carl A Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, P Morris, and Krina T Zondervan. Europe PMC Funders Group Data quality control in genetic case-control association studies. 5(9):1564–1573, 2011.

[16] Y Aoki, Y Nakayama, K Saigusa, M Nata, and M Hashiyada. Comparison of the Likelihood Ratio and identity-by-State Scoring Methods for Analyzing Sib-Pair Test Cases: A Study Using Computer Simulation. *Tohoku J. Exp. Med.*, 194:241–250, 2001.

[17] K.B. Athreya and P.E. Ney. *Branching Processes*. Springer, 1972.

[18] M.N. Bainbridge, M. Wang, Y. Wu, I. Newsham, D.M. Muzny, J.L. Jefferies, T.J. Albert, D.L. Burgess, and R.A. Gibbs. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome Biology*, 12:R68, 2011.

[19] 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 Publishing Group*, 12(11):745–755, 2011.

[20] MichaelJ Becich, Lucas Santana-Santos, RamaR Gullapalli, KetakiV Desai, and JeffreyA Kant. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. *Journal of Pathology Informatics*, 3(1):40, 2012.

[21] Callum J Bell, Darrell L Dinwiddie, Neil a Miller, Shannon L Hateley, Elena E Ganusova, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Science translational medicine*, 3(65):65ra4, 2011.

[22] D Bentley, S Balasubramanian, H Swerdlow, G Smith, J Milton, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.

[23] P. Billingsey. *Probability and Measure.* John Wiley & Sons, 3 edition, 1995.

[24] Michael S Blouin. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology and Evolution*, 18(10):503–511, 2003.

[25] DC Brock. *Understanding Moore's Law: Four Decades of Innovation.* 2006.

[26] Andrew R Carson, Erin N Smith, Hiroko Matsui, Sigrid K Bræ kkan, Kristen Jepsen, John-Bjarne Hansen, and Kelly a Frazer. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC bioinformatics*, 15:125, 2014.

[27] Shihyen Chen, Bin Ma, and Kaizhong Zhang. On the similarity metric and the distance metric $. *Theoretical Computer Science*, 410(24-25):2365–2376, 2009.

[28] Ananyo Choudhury, Scott Hazelhurst, Ayton Meintjes, Ovokeraye Achinike-oduaran, Shaun Aron, Junaid Gamieldien, Mahjoubeh Jalali, Sefid Dashti, Nicola Mulder, Nicki Tiffin, and Michèle Ramsay. Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance. pages 1–20, 2014.

[29] Deanna M Church, Valerie a Schneider, Karyn Meltz Steinberg, Michael C Schatz, Aaron R Quinlan, Chen-Shan Chin, Paul a Kitts, Bronwen Aken, Gabor T Marth, Michael M Hoffman, Javier Herrero, M Lisandra Zepeda Mendoza,

Richard Durbin, and Paul Flicek. Extending reference assembly models. *Genome biology*, 16(1):13, 2015.

[30] Fs Collins, Fs Collins, Ld Brooks, Ld Brooks, a Chakravarti, and a Chakravarti. A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation. *Genome Research*, 8(12):1229–1231, 1998.

[31] The 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. *Nature*, 476:1061–1073, 2010.

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

[33] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526:68–74, 2015.

[34] The 1000 Genomes Project Consortium. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526:75–81, 2015.

[35] TF Cox and MAA Cox. *Multidimensional Scaling.* 2001.

[36] 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*, 27(15):2156–2158, 2011.

[37] Jacob F. Degner, John C. Marioni, Athma a. Pai, Joseph K. Pickrell, Everlyne Nkadori, Yoav Gilad, and Jonathan K. Pritchard. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–3212, 2009.

[38] M. DePristo, E. Banks, R. Poplin, K. Garimella, J. Maguire, C. Hartl, A. Philippakis, G. del Angel, M.A. Rivas, M. Hanna, A. McKenna, T. Fennell, A. Kernytsky, A. Sivachenko, K. Cibulskis, S. Gabriel, D. Altshuler, and M. Daly. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature Genetics*.

[39] Faà di Bruno. Sullo sviluppo delle Funzioni. *Annali di Scienze Matematiche e Fisiche*, 6:479–480., 1855.

[40] Juliane C. Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16), 2008.

[41] JT Dudley and KJ Karczewski. *Exploring Personal Genomics*. 2013.

