Technische Universität München

Lehrstuhl für Bioinformatik

Identification of genetic variation using Next-Generation Sequencing

Sebastian Hubert Eck

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:

Univ.-Prof. Dr. H.-R. Fries

Prüfer der Dissertation:

1. Univ.-Prof. Dr. H-W. Mewes
2. Univ.-Prof. Dr. R. Zimmer, Ludwig-Maximilians Universität München
3. Priv.-Doz. Dr. T. M. Strom

Die Dissertation wurde am 30.07.2013 bei der Technischen Universität München eingereicht und am 27.01.2014 durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt angenommen.

Summary

The field of genetic research was drastically changed by the advent of next-generation DNA sequencing technologies during the last three years. These technologies are able to generate unparalleled amounts of DNA sequence information at a cost that is several orders of magnitude lower than standard sequencing techniques. Employing this technology allows identification of disease causative and disease associated mutations from a frequency spectrum previously not accessible with standard techniques. Even though showing great promise, researchers are also facing new challenges. In particular the handling and analysis of unprecedented amounts of data.

The goal of this thesis was the conception and implementation of an automated, computational analysis pipeline for next generation sequencing data in general and whole exome sequencing data in particular in order to robustly identify disease related mutations. The pipeline covers all key analysis steps including alignment to a reference genome, variant calling, quality filtering and annotation of identified variants, storage of all variants in a relational database and identification of putatively causative mutations. The pipeline is a combination of public available software and custom developed programs implemented primarily in Perl. Identification of candidate mutations for specific disorders is facilitated through preformulated database queries via a web interface. The database integrates external ressources such as dbSNP, the Human Gene Mutation Database and Polyphen functional consequences prediction in order to provide key information to select putatively causative mutations. The user may specify regions of particular interest, for instance from a previous linkage analysis, which are then highlighted during analysis.

The pipeline was employed in four distinct projects to show the applicability for mutation identification from a broad frequency spectrum. In the following all components of the analysis pipeline and the complementary database are discussed in detail. Additionally, example results of the four projects, involving the identification of disease associated variants in Leukemia, Intelectual Disability, Parkinson's Disease and Myocardial Infarction, employing the pipeline are presented.

Zusammenfassung

In den letzten drei Jahren wurde das Feld der Genetik durch die Einführung von Next-Generation Sequencing Technologien drastisch revolutioniert. Diese Technologien ermöglichen es noch nie dagewesene Mengen an DNA Sequenzdaten zu generieren, bei Kosten, die mehrere Grössenordnungen niedriger sind als bei den bisherigen Standardtechniken. Der Einsatz dieser Techniken erlaubt die Identifikation von Krankheitsverusachenden und Kranheitsassoziierten Mutationen aus einem Frequenzspektrum, das bisher nicht erreichbar war. Obwohl dieser Ansatz vielversprechend ist sehen sich Wissenschafftler auch mit neuen Herausforderung konfrontiert, insbesondere der Umgang und die Analyse von beispielosen Mengen an Daten.

Das Ziel dieser Arbeit war die Entwicklung und Implementierung einer automatisierten, rechnergestützen Analysepipeline für Next-Generation Sequencing Daten im allgemeinen und Whole-Exome Sequencing Daten im besonderen, um krankheitsassoziierte Mutationen robust und reproduzierbar detektieren zu können. Die Pipeline deckt alle notwendigen Auswerteschritte ab: Alignment der Sequenzierdaten an ein Referenzgenom, Identifikation von Varianten, Qualitätsfiltering und Annotation der identifizierten Varianten, Abspeicherung der Varianten in einer relationalen Datenbank und Detektion von möglichen Krankheitsverusachenden Mutationen. Die Pipeline ist eine Kombination von frei verfügbarer und spezifisch entwickelter Software, die vorrangig in Perl implementiert ist. Die Detektion von kranheitsrelevanten Varianten wird durch Datenbankabfragen über ein Webinterface ermöglicht. Die Datenbank integriert externe Ressourcen wie dbSNP, die Human Gene Mutation Database und Polyphen Vorhersagen für funktionelle Konsequenzen, um essentielle Informationen für die Selektion von kranheitsrelevanten Mutationen zur Verfügung zu stellen. Benutzer können optional Regionen von besonderer Signifikanz spezifieren, zum Beispiel von vorherigen Linkage Analysen, die dann während der Analyse besonders hervorgehoben werden.

Die Analysepipeline wurde Anhand von vier Projekten eingesetzt, um die Anwendbarkeit zur Identifikation von Mutationen aus einem breiten Frequenzspektrum zu demonstrieren. Im Folgenden werden alle Komponenten der Analysepipeline und der zugehörigen Datenbank im Detail erläutert und diskutiert. Die Ergebnisse des Einsatzes der Pipeline in vier Projekten, in denen neue, krankheitsassoziierte Varianten bei Patienten mit Leukämie, geistiger Behinderung, der Parkinson Krankheit und Myokardinfarkt erfolgreich identifiziert wurden, wird beispielhaft dargestellt.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Human Genetic Disorders

A genetic disorder is defined as an illness which is caused by abnormalties or defects in the genome. These abnormalties may affect genes on the DNA sequence level or constitute defects in the chromosomes (chromosomal rearangements or abnormal chromosomal counts). A genetic disorder may either be caused by new mutations in an affected individual or may be passed on through the parents genetic material. In the latter case the disorder is referred to as heritable. In some cases, the same disease may exist as heritable form in some individuals, while the same disorder is caused by new mutations in other individuals. Certain forms of cancer, for instance Acute Myoloid Leukemia (AML) are an example for this type of disorder. Most genetic disorders are very rare and typically affect one person in several thousands or millions. Genetic disorders may be classified according to the number of genes involved in contributing to disease genesis and, in the case of heritable disorders, the underlying inheritance mode.

Single gene disorders: A Single gene, or monogenic disorder, is caused by an genetic abberation in a single gene. There are estimated to exist over 4000 monogenic human disorders. Spontaneous de-novo mutations that affect single individuals may give rise to monogenic disorders, a prominent example of such disorders are types of Intelectual Disability (ID). On the other hand, heritable monogenic disorder may be further classified by the inheritance mode in which they are passed on to subsequent generations.

Autosomal dominant: For an autosomal dominant disorder it is sufficient to be affected if one mutated copy of a gene is present in an individual. Each affected person thus ususally has one affected parent, where the chance of the children inherting the mutated allele is 50%. Autosomal dominant disorders sometimes have reduced penetrance, which means that even though only one mutated copy is sufficient for disease progression, not all individual carrying the mutation are affected. Examples of this type of disorders include Huntington's disease, Marfan syndrome and neurofibromatosis type 1 and type 2, which are all highly penetrant dominant disorders.

Autosomal recessive: In contrast to autosomal dominant disorders, two copies of the genes have to carry mutated alleles in order to develop an autosomal recessive disorder. Affected individuals typically have unaffected parents who each carry a single mutated allele. The unaffected parents are referred to as carriers. The chance of children of two carriers to inherit both mutated alleles is 25%. The mutated alleles typically differ from one another, i.e. two different gene affecting mutations are present. If a child inherits these two distinct mutations the gene is referred to as compound heterozygous. A different case arises if the parents carry the same mutation. In this case the affected child will have a single homozygous mutation in the gene. This scenario is extremly rare, except in the case of a consanginous union, where this is the to be expected case. Examples of this type of disorders are sickle-cell disease, cystic fibrosis and spinal muscular athropy.

X-linked disorders: A special case arises if the mutated gene resides on the X-chromosome. X-linked dominant disorders affect both males and females, with males typically being more severly affected than females. The risk of passing on the disease to children differs with the gender of the affected parent: The sons of a man with an X-linked disorder will all be unaffected by the disease (since they inherit his Y chromosome) whereas all his daughters will be affected. A woman with an X linked dominant disorder has a 50% chance of passing on the mutated allele to her offspring, regardless of gender. Only few disorders follow this inheritance pattern, X-linked hypophosphatemic rickets and Rett syndrome being two examples. X-linked recessive conditions more frequently affect males than females. Again, the chance of passing on the disorder differs between men and women. Sons of a man with an X-linked recessive disorder will not be affected, daughters will

carry one mutated allele, thus will only be affected if they also inherit a mutated allele from the mother. Women who are carrier of a X-linked recessive disorder have a 50% chance of having affected sons and a 50% chance of having daughters who carry one copy of the mutated gene. X-linked recessive disorders include Duchenne muscular dystrophy, Hemophila A and red-green color blindness.

Y-linked and Mitochondrial disorders: A unique form of inheritance occurs if mutations causative for genetic disorders arise on the Y- or mitochondrial chromosome. These chromosomes are soley inherited from one parent, the Y-chromosome from the father, the mitochondrial chromosome from the mother. In the case of Y-linked disorders every son of an affected father will inherit the mutation and all daughters will never inherit the mutation. Since the Y chromosome is relatively small and contains very few genes, there are relatively few Y-linked disorders, examples being male infertility and hypertrichosis pinnae. The opposite case presents itself with mutations on the mitochondrial chromosome. Since only egg cells contribute mitochondria to the developing embryo, only mothers can pass on mitochondrial conditions to their children. An example of this type of disorder is Leber's hereditary optic neuropathy.

1.2 Identification of Causative Mutations in Genetic Disorders

The identification of causative variants and genes involved in the development of genetic disorders has been a major goal in genetic research since the discovery that genetics contribute significantly to human disease. Several strategies have been applied to directly localize the causally related genetic variants or identifying implied genes by mapping them in the genome based on a correlation of the disease phenotype with naturally occurring DNA variants, termed markers [3]. The most widely employed successful approaches include candidate gene studies, linkage analysis and genome wide association studies (GWAS).

Candidate Gene Approach: Candidate gene studies are typically hypothesis driven approaches utilizing prior biological knowledge about the

examined genes. A single gene or several genes that may be involved in a particular disorder due to their specific function or location in a certain pathway known to be impaired in the specific disorder. These genes are then sequenced in a group of affected individuals. If one or more putatively damaging variants are identified the sequencing is extended to a larger control group to prove the absence of these variants in unaffected individuals or show a statistically significant difference in variant frequency between the two groups. This kind of genetic studies will focus on a limited number of variants in a small number of genes due to the cost and labor involved in sequencing.

Linkage Analysis: Genetic linkage is the term used to describe the tendency of loci or alleles to be inherited together. Alleles that are located on the same chromosome tend to stay together during meiosis and are thus genetically linked. Genetic linkage is more likely the closer the loci are located on the chromosome. Chromosomal recombination (i.e. crossing over) during meiosis may prevent alleles originally located on the same chromosome from being permanently linked. Linkage analysis is mainly employed to narrow down the region in which causative mutations reside in pedigrees with heritable genetic disorders. A so called linkage map is computed to determine all chromosomal regions that are shared between multiple affected patients. The LOD score (logarithm (base 10) of odds), developed by Newton E. Morton, is a statistical test often used for linkage analysis in human. The LOD score compares the likelihood of obtaining the test data if the two loci are indeed linked, to the likelihood of observing the same data purely by chance. Positive LOD scores indicate the presence of genetic linkage. LOD score analysis is a simple way to analyze complex family pedigrees in order to determine the linkage between inherited traits (or between a trait and a genetic marker). Linkage analysis is then subsequently combined with sequencing of candidate genes from the linkage regions in order to identify causative mutations in the studied pedigree.

Genome Wide Association Studies: A genome wide association study is an examination of many common genetic variants in different individuals to determine wether any of these variants are associated with a genetic trait [61, 41, 97]. GWAS typically focus on associations of genetic disorders with single-nucleotide polymorphisms (SNPs) as follows: In order to determine association with a genetic disorder DNA markers from two groups of partici-

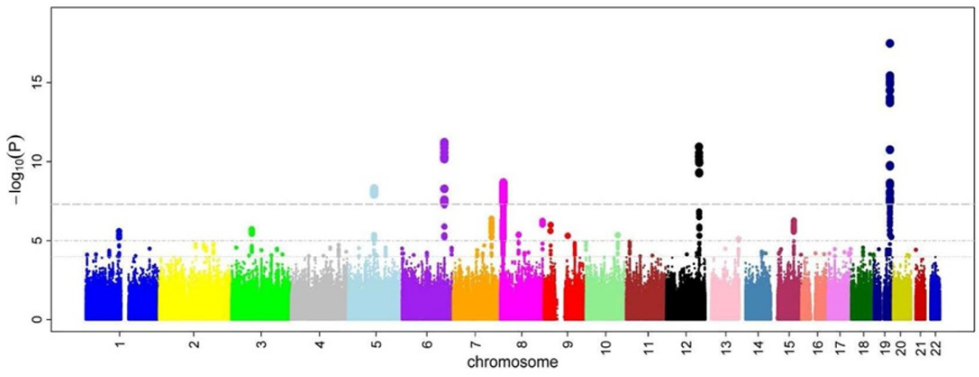


Figure 1.1: Manhatten Plot: An illustration of a Manhattan plot depicting several strongly associated risk loci. Each dot represents a SNP, with the X-axis showing genomic location and Y-axis showing association level. Example taken from Ikram *et al.* 2010 [41].

pants are compared, people with the disease (cases) and similar people from the same ethnological background without the disorder (controls). Millions of genetic variants are tested typically using high-density SNP arrays. If a variant (one allele) is more frequent in people with the disease, the SNP is said to be 'associated' with the disease. The probabilzy of association is depicted in a so called Manhatten plot (example shwon in Figure 1.1). The associated SNPs are then considered to mark a region of the human genome which influences the risk of disease. In contrast to methods which specifically test one or a few genetic regions, the GWAS investigates the entire genome. The approach is therefore said to be non-candidate-driven in contrast to candidate gene studies. GWAS identify SNPs and other variants in DNA which are associated with a disease, but cannot on their own specify which genes are causal.

1.3 Common and Rare Variants in Genetic Disorders

The underlying rationale for GWAS is the 'common disease, common variant hypothesis' [80]. The hypothesis postulates that in all populations which manifest a give common polygenic disease there are common variants that

confer susceptibility to this disorder. The individual risk convered will be very small, but a potentially large number of common variants with additive or multiplicative effects may explain the heritability of the disease. Many associated loci for common disorders and traits in general have been identified through successful GWAS, yet the majority of heritability of these traits and disorders remains unexplained. For example, human height is a complex trait with an estimated heritability of 80%. Numerous GWAS involving tens of thousands of participants have been conducted and identified over 40 height-associated loci. However, altogether they explain barely 5% of phenotypic variance, leaving the question how the remainder of heritabilty can be explained. There are several possible explanation currently discussed [62, 50]:

1. Much larger number undiscovered variants with even smaller effect sizes.

2. Structural variants currently poorly detected by SNP arrays employed in GWAS.

3. Gene-gene interactions that are difficult to detect using the GWAS approach.

4. Rarer variants with potentially larger effect sizes that are undetected by genotyping arrays which focus on variants with a population frequency of more than 5%.

The main focus of discussion has been the last point, the contribution of rare ($0.5\% < \text{MAF} < 5\%$) and very rare ($\text{MAF} < 0.5\%$) variants. These variants are not sufficiently captured by genotyping arrarys nor do they amount sufficiently large effect sizes for detection with linkage analysis. However, the moderate effect sizes of multiple rare variants combined can account for a large amount of the yet missing heritability. This up to date elusive point can now be approached with the emerge of next-generation sequencing as a new technology in variant discovery. To add to this point a recent study has identified an excess of low frequency non-synonymous variants in the re-sequencing of 200 human exomes from Denmark [58]. These low-frequency non-synonymous variants are prime subjects for causative mutations in rare and common genetic disorder. The goal of this work was to implement a computational analysis pipeline to identify causative mutations from a wide

spectrum of genetic disorders in next generation sequencing data in general and exome sequencing data in particular. Figure 1.2 shows the frequency spectrum of variants in genetic disorders depending on the expected effect sizes. Next generation sequencing allows to mine previously unaccessible areas from this spectrum, as is demonstrated on four distinct projects. The first project, Identification of recurring tumor-specific somatic mutations in acute myeloid leukemia by transcriptome sequencing, serves as prove of principle showing that the accurate identification of *somatic* mutations is possible by next-generation sequencing. Most parameters for alignment, variant calling and filtering that were later incoporated in the analysis pipeline implementation were evaluated on this project. The second project, Identification of mutations in Adaptor Protein Complex 4 proteins as cause of Intellectual Disability, is an example for the identification of a causative mutation in a monogenic disorder where conventional analysis methods (linkage in this case) failed. The third project, Identification of a mutation in VPS35 causing late-onset Parkinson, shows that by employing this new technology the identification of causative mutations in a complex and common disorder like Parkinson's disease becomes feasible. Many different genetic and epigentic factors can contribute to the phenotype of complex disorder, but there are always a certain percentage of familial cases as opposed to sporadic incidence. The project shows that these familial cases can be effectively mined with this appraoch. The last project, Dysfunctional nitric oxide signaling increases risk of myocardial infarction, goes one step further by identifying strongly disease associated variants in myocardial infarction. The approximate location of variants identified in each of the projects in the frequency spectrum of mutations is depicted in Figure 1.2.

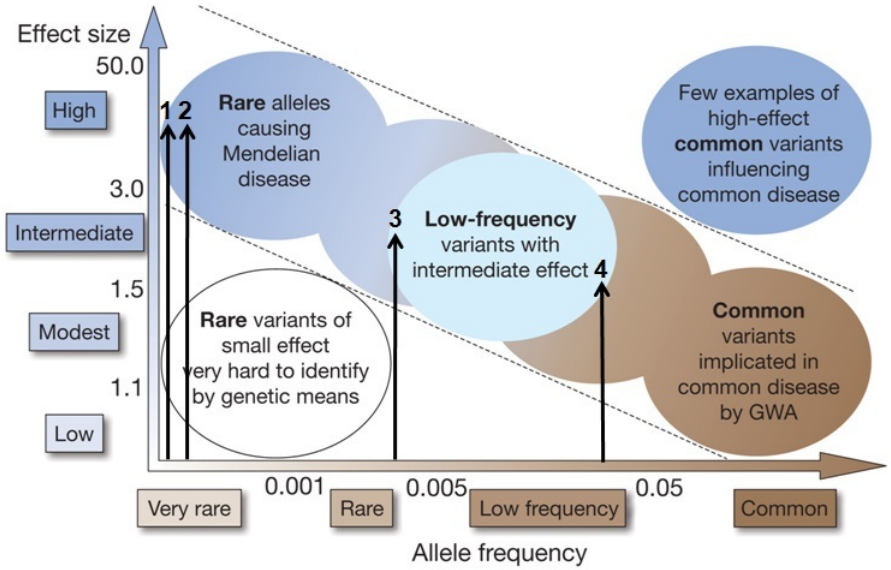


Figure 1.2: Spectrum of Genetic Disorders: Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect. Arrows denote projects approached in this work and their approximate placement in the spectrum of genetic disorders. Rare alleles: Identification of recurring tumor-specific somatic mutations in acute myeloid leukemia by transcriptome sequencing (1) and Identification of mutations in Adaptor Protein Complex 4 proteins as cause of Intellectual Disability (2). Low-frequency variants: Identification of a mutation in VPS35 causing late-onset Parkinson (3) and Dysfunctional nitric oxide signaling increases risk of myocardial infarction (4). Figure adapted from Manolia *et al.* [62].

1.4 DNA Sequencing

The modern basis of genetic research was laid by Watson and Crick in 1953 with the elucidation of the molecular structure of Deoxyribonucleic acid (DNA) [100, 101]. Yet, it took further 20 years until methods were developed for DNA sequencing that could actually determine the order of the nucleotides adenosin, thymin, cytosin and guanine in the DNA, opening a plethora of new opportunities and research fileds. Two different methods of DNA sequencing were developed almost in parallel.

In 1977, Allan Maxam and Walter Gilbert developed a DNA sequencing

method based on chemical modification of DNA and subsequent cleavage at specific bases [64]. In brief, the method is based on radioactive labeling of the DNA at the 5' end. Afterwards, the to be sequenced DNA fragment is purified. Chemical treatment introduces breaks at a small fraction of one or two of the four nucleotide bases in each of four reactions (G, A+G, C, C+T). Subsequently, a series of labeled fragments is generated, from the radiolabeled end to the first 'cut' site in each molecule. These fragments can then be size-separated side by side by electrophoresis in denaturing acrylamide gels. In order to visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each corresponding to a radiolabeled DNA fragment. The actual nucleotide sequence can then be inferred from these bands. This method is also known as 'chemical sequencing'. Although Maxam and Gilbert published their chemical sequencing method two years after the ground-breaking paper of Sanger and Coulson on plus-minus sequencing, MaxamGilbert sequencing rapidly became more popular, since purified DNA could be used directly, while the initial Sanger method required that each read start is cloned for production of single-stranded DNA. However, with the improvement of the chain-termination method (see below), Maxam-Gilbert sequencing has fallen out of favour due to its technical complexity prohibiting its use in standard molecular biology kits, extensive use of hazardous chemicals, and difficulties with scale-up.

The second technique to sequence DNA was invented by Frederick Sanger and colleagues in 1975 [87]. It became known after it's inventor as Sanger sequencing or chain terminator method. The key component of Sanger sequencing are dideoxynucleotide triphosphates (ddNTPs), which are employed as chain terminators. Whenever a DNA polymerase incoporates one of the ddNTPs into a growing strand of DNA, the DNA strand is effectively blocked and further elongation is prevented due to the ddNTPs lacking a 3'-OH group, which is neccessary for the formation of a phosphodiester bond between two nucleotides. The method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, normal deoxynucleotidetriphosphates (dNTPs), and modified ddNTPs, which additionally carry either a radioacive or flourescent label. Radioactive labels were used at the start of the techinique, but were later replaced by flourescently labeled groups. For the sequencing reaction, the DNA sample is first divided into four distinct sequencing reactions. Each reaction setup contains all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase. In contrast, only one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP) is added

to each reaction. The DNA polymerase randomly incorporates ddNTPs instead of normal dNTPs, stopping further elongation of the strand and thus resulting in DNA fragments of varying length. The newly synthesized and labelled DNA fragments are then heat denatured and separated according to size by gel electrophoresis on a denaturing polyacrylamide-urea gel with the resolution of one nucleotide. Each of the four reactions is run in one of four individual lanes (one for each nurcleotide). The resulting DNA bands are then visualized by autoradiography or UV light, allowing the DNA sequence to be directly read off the X-ray film or gel image. This techinque was later refined for automated sequencing machines [92]. As already previously mentioned, the radioactive labels got replaced by flourescent labels, suitable for automated signal detection using an optics system. The sequencing instruments employ capillary gel electrophoresis for the size selection and detection of the DNA fragments. The process can be parallelized to 384 DNA samples in a single run. After more than three decades of gradual improvement and optimizing, the Sanger biochemistry can be applied to achieve read-lengths of up to ∼1,000 bp, and per-base accuracies as high as 99.999%. In the context of highthroughput shotgun genomic sequencing, Sanger sequencing costs in the order of 0.50$ per kilobase.

During the last three years the field of DNA sequencing changed dramatically with the advent of 'next-generation' sequencing methods. These methods, in contrast to traditional Sanger sequencing, rely on the massively parallel sequencing of very short DNA fragments [91]. Reaching an unpredecessed throughput the cost of DNA sequencing decreased by several orders of magnitude (Figure 1.3). This decrease in sequencing cost is even more remarkable when translated into a cost for sequencing a complete human genome. When the first human genome was published in 2002 the project took about 11 years from the start of sequencing in 1990 to the first working draft sequence [21]. The total cost amounted to approximately over $100,000,000. In contrast, using next generation sequencing instruments it is possible to determine the nucleotide order of a complete human genome for about $10,000 (Figure 1.4). The decrease in sequencing cost is so rapid that it even surpasses Moore's law, enabling new venues of research previously not thought possible. This includes large scale endeavors like the 1000 Genomes Project [17] which aims to detect rare polymorphisms in the frequency range of 0.5-5% by population scale sequencing, or the International Cancer Genome project aiming to determine the sequence of 50 different tumor types by sequencing 500 samples of every tumor, totaling over 25,000 complete human genomes.

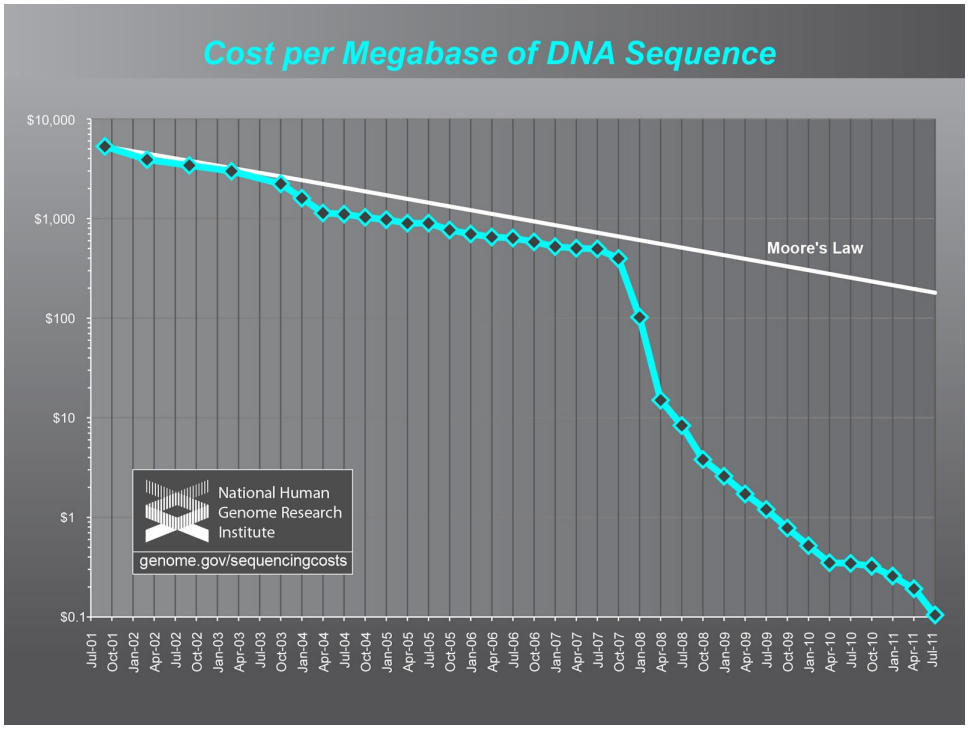


Figure 1.3: Development of DNA sequencing cost per megabase. Figure available at NIH (http://www.genome.gov/sequencingcosts/)

Advantages of second-generation or next-generation sequencing, in contrast to Sanger sequencing, include the following:

1. In vitro construction and amplification of a sequencing library, circumventing previous bottlenecks that restrict the parallelism of Sanger sequencing (i.e. transformation of E. coli and colony picking)

2. Array-based sequencing, enabling a much higher degree of parallelism than capillary-based sequencing by omiting gel electrophoresis steps

3. Significantly lower cost in sequence production due to the high parallelism

Although second-generation sequencing platforms offer several major improvements there are currently certain disadvantages. The most prominent of these include read-length, which is with reads of length 36-100 bp drasctically lower than Sanger sequencing. Secondly raw base accuracy is about

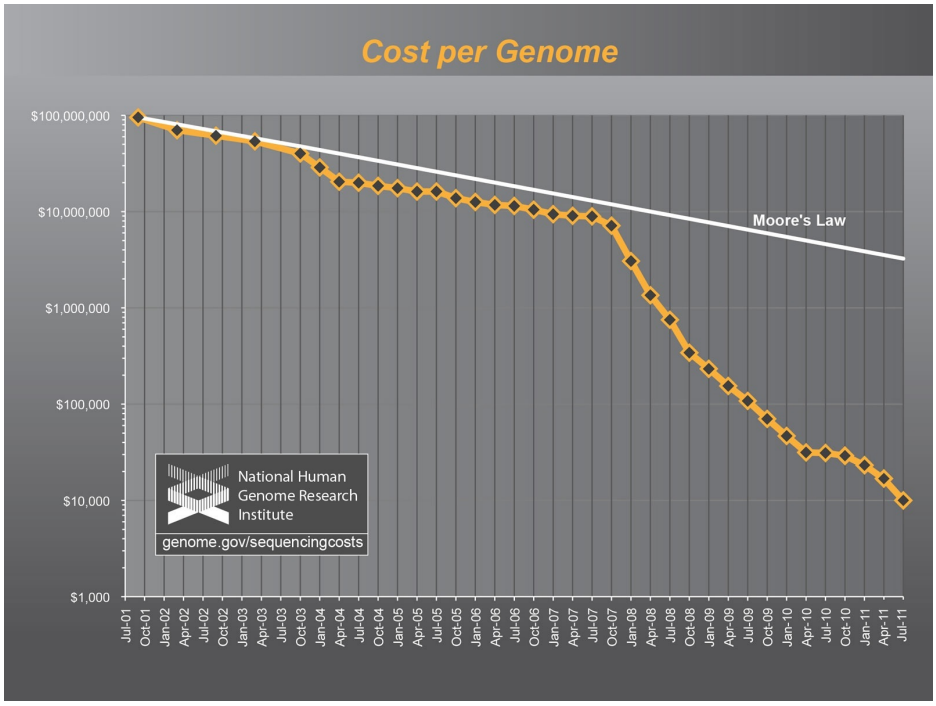


Figure 1.4: Decrease in sequencing cost for a complete human genome. Different coverage levels are assumed for a high quality human genome (size for calculation 3 GB): 6-fold coverage for Sanger-based sequencing (average read length 500-600 bp), 10-fold coverage for 454 sequencing (average read length 300-400 bp) and 30-fold coverage Illumina and SOLiD sequencing (average read length 50-100 bp). Figure available at NIH (http://www.genome.gov/sequencingcosts/).

ten-fold lower than base-calls generated by Sanger sequencing, although this disadvantage can be alleviated by redundant sequencing, as each base is sequenced multiple times and a consensus genotype call is then generated.

1.5 Next-Generation Sequencing

Initially, the term 'next-generation sequencing' arouse to describe the first sequencing technologies that took alternative approaches to Sanger sequencing to determine the DNA nucleotide order. Three distinct platforms where introduced in a relativley short time span from 2005-2007. The first technology

commercially available was the Roche/454 instrument (introduced in 2005) [102]. It was closely followed by two further sequencing instruments, the Illumina/Solexa technology (2006) [6] and the SOLiD System manufactured by LifeTechnologies (former ABI, 2007) [65]. While all of these platforms are commonly summarized as second-generation sequencing machines, each employs a distinct and unique sequencing approach. In this work the Illumina/Solexa technology was employed and the sequencing and data analysis workflow of this system will be discussed in detail in the following, while briefly introducing the defining characteristics of the remaining two technologies.

1.6 Roche/454 Sequencing

454 Life Sciences was founded by Jonathan Rothberg originally as 454 Corporation. The company has experienced rapid growth since its acquisition by Roche Diagnostics and release of the GS20 sequencing machine in 2005. 454 sequencing employs a large scale pyrosequencing approach that is capable of sequencing 400-600 megabases of DNA in the course of a 10 hour instrument run, while achieving an average read length of 300-400 bp. In 2008, 454 Sequencing launched the GS FLX Titanium series reagents for use on the Genome Sequencer FLX instrument, improving both in read length (400-500 bp) and overall throughput. The key steps of library preparation and sequencing are now briefly presented [102, 85].

Library preparation and emPCR: The to be sequenced genomic DNA is first fractioned to fragments of length 300-800 bp. Specific adaptors that provide priming sequences for both amplification and sequencing are then ligated to the fragments. One adaptor contains a 5'-biotin tag enabling immobilization of the DNA library onto streptavidin-coated beads. The non-biotinylated strand is released and used as a single-stranded template DNA library. Following a quality assessment of the library, the optimal amount (DNA copies per bead) needed for emulsion PCR (emPCR) is determined by titration. In the next step the library is immobilized onto beads and the bead-bound library is then emulsified with the amplification reagents in a water-in-oil mixture. Each bead is captured within its own microreactor where PCR amplification occurs, resulting in bead-immobilized, clonally amplified DNA fragments (Figure 1.5).

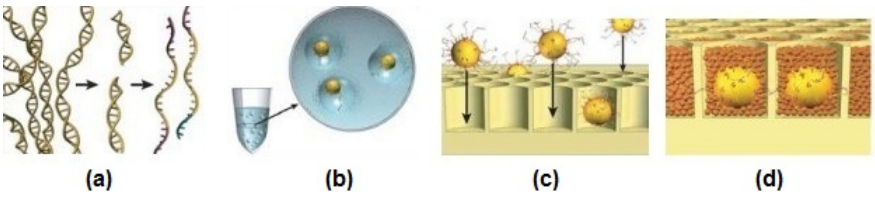


Figure 1.5: 454 Sequencing. (a) Genomic DNA is isolated, fragmented, ligated to adapters and separated into single strands. (b) Fragments are bound to beads under conditions that favor one fragment per bead. (c) Beads carrying single-stranded DNA templates deposited into wells of a PicoTiterPlate. (d) Smaller beads carrying immobilized enzymes are deposited into each well. Figure available at http://www.roche.com.

Sequencing: Prior to sequencing the library beads are distributed onto a PicoTiterPlate, each library bead positioned in a specific well on the plate. The loaded PicoTiterPlate is placed into the Genome Sequencer FLX Instrument for sequencing. The fluidics system delivers sequencing reagents (containing buffers and nucleotides) across the wells of the plate. The four DNA nucleotides are added sequentially in a fixed order over the course of a sequencing run. During the nucleotide flow, millions of copies of DNA bound to each of the beads are sequenced in parallel. When a nucleotide complementary to the template strand is added into a well, the polymerase extends the existing DNA strand by adding the complementary nucleotide(s). Addition of one (or more) nucleotide(s) generates a light signal that is recorded by the optics system in the instrument. The signal strength is proportional to the number of nucleotides incorporated. If multiple identical nucleotides occur consecutively in the DNA template, i.e. homopolymer stretches, a greater signal is emited. However, the signal strength for homopolymer stretches is not linear, resulting in difficulties to determine the exact length of a homopolymer.

1.7 SOLiD - Sequencing by Ligation

SOLiD (Sequencing by Oligonucleotide Ligation and Detection) is a next-generation sequencing technology developed by Life Technologies (fromerly Apllied Biosystems Inc., ABI) and has been commercially available since

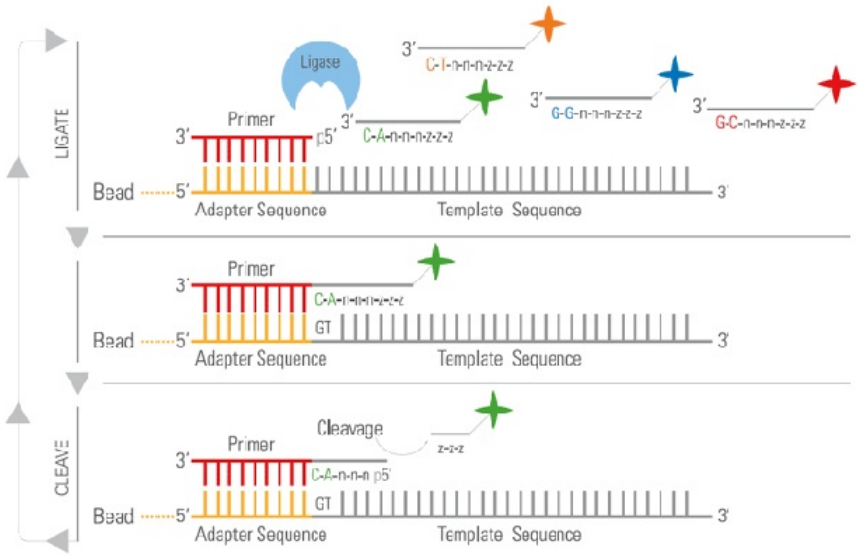


Figure 1.6: SOLiD sequencing. Schematic overview of steps involved in SOLiD sequencing: Addition of flourescently labeled di-nucleotides, ligation of complementary probe and cleavage of the flourescent label for color detection. Figure available at www.appliedbiosystems.com.

2008. The SOLiD technology is a short read sequencing system initially capable of generating reads of length 35 bp and achieving a throughput of 3 GB per x-day run. In comparison to other sequencing approaches, SOLiD sequencing bears some specific characteristics that will be briefly highlighted in the following [65].

Library preparation: A library of DNA fragments is prepared from the sample to be sequenced, and is used to generate clonal bead populations. These fragments are bound to magnetic beads such that only a single species of fragment is present at each bead. In a comparable fashion to 454 sequencing emPCR is employed to amplify the fragments. The resulting PCR products attached to the beads are then covalently bound to a glass slide. Each fragment contains a known adaptor sequence that is used to initiate the sequencing reaction.

Sequencing-by-ligation: In the sequencing reaction a set of four fluorescently labeled di-base probes compete for ligation to the sequencing primer

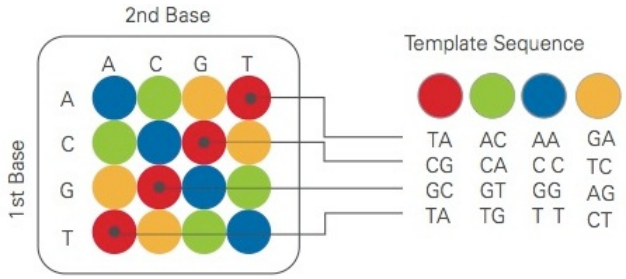


Figure 1.7: SOLiD colorspace encoding. Table of color space encoding of the 16 possible dinucleotides and colors grouped by template sequences that generate the same color during sequencing. Figure available at www.appliedbiosystems.com.

(Figure 1.6). Specificity of the di-base probe is achieved by interrogating every 1st and 2nd base in each ligation reaction. Multiple cycles of ligation, detection and cleavage are performed with the number of cycles determining the eventual read length. As only the first two nucleotides are interrogated at each cycle it is necessary to remove the extension product and to reset the template with a primer complementary to the n-1 position for a second round of ligation cycles. Five rounds of primer reset are completed for each sequence tag. Through the primer reset process, each base is interrogated in two independent ligation reactions by two different primers. If a probe is complementary to the interrogated di-nucleotide a ligation reaction takes place during which course the flourescent label is excited and detected via an optics system. The color system is redundant as four distinct colors are used to detect the 16 different dinucleotide combinations. This lead to the term 'color-space sequencing' and even though the system may not be as intuitive as traditional nucleotide space sequencing it has certain advantages that may be exploited during data analysis.

Color space: A complete table of all dinucleotides and their respective color encoding can be seen in Figure 1.7. It is not advisable to convert the color space reads to nucleotide space for alignment to a reference sequence as a single error in color space (i.e. the detection of a 'wrong' color) leads to a frameshift in the corresponding nucleotide sequence. Therefor the reference sequence is converted to color space and the alignment is performed by matching color spaced reads to color spaced reference sequence. In doing

so specific color space featured may be exploited during alignment: As each base is interrogated twice during sequencing a true single nucleotide polymorphism is always indicated by two adjacent color changes. This enables true SNVs to be distiguished from sequencing errors which only show an isolated, single color change in comparison to the reference sequence. More specifically, this method is not an error correction tool but an error transformation tool. Color-space transforms your most common error mode (single measurement errors) into a different frequency than your most common form of DNA variation (single base changes). On the other hand, color space sequencing has some sgnificant disadvantages depending on the application. For example when performing de novo assembly one is left with the raw instrument error rate as errors may not be corrected in the way described above without a reference sequence.

1.8 Solexa Technology

The company Illumina was founded in April 1998 and started offering its first SNP genotyping services in 2001. The first system, the Illumina BeadLab based on GoldenGate Technology was launched 2002. Illumina entered the Next-generation sequencing sector in 2006 by acquiring the company Solexa, which developed the first 1G Genome Analyzer. This instrument was the first to sequence over 1 GB in a single run. The system matured in 2007 with the introduction of 'paired-end' sequencing, allowing to sequence short DNA fragments from both ends. This sequencing method was termed 'sequencing-by-synthesis' and is described in detail in the following paragraphs. .

1.8.1 Sequencing-by-Synthesis

The Illumina 'sequencing by synthesis' approach uses reversible terminator chemistry to sequence billions of short, fragmented DNA molecules covalently attached to a glass slide, the so-called flow cell [6]. The flow cell acts as reaction chamber and has eight distinct compartements, referred to as lanes. The sequencing is a stepwise process which subsequently alternates between incorporation of a single nucleotide, complementary to the base in the DNA template, and the detection of the incorporated base, the imaging. Sequencing is performed cycle-wise, where the cleavage of the terminator group marks the end of the cycle. Prior to the sequencing the DNA sample

has to be prepared for sequencing in the so-called library preparation. All steps involved in the library preparation and sequencing using the Illumina technology are now further discussed.

Library preparation: In this step the sample, which is to be sequenced, typically genomic DNA, is prepared for the sequencing with the Illumina technology. In the following the standard sample preparation for genomic DNA is discussed. The library preparation for transcriptome sequencing, chromatin immuno-precipitation followed by sequencing (ChIP-Seq) and targeted enrichment of specific regions of the genome followed by sequencing all have modified library preparation procedure (for exome sequencing see Section 1.4). The processing of the sample DNA begins with the fragmentation of the DNA into smaller fragments. The DNA may be sheared into the desired fragments using different approaches, for example nebulization or sonic waves. Fragmentation of all libraries used in this work was performed with the sonic wave based Covaris technology, either employing the single tube S2 system or the E210 system, which enables the processing of 24 samples in parallel. The shearing process is followed by an end repair step of the fragments. A single Adenosin is added to the 3' end of the fragments to enable the ligation of specific adapters, which have a single Thymin overhang. The adapters are required later to attach the fragments to the flow cell (Figure 1.8a). Since the shearing generates fragments from a wide size distribution, the fragments of the desired length of 350-450 bp have to be selected by gel extraction. The quality of the finished library is reviewed with a 2100 Agilent Bioanalyzer. The following steps of attachment of the fragments to the flow cell surface and cluster generation are carried out on the Illumina Cbot.

Attachment to flow cell: A small amount (1-5 mug) of the library is inserted in one or multiple lanes of the flow cell. The ligated adapters will bind to complementary sequences, which are covalently attached to the flow cell surface. The adapters are then used as sequencing primers to synthesize a complementary DNA strand at the flow cell adapters, in order to covalently bind the DNA fragments to the flow cell surface. The hybridized strands are then removed in a washing step, leaving single stranded, covalently bound DNA fragments in the flow cell (Figure 1.8b).

Cluster generation: Since the detection of a nucleotide incorporation event

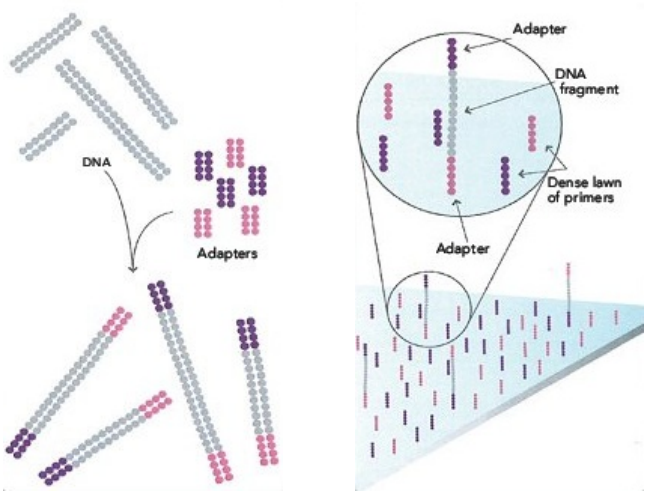


Figure 1.8: Library preparation: Ligation of adaptor sequences (a) and attachement of DNA fragments to flow cell (b). Figure available at www.illumina.com.

in a single molecule is not possible with the Illumina technology, the last step prior to sequencing is the generation of millions of identical molecules, i.e. clusters, that are densely packed at the flow cell surface. Cluster generation is facilitated through so-called bridge amplification: By increasing the temperature from 4C to 40C the free adapter of the single stranded DNA fragment hybridizes to the complementary oligonucleotide on the flow cell surface, forming the characteristic 'bridge' (Figure 1.9a). This single stranded DNA bridge is then processed to double stranded DNA via addition of a polymerase and free dNTPs. The original DNA fragment now forms a double stranded bridge (Figure 1.9b), which is then denaturated using a thermal step. This results in the, again, single stranded original DNA fragment and its complementary strand being covalently attached to the flow cell (Figure 1.10a). This process, in effect a modified PCR reaction, is then repeated to build dense clusters of identical molecules (and their reverse complements) that can now be sequenced (Figure 1.10b).

Sequencing by synthesis: Following the cluster generation step, the flow cell is transferred to the sequencing instrument. The first step of the actual sequencing process is the linearization of clusters: The reverse complement of the original DNA fragments gets enzymatically cleaved at the adapter

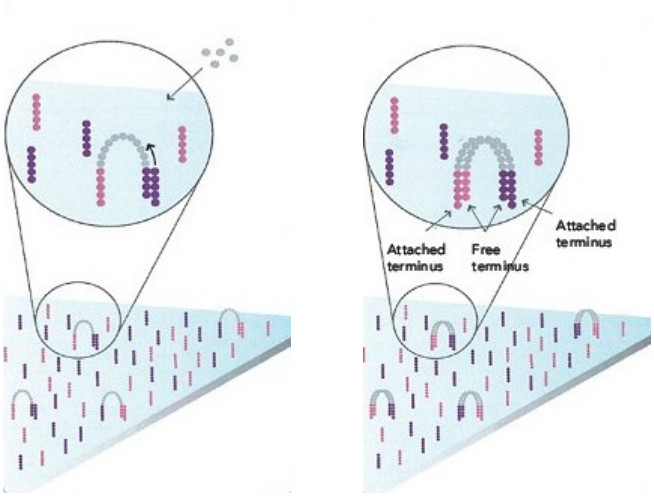


Figure 1.9: Bridge Amplification: Formation of single stranded (a) and double stranded DNA bridges (b). Figure available at www.illumina.com.

sequences to generate clusters that are now in fact composed of identical nucleotides only. After the linearization, modified nucleotides and sequencing primers are washed through the flow cell (Figure 1.11a). The nucleotides carry both a flourescent label and a reversible terminator group that ensures that only a single nucleotide gets incorporated during each sequencing cycle (Figure 1.11b). Subsequently the flourescent lable is excised by a laser and the signal is detected during the imaging step (Figure 1.11c). The images can then be used to determine the DNA sequence in the base-calling step. The sequencing is performed cycle-wise up to a predefined sequencing length. The number of possible bases that can be accurately determined started at 36 bp, but subsequently increased to a read length of up to 150 bp through improvements in the underlying chemistry and reagents.

Paired-End Sequencing: A particular feature of the Illumina instruments is the capability of paired-end sequencing. The 350-450 bp long DNA fragments are hereby sequenced from both ends with the desired read length of up to 100 bp. The two reads originating from the ends of one DNA fragment are called mates and are seperated by an unsequenced stretch of DNA, termed insert, of size 150-200 bp. Paired-end sequencing has advantages in further downstream analysis steps, in particular in the read mapping and variant

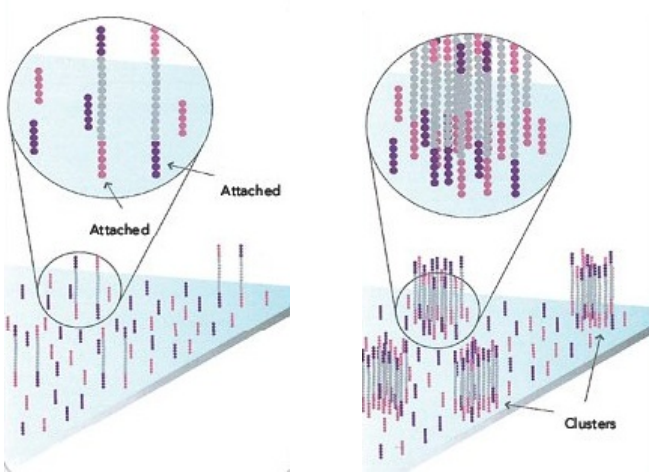


Figure 1.10: Cluster generation: Denaturation of DNA bridges (a) and amplification to densely packed clusters (b). Figure available at www.illumina.com.

calling [6, 56, 53]. Additionally, the detection of larger structural variation becomes feasible by exploiting paired-end sequencing properties [12]. The benefits of paired-end sequencing will be discussed in the respective chapters (1.3 Alignment algorithms and 2.2.5 Variant calling). Paired-end sequencing is facilitated through a cluster resynthesis after the completion of all sequencing cycles of the first read. The process is similar as described in this chapter, with the exception that this time the forward strands, instead of the reverse strands, are cleaved and washed away during the cluster linearization step. On the Illumina Genome Analyzer IIx this step was performed using an add-on machine to the sequencing instrument, the Illumina Paired-End Module. This circumvents the need to remove the flow cell from the instrument for use with the CBot. In the newer generation of sequencing machines, the HiSeq2000 series, the paired-end capability is directly integrated into the core instrument.

1.8.2 Illumina Data Analysis Pipeline

The primary Illumina data analysis pipeline consists of three key steps:

1. Image Analysis, performed by the Firecrest module
2. Base Calling, performed by the Bustard module

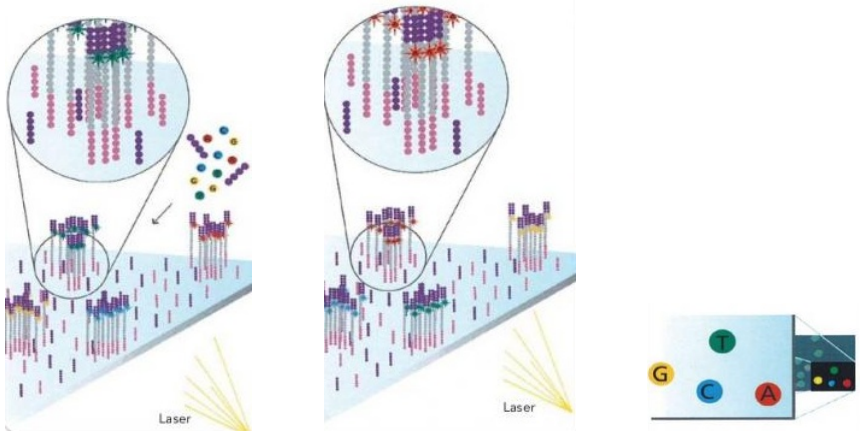


Figure 1.11: Sequencing by synthesis: Incorporation of complentary, flourescently labeled nucleotides into cluster (a), signal detection by laser excition of lables (b and c). Figure available at www.illumina.com.

3. Alignment to the reference sequence using the ELAND module

Image Analysis: The first step of the Illumina analysis pipeline is the image analysis, the extraction of actual intensitiy levels from the recorded images. This step has two main difficulties: Overlapping clusters and signal decay due to phasing. If the flowcell is too densely loaded with sample DNA, the clusters are in very close proximity and the image analysis algorithm has difficulties distinguishing between the single clusters. This leads to the problem that some intensity signals are mixtures of different clusters with overlapping borders. The second key problem is phasing, or rather de-phasing of clusters. With growing number of sequencing cycles some molecules of the cluster get out of phase, meaning that the nucleotide incorporation fails or that more than one nucleotide gets incorporated in a single cycle. These molecules are then out of phase with the majority of the cluster for the remaining sequencing cycles and give a 'wrong' intensity signal during each further incorporation. Since the clusters are composed of thousands of identical molecules, the majority of them still yield the correct signal. While the problem is negligible during the first cycles it aggravates with growing number of sequencing cycles, resulting in the phenomenon that the base quality steadily decays during a sequencing run, thus limiting the maximum number of sequencing cycles that may be performed. Although the maximum read length steadily increased due to improvements in the reagents and underlying

chemistry, the core probelm of phasing yet still remains in Illumina sequencing.

Base Calling: The process of converting the intensity levels to an actual nucleotide sequence is referred to as base calling. Due to phasing problems or impure signals it is possible that a base gets miscalled or not called at all (indicated by the letter N). Thus, the base calling algorithm assigns a PHRED scaled sore [30, 31] to measure the probability that the base is called wrong. The PHRED score is defined as:

$$score = \frac{-10 \cdot log(errorprobability)}{log(10)}$$

A quality value Q10 for example means one error in 10 bases, while Q20 indicates one error in 100 bases, and so forth. This score is a key quality criteria to estimate the success of the run, currently only runs exceeding 80% Q30 bases are considered a success. Previously Illumina used a different, custom definition for the quality values which was only asymptotically equal to the PHRED scores, but since the latest software update the standard PHRED scores are employed.

Read Alignment: The read alignment to a reference sequence is performed by the ELAND algorithm [6], a hash table based alignment program. Different types of alignment algorithms and their respective strengths are discussed in Section 1.3.1.1.

1.8.3 Genome Analyzer IIx

The Illumina Genome Analyzer IIx (GA) was the first widely adopted next-generation sequencing platform, capable of sequencing up to 3 GB in one intrument run using 36 bp paired-end reads [6] (Figure 1.12). The throughput of the system gradually scaled up with the introduction of new sequencing reagents and improvements in the maximum read length. The read length was stepwise increased from 36 bp to 50 bp and 76 bp, ultimately reaching 100 bp paired-end reads in the end of 2009 for a final throughput of up to 50 GB per run.

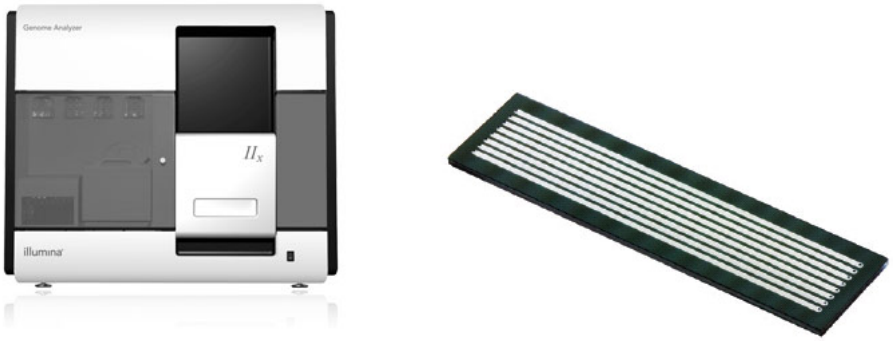


Figure 1.12: Illumina GAIIx sequencing instrument (a) and flowcell (b). Figure available at www.illumina.com.

1.8.4 HiSeq2000

While based on the same underlying chemistry, the HiSeq2000 system provides several advancements in comparison to its predecessor, the GA IIx. The first major improvement is the possibility of sequencing two flow cells in parallel. The two flow cells are independent, and the sequencing of the two flow cells need not necessarily be started together, enabling a flexible run handling. Additionally, the HiSeq2000 employs a new generation of flow cells where both the top and the bottom surface may be populated with clusters, effectively doubling the amount of clusters per flow cell. In conjunction with a maximum read length of 150 bp this results in a throughput of up to 300 GB per flowcell, or 600 HB per complete instrument run. The HiSeq2000 retains the paired-end capability without the need for a separate paired-end module, integrating this functionality in the core machine.

Table 1.1: Sequencing Throughput Development of the Illumina Platform

Instrument	Date	Read Length (BP)	Throuput per flow-cell (GB)
GAIIx	Jun 2008	36	3
GAIIx	Nov 2010	76	50
HiSeq2000	Dec 2010	76	100
HiSeq2000	Feb 2011	100	150
HiSeq2000	Jun 2011	100	300

1.9 Short-read Alignment Algorithms

A key point in the analysis of next-generation sequencing data is the sequence alignment, where millions of short reads must by aligned to a chosen reference sequence in reasonable time. A variety of different alignment algorithms and strategies have been subsequently implemented since the emerge of next-generation sequencing [55]. Due to the amount of data that is generated, next-generation sequencing aligners use auxiliary data structure, so-called indices, for the read or the reference sequences. Based on the properties of these indices alignment algorithms may be grouped into two major categories. The first group utilizes hash tables as key index structure while the second group focuses on suffix arrays. In particular they apply and exploit the Burrows-Wheeler string transformation for sequencing data. In the following the different key concepts of these aligners will be discussed and specific members of each group will be examined in Table 1.2.

Table 1.2: Sequence Alignment Algorithms

Program	Algorithm	Paired End	Gapped	Quality
ELAND	Hash based	Yes	Yes	No
MAQ	Hash based	Yes	Yes	Yes
Bowtie	BWT	No	No	No
BWA	BWT	Yes	Yes	Yes

1.9.1 Hash table based Algorithms

One of the first and most well known algorithms to employ hash tables for sequence alignment were BLAST and BLAT [2, 43]. In general all hash table based algorithms follow a seed-and-extend strategy by storing k-mer subsequences of the query in a hash table with the k-mer sequence as key. A database of sequences, i.e. the reference genome, is then scanned for k-mer exact matches, the seeds, by looking up the hash table. Speed and sensitivity is dependant on the choice of k, Blast for example uses k=11 as default parameter. The seed hits are then extended and refined using Smith-Waterman alignment [93, 37] and statistically significant hits are outputted. The basic idea of hash tables for sequence alignment has been subsequently improved and applied to short read sequencing.

One improvement was the implementation of spaced seeds, i.e. seeds allowing for internal mismatches. These non-consecutive seeds improve sensitivity [60]. For example, a template '111010010100110111' requiring 11 matches at the '1' positions is 55% more sensitive than BLASTs default template '11111111111' for two sequences of 70% similarity. The number of mismatches in the seed is called weight.

A drawback of both the consecutive and spaced seed strategies is that they disallow gaps within the seed. Since short insertion and deletion events, so called Indels, are a frequent type of genomic variation this poses a problem to read mapping if an Indel is in the seed region. Per default these types of algorithms identify gaps only during the extension step. A solution to this problem is the implementation of an index that allows gaps also within the seed, the so-called q-gram filter. The q-gram filter is based on the observation that at the occurrence of a w-long query string with at most k differences (mismatches and gaps in this case), the query and the w-long database substring share at least (w+1)-(k+1)q common substrings of length q. In contrast to spaced seed algorithms, which initiate extension from long seed matches, q-gram filter based algorithms typically initiate seed extension from multiple shorter seed matches, but both algorithm categories rely on fast lookups in a hash table [55].

Examples of hash-based algorithms are ELAND, the Illumina default alignment program [6] and MAQ, a popular open source aligment software [56] that was employed in this work. MAQ was later replacement by the BWA alignment software, yet many key concepts of aligning next-generation sequencing reads, like mapping uniqueness and mapping quality, were introduced by MAQ.

1.9.1.1 MAQ

The Mapping and Assembly with Quality software, MAQ, is the first short read aligner to adopt the concept of mapping quality. Similar to base quality score in Sanger [30, 31] and next generation sequencing, mapping quality is a measure of confidence that a read is actually derived from the position it is aligned to by the mapping algorithm. This concept adresses two main problems of short read mapping: First, mammalian genomes contain a multitude of repetitive or close to repetitive sequences on the length scale of the reads, so in consequence a read may map equally well to multiple genomic locations. Additionally, reads contain mismatches to the reference sequence, sequencing

errors or genuine mutations, which may lead to incorrectly mapped reads. A measure of mapping quality addresses these problems through allowing to keep all ambiguously mapped reads and evaluate for each read alignment the likelyhood that it has been wrongly mapped. Poor alignments may be discarded later through a minimum threshold of mapping quality, to ensure that low confidence alignments do not contribute to variant calling.

During alignment, MAQ determines the ungapped match with the lowest mismatch rate, defined as the sum of base quality values of all mismatched bases. With default parameters MAQ considers only the first 28 bases of a read, the most reliable part of Illumina reads. Gapped alignment is performed for reads which could not be mapped in the alignment stage but have a reliable mapped mate pair read. To speed up the mapping, gapped alignment is only performed for the genomic region defined by the mapped mate. MAQ then assigns a PHRED scaled score to each alignment that evaluates the probability that the alignment is wrong. If the case occurs that a read may be mapped equally well to multiple genomic locations MAQ chooses a position randomly and assigns the read a mapping quality of zero, thus always reporting a single alignment for each individual read. Reads with mapping quality of zero do not contribute to variant calling but are kept to provide valuable information about the repetitiveness of a genomic region.

The algorithm used to identify best hits is similar to ELAND. Per default, MAQ uses six hash tables, ensuring to find all alignments with up to two mismatches. For a 8 bp long read the six spaced seeds would for example consist of 11110000, 00001111, 11000011, 00111100, 11001100, and 00110011. During alignment, MAQ loads the reads into memory and applies the hash function to transform all nucleotides at '1' positions of the template into a 24 bit integer. The reads are then ordered based on the integers to group reads with similar sequences together. MAQ always indexes the reads with two complementary templates simultaneously. Each 28 bp subsequence of the reference is then scanned through the hash table to identify seed hits. If a hit is found the alignment is extended to the full length of the read and the sum of mismatching base qualities is calculated. MAQ stores the position of the best two alignments, while additionally recording the number of 0-, 1-, and 2-mismatch hits in the seed region. This process is repeated with the following two templates until all templates are scanned.

The main concept of calculating a mapping quality is to evaluate mapping uniqueness. There are several ways to define mapping uniqueness, one of the most widely adopted being the following: A read is uniquely mapped if its

best hit contains more mismatches than its second best hit [56]. In theory this simple definition works well, whereas in practice the case is not that simple. For instance consider the following two examples:

1. a read has two one-mismatch hits, one with the mismatched base having a quality of 30 and the other having a quality of 3

2. a read has one perfect hit and 100 one-mismatch hits

In the first case the read is, according to the aforementioned definition, not uniquely aligned, yet the Q30 mismatch may be reliable, representing a genuine variation. In contrast, in the latter case the read is uniquely mapped according to the defintion, although the alignment is highly unreliable. Thus MAQ uses mapping quality to asign a reliability to each read alignment that can be interpreted as the likelihood of the alignment being correct.

1.9.2 Suffix-Array-based Algorithms

Suffix-Array-based algorithms distinguish themselves from the former group through the characteristic of building an index of the reference genome using a certain representation or data structure, such as suffix tree, enhanced suffix array and FM-index. The inexact matching problem is hereby first reduced to the exact matching problem, while inexact alignments are generated from supporting exact matches. The main advantage of using a trie is that alignment to multiple identical copies of a repeat in the reference genome is only needed to be performed once. The identical copies of the repeat collapse into a single path in the trie, whereas with a typical hash-based alignment tool, the alignment has to be computed for each copy seperately. Special attention has been drawn to the BurrowsWheeler transform and its use in string matching, which will be discussed in the following.

1.9.2.1 Burrows-Wheeler Transformation

The BurrowsWheeler transform (BWT, also called block-sorting compression), is an algorithm originally invented by Michael Burrows and David Wheeler in 1994. It was first employed in data compression techniques such as bzip2. When a character string gets transformed by the BWT, a permutation of the order of characters is created. If the original string had several substrings that occurred multiple times, the transformed string will have

several places where a single character is repeated multiple times in a row, which can be exploited for compression. The BWT is performed following three steps:

1. Generate all rotations of the character string
2. Sort the rotations in lexicographic order
3. Output the last column

An example BWT of the string *.ANANAS.* is giving in Table 1.3. A remarkable property of the BWT is that the process is reversible - the original string can be recreated from the last column data. The BWT is most successful if a single character or substring occurs multiple times, an attribute most suited for use with DNA sequences which consist only of a four letter alphabet. In a new generation of alignment algorithms [57, 53, 49] the BWT is used to index the reference genome in order to reduce memory requirements.

Table 1.3: BWT of string .ANANAS

input	rotations	sorting	output
.ANANAS.	.ANANAS.	ANANAS..	.NNAAA.S
	..ANANAS	ANAS..AN	
	S..ANANA	AS..ANAN	
	AS..ANAN	NANAS..A	
	NAS..ANA	NAS..ANA	
	ANAS..AN	S..ANANA	
	NANAS..A	.ANANAS.	
	ANANAS..	..ANANAS	

1.9.2.2 BWA

The read alignment program BWA was developed by Heng Li at the Sanger Center as substitute for his prior hash-based alignment tool MAQ. Since the output of the sequencing instuments steadily increased the speed of MAQ could not scale up to the need as new large scale sequencing programs, like the 1000 Genomes Project [17], are conducted. Additionally, the read length also increased and MAQ does not support gapped alignment for single reads,

rendering it unsuitable for the alignment of longer reads where indels occur more frequently. BWA is based on backwards search using the BWT. The BWT is used to index the genome, effectively collapsing all repeat regions into a single path in the prefix trie. In consequence reads do not have to be aligned to each copy of the repeat seperately, greatly reducing alignment time. Evaluations on both simulated and real data suggest that BWA is 10 to 20 times faster than MAQ while achieving similar accuracy. Additionally BWA outputs the aligned sequences in the standard SAM format.

1.9.2.3 SAMtools

SAM (Sequence Alignment/Map) is a generic format for storing large DNA sequence alignments [54]. SAMtools is a collection of programs developed complementary to BWA to provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a standard format. To circumvent the caveats that result from sequence alignments being stored in many different formats SAM was developed with key features:

1. Generic format to store all alignment information generated by various alignment programs
2. Simple format to be easily generated by alignment programs or converted to from existing formats
3. Compact in size, achieved by the binary SAM format (BAM)
4. Allowing most operations on the alignment to work on a stream without loading the whole alignment into memory
5. Allowing to be indexed by genomic position to efficiently retrieve all reads aligning to a locus

SAMtools is employed to store and manipulate alignments. Core functionalities include sorting alignments for fast access as well as merging of alignments of the same sample that are sequenced in multiple lanes or sequencing runs. With the *mpileup* position specific information for each chromosomal coordinate can be extracted, which constitutes the basis for variant calling. Additionally, SAMtools offers utility programs such as a function to remove duplicate reads (command *rmdup*) and extract statistics of the alignment

(*flagstat*). A text alignment viewer (*tview*) for visualization of alignment is also included.

All utilities resolving around variant calling and output in the standard variant call format VCF (and its binary form BCF) [24] are included in an additional set of programs, bcftools. These tools handle variant calling based on the *mpileup* output, variant filtering on multiple criteria (which will be further discussed in Chapter 2.2.2. Variant Calling), generating output in VCF or BCF format and conversion between the two.

1.10 Exome Sequencing

Exome sequencing is the term used to describe the targeted enrichement and sequencing of all coding regions of the human genome. Despite rapidly decreasing sequencing costs, the cost for whole genome sequencing still remains prohibitive for many projects, thus requiring new methods for targeted enrichment of specific regions of the genome, in this case the complete exome. The technique was widely adopted and used to identify new causative genes in recessive [72, 32] and dominant disorders [71, 40, 106]. Different technologies for targeted enrichment are available, based on in-solution enrichment, which replaced the previous array based enrichment procedures. Currently, there are three different exome enrichment systems available:

1. The SureSelect Human All Exon 50 Mb Kit, developed by Agilent
2. The SeqCap EZ Exome library v2.0 by Roche/Nimblegen
3. The TruSeq Exome Enrichment kit from Illumina

Each system employs a form of biotinylated oligonucleotide baits that are designed complementary to exonic target regions of the human genome. These baits then hybridize to their respective targeted regions and are collected using magnetic streptavidin beads. While all technologies follow the same initial approach there are substantial differences between them in the target defintion, bait density and length and the molecule employed for the capture procedure (DNA oligos for Illumina and Nimblegen, RNA baits for Agilent [14]). The most crucial point where these platforms differ from each other is the bait density and spacing. Nimblegen uses overlapping baits, thus each targeted base is covered by more than one bait. Agilent employs directly adjacent baits to cover each targeted base exactly once, while Illumina uses

spaced out baits, relying on paired-end reads to fill in the gaps. Another major point of difference are the target definitions, i.e. which exonic bases are enriched. The targeted sequences with respect to different exome definitions (RefSeq, UCSC KnownGenes and Ensembl) are shown in Table 1.4. The total time needed for a complete library preparation and exome enrichment amounts to 3.5 days for Agilent and Illumina protocols and 7 days for the Nimblegen workflow. Of special note is that Illumina is the only platform up to date which also covers UTR regions. In the following, the SureSelect System developed by Agilent Technologies [36] will be described in detail.

Table 1.4: Comparison of exome enrichment platforms

	Nimblegen	Agilent	Illumina
Bait length (bp)	55-105	124-126	95
Total targets (bp)	44,070,352	51,542,882	61,884,224
RefSeq coding (bp)	29,918,359	32,326,914	31,817,166
RefSeq UTR (bp)	2,804,875	3,920,825	31,642,004
Ensembl CDS (bp)	29,779,535	33,472,589	31,918,846
Ensembl all exons (bp)	32,499,085	38,123,201	59,275,652

1.10.1 In-Solution Enrichment

Agilents SureSelect protocol follows a standard Illumina library preparation procedure with additional steps to capture the targeted region. The genomic library is mixed in solution with an excess of 120 bp long, biotinylated RNA oligos, termed probes or baits, that are complementary in sequence to the targeted exons. The probes are designed with an 60 bp overlap, ensuring that each exonic sequence is targeted by at least two probes. These RNA baits then anneal to their targets and are pulled down with streptavidin-coated magnetic beads. The caught sequences are then PCR amplified with universal primers and subjugated to sequencing on a next-generation sequencing instrument (see Figure 1.13). Standard kits are available for 'whole exome' sequencing that target either 38 MB or 50 MB of the human genome (assembly hg19). The definition of the exonic sequences is based on the Consensus Coding Gene Set CCDS. The enrichment is a quantitative process and depending on the experiment 50-80% of actual reads will directly overlap the

targeted area. The capture process also bears certain difficulties during the analysis of sequencing data, mainly incomplete enrichment and a non-uniform coverage distribution over the targeted bases. These caveats will be further discussed in Chapter 3.3.

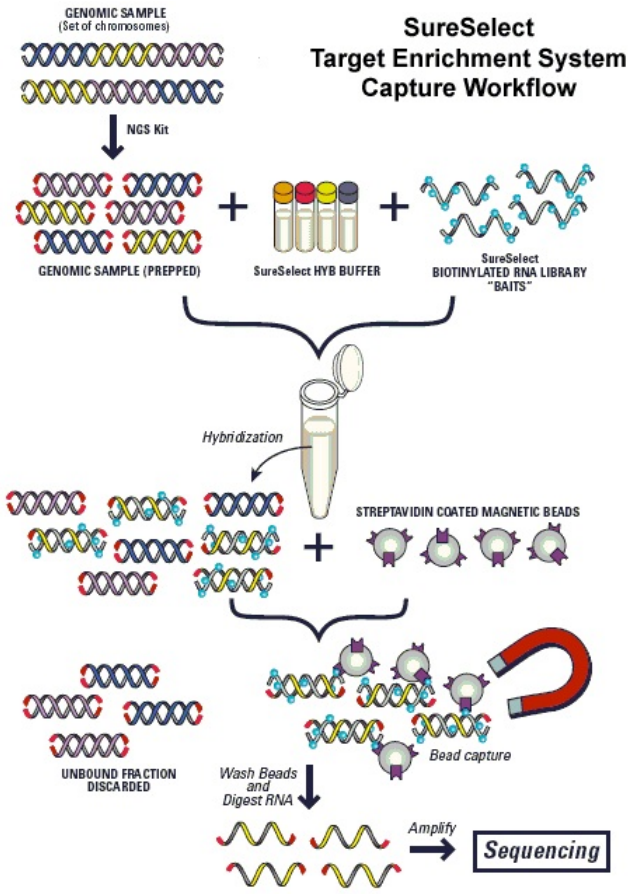


Figure 1.13: Agilent SureSelect Workflow: Genomic library is mixed in solution with capture probes. Specific probes bind to complementary exomic sequences and are captures via magnatic beads. Image available at www.agilent.com.

1.10.2 Applications

Exome sequencing has become the method of choice for many researchers in rare and common disease. The standard procedure consists of the sequencing of the complete exomes of several affected, related or unrelated, individuals and subsequent validation of candidate causative variants in larger cohorts of cases and controls. This approach identified numerous new mutations causative for a variety of diseases. One of the first successful results was the identification of *DHODH*, which encodes a key enzyme in the pyrimidine de novo biosynthesis pathway, of being causative for Miller syndrome (MIM%263750). Miller syndrome is a rare multiple malformation disorder displaying an autosomal recessive form of inheritance [72]. Another example for the discovery of a gene causative for an autosomal recessive disorder is *WDR35* involved in Sensenbrenner syndrome / cranioectodermal dysplasia (CED) [32].

Additonally, the genetic basis for several autosomal dominant disorders have been elucidated by the use of exome sequencing. The first report identified heterozygous *de novo* variants in *SETP1* as causative for Schinzel-Gideon syndrome, a disorder characterized by severe intellectual disability, distinctive facial features and multiple congenital malformations [40]. Mutations in the gene *MLL2* were proven causative for another autosomal dominant disorder, Kabuki syndrome [71]. Kabuki syndrome is a rare, multiple malformation disorder characterized by a distinctive facial appearance, cardiac anomalies, skeletal abnormalities, immunological defects and mild to moderate intellectual disability. Apart from rare disorders, there have additionally been successful applications to common disorders, including the identification of *VPS35* as being involved in Parkinson's disease [106]. All these studies have in common the general setup of sequencing relatively few affected individuals, ranging from 3 to 10 patients.

Moving beyond the single causative mutation scenarios, exome sequencing has also been applied to identify an excess amount of *de-novo* mutations in several different genes as prime cause of general Intellectual Disability [98] and Shizophrenia [34]. In both studies an increased *de novo* mutation rate could be demonstrated in the affected individuals, while the evaluation of causality of the individual variants still differs from case to case.

On the other hand there have been larger studies employing exome sequencing as the method of choice. The sequencing of 50 exomes led to the identification of *EPAS1*, a transcription factor involved in response to hy-

poxia, as key to high altitude adaption when compairing Tibetian and Han Chinese samples [105]. The discovered polymorphism in *EPAS1* had a 78% frequency difference between the two populations. In another study 200 exomes of danish origin were sequenced, revealing an excess amount of low-frequency (minor allele frequencies between 2% and 5%), non-synonymous variation [58].

These examples demonstrate that exome sequencing is established as a state-of-the-art technology in genetic research. Therefore new computational methods have to be developed for the analysis of exome sequencing data. In the following an automated anaylsis pipeline and database structure covering all steps necessary to identify causative variants in exome sequencing data and the application of this pipeline in four distinct projects is presented.

Chapter 2

Results

2.1 Analysis Pipeline for Next-Generation Sequencing Data

For the analysis of next generation sequencing data in general and whole exome sequencing in particular, an automated analysis pipeline was developed and implemented. The pipeline is a combination of public available software and custom developed scripts with the aim of generating reliable variant calls and to identify candidate genes for a variety of disorders and different study designs. It calculates quality metrics and performs read alignment to the reference sequence, variant calling, variant annotation and selection of candidate variants according to the suspected underlying genetic model. All called variants are imported in an relational database scheme and candidate gene identification is facilitated through data base query. A web front end has been implemented where users are able to search the database using preformulated queries. The pipeline has a modular composition and subsets of components may be run in an arbitrary combination. For manual inspection of the results, bed- and html-files are provided. For variant annotation external resources like dbSNP, the UCSC Genome Browser gene tables, the Human Gene Mutation Database HGMD, the 1000 Genomes Project results and polyphen predictions are integrated. The pipeline has been sucessfully employed in identifying new genes implicated in Intellectual Disability (ID) [42] and Parkinson's disease [106]. An schematic overview of the pipeline can be seen in Figure 2.1 and the individual components are discussed in detail in the following section. Table 2.1 shows the individual programs and

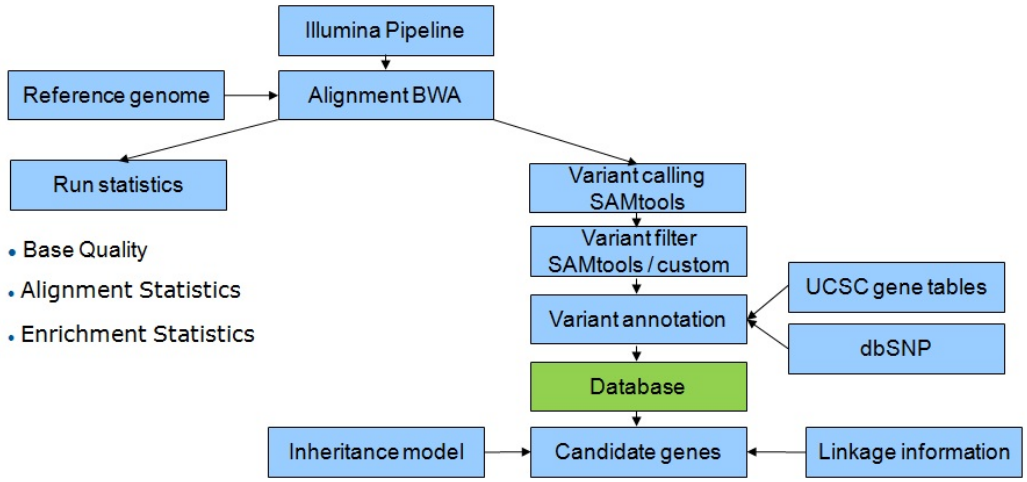


Figure 2.1: Exome sequencing analysis pipeline: Schematic overview of components.

scripts, their respective origin (public available or custom developed) and their purpose in the analysis pipeline as seen in Figure 2.1.

2.2 Components

Custom analysis starts following the completion of the standard image analysis and base calling steps of the Illumina pipeline. As a first step an initialization file is created by the user which contains all relevant information to initiate the pipeline. The following entries are mandatory in the initialization file:

1. Input folder: Run folder of the Illumina analysis pipeline (for sequence file retrival)

2. Reference genome (default human reference genome, assembly hg19)

3. Output folder: Destination of all output files

With this information a config file containing all analysis steps that will be executed is created. All parameters for the specific analysis steps may be altered by the user prior to analysis. If only certain analysis steps are to be performed, for example if specific analysis steps are repeated using

Table 2.1: Individual Pipeline Programs and Scripts

Name	Origin	Function
pipeline.pl	custom	pipeline initialization
Utilities.pm	custom	pipeline initialization
bwa	public	alignment
samtools	public	variant calling
bwa_aln.pl	custom	alignment
parse_stats.pl	custom	run statistics
varfilter.pl	custom	variant filtering
filterSNPs.pl	custom	variant filtering
checkPileup.pl	custom	variant filtering
annotateSNPs.pl	custom	variant annotation
annotateIndel.pl	custom	variant annotation
codingSNPStats.pl	custom	variant statistics
transversion.pl	custom	variant statistics
snvdbExomeInsert	custom	database import
statsdb.pl	custom	database import

different parameter settings, a list containing these analysis steps can be optionally provided in the configuration file. The pipeline then executes only the specified components.