[42] Richard M Durbin, David L Altshuler, Gonçalo R Abecasis, David R Bentley, et al. SUPP A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[43] N. Ehmke, A. Caliebe, R. Koenig, S.G. Kant, Z. Stark, D. Wieczorek, G. Gillessen-kaesbach, K. Hoff, A. Knaus, N. Zhu, V. Heinrich, C. Huber, I. Harabula, M. Spielmann, D. Horn, H. Manzke, and S. Mundlos. Homozygous and Compound-Heterozygous Mutations in TGDS Cause Catel-Manzke Syndrome. *Am J Hum Genet.*, 95(6):763–70, 2014.

[44] D. Emmerich, T. Zemojtel, J. Hecht, P. Krawitz, M. Spielmann, J. Kühnisch, K. Kobus, M. Osswald, V. Heinrich, P. Berlien, U. Müller, V. Mautner, K. Wimmer, P.N. Robinson, M. Vingron, S. Tinschert, S. Mundlos, and M. Kolanczyk. Somatic neurofibromatosis type 1 (NF1

) inactivation events in cutaneous neurofibromas of a single NF1 patient. *Eur J Hum Genet.*, 23(6):870–3, 2015.

[45] M P Epstein, W L Duren, and M Boehnke. Improved inference of relationship for pairs of individuals. *American journal of human genetics*, 67(5):1219–1231, 2000.

[46] Edward J Fox, Kate S Reid-Bayliss, Mary J Emond, and Lawrence a Loeb. Accuracy of Next Generation Sequencing Platforms. *Next generation, sequencing & applications*, 1(1):1–4, 2014.

[47] M. Franke, D.M. Ibrahim, G. Andrey, W. Schwarzer, V. Heinrich, R. Schöpflin, K. Kraft, R. Kempfer, I. Jercovic, W.L. Chan, M. Spielmann, B. Timmermann, I. Kurth, P. Cambiaso, O. Zuffardi, G. Houge, L. Lambie, F. Brancati, A. Pombo, M. Vingron, F. Spitz, and S. Mundlos. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538(7624):265–9, 2016.

[48] Jl Freeman, Gh Perry, and Lars Feuk. Copy number variation: new insights in genome diversity. *Genome research*, (617):949–961, 2006.

[49] S. Gazal. FSuite: exploiting inbreeding in dense SNP chip and exome data. *Bioinformatics*, 23(10):1289–91, 2007.

[50] Travis C. Glenn. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, 11(5):759–769, 2011.

[51] Andreas Gnirke, Alexandre Melnikov, Jared Maguire, Peter Rogov, Emily M LeProust, William Brockman, Timothy Fennell, Georgia Giannoukos, Sheila Fisher, Carsten Russ, Stacey Gabriel, David B Jaffe, Eric S Lander, and Chad Nusbaum. Solution hybrid selection with ultra-long oligonucleotides for

massively parallel targeted sequencing. *Nature biotechnology*, 27(2):182–9, 2009.

[52] J.C. Gower. Metric and Euclidean Properties of Dissimilarity Coefficients. *Journal of Classification*, 3:5–48, 1986.

[53] JJ Graw. Genetik, 4. Auflage. *Springer*, 2006.

[54] Georgii H. *Stochastik - Einfhrung in die Wahrscheinlichkeitstheorie und Statistik*. Walter de Gruyter, 1002.

[55] Olivier J Hardy, Olivier J Hardy, Nathalie Charbonnel, and Nathalie Charbonnel. Microsatellite Allele Sizes: A Simple Test to Assess Their Significance on Genetic Differentiation. *Computer*, 1482(April):1467–1482, 2003.

[56] Olivier Harismendy, Pauline C Ng, Robert L Strausberg, Xiaoyun Wang, Timothy B Stockwell, Karen Y Beeson, Nicholas J Schork, Sarah S Murray, Eric J Topol, Samuel Levy, and Kelly a Frazer. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome biology*, 10(3):R32, 2009.

[57] Ayat Hatem, Doruk Bozda, and Ümit V. Çatalyürek. Benchmarking short sequence mapping tools. *Proceedings - 2011 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2011*, pages 109–113, 2011.

[58] D. He, Z. Wang, B. Han, L. Parida, , and E. Eskin. IPED: Inheritance Path-based Pedigree Reconstruction Algorithm Using Genotype Data. *Journal of Computational Biology*, 20(10):780–791, 2013.

[59] Alexander Hruscha, Peter Krawitz, Alexandra Rechenberg, V. Heinrich, Jochen Hecht, Christian Haass, and Bettina

Schmid. Efficient CRISPR/Cas9 genome editing with low off-target effects in zebrafish. *Development*, 140(5):4982–4987, 2013.