2.2.1 Alignment

Alignment of the reads to the selected reference genome is performed using BWA [53] with mainly the default parameters (Chapter 1.5.2.2). Reads are aligned in paired-end mode (commands *bwa aln* and *bwa sampe*) and the output alignment is stored in the binary SAM format (.bam) [54]. The additional parameter -q is employed during alignment to iteratively trim reads of low quality bases at the ends of each read. The default quality for trimming is set to 15 but may be altered by the user in the initialization file. BWA trims a read as follows: Let q_i be the base quality at the i-th position of a read, l be the original read length, x a position in the read and INT the trimming quality threshold. The length of the trimmed read equals $argmax_x(\sum_{i=x+1}^{l} INT - q_i)$ if $q_l < INT$. The removal of low quality bases at

the end of reads reduces the possibilty of false positive SNV calls during the sunsequent variant calling step.

2.2.2 Variant Calling

The accurate identification of sequence variants is the key step in the analysis of exome sequencing data. In the following parameter settings and filter criteria that are employed in the pipeline are discussed in detail. The paramount objective in variant calling is to distinguish sequencing errors from genuine variants. This process has some ceveats: If a single read shows a single mismatch to the reference sequence it is X times more likely to constitute a sequencing error (approximate error rate 1 in 100) than a true variant (approximate human mutation rate 10^{-8}). Accuracy in variant calling stems from the redundant sequences as each genomic position is covered by multiple reads. If multiple reads show the same mismatch to the reference genome it likely represents a true variant as errors are mainly random. An exception are systematic errors like sequencing artifacts which are discussed in Chapter 3.2. Another difficulty is the detection of heterozygous variants in diploid organisms. Even though an equal distribution of alleles is assumed, a random sampling of alleles occurs during sequencing. This results in the need for a high amount of coverage to reliably detect heterozygotes. It is generally assumed that coverage levels of 20-30X are required to reliably detect 99% of heterozygous variants [56]. This problem is further aggrevated by the fact that the alignment is slightly biased towards reference alleles. During the alignment step all variants are treated as mismatches to the reference genome, thus reads which already contain variant alleles are more difficult to align as all alignment programs have a maximum number of mismatches to the reference genome which may be tolerated due to the time constraint on alignment. In practice this leads to the variant allele being on average slightly underrepresented. Variant calling algorithms thus have to compensate for this situation when trying to accurately fit a statistical model. In the following the variant calling and filtering of the SAMtools package and how it is employed in the analysis pipeline will be further discussed.

2.2.2.1 SNV Calling

Single nucleotide variants are called using the *mpileup* function of BWA, employing the following parameter settings.

1. -M maximum mapping quality that can be assigned to a read (default 60, only achievable with mapped mate in paired-end reads
2. -q omit alignments with a mapping quality below threshold (default 0)
3. -Q omit bases with base quality value below threshold (default 5)

Further parameters are specified for identification of indels, especially to set minimum requirements for indel calls:

1. -L maximum read depth (default 250)
2. -m minimum total number of gapped reads fpr indel candidates (default 1)
3. -F minimum fraction of gapped reads for indel candidates (default 0.002)
4. -e PHRED-scaled gap open error porbability (default 40)
5. -o PHRED-scaled gap extension error probability (default 20)

2.2.2.2 Variant Filtering

A first filtering of the variants is performed using the *varFilter* function of vcftools. The following parameters are set for the filtering process:

1. -Q Root Mean Square (RMS) mapping quality
2. -d minimum read depth
3. -D maximum read depth
4. -a minimum number of alternate alleles
5. -w window around gaps in which SNVs are filtered
6. -W window size around gaps

In addition to the vcftools variant filter a set of custom developed filters is implemented to apply auxiliary quality parameters. Three different filter criteria and default values are employed:

1. The total perecentage of reads indicating the variant allele (default 15%)

2. The ratio of forward to reverse reads showing the variant allele (default 5%)

3. The median quality of the variant bases (deafult median quality 20)

The first criteria ensures that sufficient reads indicate the variant allele. This is mainly important in regions with high coverage to ensure that sufficient distinct reads indicate a potential variant. The second and third criteria were specifically developed and configured to mitigate the problem of false positive variant calls resulting from systematic errors. This phenomenon is commonly referred to as sequencing artifacts and further described in Chapter 3.1.

2.2.3 Run Statistics

Key run parameters are recorded by the pipeline in order to assess the success of the sequencing and enrichment experiments. These parameters include:

1. Percentage of duplicate reads

2. Percentage of mapped reads

3. Number of targeted bases that are covered at levels 1X, 4X, 8X and 20X

4. Percentage of reads that are on target

The first two parameters cover the sequencing experiment itself. The amount of duplicate reads, in short duplicates, are a measure of quality for a library. Duplicates are defined as reads with identical outer coordinates, i.e. the start coordinate of the first read and the end coordinate of the second read are equal for two or more read pairs. This implies that the same DNA fragment was sequenced multiple times, denoting two problems. On the one hand, multiple sequencing of the same fragment does not yield new sequence information. On the the other hand, multiple sequencing bears the danger of generating false positive SNV calls. As the human genome is sheared randomly it is highly unlikely that fragments with the exact same start and end coordinates are generated during this step, the duplicate fragments are rather

created during the PCR amplification step. Since the PCR polymerase introduces errors with a certain probability an error during an early PCR cycle gets propagated through all further amplification rounds. If the identical fragments are then sequenced and aligned, each of them shows the same mismatch to reference sequence, thus generating a possible false positive SNV call. As a consequence, duplicates are removed after the alignment, but before the variant calling step. Duplicate removal is facilitated through the bwa *rmdup* command. The percentage of duplicates is highly dependant on the amount of starting material and the number of PCR cycles during library preparation. After duplicate removal, bwa should be able to align more than 98% of the remaining reads to the human reference genome hg19. A significant deviation of the total amount of aligned reads from this value indicates problems during the library preparation or sequencing of this sample, and the experiment is repeated if necessary.

2.3 Exome Variant Database

All identified variants are imported into a relational database. The identification of candidate genes is then facilitated through database queries. The database is accessible via a web interface where preformulated queries allow the search for candidate genes. A schematic overview of the database is seen in Figure 2.2. The scheme resolves around three central tables:

1. SNV table storing all variants and annotations
2. Sample table storing all information about the sample
3. Table snvsample resolving the n:m reslationship between samples and variants

As one sample has obviously many variants (on average 20,000 coding SNVs) and each variant may be detected in multiple samples the snvsample table becomes a necessity to link each sample to their respective variants. Variant annotation, such as occurence in dnSNP, hgmd or the 1000 Genomes Project as well as functional consequences in the corresponding gene is also stored in the variants table. Additionally, polyphen2 predictions are included here. The sample table is linked to our LIMS system to store sample specific quality measures of both sample preparation and the sequencing run. These quality values include all the parameters discussed in Chapter 2.2.3 Run Statistics.

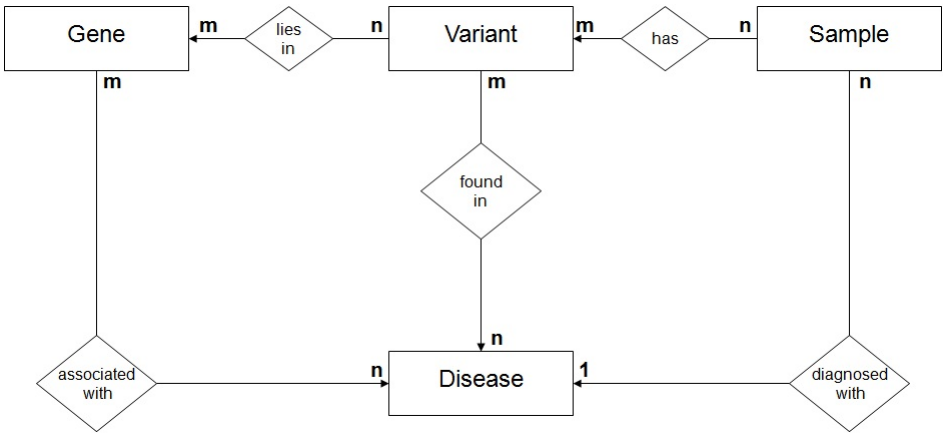


Figure 2.2: Exome variant database: Schematic overview

While a given variant may be detected in multiple samples, quality values such as coverage of the variant, bwa SNV score, percentage of reads showing the variant allele and median base quality of the variant base are stored in the snvsample table. Apart from these main tables the database contains several additional tables linking specific diseases to the affected samples and variants to their respective genes.

2.3.1 Candidate Gene identification

Figure 2.3 shows the database web front end. The specific parameters and queries to facilitate candidate gene identification for a variety of different settings is discussed in the following. In addition to candidate gene searches, the database provides further functionality. Coordinate based queries allow the display of all variants at a specific position, while gene queries report all variants in a certain gene. This functionality is especially useful to assess, for example, how many deleterious variants are present in non-affected individuals, aiding in the process of estimating how plausible the discovered genes are.

2.3.2 Query Parameters

The majority of parameters are similar between all query types and these will be discussed in the following. The first batch of parameters handle the

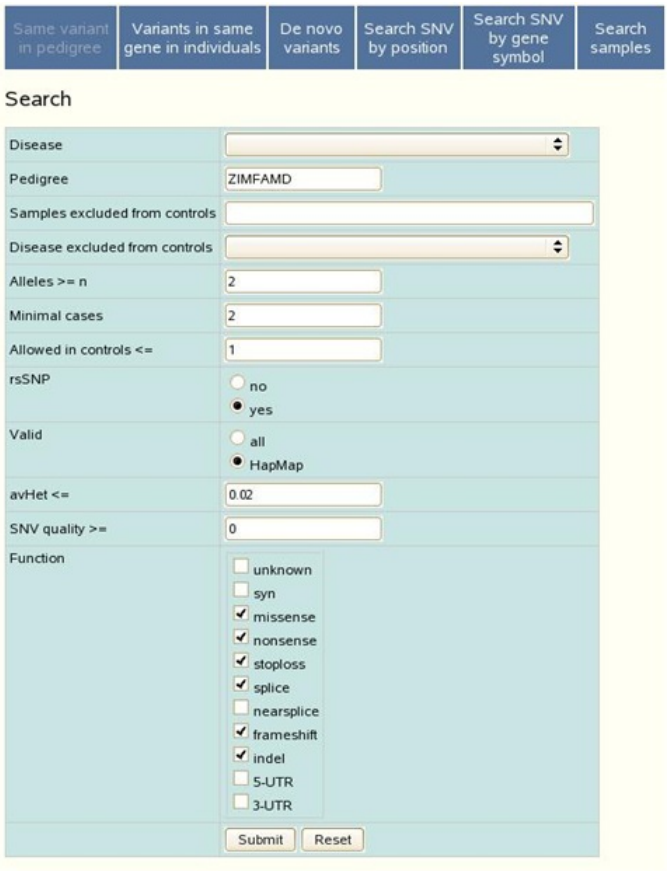


Figure 2.3: Database web front end

basic search conditions. Which samples should be used as cases and controls, especially certain samples or complete disease phenotypes may be excluded from the controls. This setting is used if multiple samples or pedigrees share the same disease phenotype, where it is possible that they also share the same causative variant. The parameter 'alleles' specifies the underlying inheritance model, dominant (alleles = 1) or recessive (alleles = 2). In the recessive setting both homozygous as well as compound heterozygous variants in a gene are identified. In the next parameters one can specify how many samples of the case group have to bear the variation and how often the variant is allowed in the controls. When querying for severe dominant disorders the 'allowed in controls' parameter is typically set to zero, whereas it is prudent to allow the variant with a low frequency in controls in the recessive setting as there

may be carriers among the sample group.

The next set of paramaters apply constraints to the variants. Variants from dbSNP and especially HapMap exceeding a certain frequency may be excluded. The default settings exclude all HapMap variants with an average heterozygosity of more than 2%, but these settings may be modified to exclude any dbSNP or HapMap variants by frequency. Additionally, a minimum SNV quality, as calculated by SAMtools, may be set. Lastly, the type of variation may be determined. In the default setting all missense, nonsense, splice-site SNVs are displayed along with all coding indels and frame-shift mutations. Synonymous, intronic, near splice site and UTR variants can be optionally displayed.

2.3.3 Query Types

There are three different query types available: Pedigree based, gene based, and de-novo centered queries.

Pedigree based queries: These queries aim at identifying causative mutations in a sample of related individuals. Typically three to four affected family members are sequenced. As the affected individuals all stem from the same family it can be assumed that they share the same causative variant. The query thus identifies all variants that are shared between the family members, but do not exceed a certain frequency in the specified control group (as discussed in the previous section). If the suspected underlying disease mechanism is recessive both homozygote and compound heterozygote mutations, i.e. two heterozygote missense mutations in the same gene, are displayed. With an additional parameter the minimum occurence of the variant(s) in the pedigree can be specified, so it can be allowed that some affected individuals do not have the variant. This is necessary to handle potential phenocopies in the pedigree, which can occur in particular when dealing with common disorders like Parkinson's disease or Cardiomyopathys. An example pedigree based query is shown below. The queries are preformulated, user defined parameters (shown in italics in the example) for the queries are parsed from the web interface.

```
SELECT result FROM
snv v RIGHT JOIN snvsample x on (v.idsnv = x.idsnv)
LEFT JOIN dgv dgv on (v.chrom = dgv.chrom AND
v.start=dgv.start)
LEFT JOIN sample s on (s.idsample = x.idsample)
LEFT JOIN snvgene y on (v.idsnv = y.idsnv)
LEFT JOIN gene g on (g.idgene = y.idgene)
INNER JOIN snv2disease f ON (v.idsnv = f.fidsnv)
LEFT JOIN hgmdpro.hg19coordsmod h ON (v.chrom =
h.chromosome AND v.start = h.coordSTART)
LEFT JOIN disease2gene dg ON (s.iddisease=dg.iddisease AND
g.idgene=dg.idgene)
LEFT JOIN pph2 pph ON v.idsnv=pph.idsnv
WHERE s.pedigree = pedigree_id
AND x.snvqual ≥ quality_threshhold
AND x.alleles ≥ 1
AND frequency_in_hapMap / frequency_in_dbSNP ≤ fre-
quency_threshhold
AND mutation_types
AND gene_not_in_blacklist
AND passed_quality_params
AND selected_disorder
GROUP BY variants
HAVING count variant_allele ≥ min_occurence_threshhold
ORDER BY chromosome, position
```

Gene based queries: Gene based queries are very similar to the previously described pedigree based searches but exhibit a key difference: In this setting unrelated individuals affected by the same disorder are investigated. Thus it is not likely that the individuals share the same causative variant(s). On the other hand it can be assumed that mutations in the same gene might be responsible. These query type outputs all variants that affect the same gene in each of the inspected samples. Likewise to gene based queries both homozygote and compound heterozygote variants are determined and a number of individuals who do not have mutations in the same gene can be allowed to cope with phencopies. An example gene based query is displayed in the following. It differs from the previously shown pedigree based query mainly

in the GROUP BY clause where in this case the result is grouped by the gene symbol first, ensuring that a specified number of affeceted individuals have variants in the same gene.

```
SELECT result FROM
snv v RIGHT JOIN snvsample x on (v.idsnv = x.idsnv)
LEFT JOIN dgv dgv on (v.chrom = dgv.chrom AND
v.start=dgv.start)
LEFT JOIN sample s on (s.idsample = x.idsample)
LEFT JOIN snvgene y on (v.idsnv = y.idsnv)
LEFT JOIN gene g on (g.idgene = y.idgene)
INNER JOIN snv2disease f ON (v.idsnv = f.fidsnv)
LEFT JOIN hgmdpro.hg19coordsmod h ON (v.chrom =
h.chromosome AND v.start = h.coordSTART)
LEFT JOIN disease2gene dg ON (s.iddisease=dg.iddisease AND
g.idgene=dg.idgene)
LEFT JOIN pph2 pph ON v.idsnv=pph.idsnv
WHERE s.pedigree = pedigree_id
AND x.snvqual ≥ quality_threshhold
AND x.alleles ≥ 1
AND frequency_in_hapMap / frequency_in_dbSNP ≤ fre-
quency_threshhold
AND mutation_types
AND gene_not_in_blacklist
AND passed_quality_params
AND selected_disorder
GROUP BY genesymbol, variant HAVING count(distinct
v.chrom,v.start) ≥ min_occurence_threshhold OR max(x.alleles) ≥
min_occurence_threshhold ORDER BY chromosome, position
```

De-novo variant queries: The third query type covers a different experimental setup, the sequencing of trios of father, mother and affected child. Usually this setup is chosen if the affected child of the healthy parents has a severe disorder where the suspected cause of the disorder is a new mutation that occured de-novo in a child. This query type consequently excludes all variants that at least one parent shares with the child as these can be eliminated as possible disease origin. All potential de-novo variants are then again checked against the control sample to further exclude variants. The

results of de-novo queries typically have a slightly elevated false positive rate as two types of errors are possible. False positive calls in the affected child, the same error that can occur in all types of experimental settings, as well as false negative, i.e. missed calls in one of the parental exomes. These missed calls can lead to variants being falsely classified as de-novo that are actually inherited. Consequently, all variants have to be investigated in the complete trio during validation step to validate the de-novo state. De-novo queries are similar to pedigree-based queries with the only difference being that none of the potential causative variants is allowed in the unaffected members of the pedigree, i.e. the parents.

2.4 Identification of Disease Causing and Disease Associated Muations

The pipeline was employed for successful identification of disease causing and disease associated mutations in four distinct projects, showing the application of the software for variant identification in different areas of the mutational spectrum.

2.5 Somatic Mutations - Identification of recurring tumor-specific somatic mutations in acute myeloid leukemia by transcriptome sequencing.

During development an alpha version of the analysis pipeline, in particular the alignment, variant calling, filtering and annotation components (with exception of the database), was employed to identify recurring tumor-specific somatic mutations in acute myeloid leukemia (AML) [38]. It had been recently shown that tumor-specific driver mutations could be identified in AML by comparing variants from a tumor and a remission sample of the same patient [52, 63]. Whole exome sequencing kits were not yet available, thus due to the prohibitive cost of whole genome sequencing a whole transcriptome (mRNA-seq) approach by sequencing all transcriptionally active genes was chosen. Mutations were detected by comparing the transcriptome sequence of

an AML sample with the corresponding remission sample, essentially replacing the database comparison that was used later for whole exome sequencing projects.

The incidence rate of a cute myeloid leukemia (AML) is approximately three to four cases per 100,000 individuals, thus representing one of the most abundant hematological malignancys. In particular the long term prognosis is is very poor, the 5 year survival rate ranges from 25 to 30%. The most frequent cause of AML are chromosmal abberations, which are encountered in approximately half of the patients with AML, whereas the other half displays a normal karyotype (cytogenetically normal-AML) [68]. Despite a growing number of studies elucidating the molecular background of the disease there still remain about 25% of AML patients where no currently known mutation can be identified.

In this study a diagnostic bone marrow sample was collected from a 69 year old patient, diagnosed with AML M1 in May 2008. A complete remission could be achieved by an induction therapy using the sequential high-dose cytosine arabinoside and mitoxantrone (S-HAM) protocol, and a remission sample from peripheral blood was taken after leukocyte recovery in July 2008. Approximately $50x10^6$ cells from each sample were used for mRNA extraction using Trizol (Invitrogen, Carlsbad, CA, USA). The sequencing library was prepared using mRNA-Seq sample preparation kit (Illumina, San Diego, CA, USA). In brief, mRNA was selected using oligo-dT beads (dynabeads, Invitrogen). The mRNA was then fragmented using metal ion hydrolysis and reversely transcribed using random hexamer primers. Following steps included end repair, adapter ligation, size selection and polymerase chain reaction enrichment [38].

Sequence alignment was performed using the analysis pipeline as previously described. In order to be able to map reads spanning splice junctions an expanded reference sequence comprising the human genome assembly (build NCBI36/hg18) and all annotated splice sites extracted from the University of California Santa Cruz (UCSC) genome browser-known gene track was used. Splice site extraction was performed using a custom Perl script (*spliceRef.pl*). In total, 127,115,919 paired-end reads of 36 bp length for the AML sample (totalling 4.35 GB of sequence), of which 95.08% aligned to the reference sequence, and 187,782,678 paired-end reads (5.54 GB) for the remission sample with 82% aligning to the reference, were generated on multiple GAII runs. During alignment, 31.27% and 39.81% apparently duplicated reads were removed from the AML and remission sample, respectively. This comparably

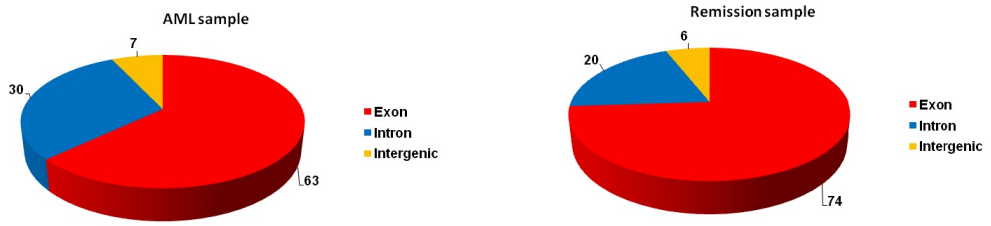


Figure 2.4: Distribution of reads mapping to exonic, intronic and intergenic regions for the AML (a) and remission sample (b). Figure reprinted from [38]

high number of duplicate reads stems from sequencing of limited input material (only expressed mRNA) to a relatively high coverage.

The success of the RNA library preparation was assessed by the percentage of reads matching to known exons from the UCSC genome browser. For the AML sample, ∼63% of reads aligned to exons, ∼28.5% to introns and ∼7.5% to intergenic regions, whereas for the remission sample, ∼73.5% of reads aligned to exons, ∼20.5% to introns and ∼6% to intergenic regions (Figure 2.4). The relatively high levels of intronic reads stem from unspliced mRNAs. The proportions of intronic and exonic reads showed high varation between different preparations from the same samples. This is an indication that minor differences in RNA concentration and quality might strongly influence the competitive binding of shorter spliced and longer incompletely spliced mRNAs to oligo dT-beads [38].

To assess the number of genes that were sufficiently covered for mutation detection a non-redundant mRNA set from the Ensembl transcripts database was compiled, resulting in a set of 35,876 genes. The average sequence read depth for every gene was calculated. The read depth per gene ranged from 0 to over 1000. A total of 10,152 genes had an average read depth of at least sevenfold and 6,989 genes had an average read depth of 20 or greater in both samples. These genes were considered as sufficiently covered for accurate mutation detection. Coverage levels varied grately due to expression bias. Nevertheless all positions, even when covered with less than 20 reads, were used for initial mutation detection, accepting a higher false positive rate for poorly covered genes. Figure 2.5 shows a histogram of the per gene average read depth for both samples, and the average read depth for the AML and the remission sample seperately.

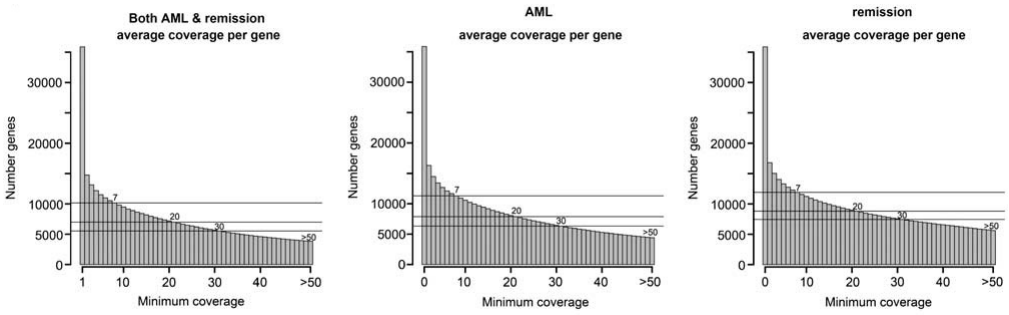


Figure 2.5: Histograms of the sequence coverage in a non-redundant gene set based on the Ensembl annotation (35,876 genes) for genes detected in both samples (left), the acute myeloid leukemia (AML, middle) and remission (right) samples. Minimum sequence coverage is plotted on the x-axis and number of genes is plotted on the y-axis. 10,152 genes were sequenced with an average coverage of 7 or greater, 6,989 genes with an average coverage of 20 or greater and 5,535 genes with an average coverage of at least 30 in both samples (left). The result obtained from the AML sample was 11,293 genes with an average coverage of 7 or greater, 7,878 genes with an average coverage of 20 or greater and 6,326 genes with an average coverage of 30 or greater (middle). The sequencing of the remission yielded 11,906 genes with an average coverage of 7 or greater, 8,805 genes with an average coverage of 20 or greater and 7,446 genes at an average coverage of at least 30 (right). The high proportion of genes detected in both samples indicates a good comparability of expression profiles. Figure reprinted from [38]

Single-nucleotide variants and indels were called, filtered and annotated in the AML sample using the analysis pipeline with the default paramters as previously described, resulting in 8,978 SNVs in coding regions. In the next step, all coding SNVs that were present in the dbSNP database version 130 or in the exomes of 8 HapMap samples [73] were excluded. The remaining 926 sites contained 612 SNVs, which led to an amino acid substitution or which disrupted canonical splice sites (the remaining 314 SNVs comprised synonymous variants). These 612 SNVs were then compared with the unfiltered calls of the remission sample at these 612 positions. We excluded all positions with any indication that the same SNV was also present in the remission sample. This strategy resulted in the identification of 11 candidate SNVs unique to the tumor sample. Capillary sequencing of genomic DNA

from both the tumor and the remission sample confirmed five SNVs, which affected the genes *RUNX1*, *TLE4*, *SHKBP1*, *XPO7* and *RRP8*. Two SNVs were false positives with the same heterozygous SNVs being also present in the genomic DNA of the remission sample, four SNVs could not be confirmed in the AML sample. Table 2.2 shows the four confirmed mutations and their respective consequence on the corresponding protein. All four mutations were heterozygous. Among the identified tumor-specific mutations were strong

Table 2.2: Confirmed tumor-specific mutations

Gene	Position (hg18)	Amino Acid Substitution	Read depth AML	Read depth Remission
TLE4	chr9:81523675	N511S	167	114
SHKBP1	chr19:45775904	V89I	81	249
RUNX1	chr21:35128760	Q208X	59	36
XPO7	chr8:21883756	R139G	52	44
RRP8	chr11:6579867	S85C	23	38

candidates for leukemogenesis in this patient. *RUNX1* (*AML1*) carried a heterozygous stop mutation in the Runt domain. *RUNX1* is the fusion partner of *RUNX1T1* in the recurring t(8;21) (q22;q22) translocation present in 813% of de novo AML cases [77]. In addition, point mutations in *RUNX1* have recently been described in AML at a frequency of 810% [74]. *TLE4* is located on chromosome 9 band q34, which is frequently deleted in AML with t(8;21) translocations, and is therefore a putative tumor suppressor gene. Additionally, the *TLE4* protein interacts with *RUNX1*, and haploinsufficiency of *TLE4* was shown to collaborate with the *RUNX1/RUNX1T1* fusion to rescue cells from apoptosis [25]. The third tumor-specific SNV resulted in a missense mutation (V89I) in *SHKBP1* (also known as *SETA* binding protein 1, *SB1*). Through *SETA*, *SHKBP1* interacts with *CBL*, a ubiquitin ligase that regulates the degradation of *FLT3*. *CBL* mutations, which result in the increased activity of *FLT3*, have been described in AML and myelodysplastic syndrome. Thus, it is possible that *SB1* mutations affect *FLT3* signaling. *SHKBP1* overexpression in cell lines has antiapoptotic effects [9, 81, 59]. The fourth and fifth AML-specific mutations were missense mutations in *XPO7* (a member of the importin beta superfamily) and *RRP8* (a methyltransferase, possibly involved in ribosomalRNA processing). These two mutations have no apperent link to AML and putatively constitute passenger mutations.

Based on these results a collective of 95 AML patients was screened for addtional mutations in the three candidate genes *RUNX1*, *TLE4* and *SHKBP1*. 11 additional patients had heterozygous mutations in *RUNX1*, while 2 additional mutations could be identified in both *TLE4* and *SHKBP1*, lending further evidence that mutations in these two genes contribute to Leukemiagenesis in these patients.

The author's contributions to this work were the implementation of the core analysis pipeline components (read mapping, variant calling, filtering and annotation), primary data analysis, candidate variant identification by tumor/remission sample comparison and writing of the manuscript.

2.6 Monogenic Disorders - Identification of mutations in Adaptor Protein Complex 4 proteins as cause of Intellectual Disability

Intellectual disability (ID), with a prevalence of approximately 2%, is the most frequent cause of severe cognitive-dysfunction disorders. A new, autosomal-recessive subtype of ID with the distinct symptoms of progressive spastic paraplegia, shy character and short statue could be identified by employing a combination of autozygosity mapping and either Sanger sequencing of candidate genes or next-generation exome sequencing. Using this approach mutations in each of the three genes encoding adaptor protein complex 4 (AP4) subunits were identified [42]. In total, three consanguineous families with eight affected individuals were investigated. The affected individuals presented with severe intellectual disability, absent speech, shy character, stereotypic laughter, muscular hypotonia that progressed to spastic paraplegia, microcephaly, foot deformity, decreased muscle mass of the lower limbs, inability to walk, and growth retardation. In the following, the discovery of a mutation in the gene *AP4S1* using exome capture and sequence anaylsis by the previously described pipeline will be discussed in detail.

In this case, DNA from a single individual of a nuclear family was enriched with the SureSelect Human All Exon Kit 38Mb kit and subsequently sequenced on a Illumina GAIIx with 76bp paired-end reads. Sequence analysis was performed using the pipeline as previously detailed. In this special case the sequencing of a single individual was sufficient due to two further filtering criteria for variants:

1. The parents are a consaguinus union, in consequence the causative mutation is expected to be homozygous

2. A previous linkage analysis is available and only variants from the linkage region were considered as causative

In total, 26,037 variants were identified. Of these, 6,655 variants were coding and homozygous, 345 of those were rare (frequency below 2% in dbSNP and the in-house exome database), and 139 of those were nonsynonymous. There were two variants identified that were located in the linkage region on chromosome 14: a missense variant in the last exon of *SLC22A17* (Val477Met) and a nonsense variant in the first coding exon of *AP4S1 (Arg42*)* [42]. A region of interest (in this case the linkage) may be optionally supplied to the pipeline during analysis. Variants residing in this region are then automatically highlighted when the results of the database query are displayed. The variants cosegregated with the affected status within the family, and were not present in a control group consisting of 740 chromosomes of similar ethnical background, providing further evidence for that one of these variants is causative for ID in this family. Additionally, both variants were subjected to an in silico analysis with MutationTaster [90] and PolyPhen [79], which showed a high probability for a pathogenic effect for both variants.

Both *SLC22A17* (highly expressed in the brain and belonging to the organic cation transporter family [MIM611461]) and *AP4S1* (encoding the small subunit of the adaptor complex 4, MIM [607243]) are thus good potential candidates for ID. It was assumed that the mutation in *AP4S1* is the true causative mutation due to two main reasons:

1. The clinical presentation of affected persons with mutations in the different AP4 subunits shows high similarity. Two additional families with mutations in this protein complex have been identified by linkage and candidate gene approaches

2. The mutation introduces a premature stop codon, truncating the protein in the first exon

Although an additional effect of the *SLC22A17* variant in the linkage region can not be ruled out with certainty. By mediating several types of vesicle formation and selecting cargo molecules to be included in these vesicles, adaptor protein complexes, like AP4, assume a pivotel role in the

signal-mediated trafficking of integral membrane proteins. These complexes have a heterotetrameric composition which is evolutionarily conserved and consist of four distinct subunits [42]. The different adaptor complexes are known to be associated with ID as mutations have already been linked to human disorders. For example, mutations of AP1S2 (encoding a subunit of the adaptor complex 1 [MIM300629]) cause an X-linked form of intellectual disability [8, 95]. Mutations in AP3B1 (encoding a subunit of the adaptor complex 3 [MIM 603401]) cause Hermansky-Pudlak syndrome type 2 (HPS [MIM 608233]), a disease characterized by hypopigmentation of the eyes and skin, prolonged bleeding, and lysosomal ceroid storage [11].

The author's contribution to this project consisted of implementation of analysis pipeline components, in particular the realization of the relational database scheme, primary data analysis of the family investigated by exome sequencing and drafting the sequencing paragraph of the manuscript.

2.7 Common Disease I - Identification of a mutation in *VPS35* causing late-onset Parkinson

One of the first whole exome sequencing projects employing the pipeline was the identification of a causative mutation in an Austrian family in which 16 members were affected by Parkinson's disease (PD)[106]. PD was suspected to be inherited in an autosomal dominant mode with high penetrance. The family consisted of seven affected members that were available for clinical and DNA investigations. The most distantly related cousins were sunjected to exome sequencing in order to identify potentially disease-causing variant. Selecting distantly related members of the pedigree minimizes the proportion of alleles shared by descent. Exome sequencing was performed on a Genome Analyzer IIx system (Illumina), exonic sequences were captured by in-solution enrichment (SureSelect Human All Exon 38 Mb kit, Agilent). Read alignment, variant calling and filtering, and selection of candidate causative mutations was performed using the analysis pipeline as previously described. The pedigree based database search identified 10 rare variants shared between the two samples. Table 2.3 shows the stepwise filtering process, only the 10 shared variants are the result of the database query. Only a single heterozygous variant in the *VPS35* gene (c.1858G>A, p.Asp620Asn) fulfilled

Table 2.3: Stepwise variant filtering for two PD cases

	Patient 1	Patient 2
Total variants (SNVs + Indels)	24,804	23,783
Coding + splice site (SS) variants	16,236	15,931
Non-synonymous (NS) + SS variants	7,717	7,517
Rare NS + SS variants (dbSNP132)	1,158	1,128
Rare NS + SS variants (exome database)	410	392
Shared variants in Patient 1 and 2	10	

two further criteria of being possibly causative:

1. it was found in all seven affected members investigated
2. it was absent in approximately 680 KORA S4 general population samples [103]

Additionally, two further families carrying this mutation were identified. All eight affected individuals were investigated in both families and the VPS35 variant was detected in all of them. On the other hand the variant was abesent in a second set of 554 Austrian controls and in an additional 1014 KORA-AGE controls. Cross-species alignment of plants, fungi, invertebrates, and vertebrates VPS35 showed complete conservation of amino acid Asp620 [106]. The likely consequence of the Asp620Asn variant was predicted to be damaging by PolyPhen2 [79], SNAP [10] and SIFT [1]. The crystal structure of the C-terminal part of *VPS35*, the area of the protein in which the variant p.Asp620Asn is located, has been resolved [39]. The impact on protein stability was investigated by molecular dynamics (MD) simulations. The mutation was manually introduced to the crystal structure and the side chains were modeled by scwrl 4.0 [47]. All MD simulations were performed via GROMACS 4.5, with the allatom force field AMBER0335 and the water model TIP3P36 as parameters. The protein is found on the edge of a helix interacting with VPS29. Wild-type residue Asp620 forms frequent hydrogen bonds (HBs) to Lys622, but these bonds are less frequent in the p.Asp620Asn variant (Figure 2.6). This change results in the loss of salt bridges and cause the protein to be locally more flexible, culminating in convincing computational evidence that this mutation may indeed be causative for Parkinson's disease in these patients [106].

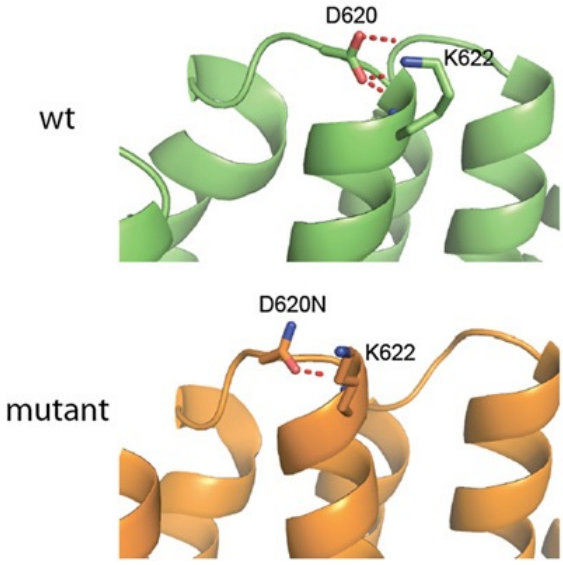


Figure 2.6: Molecular dynamics modeling: Hydrogen bonds (HB) are shown as red dashed lines. Wild type form is displayed in green, mutation Asp620Asn in orange. Asp620 forms a HB to Lys622 and shows an additional saltbridge interaction, Asp620Asn forms fewer HBs, and no electrostatic interaction is possible. Figure reprinted from [106].

Additionally, there is reason from the biological side that makes this variant a plausible choice: *VPS35* is a component of the retromer complex and is involved in retrograde transport from the endosomes back to the trans-Golgi network [7]. This multi-protein complex consists of the cargo-recognition *VPS26-VPS29-VPS35* heterotrimer and a membrane-targeting heterodimer or homodimer of *SNX1* and/or *SNX2* [7, 84]. All proteins involved are evolutionarily conserved and have been previously described in Saccharomyces cerevisiae. Most interesting in this context is the involvement of the retromer into the retrograde transport of *SORL1*, a *VPS10P*-domain receptor protein that has been implicated in Alzheimer disease [104, 83].

We therefore concluded that the variant Asp620Asn is indeed causative for PD in the investigated families [106]. Simultaneously a report was published that identified the Asp620Asn mutation in *VPS35* to be involved in Parkinson's disease in a Swiss kindred, three more families and one patient with sporadic PD [96].

The author's contribution constitued the analysis pipeline, primary data analysis and writing the sequencing paragraphs of the manuscript.

2.8 Common Disease II - Dysfunctional nitric oxide signaling increases risk of myocardial infarction

Myocardial infarction (MI) is a life-threatening disease, which results from sudden atherothrombotic occlusion of a coronary artery (coronary artery disease, CAD). Most cases of MI occur sporadically. The importance of genetic predisposition to MI/CAD is best documented by genome-wide association studies (GWAS) with more than 30 loci identified so far. Auxillary to sporadic manifestations the disease sometimes clusters in families, indicating possible monogenic subforms. In addition to such large-scale studies of sporadic cases, investigation of families with multiple affected individuals can potentially provide new insight in the pathogenesis of the disorder [86, 89].

The starting point of the present study was the identification of a MI family with 30 members diagnosed with CAD of whom 17 had early onset MI (< 60 years of age). Prior to the next-generation sequencing approach the family was already examined using traditional methods. Microsatellite-based linkage analysis failed to identify any significant single-locus logarithm-of-odds (LOD) scores. In a next step exome sequencing was performed as 54 bp paired-end runs on a Genome Analyzer IIx system (Illumina) after in-solution enrichment of exonic sequences (Agilent) for three cousins (III.21, III.25 and III.49, Figure 2.7) who represented distantly related family members. Data analysis and candidate variant identification was performed employing the analysis pipeline as described. An autosomal dominant inheritance model was assumed when querying the databse for candidate variants.

Considering only rare variants (MAF $< 0.5\%$) and a dominant model, all three affected members showed two loss-of-function (*GUCY1A3* and *ETFDH*) and two nonsynonymous variants (*CCT7* and *GCLC*). These four variants were investigated for cosegregation pattern in the family, impact on protein function, and biological links to CAD/MI by literature search, as well as presence in further control individuals. *ETFDH* and *GCLC* could be excluded from further studies, because neither a linkage signal nor pathophysiological relevance was observed. By contrast, the *GUCY1A3* and *CCT7* variants are

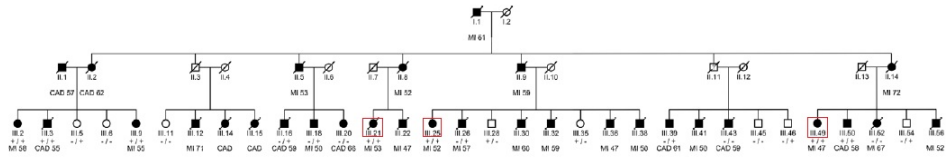


Figure 2.7: Persons III.21, III.25 and III.49 (marked in red rectangles) have been selected for exome-sequencing, +/+ denotes double mutation carriers (p.Leu163Phefs*24+/p.Ser525Leu-), +/- denotes probands carrying the p.Leu163Phefs*24 mutation in GUCY1A3 only, -/+ denotes probands carrying the p.Ser525Leu mutation in CCT7 only.

Table 2.4: Occurence of mutations in the MI extended pedigree

	Affected	Healthy
Both Mutations	7	0
One Mutations	4	5
No Mutation	2	3

linked to MI in the family in a two-locus model and mechanistically related to the disease.

A single nucleotide insertion (T) in exon 6 of *GUCY1A3* (NM_001130683.2: c.488dup, p.Leu163Phefs*24) resulted in a frameshift and a premature stop codon after 24 aberrant amino acids. This variant was present in 7/15 affected and 2/8 unaffected family members for whom DNA was available for sequencing. Moreover, a single nucleotide substitution (C>T) was found in exon 10 of *CCT7* (NM_001166284) leading to a missense mutation at amino acid position 525 (p.Ser525Leu). This variant was present in 11/15 affected and 3/8 unaffected family members. Interestingly, all 7 carriers of digenic mutations were affected (Figure 2.7, Table 2.4). Both mutations were absent in 3,150 healthy subjects and 3,842 unrelated MI cases.

These initial findings indicate a possible digenic inheritance of MI in this family. To test this hypothesis a linkage analysis was performed with assumption of the locus order M1 - D1 - D2 - M2, where M1 = marker p.Ser525Leu (chromosome 2p13.3), D1 = disease locus CCT7, D2 = disease locus GUCY1A3, and M2 = marker p.Leu163Phefs*24 (chromosome 4q31.3), with recombination fractions Θ_1, Θ_2, and Θ_3 in the three consecu-

tive intervals for the given locus order. Single-locus linkage analyses (D1 = D2) revealed a maximum LOD score of 1.11 at recombination fraction Θ_1 = 0.19 for p.Ser525Leu in CCT7, and a maximum LOD score of 0.08 at Θ_3 = 0.33 for p.Leu163Phefs*24 in GUCY1A3. However, in two-locus analyses, a maximum LOD score of 5.68 was obtained at Θ_1 = 0.21 with Θ_2 fixed at 0.0001 (Supplementary Figure 2). To estimate inheritance parameters, searching for the maximum LOD score over a range of penetrance values and recombination fractions may be used. The resulting LOD score (= MOD score [16, 82]) turned out to be 10.32 at penetrances of 0.99 and 0.20 at disease loci CCT7 and GUCY1A3, respectively, but this LOD score cannot be interpreted at face value as it has been maximized over model parameters. To see whether our evidence for linkage is simply mediated by a phenotype association between marker p.Ser525Leu in CCT7 and the disease locus cluster CCT7/GUCY1A3, we carried out pedigree segregation analysis [75] by testing whether penetrances at the disease loci depend on marker genotypes. However, results were not significant (P = 0.245) supporting the conclusion that the linkage finding is real, and not merely mediated by a phenotype association.

To further study association between rare GUCY1A3 and CCT7 variants and MI risk beyond this family, the coding exons of the two genes were sequenced in 48 patients from 22 additional MI families with more than 5 affected family members. In this analysis, one non-synonymous variant, p.Gly537Arg, was identified in GUCY1A3. This highly conserved variant (Figure 2A) was found in 3/5 affected family members and was neither present in the current 1000Genome release, or the NHLBI Exome Sequencing Project (http://evs.gs.washington.edu/EVS/), nor in 3,150 controls and 3,842 MI cases. Furthermore, we searched for rare potentially deleterious variants in GUCY1A3 and CCT7 in 252 young MI cases (age of onset between 24 and 49 years, 24% women) with a positive family history and 800 individuals affected with diseases other than cardiovascular, for whom full exome sequencing data was available through the pipeline database. In the GUCY1A3 gene, 8 rare missense mutations were identified (5 [2%] in CAD/MI cases and 3 [0.37%] in the other disease samples, Fisher exact test P = 0.023. In the CCT7 gene, 7 different missense mutations were identified (3 [1.2%] in CAD/MI cases and 5 [0.62%] in the other disease samples (Fisher exact test P = 0.12). Interestingly, p.Ser525Leu in CCT7 was found in a further patient suffering from premature MI (age of onset 43 years). No further person carrying mutations in both GUCY1A3 and CCT7 was identified.

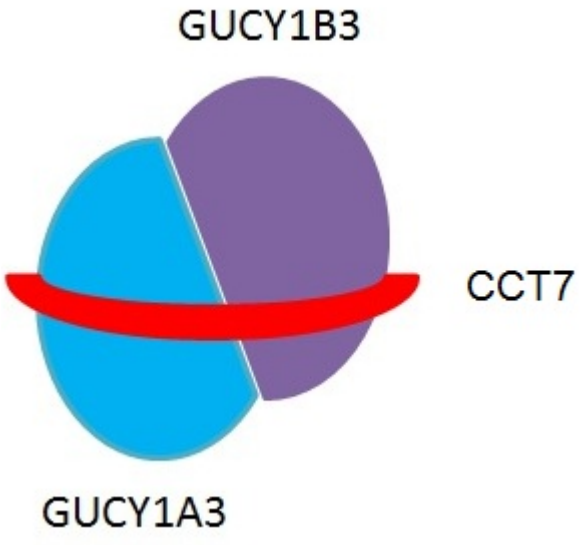


Figure 2.8: The soluble guanylyl cyclase (sGC) consists of an α1-subunit (encoded by GUCY1A3) and a β1-subunit (encoded by GUCY1B3). The protein is stabilized by the a chaperone protein encoded by CCT7.

GUCY1A3 encodes the α1-subunit of soluble guanylyl cyclase (sGC), while CCT7 a member of the chaperonin containing TCP1 complex, which, among other functions, stabilizes sGC (Figure 2.8). Mice deficient for the α1-subunit sGC protein displayed accelerated thrombus formation in the microcirculation upon local trauma. The interaction between the two genes in which mutations were identified further indicate a combined effect of these mutations in the development of MI in this family.

The author's participation in this work involved contribution of the analysis pipeline and primary data analysis.

Chapter 3

Discussion

3.1 General Mutation Identification Strategies for Exome Sequencing

Typically more than 20,000 variants are identified per sequencend exome. Quality filtering and annotation to prioritize these variants is necessary to identify causative variants. Furthermore, different studies warrant different exome sequencing strategies, i.e. which and how many samples are analyzed, in order to be successful. Likewise to traditional approaches the suspected underlying inheritance model and the individual family pedigree determine the best approach [33]. The individual exome sequencing strategies are shown in Table 3.1 (adapted from [33]) and can be briefly summarized as follows:

Homozygosity Approach: In the case of a rare recessive disorder in a child of a consanguinous union two prior assumptions can be made: (i) the causative mutation is present in homozygous state as it is inherited from both parents and (ii) that the variant resides in a larger homozygous stretch in the patient's genome. A prior homozygosity mapping can be used to pre-select these genomic regions to greatly reduce the search space. When combining this information with the standard criteria of the variant priorization steps (i.e. frequency in controls, dbSNP and the 1000 Genomes Project and effect of the variant on the protein level) the number of potentially causative variants can be on average reduced to 10 or less. In this setting the sequencing of a single affected patient is thus often sufficient to identify the causative variant. This approach was employed in the Identification of mutations in

Table 3.1: Mutation Identification Strategies

Strategy	Sequencing	Filter Criteria	Examples
Homozygosity	Single patient from consanguinous union	Homozygous mutation in homozygous region	[42, 4, 99]
Linkage	Multiple distantly related patients, same pedigree	Shared rare variant in all patients	[106, 72]
Compound Heterozygous	Single patient with recessive disorder	Rare, compound heterozygous mutations	[32]
Monogenic Dominant	Multiple unrelated patients, same phenotype	Rare variants in same gene	[40, 71]
De-novo	Trio	De-novo variants in the affected child	[98]

Adaptor Protein Complex 4 proteins as cause of Intellectual Disability (Chapter 2.4.2 [42]).

Linkage Strategy: For families with a suspected monogenic disorder the sequencing of multiple affected family members is employed to identify shared variants. To reduce the number of potentially causative variants and to keep the number of shared benign variants to a minimum it is necessary to sequence the most distantly related affected family members available. The number of sequenced patients is typically two or three, which is sufficient to narrow down the number of variants to be considered to below 10, even the identification of a single candidate variant is possible [73, 72]. In order to further reduce the number of possible causative variants it is possible to combine the sequencing data with a previous linkage analysis to further limit the search space, likewise as in the homozygosity approach. This strategy (with the incorporation of linkage information)was employed in the identification of *VPS35* variants in Parkinson's disease in this work (Chapter 2.4.3 [106]).

Compound Heterozygous Strategy: If the underlying suspected inheri-

tance mechanism is recessive with no indication of consanguity in the parents the disorder is most likely caused by two different heterozygous mutations in the same gene. Addtionally it can be assumed that both variants are rare in control populations and databases. By selecting only compound heterozygous, low-frequency mutations the search space is suffciently reduced that the sequencing of a single affected patient can lead to a successful variant identification.

Strategy for Dominant, Monogenic Disorders: In the presence of a monogenic disorder with dominant mode of inheritance the approach is very straightforward. Even though the number of non-synonymous, rare variants found in any single patient is considerably high (typically 150-250 variants), the number of candidate variants can be reduced dramatically by sequencing additional, unrelated patients exhibiting the same phenotype. Only candidate variants from genes which bear putatively damaging variants in all examined patients are retained. In this setup, the sequencing of 3-4 patients is usually sufficient for candidate gene identification. An important point to consider is that this approach relies heavily on accurate phenotyping of the patients. If a patient gets misclassified due to a phenotype closely resembling the phenotype of the investigated disorder the approach will most likely fail to identify a common candidate gene.

De-novo Strategy: For disorders that are genetically highly heterogenous it is very unklikely to identify a common gene in any number of sequenced patients. However, the increased mutational target or these kind of disorders increases the possibility of causative de-novo mutations. These mutations can be identified by a family based strategy were the trio of father, mother and affected child are sequenced. Candidate variants can be rapidly identified by using the parents' exome data as control and thus filtering out all inherited variants. On average only 1-3 candidate variants remain for follow up studies employing this approach. A drawback of this strategy is a slightly higher error rate, as a false de-novo variant may have two different causes. Firstly, a genuine de-novo variant may be missed due to an error in the affected child's variant calling or secondly, an inherited variant may be falsly classified as de-novo because it was missed in one of the parents' variant calling.

3.2 Accuracy of Variant Calls

3.2.1 Sequencing Artifacts

Variant calling is a crucial step in the analysis of next generation sequencing data and there are certain caveats that need to be discussed, in particular sequencing artifacts. Sequencing artifacts are an in literature rarely debated problem of systematic errors that cause false positive SNV calls. These false positive SNV calls are identified by following criteria:

1. Variant alleles occur only on one strand
2. Following the variant base the read has a stretch of low quality

Figure 3.1 shows a typical example of such a sequencing artifact. The figure shows an image of the SAMtools tview text alignment viewer. The reference sequence is given in the first line while the consensus sequence called from the reads is given in the second line. The following lines show the actual read sequences, where dots indicate bases similar to the reference on the foward strands and commas represent matches to the reference genome on the reverse strand. Likewise, mismatches are given as the actual base letters either in capital (forward strand) or lower case (reverse strand) letters. The marked position shows the SNV call in question: Approximately half of the reads show the variant allele G, whereas the other half shows the reference allele A. This results in a heterozygote SNV call at this position, abbreviated as R in the consensus sequence. The variant call shows the criteria previously described, the variant allele is only indicated by reads from one strand, represented by the capital letters. Additionally, the scheme in the image is color coded for base quality, white indicates the highest base quality (Q30 and more), while the other colors denote lower quality bases: Yellow Q20-Q30, green Q10-Q20 and blue the lowest quality, Q2-Q10. All reads bearing the variant allele have large stretches of very poor quality bases following the variant allele, indicated by the grey box. During candidate evaluation by Sanger sequencing we verified a variety of SNV calls displaying these characteristics as false positives. This constitutes a not to be underestimated problem in analysis. To mitigate this problem we implemented additional filter criteria.

The additional filter takes into account the median base quality all read bases following the variant psoition. If the median base quality falls below a

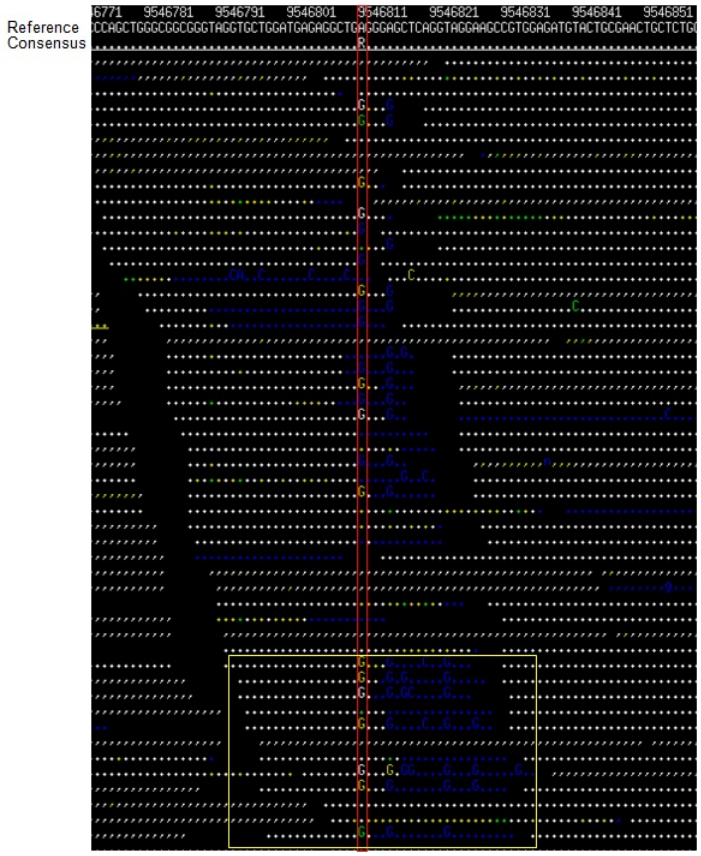


Figure 3.1: Sequencing artifact example. Red frame highlights a typical artifact, yellow box shows reads inducing the artifact.

user defined threshold (default Q10) the variant is marked as low quality and excluded for the remaining analysis. The variant is not excluded entirely, if the user chooses to also investigate low quality variants they may still be displayed in the database queries. Table 3.2 shows the results of applying this custom filter for one sample. As it is unfeasible to manually verify more than 20,000 coding variant calls, a different criteria to estimate the accuracy of varaint calls is employed: The percentage of variants in the called set that is already known from public ressources, in this case dnSNP (v132) and the International HapMap Project [18, 19, 20]. These known polymorphisms, especially the set discovered by HapMap, are more likely to represent true positives and thus a good prior for the accuracy of the calls. As one can see,

Table 3.2: Sequencing artifact filter

Calls	Coding SNVs	rsSNPs (%)	HapMap SNVs (%)
All SNVs	20924	91.1	37.3
After custom filter	19956	93.6	38.7
Filtered SNVs	968	38.7	9.5

the custom filter removes about 1,000 coding variant calls, increasing both the number of rsSNPs as well as the number of HapMap SNVs in the resulting call set. These numbers are significantly lower in the set of removed SNVs, implying that most of the removed SNVs are likely sequencing aritfacts and false positives. On the other hand, in the set of removed SNVs the percentage of HapMap SNVs is 9.5%, indicating that the additional filter also removes some true positive SNV calls. We estimate that on average the filter removes 4-7% of coding SNV calls per sample. Of these, two thirds are actual false positive calls, while 1/3 may represent genuine SNVs, constituting a problem as these are marked as low quality and are excluded from the analysis using default settings. Nevertheless, these SNVs can still be accessed by the users by modifying the search parameters. A combination of filters, or a multi-layered filter may be needed to further seperate true positive SNV calls from sequencing artifacts more accurately.

3.3 Enrichment Bias

The in-solution enrichment using oligonucleotide probes is prone to distinct biases. In particular, certain regions of the genome are easier to enrich than others. The binding affinity of the probes is influenced by sequence specific properties like GC-content and repetitivness of the target regions. These biases may result in an incomplete enrichment or an uneven coverage distribiution of certain target genomic regions. These biases are further evaluated in the following.

3.3.1 Coverage Distribution

A key parameter to evaluate the enrichment efficiency of the targeted exome capture is the percentage of bases that are covered at certain coverage levels. The average coverage of all targeted bases may convey a skewed image. For example a situation where half of the targeted bases are covered at 100X and the other half is not sequenced at all would result in an average coverage of 50X. To circumvent this problem the pipeline quantifies the percentage of targeted bases at four different coverage levels: 1X, 4X, 8X and 20X. These levels where chosen to roughly display the sequencing depths at which homozygote detection is reliable (4X), heterozygote detection is reasonably accurate (8X) and heterozygote detection is reliable (20X). In addition to the efficiency of the enrichment procedure the percentage of bases covered at certain coverage levels is also dependent on the total amount of sequence generated. The tracking of the coverage distribution over multiple samples is also used to detect possible saturation effects where more sequence does not yield further improvement in coverage levels.

Figure 3.2 shows the percentage of bases covered at the previously described levels for four different samples, dependent on the total amount of sequence generated. While almost all bases are at least covered once already at a low amount of sequence (4GB), the amount of bases suitably covered for heterozygote detection (<70% of bases at 20X for sample 54858) is still inadmissibly low. The amount of bases covered at 20X or more gradually increases with the amount of sequence generated until it saturates at ∼90% for 10-12GB of total sequence. A further increase of the sequencing amount does not yield additional bases covered at this level (sample 56302 in Figure 3.2). Interestingly, even if an excess amount of sequence is generated, regions with low coverage (1-4X) still persist. This suggests that the enrichment process is

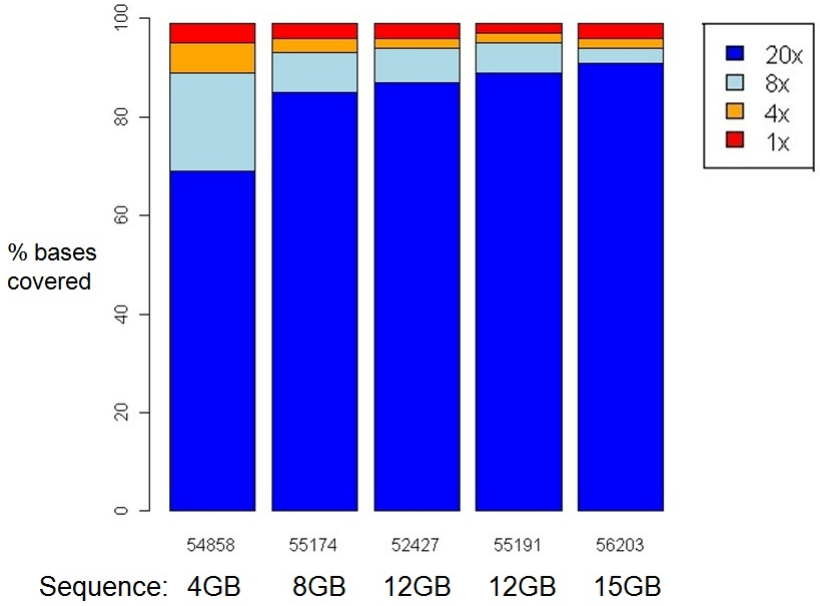


Figure 3.2: Coverage distribution at sequencing depths 1X, 4X, 8X and 20X for four different samples, depending on the amount of sequence generated.

biased in a way that certain regions are more difficult to capture than others, decreasing the capture efficiency for these regions. This effect is reproducible over multiple samples, resulting in the same regions being poorly covered in all samples.

3.3.2 Incomplete Enrichment

A further cause for inefficient capture of certain regions is the sensitivity of capture efficiency to GC content. A lower than average coverage is observed for regions from the edges of the GC content distribution, i.e. both GC rich and AT rich regions are generally poorly covered [27] in sequencing. The source of this bias is due to the PCR steps during the library preparation procedure. Regions with a very high (60% to 80%) or low (20%-40%) GC content cause reduced amplification efficiency, thus resulting in a lower sequence coverage.

This problem of general sequencing bias is further aggrevated in exome sequencing as it has been also shown that GC content influences the hybridization performance of oliginucleotides [48]. In consequence, exome se-

quencing suffers from the double burden of GC content bias in the library preparation as well as in the enrichment of the the targeted regions.

3.4 Statistical Analysis of 732 Exomes

The pipeline has been successfully employed in the analysis of 732 human exomes. In addition to the identification of potentially causative variants for a variety of disorders in these samples, aggregate analysis was performed on all of them to further characterize the 'average' human exome and the contained coding variants. It is estimated that the complete human exome contains approximately 19,000-23,000 coding variants [70]. The majority of these (18,000-22,000) constitute single nucleotide variants, small insertions and deletions comprise only a small proportion ($\sim$300) due to the high disruptive potential of these variants. Slightly more than half of the single nucleotide variants are expected to be synonymous changes, leaving on average $\sim$10,000 non-synonymous SNVs per human exome (ratio synonymous : non-synonymous changes 1.2 : 1). Figure 3.3 shows the average distribution of human exonic variation. A special case are de-novo variants that are unherited from the parents. On average only 0.8-1 de-novo variants are expected in each generation.

3.4.1 Average Exome Statistics

In total the pipeline was successfully employed for the analysis of 732 exomes. For 97 of them the targeted enrichment was performed using the Agilent 38 Mb kit, while the 50 Mb kit was used for the remaining 635 samples. On average, 7.5 GB of sequence were generated for the 38Mb samples, while 10.2 GB were sequenced for the 50 Mb samples. This results in a total amount of 7209 GB of sequence generated and analysed in the course of exome sequencing projects. About 97-98% of the reads could be mapped to the human reference genome (assembly hg19) using BWA and the pipeline as previously described. The percentage of duplicate reads that were removed prior to analysis was 3-5%. The percentage of reads counted as on target, i.e. reads with at least one base pair overlap with a targeted region, were gradually improved, starting from an average of 52.81% reads on target for the 38Mb kit to 77.67% reads on target with the 50 Mb kit. This results in an average coverage of targeted bases of 85X and 121X for the 38Mb

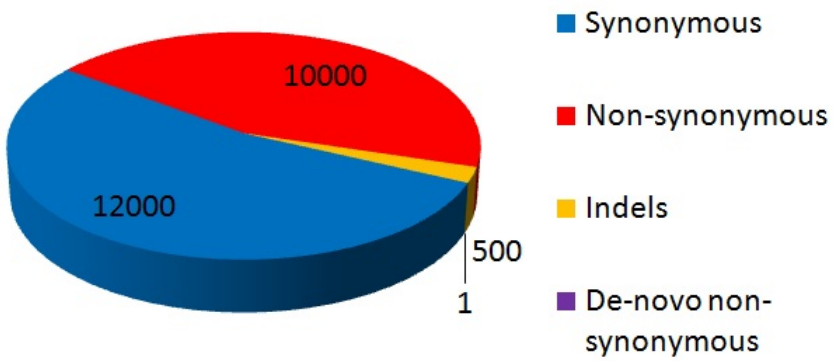


Figure 3.3: Average distribution of human exonic variation.

and 50 Mb samples, respectively. Approximately 1% of targeted bases are uncovered, while 96% / 97% are covered by at least 4 reads, 93% / 95% are covered by at least 8 reads and 82% / 89% are covered by 20 reads or more (number always given for 38Mb samples / 50 Mb samples). Detailed statistics ± standard deviation are shown in Table 3.3.

With these data we identified on average over 16,000 and 21,000 coding variants for the 38 Mb and 50 Mb samples, respectively. As expected, slightly over half of the single nucleotide variants were synonymous (8621 / 11081 on average), whereas the mean value of missene variants was 7333 and 10047 for the two kits. In addtion to the missense variants we identified 65 / 103 variants that result in premature stop codons, and 9 / 18 variants that dissolve a regular stop codon and extend the natural reading frame (stoploss variants). This results in ratios of synonymous : nonsynonymous SNVs of 1.17 : 1 and 1.10 : 1. This corresponds well to figures given in literature where the ratio is set between 1.1-1.2 : 1 [70, 73]. We further called 45 / 156 variants at canonical splice sites and 4394 / 6683 variants near splice sites (i.e. in a window of 20 bases into the intron). Variants near splice sites are annotated because a variation in close proximity to a splice site may affect splice affinity.

Table 3.3: Average exome statistics

	38 MB kit	50 MB kit
Samples	97	635
Total sequence (GB)	732.96	6572.48
Sequence per Sample ± sd	7.55 ± 2.41	10.20 ± 2.91
Mapped reads ± sd	98.36 ± 1.13	97.48 ± 4.98
Read on Target ± sd	52.81 ± 5.14	77.67 ± 5.77
Coverage ± sd	85.60 ± 26.93	121.62 ± 34.35
Targets uncovered ± sd	1.01 ± 0.48	1.35 ± 2.33
Targets covered $\geq$ 4X ± sd	96.33 ± 1.29	96.71 ± 7.28
Targets covered $\geq$ 8X ± sd	92.74 ± 2.34	94.82 ± 7.74
Targets covered $\geq$ 20X ± sd	82.11 ± 5.05	89.83 ± 7.96

Aside from single nucleotide variants 129 / 248 small insertions and deletions that result in frame-shifts of the corresponding reading frame and 113 /179 Indels that don't affect the entire reading frame (i.e. indels with a length multiple of 3) were called on average. UTR regions are generally not included in the targets for exome enrichment, nevertheless many reads and their mate-pair extend from the edges of the coding regions into UTR areas, yielding sufficient coverage for variant calling. Thus 1658 / 2831 variants are identified in 5' UTRs while 1358 / 2760 variants are called in 3' UTRs, on average.

To assess the quality of the variant calls the percentage of variants known from dbSNP and especially the HapMap project was assessed. It is expected that the majority of variants ($\sim$ 90% [70]) are already entered in dbSNP. We use HapMap variants as addtional quality measures as these variants have been called using very stringent conditions and are likely to represent true positives. 89.1% / 89.4% of the variants were previously known from dbSNP and 35.9% / 36.9% of variants constitute HapMap variants. All variant statistics are summarized in Table 3.4. The distribution of variants is not homogenoues and fluctuates greatly between different genes. The variant distribution and it's biases are further discussed in the following.

Table 3.4: Average variant statistics

	38 MB kit	50 MB kit
Coding Variants ± sd	16315 ± 1151	21803 ± 2010
Synonymous SNVs ± sd	8621 ± 347	11081 ± 1018
Missense SNVs ± sd	7333 ± 296	10047 ± 886
Nonsense SNVs ± sd	65 ± 8	103 ± 11
Stoploss SNVs ± sd	9 ± 2	18 ± 3
Frame-shift variants ± sd	129 ± 32	248 ± 51
Indels ± sd	113 ± 13	179 ± 27
Splice Site variants ± sd	45 ± 5	156 ± 14
Near Splice variants ± sd	4394 ± 493	6683 ± 640
5' UTR variants ± sd	1658 ± 158	2831 ± 332
3' UTR variants ± sd	1354 ± 192	2760 ± 296
dbSNP variants (%)	89.1	89.4
HapMap variants (%)	35.9	36.9

3.4.2 Variant Distribution per Gene

Figure 3.4 shows the distribution of potential damaging variants per gene, normalized to 1,000 amino acids. Putative damaging variants in this context comprise all non-synonymous SNVs, coding indels and splice site variants. On average each gene carries 29 of these variants per 1,000 aminoacids. The distribution shows high variation and ranges from 1,350 genes with no amino acid altering variants to genes with over 180 protein changing variants per 1,000 amino acids. Genes with an unusual high number of variants may interfere with the candidate gene identification as they appear to be candidates for a variety of disorders as all subjects have different low frequency variants in these genes. We identified several of them and these genes can be masked during initial analysis to reduce false positive candidates. These 'blacklisted' genes include for example *MUC4, MUC16, CDC27* and *PRAMEF4*.

Additionally, we investigated whether the variant distribution varies between those genes known to be involved in human heritable disease and those that are not. Figure 3.5 shows the proportion of genes from the Online Inheritance in Man database (OMIM genes) relative to the number of discovered variants from zero variants per gene to 50. The figure is capped at 50 as higher variant count classes consist of too few individual genes to be significant. Initially, one might think that OMIM genes may carry on average fewer

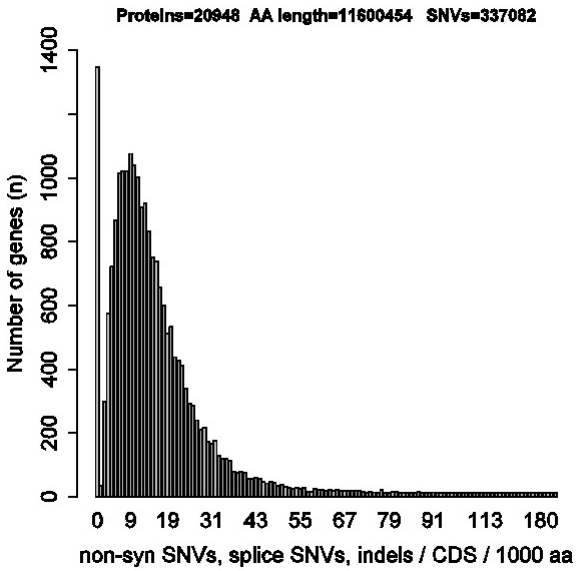


Figure 3.4: Number of non-synonymous variants per gene, normalized to 1000 amino acids

putativly damaging variants, however no significant difference between the distributions could be detected.

3.4.3 Comparison with Gene Mutation Database HGMD

The Human Gene Mutation Database (HGMD) [22, 45, 23] represents an attempt to collect all published mutations and variants responsible for heritable disease. These data comprise various types of mutation within the coding regions, splicing and regulatory regions of human genes, while somatic mutations and mutations in the mitochondrial genome are not included. Each mutation is entered only once in order to avoid confusion between recurrent and identical-by-descent lesions. The database focuses on amino acid altering mutations. Synonymous mutations are not recorded. In the case that these mutations are known to affect mRNA splicing or gene expression, or have been reported in significant association with disease, they may be included. The HGMD database aims to include only disease associated mutations. Evidence for their authenticity in a pathological context is evaluated following a catalogue of different lines of evidence:

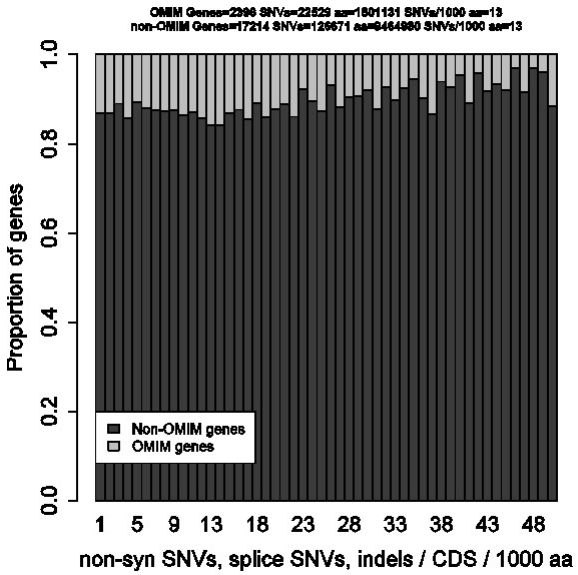


Figure 3.5: Proportion of OMIM genes and non-OMIM genes relative to the number of discovered potentially damaging variants per gene.

1. Absence in normal controls.
2. Novel appearance and subsequent cosegregation of the lesion and disease phenotype through the family pedigree.
3. Absence of any other lesion in the gene that could be responsible for the observed clinical phenotype.
4. Previous independent occurrence in an unrelated patient.
5. Non-conservative amino acid substitutions are more likely to disrupt protein function.
6. Location in a protein region of known structural or functional importance.
7. Location in an evolutionarily conserved nucleotide sequence and/or amino acid residue.
8. In vitro demonstration of reduced gene expression/mRNA splicing/activity or stability of protein product consequent to mutation.

9. Demonstration that the mutant protein has the same properties in vitro as its in vivo mutant counterpart.

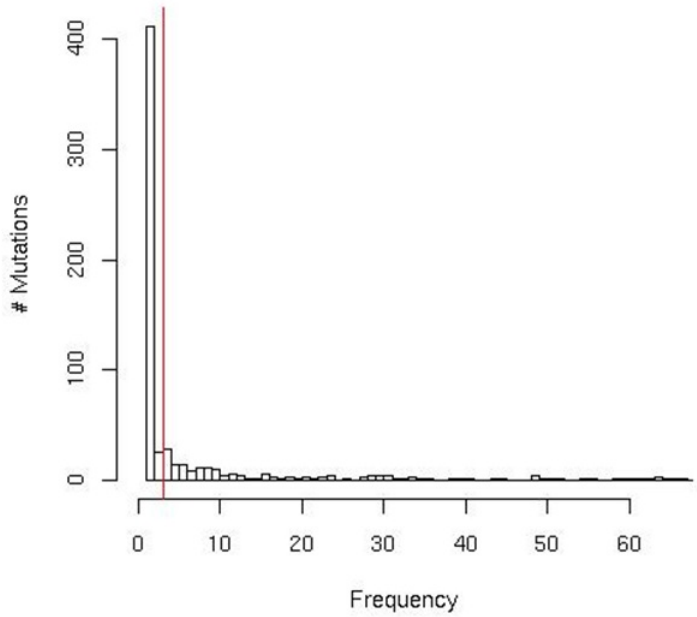
10. Reversal of the pathological phenotype in patient/cultured cells by gene replacement.

Despite the best efforts of the HGMD curators, it has to be assumed that some entries listed in HGMD are likely to include mutations that are not actually causative even though they have been reported as being so.

3.4.3.1 Carrier Burden and Identification of Literature Misannotations

The average carrier burden of severe recessive disease mutations for severe childhood recessive diseases was assessed in 327 DNA samples. All variants meeting the standard filtering criteria of the analysis pipeline described above and flagged as disease mutations in the HGMD database were enumerated. In total, we tested 327 samples for 26,890 HGMD mutations contained in 438 autosomal recessive genes [5]. We identified 699 mutations. Figure 3.6 A shows that most of them are present at a low frequency. However, 204 mutations were found in more than 2% of the samples (7 samples, indicated by the red line in Figure 3.6 A) and are therefore suspected to be not disease causing. These cases may represent misannotations in the HGMD database, however only 0.72% of entries are affected. These variants were excluded for further analysis. We then assessed the carrier burden in our samples using the remaining 495 low frequency mutations (Figure 3.6 B). The carrier burden of severe autosomal recessive mutations ranged from 0 to 12 per individual, the average carrier burden of these mutations was 3.6.

Additionally, new, putatively deleterious variants (variants in severe pediatric disease genes that create premature stop codons or coding frameshifts) were quantified: We identified 15 new non-sense mutations in the 327 samples investigated. All these mutations were heterozygous, as expected as, to our knowledge, none of our samples is affected by the corresponding disorder.



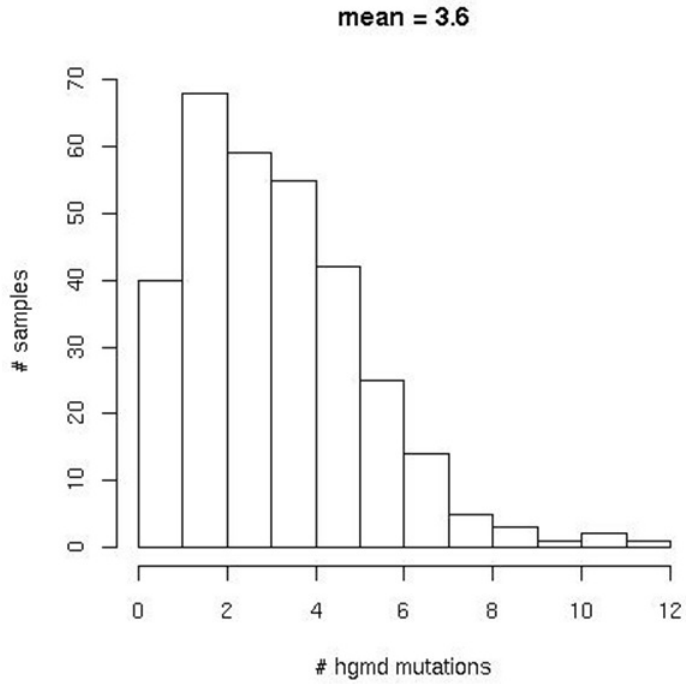


Figure 3.6: Frequency distribution of known HGMD mutations in 327 samples (A). Individual carrier burden of these mutations (B).

Chapter 4

Outlook

4.1 Third-Generation Sequencing

Next-generation sequencing technologies revolutionized the field of genomics and enabled a more complete undestanding of whole genome sequences, the transcriptome and the methylome. Yet there are still sequencing applications and aspects of genome biology that are presently beyond the scope of these sequencing technologies. These aspects could potentially be solved by a new generation of sequencing instruments, i.e. third-generation sequencing. Second-generation sequencing platforms are based on PCR-amplification of the DNA template, attaching the DNA to a solid surface and the generation of clusters of identical molecules and subsequent cycle wise sequencing and imaging to determine the nucleotide order of the template. In contrast, third-generation sequencing instruments interrogate single DNA (or RNA) molecules, circumventing the problems of PCR introduced bias and phasing. Additionally, sequencing is performed not cycle wise but in real-time by directly observing either the DNA polymerase or the nucleotides as they are incorporated in the DNA strand. As a result, run times are greatly decreased (from days to hours) and by avoiding chemistry cycles read lengths are significantly increased.

Third-generation sequencing technologies can be grouped into three distinct approaches:

1. Sequencing-by-synthesis where single DNA polymerases are observed as they synthesize a single DNA molecule. This approach is utilized by Pacific Biosciences single molecule real time sequencing.