[60] T Kamphans and P M Krawitz. GeneTalk: an expert exchange platform for assessing rare sequence variants in personl genomes. *Bioinformatics*, 28:2515–2516, 2012.

[61] T. Kamphans, P. Sabri, N. Zhu, V. Heinrich, S Mundlos, P.N. Robinson, D. Parkhomchuk, and P.M. Krawitz. Filtering for Compound Heterozygous Sequence Variants in Non-Consanguineous Pedigrees. *PLoS One*, 8(8):1–6, 2013.

[62] A.N. Kolmogorov. *Foundations of Probability Theory*. 1931.

[63] Peter M. Krawitz, Daniela Schiska, Ulrike Krger, Sandra Appelt, V. Heinrich, Dmitri Parkhomchuk, Bernd Timmermann, Jose M. Millan, Peter N. Robinson, Stefan Mundlos, Jochen Hecht, and Manfred Gross. Screening for single nucleotide variants, small indels and exon deletions with a next-generation sequencing based gene panel approach for usher syndrome. *Molecular Genetics & Genomic Medicine*, 2(5):393–401, 2014.

[64] J.B. Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29:115–29, 1964.

[65] P Y Kwok, Q Deng, H Zakeri, S L Taylor, and D a Nickerson. Increasing the information content of STS-based genome maps: identifying polymorphisms in mapped STSs. *Genomics*, 31(1):123–126, 1996.

[66] H. Lademann, B. Gerber, D.M. Olbertz, E. Darvin, L. Stauf, K. Ueberholz, V. Heinrich, J. Lademann, and V. Briese. Non-Invasive Spectroscopic Determination of the Antioxidative

Status of Gravidae and Neonates. *Skin Pharmacol Physiol.*, 28(4):189–95, 2015.

[67] B. Langmead and S. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9:357–9, 2013.

[68] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[69] H. Li. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–93, 2011.

[70] H Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv:1303.3997v1 [q-bio.GN]*, 2013.

[71] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25:1754–60, 2009.

[72] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 26:589–95, 2010.

[73] H. Li and N. Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, II(5):473–83, 2010.

[74] H. Li, J. Ruan, and R. Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18(11):1851–8, 2008.

[75] Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul M B Vitányi. The Similarity Metric. 50(12):3250–3264, 2004.

[76] 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, Alexander Hoischen, Hans Scheffer, Bert B A De Vries, Han G Brunner, and Joris A Veltman. Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. pages 1921–1929, 2012.

[77] B.H Liu. *Statistical Genomics Linkage, Mapping and QTL Analysis*. 1998.

[78] Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, 2012, 2012.

[79] D. Lupianez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J.M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S.A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, and S. Mundlos. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, 161(5):1012–25, 2015.

[80] Elaine R. Mardis. Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, 9(1):387–402, 2008.

[81] T C Marshall, J Slate, L E B Kruuk, and J M Pemberton. Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, 7:639–655, 1998.

[82] I. Mathieson and G. McVean. Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, 44(3):243–6, 2012.

[83] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M.A. DePristo. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Current Protocols in Bioinformatics.*

[84] J D McPherson and Others. A physical map of the human genome. *Nature*, 409(2):934, 2001.

[85] Ryan E Mills, Christopher T Luttig, Christine E Larkins, Adam Beauchamp, Circe Tsui, W Stephen Pittard, and Scott E Devine. An initial map of insertion and deletion (INDEL) variation in the human genome An initial map of insertion and deletion (INDEL) variation in the human genome. pages 1182–1190, 2006.

[86] Julienne M. Mullaney, Ryan E. Mills, W. Stephen Pittard, and Scott E. Devine. Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics*, 19(R2):131–136, 2010.

[87] K.F. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, and H. Erlich. pecific enzymatic amplification of DNA in vitro: The polymerase chain reaction. *Cold Spring Harbor Symposium in Quantitative Biology*, 51:263–273, 1986.

[88] Kensuke Nakamura, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, Margaret C. Linak, Aki Hirai, Hiroki Takahashi, Md Altaf-Ul-Amin, Naotake Ogasawara, and Shigehiko Kanaya. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, 39(13), 2011.

[89] Stephan Pabinger, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R Speicher, and Johannes Zschocke. A survey of

tools for variant analysis of next-generation genome sequencing data. 15(2):256–278, 2013.

[90] Stephan Pabinger, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R. Speicher, Johannes Zschocke, and Zlatko Trajanoski. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, 15(2):256–278, 2014.