2. Nanopore sequencing technologies where individual bases are detected as they pass through a nanopore, an approach currently developed by Oxford Nanopores (amongst others)

3. Direct imaging of single DNA molecules by advanced microscopy techniques.

Example technologies and instruments and their potentials and caveats are discussed in greater detail in the following.

4.1.1 Single-Molecule Real-Time Sequencing

Single molecule real time sequencing (also known as SMRT) is a parallelized single molecule DNA sequencing by synthesis technology developed by Pacific Biosciences [29]. Pacific Biosciences is a biotechnology company founded in 2004. Single molecule real time sequencing utilizes the zero-mode waveguide (ZMW), developed at Cornell University [51, 67]. The DNA sequencing is performed on a chip, comparable to the flow-cells employed by Illumina sequencing, that contains a multitude of ZMWs. The initial design comprised 3,000 ZMWs, but since then the number increased. Inside each ZMW, a single active DNA polymerase with a single molecule of single stranded DNA template is immobilized to the bottom of the ZMW. Light can penetrate the surface of the ZMW to create a visualization chamber that allows monitoring of the activity of the DNA polymerase at a single molecule level. The signal from a phospho-linked nucleotide incorporated by the DNA polymerase is detected as the DNA synthesis proceeds which results in the DNA sequencing in real time. For each of the nucleotide bases, there are four corresponding fluorescent dye molecules that enable the detector to identify the base being incorporated by the DNA polymerase as it performs the DNA synthesis. The fluorescent dye molecule is attached to the phosphate chain of the nucleotide. When the nucleotide is incorporated by the DNA polymerase, the fluorescent dye is cleaved off with the phosphate chain as a part of a natural DNA synthesis process during which a phosphodiester bond is created to elongate the DNA chain. The cleaved fluorescent dye molecule then diffuses out of the detection volume so that the fluorescent signal is no longer detected.

The zero-mode waveguide (ZMW) itself is a nanophotonic confinement structure that consists of a circular hole in an aluminum cladding film deposited on a clear silica substrate [44]. The ZMW holes are ∼70 nm in diameter and ∼100 nm in depth (Figure 4.1). When a nucleotide is incorporated

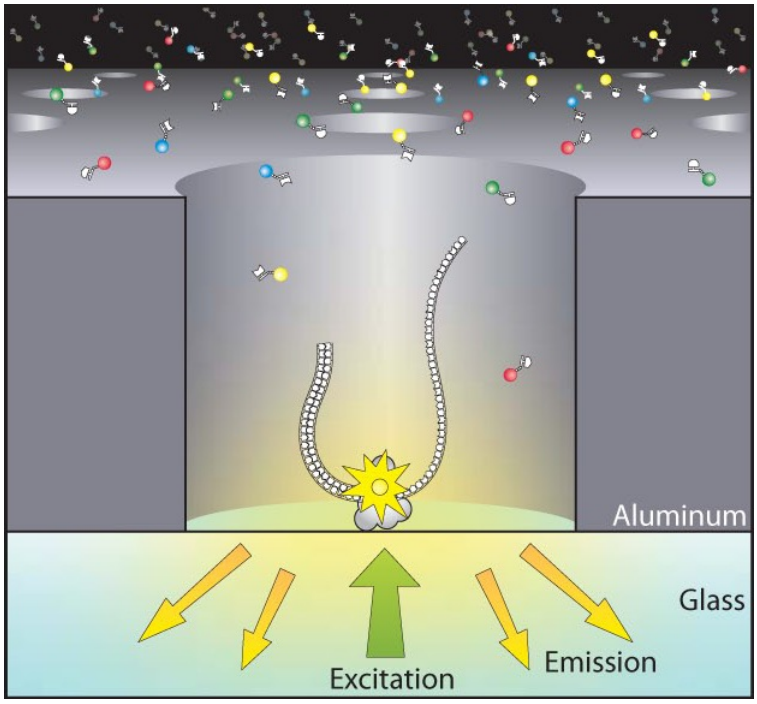


Figure 4.1: Schematic of Pacific Bioisciences SMRT sequencing technology: ZMW with polymerase and DNA template immobilzed at the surface. Image available at www.pacificbiosciences.com.

by the polymerase a flourescent 'pulse' is produced, lasting a defined time, then quickly fading from detection. This incorporation pulse is followed by a brief time of no activity, then the incorporation of the following nucleotide starts a new flourescent pulse. This pattern is detected and recorded in order to facilitate base calling (Figure 4.2).

Pacific Biosciences launched its first instrument, the PacBio *RS*, to a limited set of early access customers in 2010 and was publicly released in 2011. Currently, the instrument supports three unique modes of sequencing, standard sequencing, circular sequencing and strobe sequencing.

Standard sequencing. Standard SMRT sequencing generates single pass, long reads. This sequecing mode most resembles the second-generation sequencing standard. It mainly distinguishes iteelf from the former by considerably longer reads of up to 1,500 bp and the sequencing speed. While second-

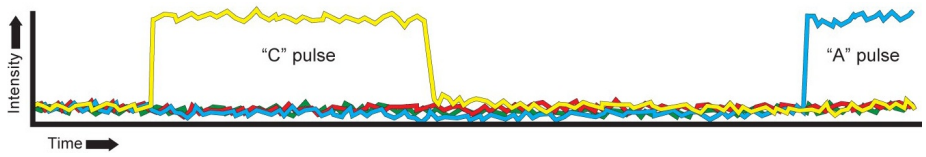


Figure 4.2: Pacific Bioisciences data analysis: Trace of flourescent pulses. Image available at www.pacificbiosciences.com.

generation sequencing instruments rely on cycle-wise sequencing where after each nucleotide incorporation the chemical reaction is halted to perform the imaging step. Here the sequencing and signal detection is performed in real-time while the polymerase continously incorporates nucleotides. This leads to a sequencing speed of approximately 10 bases per second.

Circular consensus sequencing. The circular consensus sequencing protocol uses a circular DNA template to enable multiple reads across a single molecule. This procedure significantly increases the accuracy of the consensus sequence as each template is sequenced multiple times. In comparison with established second-generation sequencing instruments the SMRT sequencing has a higher raw error rate which can be circumvented using the circular sequencing mode.

Strobe sequencing. One of the most innovative features of the PacBio *RS* is the so-called strobe sequencing. Strobe protocol increases physical coverage and effective readlength by 'strobing' the illumination on and off. This technique effectively extends readlength by minimizing polymerase damage resulting from continuous laser illumination. Data is then collected at user defined illumination intervals. When the illumination is off, sequencing continues at a predictable rate allowing to approximate the length of the unsequenced DNA stretch. This approach results in multiple sub-reads of varying lengths from a single molecule, which are interrupted by predictable long stretches of unsequenced DNA. This mode may be used primarily for de-novo sequencing and genome assembly where the great physical coverage of the strobe reads is most useful.

Specific analysis software and algorithms are provided by the developer suited to the unique nature of the sequencing reads. BLASR (Basic Local Align-

ment with Successive Refinement) maps reads to genomes by finding the highest scoring local alignment or set of local alignments between the read and the genome. The initial set of candidate alignments is found by querying a rapidly searched pre-computed index of the reference genome, and then refined until only high scoring alignments are retained. The base assignment in alignments is optimized and scored using all available quality information, such as insertion and deletion quality values.

ALLORA (A Long Read Assembler) is the PacBio de novo assembly algorithm. It is based on the open source assembly software package AMOS along with additional software components tailored to the long reads and error profile generated by the instrument. Allora uses a traditional overlap-layout-consensus approach to iteratively assemble reads into contigs, outputting these contigs as FASTA sequence.

The goal of EviCons (Evidence-based Consensus) is to produce the consensus sequence from a multiple sequence alignment. It can be employed to both mapped reads (resequencing) or contigs from de-novo assembly. Using empirical conditional probabilities and a likelihood ratio test, EviCons demarcates the multiple sequence alignment into regions of certainty and regions of uncertainty. For regions of uncertainty, EviCons uses base quality values to produce the best estimate of the local consensus sequence.

The RCCS (Reference Circular Consensus Sequencing) module is designed to call SNP and small indel variants against a reference sequence from the circular subreads for each single molecule. It uses a probability alignment algorithm testing all possible single base variants at a given location and determines the correct one using the likelihood calculated with the alignment model.

While providing several innovative features and improvements in comparison with second-generation sequencing instruments this technology also has some disadvantages when compared to the former. First, the raw accuracy of the reads is lower ($\sim$5% raw error rate). Additionally, the main source of errors are 'missed' bases which cause a false positive deletion in the reads when compared to the reference sequence. This complicates the alignment process as more gaps have to be introduced during this step. In second-generation sequencing the prominent error type are substitution errors which are considerably easier to deal with than deletion errors. On the other hand the raw sequencing throughput of second generation instruments is not yet reached.

4.1.2 Nanopore Sequencing

Nanopore sequencing refers to the approach of determining the nucleotide sequence of a template strand by detection of the bases as they move through a defined 'hole' or 'tunnel' (i.e. the nanopore) in a membrane. The base is then detected by either a measurable impact on the electric current or by an optical signal. Since this appraoch uses unmodified, single DNA molecules it circumvents previously discussed caveats of other sequencing platforms. Both biological nanopores constructed from preengineered proteins and synthetic nanopores are currently developed by different institutions and/or companies [88]. Some of the most promising approaches up to date are discussed in the following.

Oxford Nanopore. Oxford Nanopore is currently developing a sequencing system that relies on the direct, electrical detection of DNA molecules. In this approach a biological nanopore is utilized that is constructed from a modified α-hemolysin pore. The pore has an additional exonuclease attached on the extracellular surface of the pore. On the inside of the pore a synthetic cyclodextrin sensor is introduced [15, 94]. For sequencing, the system is contained in a liquid bilayer where the to be sequenced DNA template is loaded on the exonuclease side. By applying a voltage to the bilayer and changing the salt concentration the exonuclease subsequently cleaves off single nucleotides from the DNA strand. The cleaved bases are then detected by their characteristic alteration of the ionic flow as they pass one-by-one through the nanopore (Figure 4.3). The company has recently announced the development of two distinct sequencing devices employing their nanopore technology. The GridION sequencer has a modular, node based design similar to computing installations. A single node can be used as a benchtop sequencing device, while the combination of multiple nodes allows flexible scalability comparable to storage system solutions. Each node hereby operates with single-use cartridges that contain all necessary reagents. The second sequencing device, the MinIon could represent a major paradigm shift in next-generation sequencing by using the same technology in a miniaturized, USB-stick sized, portable sequencing instrument. The device contains all components to perform a single molecule experiment and transmits the generated data to a laptop or computer via USB port (Figure 4.4). Although, this approach looks very promising it remains to be seen how it translates into practice onced released.

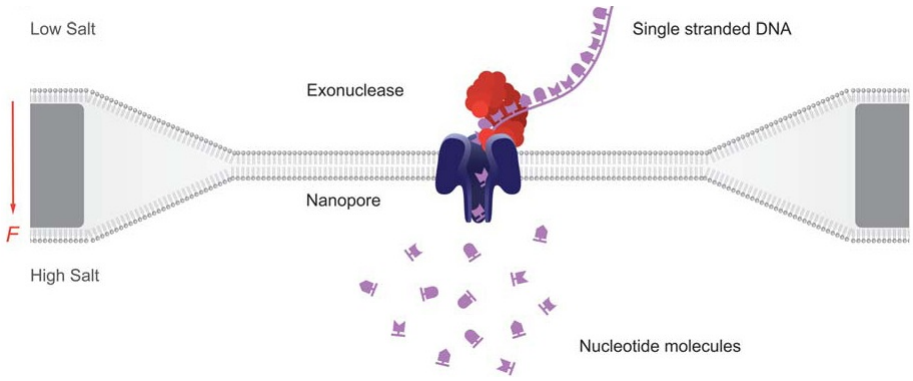


Figure 4.3: Oxford Nanopore sequencing approach. Single nucleotides are detected by electric signal as they pass through a biological nanopore. Image available at www.nanoporetech.com.

MspA Nanopore sequencing. A different approach employing a biological nanopore aims to directly sequence an intact strand of template DNA. In contrast to the Oxford Nanopore approach where each nucleotide is cleaved by an exonuclease and traverses the nanopore seperately, this approach utilizes a *Myobacterium smegatis* Porin A (MspA) protein as the core nanopore [26]. This protein allows for a complete molecule of single-stranded DNA to pass through. The current that passes the nanopore is then measured to gain a readout and determine the nucleotide sequence of the passing DNA template. Each individual base interupts the current in a distinguishable way, thus allowing the base composition to be determined. To accurately detect each base the transition of the single-stranded DNA molecule has to be slowed down as it travels through the nanopore. To achieve this travel speed reduction a region of double-stranded DNA is introduced, which up to date seems to be the crucial point in this technology and poses a significant obstacle to overcome for a finished instrument. On the other hand, if this appraoch matures and reaches production stage the possibility to directly sequence complete single-stranded DNA molecules without the need to break it down to the individual bases is a very attractive alternative to existing sequencing technologies [88].

Optical readout nanopore sequencing. The two previously described nanopore sequencing approaches both rely on a change in the current through the nanopore to gain a sequence readout. These approaches may prove dif-

Figure 4.4: Oxford Nanopore Sequencing Instruments: GridION (single node and combined rack with multiple nodes) and MinION sequencing devices. Image available at www.nanoporetech.com.

fcult to scale up to a highly parallel environment due to the need to monitore a large number of different nanopores. An alternative approach uses optical detection which is easier to achieve in a highly parallel instrument [66]. In this case each nucleotide is first biochemically converted to an ordered pair of concatenated oligonucleotides. In the next steps two different flourescent markers, so-called 'beacons' are hybridized to the converted DNA. These beacons are then subsequently removed from the DNA molecules and pass through the nanopore. Upon traversal through the nanopore the flourophores are released, generating two distinct flourescent pulses, that are subsequently detected by a high resolution optic system. In order to detect the flourescent signals of individual bases the process of beacon traversal through the nanopore is slowed by adjusting the voltage in the system. The basic prove of principle with this technology has been successful, though it remains to be seen if long read length and accurate sequencing can be achieved. One potential caveat of this system is the conversion of the DNA template to the pairs of oligonucleotides which may introduce bias if the conversion process is not flawless [88].

Transitor based nanopore sequencing. An example for a DNA sequencing technology that employs synthetic nanopores is currently under development by IBM [88]. The nanopore follows the architecture of an transistor and

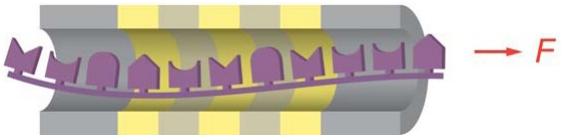


Figure 4.5: IBM nanopore sequencing approach. Single stranded DNA template strand passes through the transistor-nanopore. Individual bases are identified by unique electron signatures. Figure adapted from [88].

consists of alternating layers of metal and dielectric material. Similar to the MspA nanopore sequencing approach this 'transistor pore' allows the passing for complete, single stranded DNA molecules (Figure 4.5). The flow of the DNA molecule through the pore is controlled by modulation of the current in the electrodes of the transistor. This approach promises low cost by circumventing any advanced optics devices. Additionally, the technology uses label free, complete DNA molecules as substrate. Theoretically, unpreceded sequencing speed can be achieved though the main disadvantage up to date is the difficulty of distinguishing the singal of a single base from the signals of its surrounding nucleotides.

4.1.3 Sequencing by microscopy techniques.

Several efforts have been launched to accurately determine a DNA sequence by advanced electron micoscopy techniques. These approaches are still at an early phase of development and as to date no substantial proof of principle publication has been released. However, the basic approaches offer some promising advantages should they prove feasible in the future. These advantages consist of unbiased detection of unmodified bases (as now flouroescent groups are required) and very long read lengths, potentially in the scale of million base long reads, at a fairly low cost [88]. These approaches are driven by the principle that the best way to look at important biomolecules like DNA, RNA and proteins is to look at them directly (postulated by Richard Feynman 1959). Two example approaches are now highlighted in greater detail.

Halcyon Molecular microscopy sequencing. Halcyon Molecular is currently working on a technique to directly image and detect atoms that unambigously identify the different nucleotides in an DNA template strand

by employing transmission electron microscopy (TEM). It has already been demonstrated that different atoms can be distinguished with this approach [46]. To apply this technique to complete DNA molecules different, supporting technologies have to be developed. For example, a method is needed to stretch and attach DNA molecules to a substrate surface on which they can be imaged by TEM. The current state of this technique is still in early development. When the technology matures it remains to be seen what the specific advantages and disadvantages of this approach will be in comparison to other second and third generations sequencing instruments.

ZS genetics microscopy sequencing. A different TEM-based approach is pioneered by ZS genetics. In contrast to the former technology, ZS genetics employs labeled bases that are imaged using high-resolution, sub-angstrom electron microscopy. The individual nucleotide sequence of the DNA template is then identified based on their size and intensity differences between the labeled bases. Although no publication of proof-of-principle of this technology has surfaced, the developer claims the capability of sequencing 10,000-20,000 base reads at a rate of 1.7GB per day [88]. Whether this claim holds true remains to be seen when the technology is further developed.

4.1.4 Complete Genomics

A different entrance to the high-throughput sequencing market was made by a company called Complete Genomics [28]. In contrast to established sequencing companies like Illumina, Life Technologies or Pacific Biosciences, who each adopted the business model of developing a technology and then selling a finished instrument and the complementary reagents, Complete Genomics took a different approach. Even though they invented a unique sequencing strategy and constructed instruments to perform the sequencing, these instruments are not sold to the customer. Instead, Complete Genomics is solely a service provider where all sequencing projects are carried out on their in-house genome sequencing center. Additionally, bioinformatics analysis like mapping, assembly and variant calls are also performed by Complete Genomics and only the finished sequences, assemblies and/or variant reports are returned back to the customer.

The sequencing technology that was developed and employed by Complete Genomics is based on self-assembling DNA nanoarrays. The main aim of the approach is to reduce volumes and concentrations of the employed

reagents in comparison to existing next-generation sequencing instruments in order to reduce the total cost of the sequencing process. It has been recently demonstrated by the sequencing of three human genomes at average coverage levels of 45-fold to 87-fold that this technology can produce high quality genome sequences (1 false positive variant per 100kb) at a fairly low cost of $ 4,400 [28].

The sequencing procedure at first follows a standard library preparation with the shearing of the genomic template DNA in 400-500 bp long fragments and ligation of specific adaptor sequences to the fragments. The library preparation then follows a unique approach by circularizing the fragments, which are then replicated by rolling circle replication, resulting in many single-stranded copies of each DNA fragment. The DNA fragments then cocatenate head to tail into a single, long strand, and are folded into compact DNA nanoballs. These nanoballs are then attached to the surface of a flowcell, in this context called nano-array (Figure 4.6). The actual DNA sequence of the template is then determined by ligating flourescent probes, which interrogate specific positions in the DNA nanoball. First, oligonucleotide anchor DNA that is complementary to either the right or left end of one of the adapters is added to the flow cell. Next, T4 DNA ligase is added to a pool of four 10-mer DNA sequences that have degenerate nucleotides in all but one position and are added to the flow cell. Only the DNA probe with the complementary nucleotide in the interrogated position is able to bind to the template DNA. The flourescent group is released upon ligation of the matching probe by the T4 DNA ligase. The signal is recorded through a high-resolution optics device.

DNA nanoball sequencing technology offers several advantages over other sequencing platforms. One major advantage is the use of very high-density arrays. The array design permits one DNA nanoball to attach to each pit that is part of an ordered array, and therefore a higher concentration of DNA can be added. This allows a high percentage of the pits to be occupied by a DNA nanoball thus maximizing the number of reads per flow cell [28], compared to other sequencing technologies, for example Illumina, where the template DNA molecules are randomly attached to a flow cell. Another important advantage is that the sequencing reactions are non-progressive; after each reading of the probe, the probe and anchor are removed and a new anchor and probe set are added. Therefore, if a probe did not bind in the previous reaction, this has no effect on the next probe ligation, thus eliminating a major source of reading error [78]. The main disadvantages of this approach

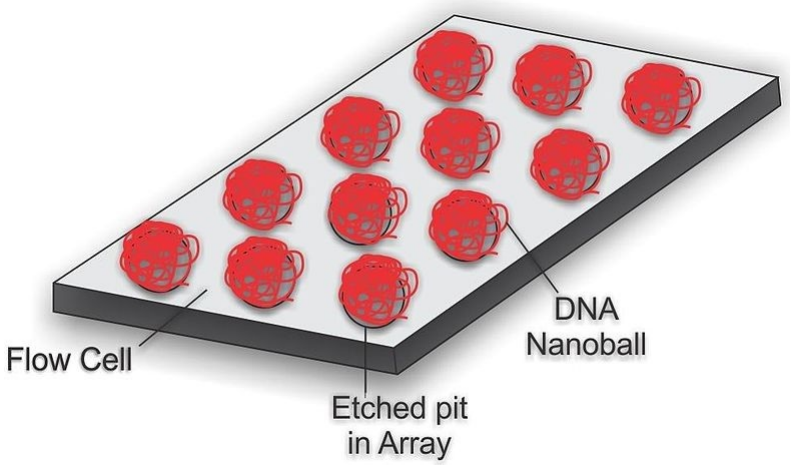


Figure 4.6: Sequencing setup for Complete Genomics technolgy: Template DNA is compacted into nanoballs which are attached to a sequencing flowcell. Figure adapted from [28].

is the fairly low read length in comparison to established instruments and the use of multiple PCR cycles during library preparation, which may introduce PCR bias and result in false positive variant calls.

4.1.5 Ion Torrent

In February 2010 Ion Torrent Inc. relaeased a novel sequencing technology based on ion semiconductors [76]. This technique is also referred to as pH-mediated sequencing or silicon sequencing. It follows a traditional sequencing-by-synthesis approach as a DNA template strand is complementarized, although the method of detecting the incorporated nucleotides differs substantially from other sequencing instruments. The incorporation of a deoxyribonucleotide (dNTP) into a growing DNA strand involves the formation of a covalent bond and the release of pyrophosphate and a positively charged hydrogen ion. The Ion Torrent system detects hydrogen ions as they are released during nucleotide incoporation by the DNA polymerase. Microwells on a semiconductor chip containing the to be sequenced template DNA strand and a DNA polymerase are subsequently flooded with a single type of nucleotide at a time. If the nucleotide is complentary to the template it is incorporated and relases a hydrogen ion. Beneath the layer of microwells

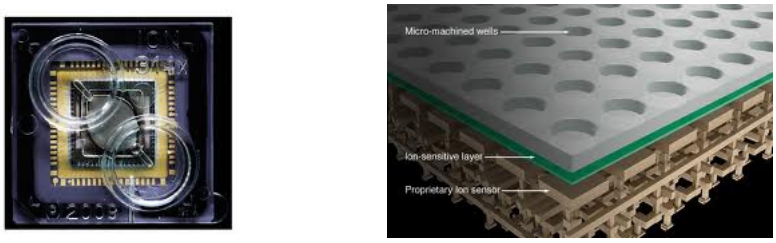


Figure 4.7: Ion Torrent sequencing technology: Semiconductor chip (version 314) (a) and schematic layout of chip design. Image available at www.lifetechnologies.com.

is an ion sensitive layer, below which is a hypersensitive ISFET ion sensor. Each released hydrogen ion triggers the ISFET ion sensor and the signal is translated into the DNA sequence of the template strand. All layers are contained within a CMOS semiconductor chip, similar to that used in the electronics industry (Figure 4.7).

The intial specifications given for the system were a throughput of 100Mb per run, with 50bp long reads, achieving a raw accuracy of 99.6%. As of February 2011 the read length was expanded to 100bp. One of the main strengths of the system is the sequencing speed as the measurements can almost be performed in real time (4 seconds per incorporation), resulting in a run time of approximately two hours. Additionally, the up-front sequencing costs are lower due to the unmodified nucleotides that are employed. In contrast to other sequencing instruments no optic devices are needed and unmodified nucleotides can be employed, as no flourescent signaling is required, circumventing potential biases. The throughput of the system is fairly low in comparison to other platforms but is alleviated by the short run time and can potentially scale up with further improvements of the semiconductor chips (as predicted by Moore's law). On the other hand, if the template DNA contains homopolymer runs multiple nucleotides get incorporated in one cycle, leading to the release of multiple hydrogen ions at once. This results in a proportionally higher signal, yet homopolymer runs are more difficult to accurately determine than with other sequencing instruments. Additionally, the throughput of the system is not yet suited for large-scale whole genome projects. If this technology aims to be employed in this kind of projects the throughput has to be increased substaintially to make it viable in such an environment.

4.1.6 Clinical Diagnostics and Personalized Medicine

As previously discussed exome sequencing can be successful in the identification of causative mutations in genetic disorder. Consequently, there is considerable interest in implementing this application for use in clinical diagnosis. The first, although unexpected, molecular diagnosis was made in early exome sequencing study [13]. The authors were able to reach an unanticipated genetic diagnosis of congenital chloride diarrhea in a patient with a suspected diagnosis of Bartter syndrome, a renal salt-wasting disease. The molecular diagnosis was based on the finding of a homozygous missense mutation in the gene *SLC26A3* in the vicinity of a known locus for congenital chloride diarrhea which was confirmed by a clinical follow-up. Recently a larger study attempted to evaluate the application of whole exome sequencing in clinical diagnosis by examining 12 patients with unexplained genetic conditions by sequencing both the patients and their unaffected parents [69]. In conclusion a likely genetic diagnosis could be made for 6 or the 12 probands, including the identification of causative mutations in 4 genes known to be related in genetic disorders and one gene which is likely related to a mendelian disorder. Additionally, the authors could identify a homozygous mutation which can explain at least a portion of the patients phenotypical features. These studies show that next-generation sequencing in particular and exome sequencing in general may become a standard tool in clinical diagnostics, yet a multitude of challenges remain for this transition to occur. These challenges include cost and reimbursement, communication of results to the patients and their families and the handling of chance findings. Especially the question of how a researcher or clinician should handle a molecular diagnosis or the identification of a risk factor for a completely different disorder from the one that was under initial investigation? Aside from these questions two main challenges are highlighted: First, in the opinion of the authors laboratory-based functional analysis of candidate variants is still a crucial part of determining causative mutations in a larger set of identified variants, and it remains unclear how this functional analysis can be incorporated into a diagnostic setting. Secondly, the authors state that even though a likely genetic diagnosis could be reached for 6 of 12 probands the identification of causative mutations still required substantial manual inspection of sequence data, alignments, variants and candidate genes, explicitly illustrating the need to further develop and implement standardized, automated analysis software suited for use in a clinical setting [69]. In addition the multitude of sequencing instruments

available renders the selection of the approriate instrument for each application increasingly difficult, in particular since unbiased comparisons between different devices are scarce, and it remains to be seen how the introduction of new methodologies further changes this fast-paced field.

4.1.7 Summary and Conclusion of Next-Generation Sequencing Instruments

Since the introduction of the second-generation sequencing instruments the field of DNA sequencing is constantly and rapidly evolving as even more technologies and sequencing approaches (as discussed in the previous chapter) become available on the market or near completion. An era of a single dominant sequencing technology as gold standard which is constantly improved and refined, like Sanger sequencing, seems highly unlikely today. The development is rather going in many different directions where multiple sequencing technologies have their own niche applications where they excel in comparison to other techniques.

Unbiased comparisons between different platforms is a difficult task as defined standards are non-existent and all companies release data and statements that cast their systems in the best possible light. There are no accepted standards for what measures the companies need to report, let alone particular protocols and procedures of how the data is analysed. The templates used, types of pre-analysis data filters used and number of runs used (e.g. best single run, average of many runs, etc.) can have significant impacts. Independent testing of NGS platforms to determine yield, error rates, etc. would be ideal, but is expensive and problematic. These problems are further aggrevated by the pace of development in this sector. Chemistry, software and other key components of the systems are frequently updated, rendering comparisons obsolete in a short time. A minimum set of characteristics include run time, number of reads generated, length and type of reads (i.e. single or paired-end reads), total yield per run, cost per run and cost per megabase of sequence. Table 4.1 summarizes key parameters for the different sequencing instruments that are currently available on the market [35]. Everyone using NGS data would benefit from the development of a standard set of conditions, analysis and a template (possibly Escherichia coli genomicDNA) or set of templates (e.g. specific clones, E. coli genomic DNA, human or mouse genomic DNA) that could be adopted and used for

testing of all platforms. Results from these templates could then be used to determine values that would allow direct comparison of NGS platforms, chemistry and software upgrades.

It is already obvious that each technology has advantages and disadvantages and thus focuses on diverse main fields of application. Some techniques have their strengths in very long reads, especially the Roche 454 series and the recently launched PacBio RS, which will be primarily employed in denovo sequencing where the longer read length is thought to help with the assembly of unknown genomes. On the other hand there are the sequencing machines that excel in pure throughput and thus reach an unprecedented low cost per sequencend base which should be the prime choice for large scale genome resequencing projects like the 1000 Genomes project or the International Cancer Genome Project. The most cost efficient machines currently are the Illumina HiSeq and LifeTechnologies SOLiD series. Another set of instuments focus on very short runtimes and aim at smaller labs which don't have the need and the capability to invest in a genome scale sequencing machines. These sequencers are generally smaller bench-top instuments and the run time is typically as fast as one day or even several hours. This enables labs a short turn around time for samples and experiments that don't require whole genome sequencing. Example applications include targeted amplicon sequencing of a moderate number of genes (50-200), RNA expression profiling and ChIP-Seq. Instuments in this category are the Roche 454Jr., the Illumina MiSeq and the Ion Torrent system of Life Technologies.

Table 4.1: Summary of Sequencing Instuments Performance

Instrument	Amplification	Run time	Reads (mil)	Read length	Yield (Mb)	Cost/run ($)	Cost/Mb ($)
3730xl (capillary)	PCR	2 h	0.000096	∼650	0.06	96	1500
454 GSJr. Titanium	emPCR	10 h	0.1	400	50	500	50
454 FLX Titanium	emPCR	10 h	1	400	500	6200	12.4
Illumina GAIIx	bridgePCR	14 d	320	150+150	96000	11524	0.12
Illumina MiSeq	bridgePCR	26 h	3.4	150+150	1020	750	0.74
Illumina HiSeq1000	bridgePCR	8 d	500	100+100	100000	10220	0.1
Illumina HiSeq2000	bridgePCR	8 d	1000	100+100	200000	20120	0.1
Illumina HiSeq2000 (v3)	bridgePCR	10 d	3000	100+100	600000	23500	0.04
SOLiD 4	emPCR	8 d	840	50+35	71400	8128	0.11
SOLiD 5500 (PI)	emPCR	8 d	700	75+35	77000	6101	0.08
SOLiD 5500 (4hq)	emPCR	8 d	1410	75+35	155000	10500	0.07
Ion Torrent 314	emPCR	2 h	0.1	100	10	500	50
Ion Torrent 316	emPCR	2 h	1	100-200	100	750	7.5
Ion Torrent 318	emPCR	2 h	4-8	100-200	1000	925	0.74
PacBio RS	none	0.5-2 h	0.01	800-1100	5-10	110-900	11-180

The continous development of these diverse instruments and application, each with specific strength and weaknessed/biases, also create a challenging environment for Bioinformatics that calls for the improvement of existing and implementation of new algorithms to facilitate data analysis. Problems and challenges that need to be addressed in the future include:

1. Primary data analysis: New methodology of sequencing warrants implementation of new analysis algorithms (for example base callers) which are tailored to the unique properties of the different sequencing platforms.

2. Data integration: Many projects may employ a variety of different instruments and the different data needs to be converted to standard formats in order to be combinable and exchangable between different labs.

3. Data storage: Efficient techniques of data storage to cope with the tremendous amount of raw (base calls) and secondary (variant calls) sequencing data generated need to be developed to make the results of large sequencing projects available to the scientific community.

4. Data comparison: Standard algorithms, procedures and filters for analysis need to be formulated in order to enable unbiased comparison between different platforms.

4.2 Appendix - Manuscripts

4.2.1 Identification of recurring tumor-specific somatic mutations in acute myeloid leukemia by transcriptome sequencing.

Leukemia (2011), 1–7

www.nature.com/leu

ORIGINAL ARTICLE

Identification of recurring tumor-specific somatic mutations in acute myeloid leukemia by transcriptome sequencing

PA Greif[1,2,5], SH Eck[3,5], NP Konstandin[1,2,5], A Benet-Pagès[3], B Ksienzyk[1,2], A Dufour[2], AT Vetter[1], HD Popp[2], B Lorenz-Depiereux[3], T Meitinger[3,4], SK Bohlander[1,2,6] and TM Strom[3,4,6]

[1]Clinical Cooperative Group 'Leukemia', Helmholtz Zentrum München, German Research Center for Environmental Health, Munich, Germany; [2]Department of Medicine III, Universität München, Munich, Germany; [3]Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany and [4]Institute of Human Genetics, Technische Universität München, Munich, Germany

Genetic lesions are crucial for cancer initiation. Recently, whole genome sequencing, using next generation technology, was used as a systematic approach to identify mutations in genomes of various types of tumors including melanoma, lung and breast cancer, as well as acute myeloid leukemia (AML). Here, we identify tumor-specific somatic mutations by sequencing transcriptionally active genes. Mutations were detected by comparing the transcriptome sequence of an AML sample with the corresponding remission sample. Using this approach, we found five non-synonymous mutations specific to the tumor sample. They include a nonsense mutation affecting the *RUNX1* gene, which is a known mutational target in AML, and a missense mutation in the putative tumor suppressor gene *TLE4*, which encodes a RUNX1 interacting protein. Another missense mutation was identified in *SHKBP1*, which acts downstream of FLT3, a receptor tyrosine kinase mutated in about 30% of AML cases. The frequency of mutations in *TLE4* and *SHKBP1* in 95 cytogenetically normal AML patients was 2%. Our study demonstrates that whole transcriptome sequencing leads to the rapid detection of recurring point mutations in the coding regions of genes relevant to malignant transformation.
Leukemia advance online publication, 22 February 2011; doi:10.1038/leu.2011.19
Keywords: acute myeloid leukemia; point mutations; TLE4; SHKBP1; RUNX1; transcriptome sequencing

Introduction

Acute myeloid leukemia (AML) is the most frequent hematological malignancy in adults, with an annual incidence of 3 to 4 cases per 100 000 individuals. Despite the increasing knowledge about the molecular pathology of AML, the prognosis remains poor, with a 5-year survival of only 25–30%. Chromosomal aberrations in tumor cells are found in approximately half of the AML patients, whereas the other half of the patients has a normal karyotype (cytogenetically normal-AML).[1] Even though a growing number of submicroscopic genetic lesions is identified in AML, about 25% of cytogenetically normal-AML patients do not carry any of the currently known mutations. The list of frequently affected genes includes the receptor tyrosine kinase *FLT3*, the transcription factor *CEBPA*, the human trithorax homolog and histone methyltransferase *MLL* and nucleophosmin (*NPM1*).[2–7]

So far, most of the genes that were found mutated in AML were found through a candidate gene approach, because of their involvement in translocations or in hematopoetic differentiation. For example, *CEBPA* knockout mice show a block in myeloid differentiation, and both *MLL* and *NPM1* were initially found to be involved in fusion genes that resulted from chromosomal translocations in leukemia patients.[5–9]

With the advent of next generation sequencing technology, the unbiased detection of tumor-specific somatic mutations became possible.[10–15] Sequence analysis of an AML genome resulted in the identification of recurring mutations in the gene *IDH1*, encoding the enzyme isocitrate dehydrogenase 1.[11] Metabolite screening of AML samples revealed that the related enzyme IDH2 is another mutational target.[16] Despite its technical feasibility, whole genome sequencing is still cost intensive, and therefore several alternative approaches of targeted sequencing have been proposed, like the sequencing of coding regions. Although the size of a diploid human genome is about 6 Gbp, the transcriptome, as defined by the combined length of all mRNAs in a cell, is only 0.6 Gbp in size. This figure is based on the estimate that a cell contains about 300 000 transcripts, with an average length of 2000 bases.[17,18] Sequencing of only a few gigabases of the transcriptome should allow mutation detection in a large proportion of transcribed genes. Here we report that sequencing of an AML tumor and the corresponding remission transcriptome allowed us to analyze approximately 10 000 genes and to identify five tumor-specific somatic mutations.

Materials and methods

Case information

A diagnostic bone marrow sample was collected from a 69-year-old patient, diagnosed with AML M1 in May 2008. The patient was included in the AML Cooperative Group clinical trial, and informed consent and ethical approval for scientific use of the sample including genetic studies were obtained. After induction therapy using the sequential high-dose cytosine arabinoside and mitoxantrone (S-HAM) protocol, complete remission was achieved. After leukocyte recovery in July 2008, a remission sample from peripheral blood was taken.

Sample preparation

Approximately 50×10^6 cells from each sample were used for mRNA extraction using Trizol (Invitrogen, Carlsbad, CA, USA).

Correspondence: Professor SK Bohlander, Department of Medicine III, University of Munich, Marchioninistr 15, Klinikum Grosshadern, Munich, Bavaria 81377, Germany.
E-mail: bohlander@helmholtz-muenchen.de
[5]These authors contributed equally to this work.
[6]Joint senior and corresponding authors.
Received 17 October 2010; revised 28 November 2010; accepted 10 December 2010

The sequencing library was prepared using mRNA-Seq sample preparation kit (Illumina, San Diego, CA, USA). In brief, mRNA was selected using oligo-dT beads (dynabeads, Invitrogen). The mRNA was then fragmented using metal ion hydrolysis and reversely transcribed using random hexamer primers. Following steps included end repair, adapter ligation, size selection and polymerase chain reaction enrichment.

Sequence alignment

Short-read alignment and consensus assembly were performed using the BWA (v.0.5.5) sequence-alignment program,[19] with the default parameters and interactive trimming of low quality bases at the end of reads (cut-off quality value $q=15$). We used an expanded reference sequence comprising the human genome assembly (build NCBI36/hg18) and all annotated splice sites extracted from the University of California Santa Cruz (UCSC) genome browser-known gene track. In total, we generated 127 115 919 paired-end reads of 36 bp length for the AML sample, of which 95.08% aligned to the reference sequence, and 187 782 678 paired-end reads for the remission sample with 82 % aligning to the reference. Read mapping, subsequent assembly and variant calling were performed using the resequencing software packages BWA and SAMtools.[19,20] During alignment, 31.27 and 39.81% apparently duplicated reads were removed from the AML and remission sample, respectively.

Distribution of reads across exonic and non-exonic regions

To determine the success of the RNA library preparation, we calculated the percentage of reads matching to known exons from the UCSC genome browser. For the AML sample, ~63% of reads aligned to exons, ~28.5% to introns and ~7.5% to intergenic regions, whereas for the remission sample, ~73.5% of reads aligned to exons, ~20.5% to introns and ~6% to intergenic regions (Figure 1b). The relatively high proportion of intronic reads may stem from unspliced mRNAs. Variable proportions of intronic and exonic reads were observed between different preparations from the same samples, indicating that minor differences in RNA concentration and quality might strongly influence the competitive binding of shorter spliced and longer incompletely spliced mRNAs to oligo dT-beads. The values varied between the different chromosomes and the number of reads mapping to exons were correlated with overall gene density on the chromosome (Supplementary Figure S1).

Expression analysis

Expression values were calculated as RPKM (reads per kilo-base of gene model per million mapped reads.[21] In brief, the number of uniquely mapping reads (BWA mapping quality >0, ~75 to 85% of reads for both samples) for each gene was counted and then normalized by gene length and the total number of reads generated in the experiment. As the reference set,

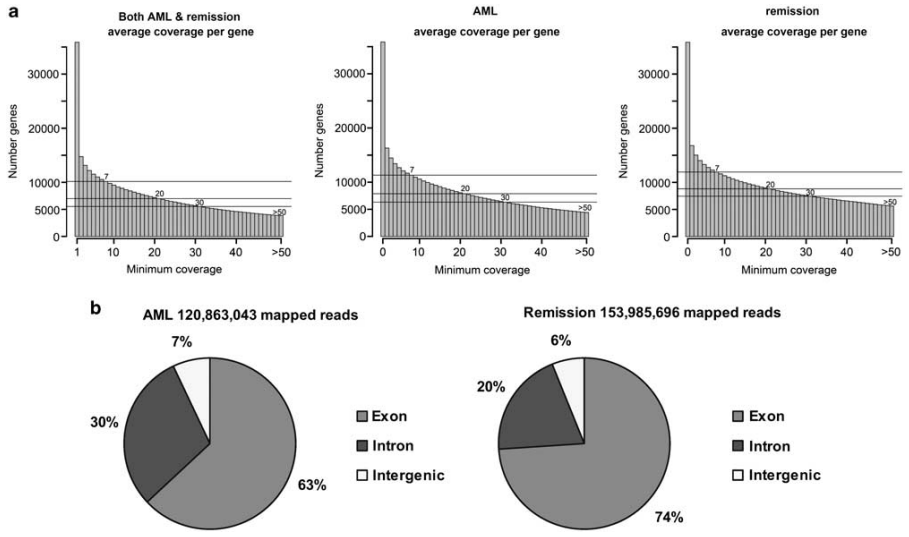


Figure 1 (**a**) Histograms of the sequence coverage in a non-redundant gene set based on the Ensembl annotation (35 876 genes) for genes detected in both samples (left), the acute myeloid leukemia (AML, middle) and remission (right) samples. Minimum sequence coverage is plotted on the x-axis and number of genes is plotted on the y-axis. We sequenced 10 152 genes with an average coverage of 7 or greater, 6989 genes with an average coverage of 20 or greater and 5535 genes with an average coverage of at least 30 in both samples (left). The result obtained from the AML sample was 11 293 genes with an average coverage of 7 or greater, 7878 genes with an average coverage of 20 or greater and 6326 genes with an average coverage of 30 or greater (middle). The sequencing of the remission yielded 11 906 genes with an average coverage of 7 or greater, 8805 genes with an average coverage of 20 or greater and 7446 genes at an average coverage of at least 30 (right). The high proportion of genes detected in both samples indicates a good comparability of expression profiles. (**b**) Two pie charts showing the percentage of reads from the AML (left) and remission samples (right) that map to exons, introns or intergenic regions (see also Supplementary Figure S1).

we used a non-redundant gene set based on the Ensembl gene annotations by merging all annotated transcripts from the same gene into a single 'maximum coding sequence'. This set contained 35 876 genes. Exonic regions that were shared by two or more different genes (for instance sense and anti-sense transcripts or non-coding RNAs within exons) were excluded and not used for RPKM calculation as reads from these regions can not be unambiguously assigned to single genes.

Spearman's rank correlation coefficient

Spearman's rank correlation coefficient was calculated from the $\log_2$ RPKM values of the tumor and remission sample, using the R package for statistical computing.

SNP calling

Variant calling was performed using the SAMtools package (v.0.1.5c).[20] For the variant filter of SAMtools, we used the following settings: minimum read depth = 3; maximum read depth = 9999; minimum root mean square mapping quality for single-nucleotide polymorphisms (SNPs) = 25; minimum mapping quality of gaps = 10; minimum indel score for filtering = 25; window size around potential indels = 10; window size for filtering dense SNPs = 10; maximum number of SNPs allowed in window = 2.

Subsequently, we applied additional filters. We required each putative SNP to have (i) a median quality value of the variant bases of at least 20 (ii) that at least 15% of all reads covering the position show the variant allele and (iii) that at least 10% of reads showing the variant allele are from opposite strands.

Functional analysis of SNPs was performed with custom Perl scripts using data sets from Ensembl and the UCSC genome browser. Known SNP locations, Ensembl and known gene annotations were used as provided by the UCSC genome browser.

Results

To demonstrate the feasibility of this approach, we selected an AML sample (bone marrow aspirate) and a corresponding remission sample (peripheral blood) for transcriptome sequencing. The patient, a 69-year-old female, presented with *de novo* AML, with blood counts and bone marrow morphology being consistent with the diagnosis of AML without maturation according to the French-American-British classification (FAB AML M1). After induction therapy, complete remission was achieved. One year after initial diagnosis, the patient relapsed and received an allogenic bone marrow transplant.

Conventional cytogenetic analysis revealed a normal female karyotype (46, XX[20]). An internal tandem duplication of *FLT3*, an *NPM1* mutation and a partial tandem duplication in the *MLL* gene were excluded in a routine diagnostic screen. We further investigated whether the tumor sample contained somatic copy number variations using the HumanOmni1-Quad chip (Illumina), containing probes for approximately 1 million loci. We found no evidence of somatic loss-of-heterozygosity indicating the presence of a normal diploid genome. A total of 29 copy number changes were present in both the tumor and remission sample. We compared the copy number variations with those contained in the database of genomic variants and 1600 controls from a population-based study. All the copy number variations were present at least once in these cohorts.

We sequenced 4.35 and 5.54 Gbp of the tumor and remission sample, respectively, on an Illumina GA IIx sequencer (Illumina). We used the NCBI36/hg18 genome assembly as reference sequence and compiled a non-redundant mRNA set from the Ensembl transcripts database resulting in a set of 35 876 genes. Read mapping to the reference genome was performed with the BWA software.[19] Approximately 95 and 82% of the reads mapped to the reference, of which 63 and 74% mapping to exonic sequences in the tumor and remission sample, respectively (Figure 1b, Supplementary Figure S1).

The average sequence read depth for every gene was first calculated to obtain the number of genes suitable for mutation detection. The read depth per gene ranged from 0 to over 1000. A total of 10 152 genes had an average read depth of at least sevenfold and 6989 genes had an average read depth of 20 or greater in both samples. These numbers were only slightly higher when the tumor and remission samples were analyzed individually, indicating that the gene expression pattern was comparable even though the tumor sample was a bone marrow aspirate with more than 90% blasts, whereas the remission sample was from peripheral blood with a normal white blood cell count (Figure 1a). The comparability was supported by a high correlation of the gene expression levels between the samples as shown by a Spearman Rank correlation coefficient of 0.82 (Figure 2a).

Single-nucleotide variants (SNV) were called with the SAMtools software package,[20] using mainly the default parameters and custom filters applied at later stages. To achieve a low false-positive rate, we required a minimum read depth of 7 in both samples. We set this threshold because there is a detection rate of approximately 70% at this read depth.[22] For the same purpose, we quality filtered the SNV set of the tumor sample, but used an unfiltered set of the remission sample for comparison (Figure 2b).

Quality filtering in the tumor resulted in a set of 8978 SNVs in coding regions. This compares favorably with approximately 20 000 SNVs that can be found in the entire coding sequence using exome sequencing.[23] In the next step, we excluded all coding SNVs that were present in the dbSNP database version 130 or in the exomes of 8 HapMap samples. The remaining 926 sites contained 612 SNVs, which led to an amino acid substitution or, which disrupted canonical splice sites. These 612 SNVs were then compared with the unfiltered calls of the remission sample at these 612 positions. We excluded all positions with any indication that the same SNV was also present in the remission sample.

This strategy resulted in the identification of 11 candidate SNVs unique to the tumor sample. Capillary sequencing of genomic DNA from both the tumor and the remission sample confirmed five SNVs, which affected the genes *RUNX1*, *TLE4*, *SHKBP1*, *XPO7* and *RRP8*. (Table 1, Figure 3). Two SNVs were false positives with the same heterozygous SNVs being also present in the genomic DNA of the remission sample, four SNVs could not be confirmed in the AML sample.

RUNX1 (AML1) carried a heterozygous stop mutation in the *Runt* domain. *RUNX1* is the fusion partner of *RUNX1T1* (eight twenty one (ETO)) in the recurring t(8;21) (q22;q22) translocation present in 8–13% of *de novo* AML cases.[24] In addition, point mutations in *RUNX1* have recently been described in AML, in particular AML secondary to myelodysplastic syndrome, radiation exposure or chemotherapy, at a frequency of 8–10%.[25]

TLE4 carried a missense mutation at position 511 (N511S). *TLE4* is located on chromosome 9 band q34, which is frequently deleted in AML with t(8;21) translocations, and is therefore a

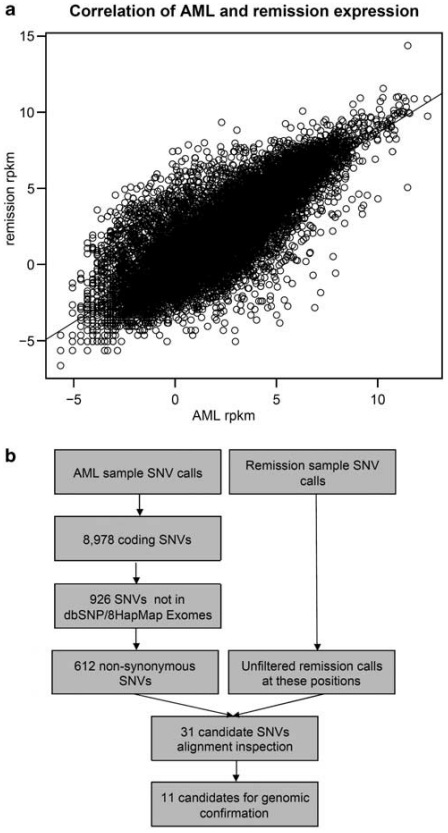


Figure 2 (**a**) Correlation between gene expression values (log2 RPKM) for the AML and remission sample for robustly detected genes (⩾20 unique reads in each sample): The Spearman Rank correlation coefficient between the two samples was 0.82. (**b**) Work flow overview for the mutation detection. AML SNVs were stringently quality filtered to a set of high confidence, non-synonymous SNVs ($n=612$). This set was then compared with the unfiltered remission results at these positions to exclude positions with evidence for the same SNV in the remission sample. The alignments of candidate SNVs (31 SNVs) were manually inspected, yielding the final candidate list (11 SNVs). Finally, five heterozygous point mutations could be confirmed by capillary sequencing.

putative tumor suppressor gene. Interestingly, the TLE4 protein interacts with RUNX1, and haploinsufficiency of TLE4 was shown to collaborate with the RUNX1/RUNX1T1 fusion to rescue cells from apoptosis.[26]

The third tumor-specific SNV resulted in a missense mutation (V89I) in *SHKBP1* (also known as SETA binding protein 1, *SB1*). Through SETA, SHKBP1 interacts with CBL,[27] a ubiquitin ligase that regulates the degradation of FLT3. *CBL* mutations, which result in the increased activity of FLT3, have recently been described in AML and myelodysplastic syndrome.[28] Thus, it is likely that *SB1* mutations affect FLT3 signaling. *SHKBP1* overexpression in cell lines has antiapoptotic effects.[29]

The fourth and fifth AML-specific mutations were missense mutations in *XPO7* (a member of the importin beta superfamily) and *RRP8* (a methyltransferase, possibly involved in ribosomal RNA processing).

Although recurring mutations in *RUNX1* are known to occur in AML, mutations in *TLE4* or *SHKBP1* have not been described before. We therefore screened the complete coding sequence of *TLE4* and *SHKBP1*, as well as of *RUNX1* in 95 cytogenetically normal-AML patients by capillary sequencing of genomic DNA (Table 2). As expected, we found several patients with *RUNX1* mutation (9/95; 9.5%): nine missense mutations (two patients with two mutations each), one nonsense mutation and a 5 bp insertion. We also discovered two missense mutations in *TLE4* and two missense mutations in *SHKBP1* (Table 2), strongly suggesting that both *TLE4* and *SHKBP1* are mutational targets in AML at a frequency of about 2%. Mutations in *TLE4, SHKBP1* and *RUNX1* were mutually exclusive in the cohort of 95 cytogenetically normal-AML patients. *TLE4* mutations were found in patients with mutations in *NPM1* and *C/EBPA*, whereas *SHKBP1* mutations were found in combination with mutations in *NPM1* and *FLT3* (Table 2).

Discussion

Our results demonstrate that whole transcriptome sequencing is an efficient method to discover point mutations in AML. Using stringent filtering criteria, we were able to identify just 11 candidate mutations from a total of almost 10 Gbp of primary transcriptome sequence. Five of these mutations were confirmed by sequencing of genomic DNA. Three of these mutations affect genes in pathways involved in AML pathogenesis (Figure 4). Although *RUNX1* mutations are known to occur at a frequency of about 8% in AML patients, we describe for the first time *TLE4* and *SHKBP1* as recurring mutational targets in AML. In summary, our approach proved to be extremely efficient in identifying recurring mutations with a high likelihood of contributing to the pathogenesis of AML.

Overexpression of *TLE4* in the *RUNX1/RUNX1T1 (AML1/ETO)* fusion-positive Kasumi cell line was reported to cause apoptosis and cell death, suggesting that *TLE4* may act as a tumor suppressor gene.[26] The missense mutations we identified may diminish the function of TLE4 or even act in a dominant negative fashion. The point mutations in *SHKBP1*, on the other hand, may result in a gain of function, because the antiapoptotic effects of its overexpression classify SHKBP1 as a putative proto-oncogene.[29] In AML, mutations in *SHKBP1* may disturb the degradation of the FLT3 tyrosine kinase through the interaction of SHKBP1 with SETA and indirectly with the ubiquitin-ligase CBL.[27] Although little is known about the protein structure of both TLE4 and SHKBP1, all point mutations found in the present study affect evolutionarily highly conserved domains encoded by neighboring exons (Table 2). Although biochemical assays are required to test whether these missense mutations influence the protein interactions between TLE4 and RUNX1 or SHKBP1 and CBL, *in vivo* transformation assays are required to elucidate the potential role of these mutations during the onset and progression of AML. Considering the increasing number of recurring mutations that have been identified in AML, it will be very challenging to understand their complex interplay.

Table 1 Confirmed tumor-specific somatic mutations identified by transcriptome sequencing

Gene	*Position (hg18)*	*Reference genotype*	*Tumor genotype*	*Amino acid*	*Ensembl protein*	*Read depth tumor*	*Read depth remission*
TLE4	chr9:81523675	A/A	A/G	Asn→Ser (N511S)	ENSP00000365735	167	114
SHKBP1	chr19:45775904	G/G	G/A	Val→Ile (V89I)	ENSP00000291842	81	249
RUNX1	chr21:35128760	G/G	G/A	Gln→Stop (Q208X)	ENSP00000300305	59	36
XPO7	chr8:21883756	A/A	A/G	Arg→Gly (R139G)	ENSP00000252512	52	44
RRP8	chr11:6579867	G/G	G/C	Ser→Cys (S85C)	ENSP00000254605	23	38

Abbreviation: AML, acute myeloid leukemia.
The table shows details of the five point mutations detected in the AML patient including affected genes, genomic position, resulting amino-acid change and sequence coverage of the affected sites (see also Figure 3).

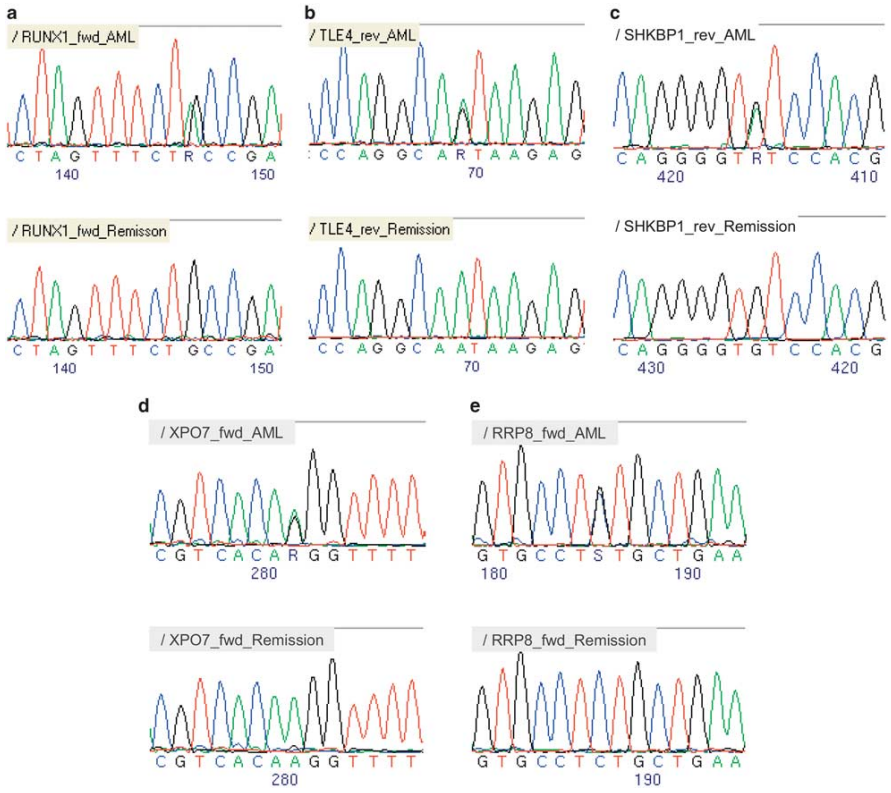


Figure 3 Sequencing of genomic DNA from the patient confirms five point mutations detected by whole transcriptome sequencing. The genes affected are *RUNX1*, (**a**) *TLE4*, (**b**) *SHKBP1*, (**c**) *XPO7* (**d**) and *RRP8* (**e**) Chromatograms show sequences from AML (upper panels) and remission (lower panels) for each gene.

Apparently, many subtle genetic changes may contribute to the disease through multiple interactions.

Although analysis of the two AML genomes required sequencing of over 120 Gbp for each patient and resulted in the detection of 10 to 12 tumor-specific mutations in the gene coding regions in each case,[10,11] our analysis of an AML transcriptome required only the sequencing of 10 Gbp and resulted in the identification of five tumor-specific mutations in the gene coding regions. Thus, our findings demonstrate that whole transcriptome sequencing might be an order of magnitude, faster and more cost effective than whole genome sequencing for the detection of point mutations in coding

Table 2 Mutations identified by capillary sequencing of 95 CN-AML patients

Gene	*Position (hg18)*	*Reference genotype*	*Tumor genotype*	*Amino acid*	*Ensembl protein*	*Number of pat. (ID#)*	*Additional mutations*
TLE4	chr9:81511889	G/G	G/A	Arg→His (R323H)	ENSP00000365710	1 (1)	NPM1
TLE4	chr9:81514412	A/A	A/G	Thr→Ala (T431A)	ENSP00000365735	1 (2)	C/EBPA
SHKBP1	chr19:45786394	G/G	G/A	Arg→Gln (R454Q)	ENSP00000291842	1 (3)	FLT3-ITD
SHKBP1	chr19:45788844	G/G	G/A	Arg→Gln (R672Q)	ENSP00000291842	1 (4)	NPM1
RUNX1	chr21:35181194	A/A	A/G	Leu→Ser (L56S)	ENSP00000300305	1 (5)	FLT3-ITD MLL-PTD
RUNX1	chr21:35181042	G/G	A/A	Arg→Cys (R107C)	ENSP00000300305	1 (6)	FLT3-ITD
RUNX1	chr21:35181032	T/T	T/C	Lys→Arg (K110R)	ENSP00000300305	1 (7)	—
RUNX1	chr21:35174747	C/C	T/T	Arg→Lys (R162K)	ENSP00000300305	1 (8)	MLL-PTD
RUNX1	chr21:35174739	C/C	C/A	Gly→Cys (G165C)	ENSP00000300305	1 (9)	FLT3-ITD MLL-PTD
RUNX1	chr21:35174846 chr21:35174847	A/A G/G	−/5bp Insertion	Frame shift (L129fs)	ENSP00000300305	1 (10)	—
RUNX1	chr21:35153730	A/A	A/G	Leu→Pro (L175P)	ENSP00000300305	1 (11)	FLT3-ITD NRAS
RUNX1	chr21:35153661	T/T	T/A	Asp→Val (D198V)	ENSP00000300305	1 (12)	—
RUNX1	chr21:35153662	C/C	C/T	Asp→Asn (D198N)	ENSP00000300305	1 (7)	—
RUNX1	chr21: 35153652	C/C	C/T	Arg→Gln (R201Q)	ENSP00000300305	1 (12)	—
RUNX1	chr21: 35153653	G/G	G/A	Arg→Stop (R201X)	ENSP00000300305	1 (13)	FLT3 D324

Abbreviation: CN-AML, cytogenetically normal acute myeloid leukemia.
The table shows details of the 15 mutations (affecting 13 patients) identified by capillary sequencing of *RUNX1*, *TLE4* and *SHKBP1* in 95 CN-AML patients, including affected genes, genomic position, resulting amino-acid change and Ensembl protein.

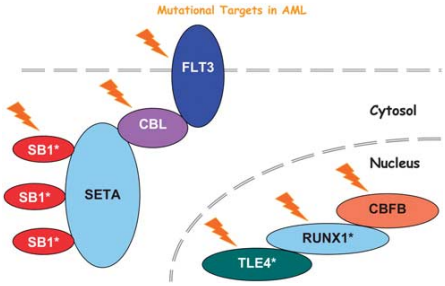


Figure 4 Illustration of proteins encoded by genes that are mutational targets in AML. The RUNX1 protein interacts with CBFB and TLE4. RUNX1 is the target of mutations in AML. Both RUNX1 and CBFB are involved in common chromosomal translocations in AML (t(8;21)(q22;q22) and inv(16)(p13q21)). In addition, one allele of TLE4 is frequently deleted in AML samples with a RUNX1/RUNX1T1(ETO) fusion. SB1 (SHKBP1) interacts through SETA with the ubiquitin-ligase CBL that mediates degradation of the receptor tyrosine kinase FLT3. Both FLT3 and CBL are frequently mutated in AML and/or myelodysplastic syndrome. The stars indicate the mutations in RUNX1, TLE4 and SB1, which were found in our study.

regions of expressed genes. The main limitation of transcriptome sequencing is the representational bias of transcripts. Considering recent reports of alternative cleavage and polyadenylation of oncogenic transcripts,[30] sequencing of reversely transcribed poly-A selected transcripts may not always correctly reflect the original expression levels in the leukemia cells. Moreover, mutations that lead to increased-mRNA decay might be missed in the present study. As only expressed mRNAs are sequenced, non-expressed and extremely rare transcripts are not sequenced at all or are not sequenced to sufficient coverage levels for reliable mutation detection. However, this limitation might not greatly affect the ability of this method to detect activating mutations in oncogenes, as these genes would have to be transcribed and translated to mediate their oncogenic effect. Recently, exon-capture techniques became available providing a more even read depth across protein coding regions, thus allowing an exhaustive mutation analysis. In contrast to whole exome or genome sequencing, transcriptome sequencing provides valuable additional information on gene expression levels and exon-composition of transcripts. Apart from mutation detection, transcriptome sequencing could also be used to detect tumor-specific fusion genes and splice variants.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

This work was funded by a *Deutsche Krebshilfe* grant 109031 to PA Greif and SK Bohlander, and by grants from the German Ministry of Research and Education (BMBF; 01GS0876) and the Deutsche Forschungsgemeinschaft (SFB 684-A6) to SK Bohlander.

References

1 Mrozek K, Marcucci G, Paschka P, Whitman SP, Bloomfield CD. Clinical relevance of mutations and gene-expression changes in adult acute myeloid leukemia with normal cytogenetics: are we ready for a prognostically prioritized molecular classification? *Blood* 2007; **109**: 431–448.
2 Yamamoto Y, Kiyoi H, Nakano Y, Suzuki R, Kodera Y, Miyawaki S *et al.* Activating mutation of D835 within the activation loop of FLT3 in human hematologic malignancies. *Blood* 2001; **97**: 2434–2439.
3 Nakao M, Yokota S, Iwai T, Kaneko H, Horiike S, Kashima K *et al.* Internal tandem duplication of the flt3 gene found in acute myeloid leukemia. *Leukemia* 1996; **10**: 1911–1918.
4 Reindl C, Bagrintseva K, Vempati S, Schnittger S, Ellwart JW, Wenig K *et al.* Point mutations in the juxtamembrane domain of FLT3 define a new class of activating mutations in AML. *Blood* 2006; **107**: 3700–3707.
5 Pabst T, Mueller BU, Zhang P, Radomska HS, Narravula S, Schnittger S *et al.* Dominant-negative mutations of CEBPA, encoding CCAAT/enhancer binding protein-alpha (C/EBP alpha), in acute myeloid leukemia. *Nat Genet* 2001; **27**: 263–270.
6 Yu M, Honoki K, Andersen J, Paietta E, Nam DK, Yunis JJ. MLL tandem duplication and multiple splicing in adult acute myeloid leukemia with normal karyotype. *Leukemia* 1996; **10**: 774–780.

7 Falini B, Mecucci C, Tiacci E, Alcalay M, Rosati R, Pasqualucci L *et al*. Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *N Engl J Med* 2005; **352**: 254–266.
8 Zhang DE, Zhang P, Wang ND, Hetherington CJ, Darlington GJ, Tenen DG. Absence of granulocyte colony-stimulating factor signaling and neutrophil development in CCAAT enhancer binding protein alpha-deficient mice. *Proc Natl Acad Sci USA* 1997; **94**: 569–574.
9 Caligiuri MA, Schichman SA, Strout MP, Mrozek K, Baer MR, Frankel SR *et al*. Molecular rearrangement of the ALL-1 gene in acute myeloid leukemia without cytogenetic evidence of 11q23 chromosomal translocations. *Cancer Res* 1994; **54**: 370–373.
10 Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K *et al*. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 2008; **456**: 66–72.
11 Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, Chen K *et al*. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* 2009; **361**: 1058–1066.
12 Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD *et al*. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 2010; **463**: 191–196.
13 Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D *et al*. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 2010; **463**: 184–190.
14 Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT *et al*. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* 2009; **462**: 1005–1010.
15 Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009; **458**: 719–724.
16 Gross S, Cairns RA, Minden MD, Driggers EM, Bittinger MA, Jang HG *et al*. Cancer-associated metabolite 2-hydroxyglutarate accumulates in acute myelogenous leukemia with isocitrate dehydrogenase 1 and 2 mutations. *J Exp Med* 2010; **207**: 339–344.
17 Hurowitz EH, Drori I, Stodden VC, Donoho DL, Brown PO. Virtual Northern analysis of the human genome. *PLoS ONE* 2007; **2**: e460.
18 Velculescu VE, Madden SL, Zhang L, Lash AE, Yu J, Rago C *et al*. Analysis of human transcriptomes. *Nat Genet* 1999; **23**: 387–388.
19 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**: 1754–1760.
20 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078–2079.
21 Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008; **5**: 621–628.
22 Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 2008; **18**: 1851–1858.
23 Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM *et al*. Exome sequencing identifies the cause of a Mendelian disorder. *Nat Genet* 2010; **42**: 30–35.
24 Peterson LF, Zhang DE. The 8;21 translocation in leukemogenesis. *Oncogene* 2004; **23**: 4255–4262.
25 Osato M. Point mutations in the RUNX1/AML1 gene: another actor in RUNX leukemia. *Oncogene* 2004; **23**: 4284–4296.
26 Dayyani F, Wang J, Yeh JR, Ahn EY, Tobey E, Zhang DE *et al*. Loss of TLE1 and TLE4 from the del(9q) commonly deleted region in AML cooperates with AML1-ETO to affect myeloid cell proliferation and survival. *Blood* 2008; **111**: 4338–4347.
27 Borinstein SC, Hyatt MA, Sykes VW, Straub RE, Lipkowitz S, Boulter J *et al*. SETA is a multifunctional adapter protein with three SH3 domains that binds Grb2, Cbl, and the novel SB1 proteins. *Cell Signal* 2000; **12**: 769–779.
28 Reindl C, Quentmeier H, Petropoulos K, Greif PA, Benthaus T, Argiropoulos B *et al*. CBL exon 8/9 mutants activate the FLT3 pathway and cluster in core binding factor/11q deletion acute myeloid leukemia/myelodysplastic syndrome subtypes. *Clin Cancer Res* 2009; **15**: 2238–2247.
29 Liu JP, Liu NS, Yuan HY, Guo Q, Lu H, Li YY. Human homologue of SETA binding protein 1 interacts with cathepsin B and participates in TNF-Induced apoptosis in ovarian cancer cells. *Mol Cell Biochem* 2006; **292**: 189–195.
30 Mayr C, Bartel DP. Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* 2009; **138**: 673–684.

Supplementary Information accompanies the paper on the Leukemia website (http://www.nature.com/leu)

4.2.2 Adaptor Protein Complex 4 Deficiency Causes Severe Autosomal-Recessive Intellectual Disability, Progressive Spastic Paraplegia, Shy Character, and Short Stature

REPORT

Adaptor Protein Complex 4 Deficiency Causes Severe Autosomal-Recessive Intellectual Disability, Progressive Spastic Paraplegia, Shy Character, and Short Stature

Rami Abou Jamra,[1,8,*] Orianne Philippe,[2,8] Annick Raas-Rothschild,[3] Sebastian H. Eck,[4] Elisabeth Graf,[4] Rebecca Buchert,[1] Guntram Borck,[2] Arif Ekici,[1] Felix F. Brockschmidt,[5,6] Markus M. Nöthen,[5,6] Arnold Munnich,[2] Tim M. Strom,[4,7] Andre Reis,[1,9] and Laurence Colleaux[2,9,*]

Intellectual disability inherited in an autosomal-recessive fashion represents an important fraction of severe cognitive-dysfunction disorders. Yet, the extreme heterogeneity of these conditions markedly hampers gene identification. Here, we report on eight affected individuals who were from three consanguineous families and presented with severe intellectual disability, absent speech, shy character, stereotypic laughter, muscular hypotonia that progressed to spastic paraplegia, microcephaly, foot deformity, decreased muscle mass of the lower limbs, inability to walk, and growth retardation. Using a combination of autozygosity mapping and either Sanger sequencing of candidate genes or next-generation exome sequencing, we identified one mutation in each of three genes encoding adaptor protein complex 4 (AP4) subunits: a nonsense mutation in *AP4S1* (NM_007077.3: c.124C>T, p.Arg42*), a frameshift mutation in *AP4B1* (NM_006594.2: c.487_488insTAT, p.Glu163_Ser739delinsVal), and a splice mutation in *AP4E1* (NM_007347.3: c.542+1_542+4delGTAA, r.421_542del, p.Glu181Glyfs*20). Adaptor protein complexes (AP1-4) are ubiquitously expressed, evolutionarily conserved heterotetrameric complexes that mediate different types of vesicle formation and the selection of cargo molecules for inclusion into these vesicles. Interestingly, two mutations affecting *AP4M1* and *AP4E1* have recently been found to cause cerebral palsy associated with severe intellectual disability. Combined with previous observations, these results support the hypothesis that AP4-complex-mediated trafficking plays a crucial role in brain development and functioning and demonstrate the existence of a clinically recognizable syndrome due to deficiency of the AP4 complex.

With a worldwide prevalence of around 2%, early-onset cognitive impairment, commonly referred to as intellectual disability (ID), is the most frequent cause of severe disability and a leading socioeconomic healthcare problem in Western countries.[1] For the last two decades, remarkable progress has been made in the elucidation of ID conditions. About 30% percent of severe ID cases have been ascribed to chromosomal imbalances.[2] Defects in X-linked genes account for about 10% of male ID cases.[3] Yet despite these recent advances, the cause of ID remains unexplained in the majority of cases, and this leaves families without accurate diagnosis or genetic counseling. In particular, very little is known about the autosomal-recessive forms of ID (ARID). The broad genetic heterogeneity of ARID precludes any possibility of pooling families, and the scarcity of large pedigrees suitable for linkage analyses have hitherto hampered identification of the genes responsible for most of these cases. This problem has been successfully circumvented by the use of autozygosity mapping in large consanguineous families; in nonspecific ARID ten genes have been identified so far.[1,4–12] Nevertheless, mutations in each of these genes only account for one or very few families, suggesting that many genes remain to be identified.

Here, we present linkage analyses, mutation discovery, and clinical characterization of a recognizable ARID syndrome in eight affected individuals from three consanguineous families.

Family ID01 is a sibship of three affected and two healthy siblings born to healthy parents, who are second cousins of Israeli–Arab descent (Figure 1A and Table 1). Pregnancy and delivery were unremarkable in all three affected cases. At birth all three siblings presented with microcephaly and muscular hypotonia, which later developed to hypertonia. At the time of examination, a clinical assessment showed hyperreflexia, spastic paraplegia, and an inability to walk unaided. All affected individuals revealed a severe cognitive deficit, marked speech delay, and adaptive impairment. Furthermore, they presented with microcephaly, a high palate, mildly remarkable facial gestalt with a wide nasal bridge, short stature, hyperlaxity, genu recurvatum, pes planus, and a waddling gait. All three affected individuals had stereotypic laughter and markedly shy character. None of the patients had seizures, vision or

[1]Institute of Human Genetics, University of Erlangen, D-91054 Erlangen, Germany; [2]INSERM U781, Fondation IMAGINE, Département de Génétique and Département de Radiologie Pédiatrique, Université Paris Descartes, Hôpital Necker-Enfants Malades, 75015 Paris, France; [3]Department of Human Genetics and Metabolic Diseases, Hadassah Hebrew University Medical Center, 91120 Jerusalem, Israel; [4]Institute of Human Genetics, Helmholtz Center Munich, German Research Center for Environmental Health, D-85764 Neuherberg, Germany; [5]Department of Genomics, Life and Brain Center, University of Bonn, D-53127 Bonn, Germany; [6]Institute of Human Genetics, University of Bonn, D-53127 Bonn, Germany; [7]Institute of Human Genetics, Klinikum rechts der Isar, Technische Universität München, D-80634 München, Germany
[8]These authors contributed equally to this work
[9]These authors contributed equally to this work
*Correspondence: rami.aboujamra@uk-erlangen.de (R.A.J.), laurence.colleaux@inserm.fr (L.C.)
DOI 10.1016/j.ajhg.2011.04.019.

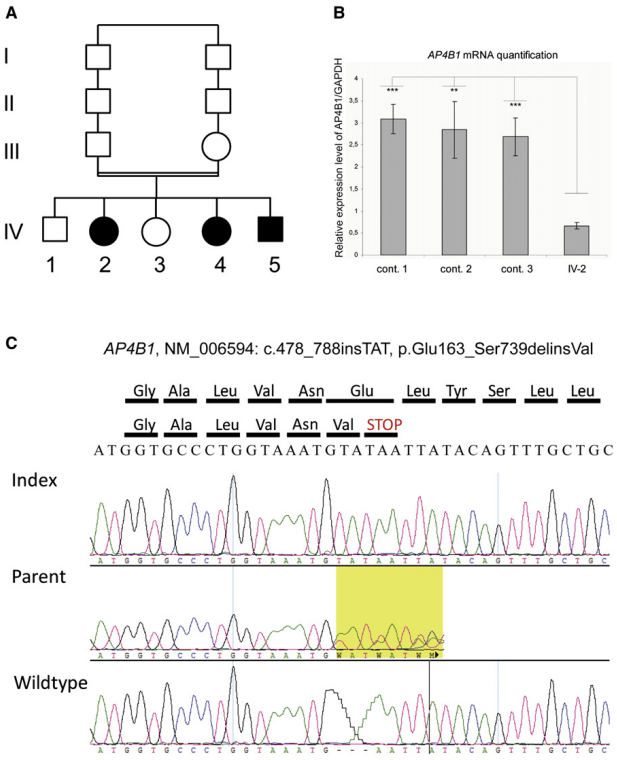


Figure 1. Genetic Analysis of Family ID01

(A) Pedigree of the family.

(B) Electrophoregrams illustrating the c.487_488insTAT, p.Glu163_Ser739delinsVal variant in exon 5 of *AP4B1*. Data are shown for homozygous affected individuals, heterozygous healthy parents, and homozygous wild-type healthy siblings.

(C) Quantitative RT-PCR analysis of *AP4B1* mRNA. *AP4B1* expression in fibroblast cells from three controls and from patient IV-2.

Data are normalized to glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*). Means ± standard deviation are given (n = 3 independent experiments). ***p value of < 0.01 (Student's test) for the difference of expression. **p value of < 0.05 (Student's test) for the difference of expression.

hearing impairments, or any anomalies of inner organs (Table 1).

Family MR061 is composed of three nuclear families from middle Syria; the families have one affected child each: the index, V-14, his nephew VI-3, and his cousin once removed V-28 (Figure 2). Pregnancy, delivery, and neonatal period were unremarkable. Affected individuals sat at age one and walked at age two to four but lost their ability to walk six to 24 months later. At the time of examination, clinical assessment showed muscular hypertonia, especially of lower limbs; contractures; talipes equinovarus; weak and decreased muscle mass of the shanks; and a clinical suspicion of peripheral neuropathy. All affected individuals revealed a severe cognitive deficit and absent speech and could only express basic needs (i.e., thirst, hunger, and strangury). Further, they presented with microcephaly (VI-3 and V-28); mild facial dysmorphisms, including a prominent and bulbous nose, a wide mouth, and coarse features; short stature; and mild spasticity in flection of upper limbs, which could be used only for simple tasks (e.g., holding a bottle of water). Similar to the former family, all affected individuals were markedly shy, amicable, and calm and kept smiling or laughing for no obvious reason but did not have laughter bursts. None of the patients had seizures, a hearing impairment, or any anomalies of inner organs (Table 1). The nephew of the index, VI-3, has an unclassified vision impairment. In addition, V-9 and VI-7 presented with moderate intellectual disability, could walk and speak, and did not have growth retardation and thus presented clinically with a different form of disability.