[91] Shraddha Pandit and Suchita Gupta. A comparative study on distance measuring. 2(1):29–31, 2011.

[92] Zubin H. Patel, Leah C. Kottyan, Sara Lazaro, Marc S. Williams, David H. Ledbetter, Gerard Tromp, Andrew Rupert, Mojtaba Kohram, Michael Wagner, Ammar Husami, Yaping Qian, C. Alexander Valencia, Kejian Zhang, Margaret K. Hostetter, John B. Harley, and Kenneth M. Kaufman. The struggle to find reliable results in exome sequencing data: Filtering out Mendelian errors. *Frontiers in Genetics*, 5(FEB):1–13, 2014.

[93] J M Pemberton. Wild pedigrees: the way forward. *Proceedings. Biological sciences / The Royal Society*, 275(1635):613–621, 2008.

[94] François Pompanon, Aurélie Bonin, Eva Bellemain, and Pierre Taberlet. Genotyping errors: causes, consequences and solutions. *Nature reviews. Genetics*, 6(11):847–859, 2005.

[95] Kim D Pruitt, Jennifer Harrow, Rachel A Harte, Craig Wallin, Mark Diekhans, Donna R Maglott, Steve Searle, Catherine M Farrell, Jane E Loveland, Barbara J Ruef, Elizabeth Hart, Marie-marthe Suner, Melissa J Landrum, Bronwen Aken, Sarah Ayling, Robert Baertsch, Julio Fernandez-banet, Joshua L Cherry, Val Curwen, Michael Dicuccio, Manolis

Kellis, Jennifer Lee, Michael F Lin, Michael Schuster, Andrew Shkeda, Clara Amid, Garth Brown, Oksana Dukhanina, Adam Frankish, Jennifer Hart, Bonnie L Maidak, Jonathan Mudge, Michael R Murphy, Terence Murphy, Jeena Rajan, Bhanu Rajput, Lillian D Riddick, Catherine Snow, Charles Steward, David Webb, Janet A Weber, Laurens Wilming, Wenyu Wu, Ewan Birney, David Haussler, Tim Hubbard, James Ostell, Richard Durbin, and David Lipman. The consensus coding sequence (CCDS) project : Identifying a common protein-coding gene set for the human and mouse genomes. pages 1316–1323, 2006.

[96] S.M. Pulst. Genetic Linkage Analysis. 56(6), 1999.

[97] Anthony Rhoads and Kin Fai Au. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, 13(5):278–289, 2015.

[98] Kimberly Robasky, Nathan E Lewis, and George M Church. The role of replicates for error mitigation in next-generation sequencing. *Nature reviews. Genetics*, 15(1):56–62, 2014.

[99] J. Robinson, J.A. Halliwell, J.H. Hayhurst, P. Flicek, P. Parham, and S.G.E. Marsh. The IPD and IMGT/HLA Database: allele variant databases. *Nucleic Acids Research*, 43:D423–431, 2015.

[100] P.N. Robinson, S. Kohler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *American Journal of Human Genetics*, 83:610–5, 2008.

[101] Christian Rödelsperger, Peter Krawitz, Sebastian Bauer, Jochen Hecht, Abigail W. Bigham, Michael Bamshad, Birgit Jonske de Condor, Michal R. Schweiger, and Peter N. Robinson. Identity-by-descent filtering of exome sequence

data for disease-gene identification in autosomal recessive disorders. *Bioinformatics*, 27(6):829–836, 2011.

[102] Michael G Ross, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J Lennon, Ryan Hegarty, Chad Nusbaum, and David B Jaffe. Characterizing and measuring bias in sequence data. *Genome biology*, 14(5):R51, 2013.

[103] Matthew Ruffalo, Mehmet Koyutürk, Thomas LaFramboise, and Mehmet Koyutürk. Comparative analysis of algorithms for next-generation sequencing read alignment Background : NGS. *Bioinformatics (Oxford, England)*, 27(20):2790–6, 2011.

[104] F Sanger, S Nicklen, and a R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–7, 1977.

[105] Melanie Schirmer, Umer Z. Ijaz, Rosalinda D'Amore, Neil Hall, William T. Sloan, and Christopher Quince. Insight into biases and sequencing errors for amplicon sequencing with the illumina miseq platform. *Nucleic Acids Research*, 43(6), 2015.

[106] J.M. Schwarz, D.N. Cooper, M. Schuelke, and D. Seelow. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods*, 11(4):361–2, 2014.

[107] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, and K. Sirotkin. bSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29:308–11, 2001.