Family MR071 is composed of two nuclear families who are from the north of Syria and had one affected child each: the index, III-5, and his cousin, III-11 (Figures 3A and 3B). The pregnancy and birth were unremarkable in both cases. The parents reported muscular hypotonia in the neonatal period; this later developed to muscular hypertonia, especially of the lower limbs. At the time of examination, clinical assessment revealed contractures; talipes equinovarus; and decreased muscle mass of the shanks; together these findings resembled peripheral neuropathy. All affected individuals presented with a severe cognitive deficit and absent speech. Furthermore, they had microcephaly; short stature; and a mildly remarkable facial gestalt that included a prominent and bulbous nose, a wide mouth, and coarse features. Individual III-11 had epilepsy. Both patients showed a shy, amicable, and calm character. They smiled or laughed for no obvious reason but did not have laughter bursts. None of the patients had vision or hearing impairments or any anomalies of inner organs (Table 1). In this family one other cousin, III-1, presented clinically with a different form of

Table 1. Clinical Findings in Patients with Mutation in Distinct AP4 Subunits

	This Report								Moreno-De-Luca et al.[32]			Verkerk et al.[31]			
Family	**ID01**			**MR061**			**MR071**								
Patient	IV-2	IV-4	IV-5	V-14	V-28	VI-3	III-5	III-11	IV-4	IV-5	IV-1	IV-3	IV-4	IV-5	IV-6
AP4 subunit disrupted	B1	B1	B1	S1	S1	S1	E1	E1	E1	E1	M1	M1	M1	M1	M1
Sex	F	F	M	M	F	M	M	F	F	M	F	M	F	M	M
Age at evaluation (years)	23	15	11	22	20	18	11	6	23	22	24	23	22	1.5	21
Severe ID	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Normal speech	+	−	−	−	−	−	−	−	−	−	−	−	−	NA	−
Stereotypic laughter	+	+	+	+	+	+	−	+	+	+	+	+	+	NA	+
Shy character	+	+	−	+	+	+	−	−	NA	NA	NA	NA	NA	NA	NA
Neonatal hypotonia	+	+	+	+	NA	+	+	+	+	+	+	+	+	+	+
Progressing to hypertonia	+	+	+	+	+	+	+	+	+	+	+	+	+	NA	+
Hyperreflexia	+	+	+	+	NA	+	NA	NA	+	+	+	+	+	NA	+
Babinski sign	+	−	+	NA	NA	NA	NA	NA	+	+	+	+	+	NA	+
Spasticity	+	+	+	+	−	+	+	+	+	+	+	+	+	NA	+
Drooling	−	−	+	+	+	+	−	−	+	+	+	+	+	NA	+
Walk independently (years)	2.5	2.5	2.5	2	2	2.5	−	−	−	−	−	−	−	−	−
Ambulation	wheelchair	+	wheelchair	−	−	−	crawling	crawling	−	−	−	−	−	NA	−
Foot deformity	−	−	−	+	+	+	+	+	NA	NA	−	+	−	−	−
Head circumference	−2 SD	−2.5 SD	−3 SD	−1 SD	−4 SD	−2 SD	−3 SD	−4 SD	−3 SD	−3 SD	−1 SD	0 SD	−2 SD	NA	−2.5 SD
Height (cm)	↓	↓	↓	145	130	140	125	105	NA	NA	NA	NA	NA	NA	NA
Epilepsy	−	−	−	−	−	−	−	+	+	+	−	−	−	−	−
Sphincter control	−	−	−	+	+	+	−	−	−	−	−	−	−	NA	−
Eye evaluation	normal	normal	normal	normal	normal	amblyopia	normal	normal	normal	normal	NA	normal	NA	NA	POD
Hearing evaluation	normal	normal	normal	normal	normal	normal	normal	normal	normal	normal	NA	NA	NA	NA	NA
Overweight	+	+	normal	normal	normal	normal	normal	normal	NA	NA	NA	NA	NA	NA	NA

The following abbreviations are used: F, female; M, male, NA, not available POD, pale optic disc. The following symbols are used: +, present; −, absent; ↓, short stature, but not exactly measured.

disability. She had mild ID, could walk and speak, and did not have growth retardation.

Chromosome analysis and metabolic screening (including plasma amino acid, lactate, carnitine and urinary oligosaccharides, mucopolysaccharidoses, and organic acids) as well as biochemical screening for GM1 Gangliosidosis, Tay Sachs, and Krabbe diseases were all normal. No brain magnetic resonance imaging was available for any of the patients.

All procedures followed in this study were approved by the local ethical committee at the contributing Universities and proper informed consent was obtained. Additional consent was obtained from parents of affected persons whose photographs are presented in this work. Blood samples were collected from all affected and most unaffected siblings and parents. Blood lymphoblasts (families MR061 and MR071) and skin fibroblasts (patient IV-2 of family ID01) were cultured. Genomic DNA was extracted by standard methods and analyzed with the Affymetrix GeneChip Mapping 250K array (family ID01) or 6.0 array (families MR061 and MR071) (Affymetrix, Santa Clara, CA, USA). Analysis did not reveal pathogenic deletions or duplications. Mendelian segregation was calculated with PedCheck software and was confirmed in all instances.[13–15]

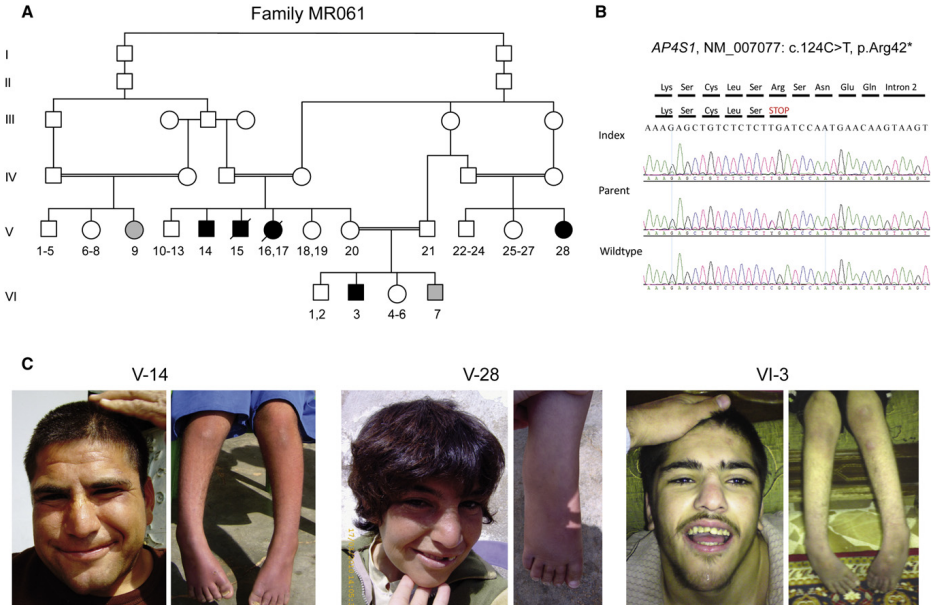


Figure 2. Genetic Analysis of Family MR061
(A) Pedigree of family; arrows indicate index. Family MR061 is large with multiple affected individuals with variable phenotypes. Grey symbols denote individuals in whom clinical presentation is markedly different and who have no *AP4S1* mutation (heterogeneity within the family).
(B) Electropherograms illustrating the mutation in exon 2 of *AP4S1*.
(C) Facial appearance of affected individuals with discreet remarkable facial gestalt, including a prominent and bulbous nose, a wide mouth, and coarse features and photographs of lower limbs with foot deformity and decreased muscle mass of the shanks.

We performed linkage analysis with ALLEGRO under an autosomal-recessive mode of inheritance with 99% penetrance and a disease allele frequency of 0.001 by using the EasyLinkage interface software.[14–16] Multipoint linkage analysis resulted in significant linkage peaks at 1p13.2-q21.2 for family ID01, at 14q11-q12 for MR061,[17] and at 15q21.1-q25.1 for MR071. Further genotype and haplotype analyses confirmed homozygosity by descent and defined critical intervals of 34 Mb, 9.1 Mb, and 32.9 Mb, respectively (Table S1, available online). Individuals from families MR061 and MR071, who presented with a clinically distinct form of ID, neither showed linkage to the above-mentioned regions nor, in the case of individuals V-9 and VI-7, showed a common locus. This means that in families MR061 and MR071 three and two different ID disorders exist, respectively.

After prioritizing genes according to their expression in the brain and their putative role in the central nervous system, we screened eight genes mapping to 1p13.1-q21.2 in family ID01 by using direct sequencing of all coding exons and exon-intron junctions. We identified a 3 bp homozygous insertion (NM_006594.2: c.487_488insTAT) within exon 5 of *AP4B1* (encoding the β subunit of the adaptor complex 4 [MIM 607245]) (Figure 1B). This variant cosegregated with the disease and was not present in NCBI dbSNP (build 131) or detected in any of 796 control chromosomes, including 160 chromosomes from individuals of Israeli-Arab origin. This frameshift mutation (p.Glu163_Ser739delinsVal) was predicted to cause premature termination of translation. Quantitative real-time PCR with the ΔCT method showed significantly less *AP4B1* transcript in patient skin fibroblasts than in three controls ($p = 0.0084–0.013$ with Student's test, Figure 1C). These results suggest that the primary effect of this frameshift mutation is nonsense-mediated decay.

DNA from individual V-14 of family MR061 was enriched with the SureSelect Human All Exon Kit, which targets approximately 38 Mb, that is 1.22%, of the human genome (Agilent, Santa Clara, Ca, USA). Sequencing was carried out on an Illumina Genome Analyzer IIx (Illumina, San Diego, CA, USA) as 54 bp or 76 bp paired-end runs. Image analysis and base calling were performed with the Genome Analyzer Pipeline version 1.5 with default

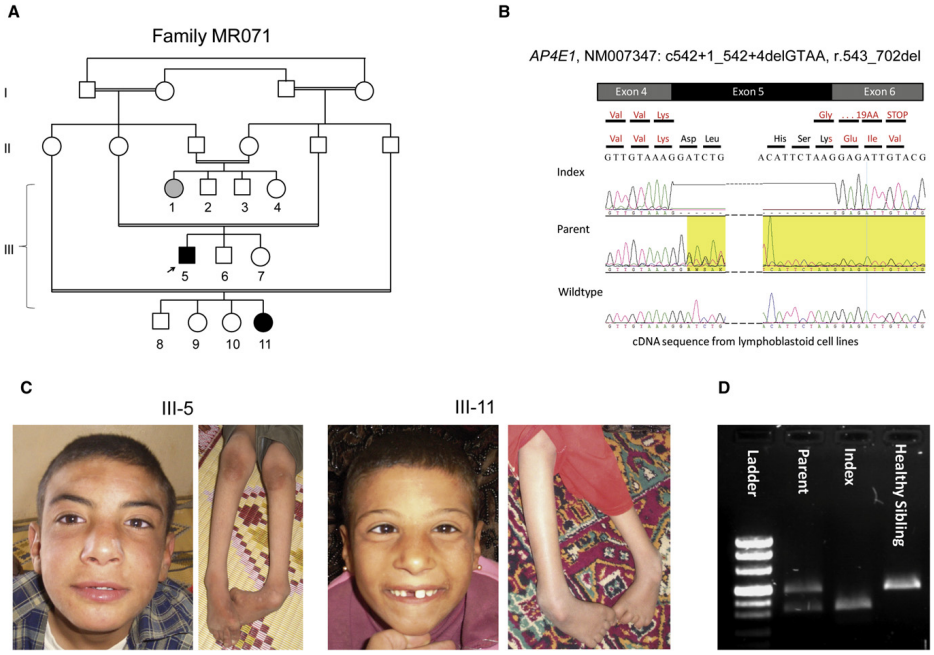


Figure 3. Genetic Analysis of Family MR071
(A) Pedigree of family MR071.
(B) Representative sequence traces from cDNA showing skipping of exon 5.
(C) Facial appearance of affected individuals includes discreet remarkable facial gestalt with prominent and bulbous nose, wide mouth, and coarse features. Also shown are photographs of the lower limbs with foot deformity and decreased muscle mass of the shanks.
(D) RT-PCR products of mRNA from homozygous affected individuals, heterozygous healthy parents, and homozygous wild-type healthy siblings; the expected size from the normal *AP4E1* allele (512 bp) as well as a smaller band corresponding to aberrant splicing of the mutated allele with skipping of exon 5 (389 bp) is shown.

parameters. We performed read alignment with BWA (v 0.5.8) by using the default parameters[18] with the human genome assembly hg19 (GRCh37) as reference. Single-nucleotide variants and small insertions and deletions (indels) were detected with SAMtools (v 0.1.7).[19] Variant annotation was performed with custom Perl scripts integrating data from dbSNP (v131) and the UCSC Genome Browser knownGene track. Additionally, we compared variants to 80 sequenced exomes in an in-house database to identify further common variants that are not present in dbSNP. We captured 26,037 variants; 6,655 variants were coding and homozygous, 345 of those were not annotated, and 139 of those were nonsynonymous. Only two variants were located in the linked region on chromosome 14: a missense variant in the last exon of *SLC22A17* (NM_020372.2: c.1429G>A, p.Val477Met) and a nonsense variant in the first coding exon of *AP4S1* (NM_007077.3: c.124C>T, p.Arg42*). In silico analysis with MutationTaster[20] and PolyPhen[21] indicated a high probability for a pathogenic effect for both variants. The variants were also absent in 740 Syrian control chromosomes and cosegregated with the affected status within the family. Both *SLC22A17* (highly expressed in the brain and belonging to the organic cation transporter family [MIM 611461]) and *AP4S1* (encoding the small subunit of the adaptor complex 4, MIM [607243]) are thus good a priori candidates for ID. Because the clinical presentation of affected persons with mutations in the different AP4 subunits shows high similarity (Table 1), we assume that the *AP4S1* mutation, which truncates the protein at its very beginning, is the main determinant of the phenotype in this family, although we cannot exclude an additional effect of the *SLC22A17* or other variant in the linked region.

Finally, we tested homozygosity at the four loci coding for the subunits B1, E1, S1, and M1 of the AP4 complex by genotyping cohorts of 22 French and 62 Syrian[17] ARID families with Affymetrix GeneChip Mapping 6.0,

Affymetrix GeneChip 250K (Affymetrix, Santa Clara, CA, USA) or Illumina 610K arrays (Illumina, San Diego, CA, USA). One Syrian family (MR071) showed linkage to chromosome 15 containing *AP4E1*. Direct sequencing of this gene (encoding the ε subunit of the adaptor complex 4 [MIM 607244]) identified a homozygous splice-donor site mutation in intron 5 (NM_007347.3, c.542+1_542+4delGTAA) (Figure 3). This variant cosegregated with the phenotype, was not present in dbSNP build 131, and was absent in 740 control chromosomes from healthy individuals of Syrian descent. This mutation was predicted to abolish the intron 5 splice-donor site. Consistently, analysis of *AP4E1* transcript from patient lymphoblastoid cell lines revealed skipping of exon 5 (Figure 3). This deletion was predicted to cause a frameshift and a premature termination of translation in exon 6 (NM_007347.3: r.421_542del, p.Glu181Glyfs*20). In an additional Syrian family, an individual who had a homozygous 11 bp deletion in intron 11 of *AP4E1* (c.1346+44_1346+54delGCAGTGACTTT) was identified . In silico analysis indicated that this deletion is not pathogenic and it was present in 20% of 320 control chromosomes from healthy individuals of Syrian descent. No additional, not annotated variants were identified.

Adaptor protein complexes (AP1, AP2, AP3, and AP4) play a key role in signal-mediated trafficking of integral membrane proteins. They mediate different types of vesicle formation and the selection of cargo molecules for inclusion into these vesicles. These evolutionarily conserved heterotetrameric complexes share a common structural organization and are composed of four subunits: two large subunits or adaptins (~100 kD; α-ε/β1-4), one medium (~50 kD; μ1-4), and one small subunit (~17 kD; σ1-4).[22] AP1-3 complexes are widely distributed among eukaryotes from yeast to humans. By contrast, the AP4 complex is absent in organisms such as *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Drosophila melanogaster*.[23] Moreover, although AP1, AP2, and AP3 complexes have been shown to be associated with clathrin vesicles, the AP4 complex functions independently of clathrin and is preferentially linked to membranes regulated by the small GTPase ARF, another coat-vesicle protein.[24–26]

As shown in Figure S1, expression analysis of *AP4B1*, *AP4E1*,and *AP4S1* transcript levels by quantitative RT-PCR in various adult and fetal tissues revealed ubiquitous expression in all fetal and adult brain structures tested. It has been suggested that the AP4 complex is involved in various sorting processes. In HeLa cells, the μ4 subunit interacts with the cytoplasmic tyrosine motif of lysosomal cargo proteins such as the lysosome-associated membrane protein 2 and mediates their direct transport to lysosomes without passing via the plasma membrane. In Madin-Darby canine kidney cells (MDCK), the μ4 subunit interacts with different cargo proteins destined for the basolateral membrane. In the brain, the AP4 complex has been involved in the transport of amyloid precursor protein (APP) from the trans-Golgi network to early endosomes.[27] Ap4b1-deficient mice were fertile, exhibited no anatomical brain abnormalities, and had normal life spans, body weight, and grip power. They exhibited no ataxia but a significantly poorer rotorod performance than wild-type mice did. There was no information about learning ability of the Ap4b1-deficient mice. Analysis of those mice demonstrated that Ap4 mediates the trans-Golgi network to the postsynaptic somatodendritic domain transport of d2 and α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) glutamate receptors in both cerebellar Purkinje cells and hippocampal neurons.[28] We thus propose that motor disturbances observed in patients with mutations in the AP4 complex might be because of cerebellar dysfunction caused by mislocalization of glutamate receptors. Similarly, defective AMPA receptor sorting might impact synaptic plasticity in hippocampal neurons and cause ID.

Mutations in other AP complexes have already been linked to human disorders. Mutations of *AP1S2* (encoding the σ small subunit of the adaptor complex 1 [MIM 300629]) cause an X-linked intellectual disability syndrome associating muscular hypotonia, delayed walking, speech delay, aggressive behavior, brain calcifications, and elevated cerebrospinal fluid protein levels (MIM 300630).[29,30] Mutations in *AP3B1* (encoding the β subunit of the adaptor complex 3 [MIM 603401]) cause Hermansky-Pudlak syndrome type 2 (HPS [MIM 608233]), a disease characterized by hypopigmentation of the eyes and skin, prolonged bleeding, and lysosomal ceroid storage.[27]

Interestingly, homozygosity for a splice-donor site mutation in *AP4M1* (encoding the μ subunit of the AP4 complex [MIM 602296]) was found to be associated with an autosomal-recessive spastic tetraplegia with intellectual disability and white matter anomalies (MIM 612936).[31] Moreover, a 192-kb-long deletion encompassing *SPPL2A* (MIM 608238) and *AP4E1* (MIM 607244) was also found to be associated with autosomal-recessive cerebral palsy, microcephaly, and intellectual disability (no MIM number has been assigned).[32] Patients carrying mutations in *AP4M1* and *AP4E1* share many clinical features with those patients described in our study (Table 1). Primarily, all patients had an infantile muscular hypotonia that progressed to spasticity, hypertonia, and paralysis and became unable to walk. They presented with a severe ID, absent or markedly delayed speech, stereotypic laughter, and growth retardation. In addition to those common symptoms, several other features were observed in some of the patients: microcephaly, epilepsy, waddling gait, joint hyperlaxity (in patients with *AP4B1* mutation), and feet deformity (Table 1). Yet, despite this phenotypic variability, our data suggest that disruption of any of the four AP4 subunits causes a clinically recognizable AP4-deficiency syndrome. This was also supported by the fact that three other individuals in families MR061 and MR071 have a clinical phenotype of ID different from the AP4 syndrome and consistently do not carry the familial mutations homozygously. Heterozygous carriers of the mutations have no ID.

In conclusion, our findings illustrate the power of combining systematic autozygosity mapping with large-scale sequencing for unraveling the molecular bases of autosomal-recessive ID. More importantly, they suggest the existence of a clinically recognizable AP4-deficiency syndrome characterized by the association of severe ID, growth retardation, stereotypic laughter, progressive spasticity, and inability to walk. Finally, this study provides further support to the hypothesis of a crucial role of AP4-mediated trafficking in brain development and functioning.

Supplemental Data

Supplemental Data include one figure and one table and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

We are grateful to the families for their participation in the study. We also thank Angelika Diem, Petra Rothe, Bianca Schmick, and Anna Benet-Pagès for the excellent technical support. This study was supported by the Centre National de la Recherche Scientifique (CNRS), the Agence Nationale de la Recherche (ANR-08-MNP-010), the Ministère de la Recherche et de l'Enseignement Supérieur, the German Intellectual disability Network (MRNET) through a grant from the German Ministry of Research and Education to A. Reis and T. Strom (01GS08160 and 01GR0804-4), the European Commission 7th Framework Program, Project N. 261123, GEUVADIS, and by the Deutsche Forschungsgemeinschaft (DFG) (AB393/1-2).

Received: March 31, 2011
Revised: April 21, 2011
Accepted: April 26, 2011
Published online: May 26, 2011

Web Resources

The URLs for data presented herein are as follows:

Ensembl Genome Browser, http://www.ensembl.org/
GenBank, http://www.ncbi.nlm.nih.gov/Genbank/
National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org

References

1. Ropers, H.H. (2010). Genetics of early onset cognitive impairment. Annu. Rev. Genomics Hum. Genet. *11*, 161–187.
2. Rauch, A., Hoyer, J., Guth, S., Zweier, C., Kraus, C., Becker, C., Zenker, M., Hüffmeier, U., Thiel, C., Rüschendorf, F., et al. (2006). Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation. Am. J. Med. Genet. A. *140*, 2063–2074.
3. Ropers, H.H., and Hamel, B.C. (2005). X-linked mental retardation. Nat. Rev. Genet. *6*, 46–57.
4. Çalışkan, M., Chong, J.X., Uricchio, L., Anderson, R., Chen, P., Sougnez, C., Garimella, K., Gabriel, S.B., dePristo, M.A., Shakir, K., et al. (2011). Exome sequencing reveals a novel mutation for autosomal recessive non-syndromic mental retardation in the TECR gene on chromosome 19p13. Hum. Mol. Genet. *20*, 1285–1289.
5. Garshasbi, M., Hadavi, V., Habibi, H., Kahrizi, K., Kariminejad, R., Behjati, F., Tzschach, A., Najmabadi, H., Ropers, H.H., and Kuss, A.W. (2008). A defect in the TUSC3 gene is associated with autosomal recessive mental retardation. Am. J. Hum. Genet. *82*, 1158–1164.
6. Higgins, J.J., Pucilowska, J., Lombardi, R.Q., and Rooney, J.P. (2004). A mutation in a novel ATP-dependent Lon protease gene in a kindred with mild mental retardation. Neurology *63*, 1927–1931.
7. Mir, A., Kaufman, L., Noor, A., Motazacker, M.M., Jamil, T., Azam, M., Kahrizi, K., Rafiq, M.A., Weksberg, R., Nasr, T., et al. (2009). Identification of mutations in TRAPPC9, which encodes the NIK- and IKK-beta-binding protein, in nonsyndromic autosomal-recessive mental retardation. Am. J. Hum. Genet. *85*, 909–915.
8. Mochida, G.H., Mahajnah, M., Hill, A.D., Basel-Vanagaite, L., Gleason, D., Hill, R.S., Bodell, A., Crosier, M., Straussberg, R., and Walsh, C.A. (2009). A truncating mutation of TRAPPC9 is associated with autosomal-recessive intellectual disability and postnatal microcephaly. Am. J. Hum. Genet. *85*, 897–902.
9. Molinari, F., Foulquier, F., Tarpey, P.S., Morelle, W., Boissel, S., Teague, J., Edkins, S., Futreal, P.A., Stratton, M.R., Turner, G., et al. (2008). Oligosaccharyltransferase-subunit mutations in nonsyndromic mental retardation. Am. J. Hum. Genet. *82*, 1150–1157.
10. Molinari, F., Rio, M., Meskenaite, V., Encha-Razavi, F., Augé, J., Bacq, D., Briault, S., Vekemans, M., Munnich, A., Attié-Bitach, T., et al. (2002). Truncating neurotrypsin mutation in autosomal recessive nonsyndromic mental retardation. Science *298*, 1779–1781.
11. Motazacker, M.M., Rost, B.R., Hucho, T., Garshasbi, M., Kahrizi, K., Ullmann, R., Abedini, S.S., Nieh, S.E., Amini, S.H., Goswami, C., et al. (2007). A defect in the ionotropic glutamate receptor 6 gene (GRIK2) is associated with autosomal recessive mental retardation. Am. J. Hum. Genet. *81*, 792–798.
12. Philippe, O., Rio, M., Carioux, A., Plaza, J.M., Guigue, P., Molinari, F., Boddaert, N., Bole-Feysot, C., Nitschke, P., Smahi, A., et al. (2009). Combination of linkage mapping and microarray-expression analysis identifies NF-kappaB signaling defect as a cause of autosomal-recessive mental retardation. Am. J. Hum. Genet. *85*, 903–908.
13. O'Connell, J.R., and Weeks, D.E. (1998). PedCheck: A program for identification of genotype incompatibilities in linkage analysis. Am. J. Hum. Genet. *63*, 259–266.
14. Hoffmann, K., and Lindner, T.H. (2005). easyLINKAGE-Plus—automated linkage analyses using large-scale SNP data. Bioinformatics *21*, 3565–3567.
15. Lindner, T.H., and Hoffmann, K. (2005). easyLINKAGE: A PERL script for easy and automated two-/multi-point linkage analyses. Bioinformatics *21*, 405–407.
16. Gudbjartsson, D.F., Thorvaldsson, T., Kong, A., Gunnarsson, G., and Ingolfsdottir, A. (2005). Allegro version 2. Nat. Genet. *37*, 1015–1016.
17. Abou Jamra, R., Wohlfart, S., Zweier, M., Uebe, S., Priebe, L., Ekici, A., Giesebrecht, S., Abboud, A., Al-Khateeb, M., Fakher, M., et al. (2011). Homozygosity mapping in 64 Syrian consanguineous families with non-specific intellectual disability reveals 11 novel loci and high heterogeneity. Eur. J. Hum. Genet., in press.

18. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
19. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
20. Seelow, D., Schuelke, M., Hildebrandt, F., and Nürnberg, P. (2009). HomozygosityMapper—an interactive approach to homozygosity mapping. Nucleic Acids Res. *37* (Web Server issue), W593–599.
21. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
22. Boehm, M., and Bonifacino, J.S. (2002). Genetic analyses of adaptin function from yeast to mammals. Gene *286*, 175–186.
23. Boehm, M., Aguilar, R.C., and Bonifacino, J.S. (2001). Functional and physical interactions of the adaptor protein complex AP-4 with ADP-ribosylation factors (ARFs). EMBO J. *20*, 6265–6276.
24. Dell'Angelica, E.C., Mullins, C., and Bonifacino, J.S. (1999). AP-4, a novel protein complex related to clathrin adaptors. J. Biol. Chem. *274*, 7278–7285.
25. Hirst, J., Bright, N.A., Rous, B., and Robinson, M.S. (1999). Characterization of a fourth adaptor-related protein complex. Mol. Biol. Cell *10*, 2787–2802.
26. Wang, X., and Kilimann, M.W. (1997). Identification of two new mu-adaptin-related proteins, mu-ARP1 and mu-ARP2. FEBS Lett. *402*, 57–61.
27. Burgos, P.V., Mardones, G.A., Rojas, A.L., daSilva, L.L., Prabhu, Y., Hurley, J.H., and Bonifacino, J.S. (2010). Sorting of the Alzheimer's disease amyloid precursor protein mediated by the AP-4 complex. Dev. Cell *18*, 425–436.
28. Matsuda, S., Miura, E., Matsuda, K., Kakegawa, W., Kohda, K., Watanabe, M., and Yuzaki, M. (2008). Accumulation of AMPA receptors in autophagosomes in neuronal axons lacking adaptor protein AP-4. Neuron *57*, 730–745.
29. Borck, G., Mollà-Herman, A., Boddaert, N., Encha-Razavi, F., Philippe, A., Robel, L., Desguerre, I., Brunelle, F., Benmerah, A., Munnich, A., and Colleaux, L. (2008). Clinical, cellular, and neuropathological consequences of AP1S2 mutations: Further delineation of a recognizable X-linked mental retardation syndrome. Hum. Mutat. *29*, 966–974.
30. Tarpey, P.S., Stevens, C., Teague, J., Edkins, S., O'Meara, S., Avis, T., Barthorpe, S., Buck, G., Butler, A., Cole, J., et al. (2006). Mutations in the gene encoding the Sigma 2 subunit of the adaptor protein 1 complex, AP1S2, cause X-linked mental retardation. Am. J. Hum. Genet. *79*, 1119–1124.
31. Verkerk, A.J., Schot, R., Dumee, B., Schellekens, K., Swagemakers, S., Bertoli-Avella, A.M., Lequin, M.H., Dudink, J., Govaert, P., van Zwol, A.L., et al. (2009). Mutation in the AP4M1 gene provides a model for neuroaxonal injury in cerebral palsy. Am. J. Hum. Genet. *85*, 40–52.
32. Moreno-De-Luca, A., Helmers, S.L., Mao, H., Burns, T.G., Melton, A.M., Schmidt, K.R., Fernhoff, P.M., Ledbetter, D.H., and Martin, C.L. (2011). Adaptor protein complex-4 (AP-4) deficiency causes a novel autosomal recessive cerebral palsy syndrome with microcephaly and intellectual disability. J. Med. Genet. *48*, 141–144.

4.2.3 A Mutation in VPS35, Encoding a Subunit of the Retromer Complex, Causes Late-Onset Parkinson Disease

REPORT

A Mutation in *VPS35*, Encoding a Subunit of the Retromer Complex, Causes Late-Onset Parkinson Disease

Alexander Zimprich,[1,14,*] Anna Benet-Pagès,[2,14] Walter Struhal,[3,14] Elisabeth Graf,[2,14] Sebastian H. Eck,[2] Marc N. Offman,[4] Dietrich Haubenberger,[1] Sabine Spielberger,[5] Eva C. Schulte,[2,6] Peter Lichtner,[2] Shaila C. Rossle,[4] Norman Klopp,[7] Elisabeth Wolf,[5] Klaus Seppi,[5] Walter Pirker,[1] Stefan Presslauer,[8] Brit Mollenhauer,[9] Regina Katzenschlager,[10] Thomas Foki,[1] Christoph Hotzy,[1] Eva Reinthaler,[1] Ashot Harutyunyan,[11] Robert Kralovics,[11] Annette Peters,[7] Fritz Zimprich,[1] Thomas Brücke,[8] Werner Poewe,[5] Eduard Auff,[1] Claudia Trenkwalder,[9,12] Burkhard Rost,[4] Gerhard Ransmayr,[3] Juliane Winkelmann,[2,6,13] Thomas Meitinger,[2,13] and Tim M. Strom[2,13,*]

To identify rare causal variants in late-onset Parkinson disease (PD), we investigated an Austrian family with 16 affected individuals by exome sequencing. We found a missense mutation, c.1858G>A (p.Asp620Asn), in the *VPS35* gene in all seven affected family members who are alive. By screening additional PD cases, we saw the same variant cosegregating with the disease in an autosomal-dominant mode with high but incomplete penetrance in two further families with five and ten affected members, respectively. The mean age of onset in the affected individuals was 53 years. Genotyping showed that the shared haplotype extends across 65 kilobases around *VPS35*. Screening the entire *VPS35* coding sequence in an additional 860 cases and 1014 controls revealed six further nonsynonymous missense variants. Three were only present in cases, two were only present in controls, and one was present in cases and controls. The familial mutation p.Asp620Asn and a further variant, c.1570C>T (p.Arg524Trp), detected in a sporadic PD case were predicted to be damaging by sequence-based and molecular-dynamics analyses. VPS35 is a component of the retromer complex and mediates retrograde transport between endosomes and the trans-Golgi network, and it has recently been found to be involved in Alzheimer disease.

Parkinson's disease (PD [MIM 168600]) is the second-most common neurodegenerative disorder; it affects 1%–2% of the population above the age of 60.[1] It is characterized by degeneration of dopaminergic neurons in the nigrostriatal pathway and other monoaminergic cell groups in the brainstem. This degeneration leads to bradykinesia, resting tremor, muscular rigidity, and postural instability as well as nonmotor symptoms. Up to 20% of cases with PD are reported to be familial,[2,3] but extended pedigrees with clear Mendelian inheritance are rare. Genetic studies have so far revealed mutations in five genes causing autosomal-recessive (*PARK2* [MIM 602544], *PINK1* [MIM 608309], *PARK7* [MIM 602533]) or autosomal-dominant (*SNCA* [MIM 163890], *LRRK2* [MIM 609007]) forms of PD.[4–9] Whereas the autosomal-recessive forms with early onset and *SNCA* missense mutations or duplications[10] are rare, a single *LRRK2* mutation (RefSeq number NM_198578.3: c.6055G>A [p.Gly2019Ser]) accounts for approximately 1% of sporadic cases of European origin.[11–13] A recent study revealed a strong association of PD with glucocerebrosidase (*GBA*) mutations in carriers for Gaucher [MIM 230800] disease, thus implicating a lysosomal enzyme in the pathogenesis of PD.[14,15] Genome-wide association studies revealed several low-risk susceptibility loci, among them *LAMP3* [MIM 605883] and *HIP1R* [MIM 605613], which have been reported to be implicated in the lysosomal pathway.[16–18]

We identified an Austrian family in which 16 members were affected by PD (family A, Figure 1). PD seemed to be inherited in an autosomal-dominant mode with high penetrance. Seven affected members were available for clinical and DNA investigations. Six of them exhibited at least three of the four cardinal signs of PD (akinesia, resting tremor, rigidity, and postural instability) and showed improvement after dopaminergic treatment. A single affected individual had displayed action tremors since childhood but developed L-Dopa-responsive resting tremors and akinesia only at the age of 62 years. The mean age of onset was 53 years (range 40–68 years) (Table 1). The clinical diagnosis of idiopathic PD was made by movement-disorder specialists who used UK brain bank criteria for PD.[19] All participants gave written informed

[1]Department of Neurology, Medizinische Universität Wien, 1090 Vienna, Austria; [2]Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany; [3]Department of Neurology and Psychiatry, Allgemeines Krankenhaus, 4021 Linz, Austria; [4]Institute of Bioinformatics, Technische Universität München, 85748 Garching, Germany; [5]Department of Neurology, Medizinische Universität Innsbruck, 6020 Innsbruck, Austria; [6]Department of Neurology, Technische Universität München, 81675 Munich, Germany; [7]Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany; [8]Department of Neurology, Wilhelminenspital, 1160 Vienna, Austria; [9]Paracelsus-Elena Klinik, 34128 Kassel, Germany; [10]Department of Neurology, Sozialmedizinisches Zentrum Ost-Donauspital, 1220Vienna, Austria; [11]Center for Molecular Medicine of the Austrian Academy of Sciences, 1090 Vienna, Austria; [12]Department of Clinical Neurophysiology, Georg-August-Universität Göttingen, 37079 Göttingen, Germany; [13]Institute of Human Genetics, Technische Universität München, 81675 Munich, Germany
[14]These authors contributed equally to this work
*Correspondence: alexander.zimprich@meduniwien.ac.at (A.Z.), timstrom@helmholtz-muenchen.de (T.M.S.)
DOI 10.1016/j.ajhg.2011.06.008.

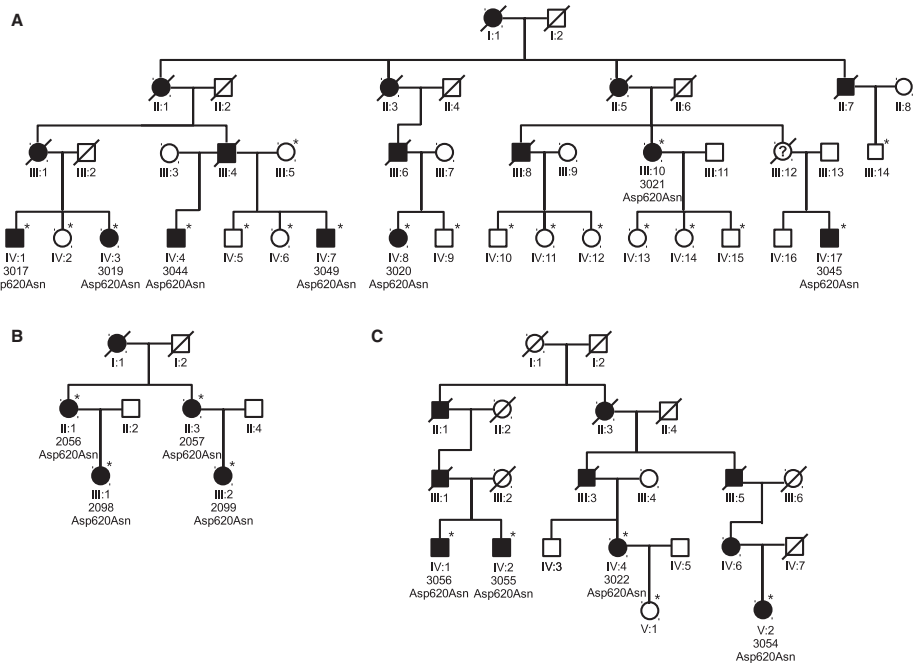


Figure 1. Pedigrees of Families A, B, and C
Unaffected family members are indicated by open symbols, affected members by closed symbols. Asterisks denote individuals genotyped for p.Asp620Asn. To maintain confidentiality, we have not shown genotypes of unaffected individuals. A question mark within a symbol denotes an unknown phenotype. Diagonal bars through symbols denote deceased individuals.

consent. The study was approved by the institutional review board of the Medizinische Universität Wien and the Hessische Landesärztekammer Wiesbaden.

To identify the disease-causing variant, we selected two second cousins (#3017 and #3020) for exome sequencing. We assumed that any rare variants common in both individuals would be disease-causing candidates. Selecting distantly related members of the pedigree should minimize the proportion of alleles shared by descent. Exome sequencing was performed on a Genome Analyzer IIx system (Illumina) after in-solution enrichment of exonic sequences (SureSelect Human All Exon 38 Mb kit, Agilent). We sequenced two lanes of a flowcell for both samples, each as 54 bp paired-end runs. Read alignment was performed with BWA (version 0.5.8) to the human genome assembly hg19 (Table S1, available online). Single-nucleotide variants and small insertions and deletions (indels) were detected with SAMtools (v 0.1.7). We filtered called variants to exclude those present in 72 control exomes from patients with other unrelated diseases. We further excluded all variants that were present in dbSNP 131 and had an average heterozygosity of more than 0.02. Variant annotation was performed with custom scripts. This approach left ten heterozygous nonsynonymous variants shared by both affected individuals (Table 2; see also Table S2).

Only a single heterozygous variant in the *VPS35* gene (RefSeq number NM_018206.4: c.1858G>A [p.Asp620Asn]) fulfilled two further criteria of being possibly causative: (1) it was found in all seven affected members investigated and (2) was absent in approximately 680 KORA S4 general-population samples (Tables 2 and 3).[20] We next screened 486 unrelated PD patients from Austria for the p.Asp620Asn variant by MALDI-TOF mass spectroscopy (Sequenom MassArray system). We detected two additional index patients carrying this mutation (families B and C; Figure 1 and Table 1). The variant was detected in all eight affected individuals investigated in both families. It was not present in a second set of 554 Austrian controls or in an additional 1014 KORA-AGE controls (Table 3). The variant was further detected in three clinically unaffected family members in families A, B, and C. Because the unaffected individuals are all younger than 60 years of age, either they are all presymptomatic or the mutation is nonpenetrant in these subjects.