[108] A.M.S. Shrestha and M.C. Frith. An approximate bayesian approach for mapping paired-end dna reads to a reference genome. *Bioinformatics*, pages 1–8, 2013.

[109] Damian Smedley and P.N Robinson. Phenotype-driven strategies for exome prioritization of human mendelian disease genes. *Genome Medicine*, 7(81):2004–15, 2015.

[110] C. Stern. The HardyWeinberg law. *Science*, 97:137–138, 1943.

[111] Eric L Stevens, Greg Heckenberg, Elisha D O Roberson, Joseph D Baugher, J Thomas, and Jonathan Pevsner. Inference of Relationships in Population Data Using Identity-by-Descent and Identity-by-State. 7(9), 2011.

[112] S C Thomas, D W Coltman, and J M Pemberton. The use of marker-based relationship information to estimatethe heritability of body weight in a natural population: a cautionary tale. *Journal of Evolutionary Biology*, 15:92–99, 2002.

[113] W.S. Torgerson. The first major MDS breakthrough. *Psychometrika*, 17:401–19, 1952.

[114] V. Heinrich, T. Kamphans, S. Mundlos, P.N. Robinson, and P.M. Krawitz. A likelihood ratio-based method to predict exact pedigrees for complex families from next-generation sequencing data. *Bioinformatics*, 33(1):72–8, 2016. doi: https://doi.org/10.1093/bioinformatics/btw550.

[115] V. Heinrich, T. Kamphans, J. Stange, D. Parkhomchuk, J. Hecht, T. Dickhaus, P.N. Robinson, and P.M. Krawitz. Estimating exome genotyping accuracy by comparing to data from large scale sequencing projects. *Genome Medicine*, 5(7):69, 2013. doi: https://dx.doi.org/10.1186%2Fgm473.

[116] V. Heinrich, J. Stange, T. Dickhaus, P. Imkeller, U. Krüger, S. Bauer, S. Mundlos, P.N. Robinson, J. Hecht, and P.M. Krawitz. The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. *Nucleic Acids Res.*,

40(6):2426–31, 2012. doi: https://dx.doi.org/10.1093%2Fnar%2Fgkr1073.

[117] G.A. Van der Auwera, M. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. Garimella, D. Altshuler, S. Gabriel, and M. DePristo. From fastq data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*.

[118] G. Vandeweyer, L. Van Laer, B. Loeys, T. Van den Bulcke, and R.F. Kooy. VariantDB: a flexible annotation and filtering portal for next generation sequencing data. *Genome Medicine*, 6(74):1–10, 2014.

[119] Joris a. Veltman and Han G. Brunner. De novo mutations in human genetic disease. *Nature Reviews Genetics*, 13(8):565–575, 2012.

[120] J Craig Venter, Mark D Adams, Eugene W Myers, et al. The Sequence of the Human Genome. 291(February), 2001.

[121] B.F. Voight and J.K. Pritchard. Confounding from cryptic relatedness in case-control association studies. *PLoS genetics*, 1(3):e32, 2005.

[122] James R. Wagner, Bing Ge, Dmitry Pokholok, Kevin L. Gunderson, Tomi Pastinen, and Mathieu Blanchette. Computational analysis of whole-genome differential allelic expression data in human. *PLoS Computational Biology*, 6(7):24, 2010.

[123] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. 38(16):1–7, 2010.

[124] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin,

Deanna M Church, Michael Dicuccio, Ron Edgar, Scott Federhen, Lewis Y Geer, Yuri Kapustin, Oleg Khovayko, David Landsman, David J Lipman, Thomas L Madden, Donna R Maglott, James Ostell, Vadim Miller, Kim D Pruitt, Gregory D Schuler, Edwin Sequeira, Steven T Sherry, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Roman L Tatusov, Tatiana A Tatusova, Lukas Wagner, and Eugene Yaschenko. Database resources of the National Center for Biotechnology Information. 35(December 2006):5–12, 2006.

[125] S Wright. Coefficients of inbreeding and relationship. *The American Naturalist*, 56:330–8, 1922.

[126] A.Y. Yakovlev and N.M. Yanev. Relative frequencies in mutltype branching processes. *The Annals of Applied Probability*.

[127] N. Zhu, V. Heinrich, T. Dickhaus, J. Hecht, P.N. Robinson, S. Mundlos, T. Kamphans, and P.M. Krawitz. Genome analysis Strategies to improve the performance of rare variant association studies by optimizing the selection of controls. *Bioinformatics*, 31(22):3577–83, 2015.

Typeset March 15, 2017