Table 1. Clinical Findings for PD Patients Carrying Variants in VPS35

Family	Patient	Variation	AaO	DD	IS	B	R	RT	PI	L-Dopa/DA	Other Features
A	3017	p.Asp620Asn	48	7	B	+	+	−	+	+	
A	3019	p.Asp620Asn	40	5	B	+	+	+	+	+	
A	3020	p.Asp620Asn	46	7	PI	+	+	−	+	+	
A	3021	p.Asp620Asn	68	16	PI	+	+	+	+	+	
A	3049	p.Asp620Asn	49	4	RT	+	+	+	−	+	
A	3044	p.Asp620Asn	64	3	PI	+	+	+	+	+	
A	3045	p.Asp620Asn	63	1	RT	+	−	+	−	+	action tremor since childhood
B	2056	p.Asp620Asn	61	15	RT	+	+	+	+	+	fluctuations, dyskinesias
B	2057	p.Asp620Asn	56	8	RT	+	+	+	+	+	fluctuations, dyskinesias
B	2098	p.Asp620Asn	46	0.5	RT	−	−	+	−	untreated	depression, action tremor, pathologic DAT SPECT
B	2099	p.Asp620Asn	51	5	B	+	+	+	−	+	fluctuations, pathologic DAT SPECT
C	3022	p.Asp620Asn	61	5	RT	+	+	+	−	+	dyskinesias
C	3055	p.Asp620Asn	46	12	RT	+	+	+	−	+	
C	3054	p.Asp620Asn	53	9	B	+	+	−	−	+	dyskinesias
C	3056	p.Asp620Asn	43	10	B	+	+	+	+	+	dyskinesias
	211	p.Arg524Trp	37	9	MG	+	+	+	−	+	mild action tremor since youth; 75% motor improvement on levodopa-test; DBS for fluctuations and dyskinesias; pathologic DAT SPECT
	524	p.Leu774Met	51	7	RT	+	+	+	−	+	marked postural tremor
	243	p.Leu774Met	73	9	RT	+	+	+	+	+	dyskinesias, pathologic DAT SPECT
	806	p.Ile241Met	72	2	Postural tremor	+	−	+	+	+	hyposmia (6/12 sniffing sticks), DAT SPECT pathologic, pathologic crying
	90/05	p.Met57Ile	62	13	RT	+	+	+	+	+	dementia (MMSE 23), dysphagia and dysarthria, hyposmia by history, depression

Abbreviations are as follows: AaO, age at onset; DD, disease duration in years; IS, initial symptoms; B, bradykiesia; R, rigidity; RT, resting tremor; PI, postural instability; L-Dopa/DA, response to L-Dopa and/or dopamine agonist; MG, micrographia; DBS, deep brain stimulation.

Cross-species alignment of VPS35 from plants, fungi, invertebrates, and vertebrates showed complete conservation of amino acid Asp620 (Figure S1). The likely consequence of the p.Asp620Asn variant was predicted to be damaging by PolyPhen2,[21] SNAP,[22] and SIFT.[23] We therefore concluded that the variant p.Asp620Asn is indeed very likely to be causative for PD in families A, B, and C.

To determine whether the variant p.Asp620Asn occurred on the same haplotype, we genotyped 20 individuals from families A–C with oligonucleotide SNP arrays (HumanOmni2.5-Quad, Illumina). Haplotyping and linkage analysis were performed with the Merlin software.[24] The haplotypes carrying the variant p.Asp620Asn in families A–C are depicted in Table S3. Family A and B shared a common haplotype across 21 Mb between markers rs1072594 and rs4444336. Family C, however, showed only a common region of 65 kb across *VPS35*. Different alleles were located at markers rs56168099 and rs74459547, 25 kb upstream and 11 kb downstream of *VPS35*, respectively (Table S3). Because the two intragenic markers did not differ, we could not determine whether the three families shared an old common haplotype or whether the mutation has recently arisen on two different haplotypes.

To assess the prevalence of other *VPS35* mutations among PD cases and the general population, we screened all 17 coding exons for variations by dye-binding/high-resolution DNA melting curve analysis (LightScanner HR I 384, Idaho Technology) in 860 cases (484 Austrian and

Table 2. Exome Sequencing: Rare, Heterozygous, Nonsynonymous Variations Shared by Two Individuals of Pedigree A

				Variations		Control Genotypes			
Gene	**Position (hg19)**	**dbSNP**	**Transcript**	**Nucleotide**	**Amino Acid**	**1/1**	**1/2**	**2/2**	**Segregation**
PLK3	chr1:45270359		NM_004073.2	c.1543T>A	p.Ser515Thr	669	0	0	4 of 7
C8A	chr1:57383357	rs41285938	NM_000562.2	c.1723C>T	p.Pro575Ser				5 of 7
ADCY10	chr1:167787479	rs41270737	NM_018417.4	c.4313A>G	p.Asn1438Ser				2 of 7
LAMB2	chr3:49166460		NM_002292.3	c.1724G>A	p.Arg575Gln	647	28	0	5 of 7
NOM1	chr7:156762317		NM_138400.1	c.2503G>A	p.Ala835Thr	670	0	0	3 of 7
KIF22	chr16:29816237		NM_007317.1	c.1780G>A	p.Asp594Asn	665	6	0	6 of 7
SEZ6L2	chr16:29899021		NM_012410.2	c.947G>A	p.Arg316His	660	4	0	7 of 7
VPS35	chr16:46696364		NM_018206.4	c. 1858G>A	p.Asp620Asn	1069[a]	0	0	7 of 7
NLRP1	chr17:5421150		NM_001033053.2	c.3985G>A	p.Val1329Ile	666	4	0	3 of 7
NEURL4	chr17:7221197		NM_001005408.1	c.4109G>A	p.Arg1370Gln				3 of 7

Rare variations revealed by exome sequencing were checked in 670 controls (KORA S4) by MALDI-TOF analysis. The variant allele was denoted as "2," the reference allele as "1."

[a] This number includes additional 554 Austrian control individuals investigated by a TaqMan genotyping assay. Segregation shows the number of affected pedigree A individuals who carry the variant allele.

376 German cases) and 1014 controls. For controls, we used a population-based cohort (KORA AGE) with a mean age of 76 years but excluded eight individuals known to be on medications for PD (Table 3). Exons 2 to 12 are located within a region that is duplicated 12 Mb upstream. Primers were designed to specifically amplify these exons (Table S4). The screening revealed six further rare coding SNVs in addition to p.Asp620Asn (Table 4). Including p.Asp620Asn, we identified four different nonsynonymous missense variants only present in cases, two only present in controls, and one present in cases and controls. Two of the variants unique to PD cases were predicted to be damaging by all three methods (c.1858G>A [p.Asp620Asn]; c.1570C>T [p.Arg524Trp]), and one was predicted by PolyPhen2 to be possibly damaging (c.723T>G, p.Ile241Met). The other variants were predicted to be benign by all methods. Family information was only available for the patient carrying the p.Arg524Trp variant. The only available family member was her mother, aged 74 years. She was found to also carry the variant and showed mild extrapyramidal signs, including intermittent resting tremor of the left fingers and mild postural tremor of both upper limbs, but no bradykinesia. However, a DAT SPECT examination showed normal striatal binding, excluding the possibility of an early stage of PD in this subject. Of note, the screening did not reveal any common nonsynonymous coding SNVs. Furthermore, common nonsynonymous coding SNVs were not found in the 72 control exomes from patients with other unrelated diseases, nor were any recorded in the dbSNP database (version 131).

Table 3. Summary of the Samples Used in This Study

Cohort	**Sample Size**	**Mean Age (SD)**	**Females/Males**
Austrian PD cases[a]	486	58.7 (11.3)	172/314
German PD cases[b]	376	71.1 (9.4)	119/257
KORA S4 controls[c]	680	54.7 (11.9)	280/400
KORA-AGE controls[d]	1014	76.0 (6.6)	508/505
Austrian controls[e]	554	46 (15.2)	254/300

Patients presenting with atypical or secondary (e.g., vascular) parkinsonian disorders as well as patients with known mutations were excluded.

[a] The Austrian cases were recruited at the Department of Neurology, Medizinische Universität Wien, Vienna, as well as in affiliated departments on a consecutive basis. A positive family history for PD was reported from 131 patients. A positive family history was defined by at least one other affected first- or second-degree related family member.

[b] The German PD population originated from the Paracelsus-Elena Klinik, Kassel, a hospital specializing in movement disorders.

[c] This control population was recruited from the KORA S4 survey, comprising individuals who were aged 25–74 years and were examined during 1999–2001.

[d] The KORA-AGE samples were collected in 2009 as a gender- and age-stratified subsample of the KORA S1–S4 studies comprising participants born before 1944. KORA S1–S4 surveys comprise four independent cross-sectional population-based studies in the region of Augsburg, Southern Germany, and were conducted in 5 year intervals. Patients for whom PD was suspected on the basis of questionnaire data were excluded.

[e] These control samples were recruited through the Department of Neurology, Medical University of Vienna, as subjects without known history of a neurological disorder and included, for example, blood donors or unrelated companions or spouses of patients.

VPS35 is a component of the retromer complex and is involved in retrograde transport from the endosomes back to the trans-Golgi network.[25] This multi-protein complex consists of the cargo-recognition VPS26-VPS29-VPS35 heterotrimer and a membrane-targeting heterodimer or homodimer of SNX1 and/or SNX2 (vps5).[25,26] All proteins involved are evolutionarily conserved and have been previously described in *Saccharomyces cerevisiae*. The best characterized cargo proteins of the retromer complex are the cation-independent mannose 6-phosphate receptor

Table 4. Rare *VPS35* Variants in Cases and Controls

ID Cases	KORA AGE Controls	Heterozygous Nucleotide Change	Amino Acid Change	Predicted Impact on Protein			Exon/ Intron	Genomic Position (hg19, chr16)	KORA S4 Controls		
				(i)	(ii)	(iii)			1/1	1/2	2/2
Nonsynonymous											
-	1	c.151G>A	p.Gly51Ser	+	+	+	3	46,716,,039			
90/05	-	c.171G>A	p.Met57Ile	+	+	+	3	46,716,019	670	0	0
-	1	c.245C>G	p.Thr82Arg	+	+	+	4	46,715,367			
806	-	c.723T>G	p.Ile241Met	±	+	+	7	46,711,308	667	0	0
[211]	-	c.1570C>T	p.Arg524Trp	–	–	–	13	46,702,919	671	0	0
[Families A-C]	-	c.1858G>A	p.Asp620Asn	–	–	–	15	46,696,364	669	0	0
243, 524	2	c.2320C>A	p.Leu774Met	+	+	+	17	46,694,455			
Synonymous											
53097	-	c.492A>G	p.Glu164Glu				5	46,714,597	671	0	0
-	1	c.954A>T	p.Gly315Gly				9	46,708,542			
53496	-	c.1881C>T	p.Ala627Ala				15	46,696,341	668	5	0
45, 117, 53626	1	c.2145A>G	p.Leu715Leu				16	46,695,696	666	2	0
53667	-	c.2241C>T	p.Ile747Ile				17	46,694,534	667	2	0
53063	-	c.2346A>G	p.Glu782Glu				17	46,694,429	671	0	0
-	1	c.2361G>A	p.Glu787Glu				17	46,694,414			
Noncoding											
2212	2	c.1-35C>T					5′UTR	46,723,080	667	2	0
-	2	c.1-29C>T					5′UTR	46,723,074			
95, 2206	3	c.3+24A>G					1	46,723,019	662	6	0
159, 528	1	c.102+33G>A					2	46,717,387	668	2	0
[157, 2023]	-	c.103-77T>C					3	46,716,164	668	0	0
-	1	c.199+9T>G					3	46,715,982			
213	-	c.506+6T>C					5	46,714,577	644	0	0
53093	-	c.720+18C>T					6	46,712,773			
-	1	c.914+38T>C					8	46,710,457			
52824	-	c.1161-87A>C					10	46,706,471			
52791	-	c.1161-70G>A					10	46,706,454	668	0	0
-	1	c.1368+16C>T					11	46,706,161			
[2028]	-	c.1369-11G>A					12	46,705,783	669	0	0
-	1	c.1525-17delT					12	46,702,985			
-	1	c.1647+14T>C					13	46,702,828			
320	-	c.2212-45T>C					16	46,694,608	670	0	0
[352]	-	c.2391+7A>G					3′UTR	46,694,377			
-	1	c.2391+8A>G					3′UTR	46,694,376			

Variants for 863 cases and 1014 KORA AGE controls were determined by dye-binding/high-resolution DNA melting curve analysis and confirmed by Sanger sequencing. The table lists the case ID and the number of detected variant alleles of the cases and KORA AGE samples, respectively. Genotypes of identified variants were further investigated by MALDI-TOF analysis in approximately 680 KORA S4 controls. For the KORA S4 samples, the variant allele was denoted as "2," the reference allele as "1." cDNA numbering is based on reference gene NM_018206.4 for *VPS35*, where +1 corresponds to the A of ATG start translation codon. Familial cases are given in square brackets. Three methods were used for predicting the impact of SNPs on the protein. (1) PolyPhen2, (2) SNAP, and (3) SIFT; "+" indicates a benign impact, "±" indicates a possibly damaging impact, and "−" indicates a damaging impact. We detected a further nonsynonymous variant (c.1093C>T [p.Arg365Cys], genomic position 46,708,293) in a patient carrying two *PARKIN* variants (c.exon3_4del and p.Arg275Trp). This variant was not present in 670 KORA S4 and 1014 KORA AGE controls. It is predicted to be possibly damaging by all three methods. This patient's brother is also affected by PD. He carries the 2 PARKIN variants but not the *VPS35* variant.

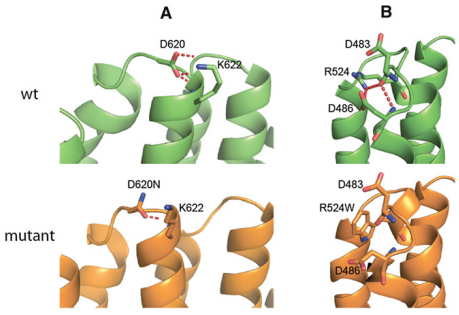


Figure 2. Hydrogen-Bonding Capacities for Wild-Type Asp620 and Arg524 and the Variants p.Asp620Asn and p.Arg524Trp
Hydrogen bonds (HB) are shown as red dashed lines. Asp60 and Arg524 are in green; p.Asp620Asn and p.Arg524Trp are in orange. (A) Asp620 forms a HB to Lys622 and shows an additional salt-bridge interaction. p.Asp620Asn forms fewer HBs, and no electrostatic interaction is possible.
(B) Arg524 forms a HB network with Asp483 and Asp486. This network is broken by the p.Arg524Trp substitution.

(CI-MPR)[27] and Vps10p in mammals and *Saccharomyces cerevisiae*, respectively; these proteins transport hydroxylases to the lysosomes or lysosomal vacuoles. Recently, additional cargo proteins and functions of VPS35 have been described.[28,29] Most interesting in our context is the involvement of the retromer into the retrograde transport of SORL1, a VPS10P-domain receptor protein that has been implicated in Alzheimer disease.[30,31] The crystal structure of the C-terminal part of VPS35 has been resolved.[32] The three variants p.Asp620Asn, p.Arg524Trp, and p.Leu774Met are located in this part of the protein, and we have investigated their impact on protein stability by using molecular dynamics (MD) simulations. We manually introduced the mutations to the crystal structure and modeled the side chains by using scwrl 4.0.[33] All MD simulations were performed via GROMACS 4.5,[34] with the all-atom force field AMBER03[35] and the water model TIP3P[36] as parameters. All three proteins are found on the edge of helices interacting with VPS29. Wild-type residue Asp620 forms frequent hydrogen bonds (HBs) to Lys622, but these bonds are less frequent in the p.Asp620Asn variant (Figure 2A). Similarly, Arg524 is involved in a triple HB network together with residues Asp483 and Asp486, but this network is broken by the introduction of p.Arg524Trp (Figure 2B). Both changes result in the loss of salt bridges and cause the protein to be locally more flexible, as shown by root-mean-square fluctuation (RMSF) profiles (Figure S2). In contrast to the effect predicted for p.Arg524Trp and p.Asp620Asn, the p.Leu774Met variant was not predicted to have a strong impact on protein stability.

In summary, we identified rare *VPS35* missense variants that are potentially pathogenic. One of these variants, p.Asp620Asn, cosegregates with late-onset PD in three unrelated families. The observation that the three families share only a small common haplotype across *VPS35*, the high conservation of VPS35, the predicted structural changes, and the protein's known involvement in lysosomal trafficking together provide strong support for the p.Asp620Asn variant's being causative for late-onset PD, although we identified only a single familial mutation. The penetrance of p.Asp620Asn is high but not complete and might be lower for the other variants. The proportion of PD caused by *VPS35* variants is expected to be low. Although exome sequencing provides perfect access to rare-variant detection, both large families and large collections of cases and controls remain a crucial resource for the identification of disease genes.

Supplemental Data

Supplemental Data include two figures and four tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

We thank all patients and their families for participating in this study. We also thank C. Fischer and B. Schmick for technical assistance and S. Schmidegg, S. Hoedl, and M. Guger for clinical examination of family A. This work was supported by a grant from the German Ministry for Education and Research (01GR0804-4). The KORA study was financed by the Helmholtz Zentrum München, the German Federal Ministry of Education and Research, the State of Bavaria, the German National Genome Research Network (NGFNplus: 01GS0823), and the Munich Center of Health Sciences (MCHealth) as part of LMUinnovativ. M.N.O., S.C.R., and B.R. were supported by the Alexander von Humboldt Foundation.

Received: May 8, 2011
Revised: June 15, 2011
Accepted: June 21, 2011
Published online: July 14, 2011

Web Resources

The URLs for data presented herein are as follows:

ExonPrimer, http://ihg.helmholtz-muenchen.de/exonprimer.html
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org
UCSC Genome Browser, http://genome.ucsc.edu

References

1. Lang, A.E., and Lozano, A.M. (1998). Parkinson's disease. First of two parts. N. Engl. J. Med. *339*, 1044–1053.
2. Bonifati, V., Fabrizio, E., Vanacore, N., De Mari, M., and Meco, G. (1995). Familial Parkinson's disease: A clinical genetic analysis. Can. J. Neurol. Sci. *22*, 272–279.
3. Payami, H., Larsen, K., Bernard, S., and Nutt, J. (1994). Increased risk of Parkinson's disease in parents and siblings of patients. Ann. Neurol. *36*, 659–661.
4. Zimprich, A., Biskup, S., Leitner, P., Lichtner, P., Farrer, M., Lincoln, S., Kachergus, J., Hulihan, M., Uitti, R.J., Calne, D.B.,

et al. (2004). Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. Neuron *44*, 601–607.
5. Bonifati, V., Rizzu, P., van Baren, M.J., Schaap, O., Breedveld, G.J., Krieger, E., Dekker, M.C., Squitieri, F., Ibanez, P., Joosse, M., et al. (2003). Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism. Science *299*, 256–259.
6. Kitada, T., Asakawa, S., Hattori, N., Matsumine, H., Yamamura, Y., Minoshima, S., Yokochi, M., Mizuno, Y., and Shimizu, N. (1998). Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. Nature *392*, 605–608.
7. Paisán-Ruíz, C., Jain, S., Evans, E.W., Gilks, W.P., Simón, J., van der Brug, M., López de Munain, A., Aparicio, S., Gil, A.M., Khan, N., et al. (2004). Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease. Neuron *44*, 595–600.
8. Polymeropoulos, M.H., Lavedan, C., Leroy, E., Ide, S.E., Dehejia, A., Dutra, A., Pike, B., Root, H., Rubenstein, J., Boyer, R., et al. (1997). Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. Science *276*, 2045–2047.
9. Valente, E.M., Abou-Sleiman, P.M., Caputo, V., Muqit, M.M., Harvey, K., Gispert, S., Ali, Z., Del Turco, D., Bentivoglio, A.R., Healy, D.G., et al. (2004). Hereditary early-onset Parkinson's disease caused by mutations in PINK1. Science *304*, 1158–1160.
10. Johnson, J., Hague, S.M., Hanson, M., Gibson, A., Wilson, K.E., Evans, E.W., Singleton, A.A., McInerney-Leo, A., Nussbaum, R.L., Hernandez, D.G., et al. (2004). SNCA multiplication is not a common cause of Parkinson disease or dementia with Lewy bodies. Neurology *63*, 554–556.
11. Gilks, W.P., Abou-Sleiman, P.M., Gandhi, S., Jain, S., Singleton, A., Lees, A.J., Shaw, K., Bhatia, K.P., Bonifati, V., Quinn, N.P., et al. (2005). A common LRRK2 mutation in idiopathic Parkinson's disease. Lancet *365*, 415–416.
12. Nichols, W.C., Pankratz, N., Hernandez, D., Paisán-Ruíz, C., Jain, S., Halter, C.A., Michaels, V.E., Reed, T., Rudolph, A., Shults, C.W., et al; Parkinson Study Group-PROGENI investigators. (2005). Genetic screening for a single common LRRK2 mutation in familial Parkinson's disease. Lancet *365*, 410–412.
13. Di Fonzo, A., Rohé, C.F., Ferreira, J., Chien, H.F., Vacca, L., Stocchi, F., Guedes, L., Fabrizio, E., Manfredi, M., Vanacore, N., et al; Italian Parkinson Genetics Network. (2005). A frequent LRRK2 gene mutation associated with autosomal dominant Parkinson's disease. Lancet *365*, 412–415.
14. Sidransky, E., Nalls, M.A., Aasly, J.O., Aharon-Peretz, J., Annesi, G., Barbosa, E.R., Bar-Shira, A., Berg, D., Bras, J., Brice, A., et al. (2009). Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. N. Engl. J. Med. *361*, 1651–1661.
15. Aharon-Peretz, J., Rosenbaum, H., and Gershoni-Baruch, R. (2004). Mutations in the glucocerebrosidase gene and Parkinson's disease in Ashkenazi Jews. N. Engl. J. Med. *351*, 1972–1977.
16. Simón-Sánchez, J., Schulte, C., Bras, J.M., Sharma, M., Gibbs, J.R., Berg, D., Paisan-Ruiz, C., Lichtner, P., Scholz, S.W., Hernandez, D.G., et al. (2009). Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nat. Genet. *41*, 1308–1312.
17. Satake, W., Nakabayashi, Y., Mizuta, I., Hirota, Y., Ito, C., Kubo, M., Kawaguchi, T., Tsunoda, T., Watanabe, M., Takeda, A., et al. (2009). Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. Nat. Genet. *41*, 1303–1307.
18. Nalls, M.A., Plagnol, V., Hernandez, D.G., Sharma, M., Sheerin, U.M., Saad, M., Simón-Sánchez, J., Schulte, C., Lesage, S., Sveinbjörnsdóttir, S., et al; International Parkinson Disease Genomics Consortium. (2011). Imputation of sequence variants for identification of genetic risks for Parkinson's disease: A meta-analysis of genome-wide association studies. Lancet *377*, 641–649.
19. Hughes, A.J., Daniel, S.E., Kilford, L., and Lees, A.J. (1992). Accuracy of clinical diagnosis of idiopathic Parkinson's disease: A clinico-pathological study of 100 cases. J. Neurol. Neurosurg. Psychiatry *55*, 181–184.
20. Wichmann, H.E., Gieger, C., and Illig, T. (2005). KORA-gen–resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen *67* (*Suppl. 1*), 26–30.
21. Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. Nucleic Acids Res. *30*, 3894–3900.
22. Bromberg, Y., and Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. *35*, 3823–3835.
23. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
24. Abecasis, G.R., Cherny, S.S., Cookson, W.O., and Cardon, L.R. (2002). Merlin—Rapid analysis of dense genetic maps using sparse gene flow trees. Nat. Genet. *30*, 97–101.
25. Bonifacino, J.S., and Rojas, R. (2006). Retrograde transport from endosomes to the trans-Golgi network. Nat. Rev. Mol. Cell Biol. *7*, 568–579.
26. Rojas, R., Kametaka, S., Haft, C.R., and Bonifacino, J.S. (2007). Interchangeable but essential functions of SNX1 and SNX2 in the association of retromer with endosomes and the trafficking of mannose 6-phosphate receptors. Mol. Cell. Biol. *27*, 1112–1124.
27. Damen, E., Krieger, E., Nielsen, J.E., Eygensteyn, J., and van Leeuwen, J.E. (2006). The human Vps29 retromer component is a metallo-phosphoesterase for a cation-independent mannose 6-phosphate receptor substrate peptide. Biochem. J. *398*, 399–409.
28. Braschi, E., Goyon, V., Zunino, R., Mohanty, A., Xu, L., and McBride, H.M. (2010). Vps35 mediates vesicle transport between the mitochondria and peroxisomes. Curr. Biol. *20*, 1310–1315.
29. Korolchuk, V.I., Schütz, M.M., Gómez-Llorente, C., Rocha, J., Lansu, N.R., Collins, S.M., Wairkar, Y.P., Robinson, I.M., and O'Kane, C.J. (2007). Drosophila Vps35 function is necessary for normal endocytic trafficking and actin cytoskeleton organisation. J. Cell Sci. *120*, 4367–4376.
30. Willnow, T.E., Petersen, C.M., and Nykjaer, A. (2008). VPS10P-domain receptors—Regulators of neuronal viability and function. Nat. Rev. Neurosci. *9*, 899–909.
31. Rogaeva, E., Meng, Y., Lee, J.H., Gu, Y., Kawarai, T., Zou, F., Katayama, T., Baldwin, C.T., Cheng, R., Hasegawa, H., et al. (2007). The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. Nat. Genet. *39*, 168–177.
32. Hierro, A., Rojas, A.L., Rojas, R., Murthy, N., Effantin, G., Kajava, A.V., Steven, A.C., Bonifacino, J.S., and Hurley, J.H.

(2007). Functional architecture of the retromer cargo-recognition complex. Nature *449*, 1063–1067.
33. Krivov, G.G., Shapovalov, M.V., and Dunbrack, R.L., Jr. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins *77*, 778–795.
34. Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. (2008). GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. J. Chem. Theory Comput. *4*, 435–447.
35. Duan, Y., Wu, C., Chowdhury, S., Lee, M.C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., et al. (2003). A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J. Comput. Chem. *24*, 1999–2012.
36. Mahoney, M.W. (2000). A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. J. Chem. Phys. *112*, 8910–8922.

4.2.4 Complete List of Publications

1. Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery.
   Eck SH*, Benet-Pags A*, Flisikowski K, Meitinger T, Fries R, Strom TM.
   Genome Biol. 2009;10(8):R82. Epub 2009 Aug 6.

2. Identification of FOXP1 deletions in three unrelated patients with mental retardation and significant speech and language deficits.
   Horn D, Kapeller J, Rivera-Brugus N, Moog U, Lorenz-Depiereux B, **Eck SH**, Hempel M, Wagenstaller J, Gawthrope A, Monaco AP, Bonin M, Riess O, Wohlleber E, Illig T, Bezzina CR, Franke A, Spranger S, Villavicencio-Lorini P, Seifert W, Rosenfeld J, Klopocki E, Rappold GA, Strom TM.
   Hum Mutat. 2010 Nov;31(11):E1851-60.

3. CpG-methylation regulates a class of Epstein-Barr virus promoters.
   Bergbauer M, Kalla M, Schmeinck A, Gbel C, Rothbauer U, **Eck SH**, Benet-Pags A, Strom TM, Hammerschmidt W.
   PLoS Pathog. 2010 Sep 23;6(9):e1001114.

4. Identification of recurring tumor-specific somatic mutations in acute myeloid leukemia by transcriptome sequencing.
   Greif PA*, **Eck SH***, Konstandin NP*, Benet-Pags A, Ksienzyk B, Dufour A, Vetter AT, Popp HD, Lorenz-Depiereux B, Meitinger T, Bohlander SK, Strom TM.
   Leukemia. 2011 May;25(5):821-7. Epub 2011 Feb 22,

5. Adaptor protein complex 4 deficiency causes severe autosomal-recessive intellectual disability, progressive spastic paraplegia, shy character, and short stature.
   Abou Jamra R, Philippe O, Raas-Rothschild A, **Eck SH**, Graf E, Buchert R, Borck G, Ekici A, Brockschmidt FF, Nthen MM, Munnich A, Strom TM, Reis A, Colleaux L.
   Am J Hum Genet. 2011 Jun 10;88(6):788-95. Epub 2011 May 27.

6. A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease.
   Zimprich A*, Benet-Pags A*, Struhal W*, Graf E*, **Eck SH**, Offman

MN, Haubenberger D, Spielberger S, Schulte EC, Lichtner P, Rossle SC, Klopp N, Wolf E, Seppi K, Pirker W, Presslauer S, Mollenhauer B, Katzenschlager R, Foki T, Hotzy C, Reinthaler E, Harutyunyan A, Kralovics R, Peters A, Zimprich F, Brcke T, Poewe W, Auff E, Trenkwalder C, Rost B, Ransmayr G, Winkelmann J, Meitinger T, Strom TM.
Am J Hum Genet. 2011 Jul 15;89(1):168-75.

7. New mouse models for metabolic bone diseases generated by genome-wide ENU mutagenesis.
Sabrautzki S, Rubio-Aliaga I, Hans W, Fuchs H, Rathkolb B, Calzada-Wack J, Cohrs CM, Klaften M, Seedorf H, **Eck SH**, Benet-Pags A, Favor J, Esposito I, Strom TM, Wolf E, Lorenz-Depiereux B, Hrabe de Angelis M.
Mamm Genome. 2012 Aug;23(7-8):416-30. Epub 2012 Apr 21.

8. Autosomal recessive cortical myoclonic tremor and epilepsy: association with a mutation in the potassium channel associated gene CNTN2.
Stogmann E, Reinthaler E, Eltawil S, El Etribi MA, Hemeda M, El Nahhas N, Gaber AM, Fouad A, Edris S, Benet-Pages A, **Eck SH**, Pataraia E, Mei D, Brice A, Lesage S, Guerrini R, Zimprich F, Strom TM, Zimprich A.
Brain. 2013 Mar 21.

9. Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage.
Jansen S, Aigner B, Pausch H, Wysocki M, **Eck SH**, Benet-Pages A, Garf E, Wieland T, Strom TM, Meitinger T, Fries R.
BMC Genomics. 2013 Jul 4.

10. Dysfunctional nitric oxide signaling increases risk of myocardial infarction.
Jeanette Erdmann*, Klaus Stark*, Philipp Moritz Rumpf*, Ulrike B. Esslinger*, Doris Koesling, Corde Wit, Frank J. Kaiser, Anja Medack, Marcus Fischer, Martina E. Zimmermann, Stephanie Tennstedt, Elisabeth Graf, **Sebastian H Eck**, Zouhair Aherrahrou, Janja Nahrstaedt, Christina Willenborg, Petra Bruse, Markus M. Nthen, Per Hofmann, Peter S. Braund, Evanthia Mergia, Wibke Reinhard, Christof Burgdorf, Stefan Schreiber, Anthony J. Balmforth, Alistair S. Hall, Lars Bertram,

Elisabeth Steinhagen-Thiessen, Shu-Chen Li, Winfried Mrz, Muredach Reilly, Sekar Kathiresan, Ruth McPherson, Ulrich Walter, CARDIoGRAM, Jurg Ott, Nilesh J. Samani, Tim M. Strom, Thomas Meitinger, Heribert Schunkert*, Christian Hengstenberg*
submitted

* - Authors contributed equally

Acknowledgments

There are a lot of people without whom this work would not have been possible and I want to express my genuine gratitude to all of them. First of all I want to thank my supervisor Dr. Tim Strom for excellent guidance and mentoring. Thank you Tim, for the all the large and small things I learned.

Next I would like to thank Prof. Dr. Thomas Meitinger for the opportunity to develop my thesis in his departement and for crucial advice concerning my work.

I also give my sincere thanks to Prof. Dr. Werner Mewes, my Ph.D. advisor, for his interest and fruiteful discussion on numerous ways to improve the thesis.

I can't thank Dr. Anna Benet-Pages enough for being the best colleague imaginable and giving me the very reason that convinced me to accept this PhD position and Bettina and Sandy for accepting their first bioinformatician. Furthermore, I thank all colleagues of the Institute of Human Genetics for the exceptionally nice working atmosphere and for being more than just co-workers. In particular I thank Anne for Vogtland, Barbara for biking, Bettina for being her favorite PhD student, Birgit for getting my name on the PhD candidate list, Carola for the parking lot, Elisabeth for the emperor of Freising, Franziska for squirreling around, good Katharina Heim for invaluable help in R, Juliane for explanations of the obvious, Thomas for speed improvements, Susanne for cookies and Simon for Canada.

Lastly, I thank my friends and family for their ongoing support without which this Thesis would never have been possible, especially Andrea for correcting my countless versions.

Bibliography

[1] Adzhubei, Ivan A. ; Schmidt, Steffen ; Peshkin, Leonid ; Ramensky, Vasily E. ; Gerasimova, Anna ; Bork, Peer ; Kondrashov, Alexey S. ; Sunyaev, Shamil R.: A method and server for predicting damaging missense mutations. In: *Nat Methods* 7 (2010), Apr, Nr. 4, 248–249. http://dx.doi.org/10.1038/nmeth0410-248. – DOI 10.1038/nmeth0410–248

[2] Altschul, S. F. ; Gish, W. ; Miller, W. ; Myers, E. W. ; Lipman, D. J.: Basic local alignment search tool. In: *J Mol Biol* 215 (1990), Oct, Nr. 3, 403–410. http://dx.doi.org/10.1006/jmbi.1990.9999. – DOI 10.1006/jmbi.1990.9999

[3] Altshuler, David ; Daly, Mark J. ; Lander, Eric S.: Genetic mapping in human disease. In: *Science* 322 (2008), Nov, Nr. 5903, 881–888. http://dx.doi.org/10.1126/science.1156409. – DOI 10.1126/science.1156409

[4] Becker, Jutta ; Semler, Oliver ; Gilissen, Christian ; Li, Yun ; Bolz, Hanno J. ; Giunta, Cecilia ; Bergmann, Carsten ; Rohrbach, Marianne ; Koerber, Friederike ; Zimmermann, Katharina ; Vries, Petra de ; Wirth, Brunhilde ; Schoenau, Eckhard ; Wollnik, Bernd ; Veltman, Joris A. ; Hoischen, Alexander ; Netzer, Christian: Exome sequencing identifies truncating mutations in human SERPINF1 in autosomal-recessive osteogenesis imperfecta. In: *Am J Hum Genet* 88 (2011), Mar, Nr. 3, 362–371. http://dx.doi.org/10.1016/j.ajhg.2011.01.015. – DOI 10.1016/j.ajhg.2011.01.015

[5] Bell, Callum J. ; Dinwiddie, Darrell L. ; Miller, Neil A. ; Hateley, Shannon L. ; Ganusova, Elena E. ; Mudge, Joann ;

Langley, Ray J. ; Zhang, Lu ; Lee, Clarence C. ; Schilkey, Faye D. ; Sheth, Vrunda ; Woodward, Jimmy E. ; Peckham, Heather E. ; Schroth, Gary P. ; Kim, Ryan W. ; Kingsmore, Stephen F.: Carrier testing for severe childhood recessive diseases by next-generation sequencing. In: *Sci Transl Med* 3 (19932011), Jan, Nr. 65, 65ra4. http://dx.doi.org/10.1126/scitranslmed.3001756. – DOI 10.1126/scitranslmed.3001756

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

; Heyer, Narinder I. ; Hims, Matthew M. ; Ho, Johnny T. ; Horgan, Adrian M. ; Hoschler, Katya ; Hurwitz, Steve ; Ivanov, Denis V. ; Johnson, Maria Q. ; James, Terena ; Jones, T. A. H. ; Kang, Gyoung-Dong ; Kerelska, Tzvetana H. ; Kersey, Alan D. ; Khrebtukova, Irina ; Kindwall, Alex P. ; Kingsbury, Zoya ; Kokko-Gonzales, Paula I. ; Kumar, Anil ; Laurent, Marc A. ; Lawley, Cynthia T. ; Lee, Sarah E. ; Lee, Xavier ; Liao, Arnold K. ; Loch, Jennifer A. ; Lok, Mitch ; Luo, Shujun ; Mammen, Radhika M. ; Martin, John W. ; McCauley, Patrick G. ; McNitt, Paul ; Mehta, Parul ; Moon, Keith W. ; Mullens, Joe W. ; Newington, Taksina ; Ning, Zemin ; Ng, Bee L. ; Novo, Sonia M. ; O'Neill, Michael J. ; Osborne, Mark A. ; Osnowski, Andrew ; Ostadan, Omead ; Paraschos, Lambros L. ; Pickering, Lea ; Pike, Andrew C. ; Pike, Alger C. ; Pinkard, D. C. ; Pliskin, Daniel P. ; Podhasky, Joe ; Quijano, Victor J. ; Raczy, Come ; Rae, Vicki H. ; Rawlings, Stephen R. ; Rodriguez, Ana C. ; Roe, Phyllida M. ; Rogers, John ; Bacigalupo, Maria C R. ; Romanov, Nikolai ; Romieu, Anthony ; Roth, Rithy K. ; Rourke, Natalie J. ; Ruediger, Silke T. ; Rusman, Eli ; Sanches-Kuiper, Raquel M. ; Schenker, Martin R. ; Seoane, Josefina M. ; Shaw, Richard J. ; Shiver, Mitch K. ; Short, Steven W. ; Sizto, Ning L. ; Sluis, Johannes P. ; Smith, Melanie A. ; Sohna, Jean Ernest S. ; Spence, Eric J. ; Stevens, Kim ; Sutton, Neil ; Szajkowski, Lukasz ; Tregidgo, Carolyn L. ; Turcatti, Gerardo ; Vandevondele, Stephanie ; Verhovsky, Yuli ; Virk, Selene M. ; : Accurate whole human genome sequencing using reversible terminator chemistry. In: *Nature* 456 (2008), Nov, Nr. 7218, 53–59. `http://dx.doi.org/10.1038/nature07517`. – DOI 10.1038/nature07517

[7] Bonifacino, Juan S. ; Rojas, Raul: Retrograde transport from endosomes to the trans-Golgi network. In: *Nat Rev Mol Cell Biol* 7 (2006), Aug, Nr. 8, 568–579. `http://dx.doi.org/10.1038/nrm1985`. – DOI 10.1038/nrm1985

[8] Borck, Guntram ; Molla-Herman, Anahi ; Boddaert, Nathalie ; Encha-Razavi, Ferechte ; Philippe, Anne ; Robel, Laurence ; Desguerre, Isabelle ; Brunelle, Francis ; Benmerah, Alexandre ; Munnich, Arnold ; Colleaux, Laurence: Clin-

ical, cellular, and neuropathological consequences of AP1S2 mutations: further delineation of a recognizable X-linked mental retardation syndrome. In: *Hum Mutat* 29 (2008), Jul, Nr. 7, 966–974. http://dx.doi.org/10.1002/humu.20531. – DOI 10.1002/humu.20531

[9] Borinstein, S. C. ; Hyatt, M. A. ; Sykes, V. W. ; Straub, R. E. ; Lipkowitz, S. ; Boulter, J. ; Bogler, O.: SETA is a multifunctional adapter protein with three SH3 domains that binds Grb2, Cbl, and the novel SB1 proteins. In: *Cell Signal* 12 (2000), Dec, Nr. 11-12, S. 769–779

[10] Bromberg, Yana ; Rost, Burkhard: SNAP: predict effect of non-synonymous polymorphisms on function. In: *Nucleic Acids Res* 35 (2007), Nr. 11, 3823–3835. http://dx.doi.org/10.1093/nar/gkm238. – DOI 10.1093/nar/gkm238

[11] Burgos, Patricia V. ; Mardones, Gonzalo A. ; Rojas, Adriana L. ; daSilva, Luis L P. ; Prabhu, Yogikala ; Hurley, James H. ; Bonifacino, Juan S.: Sorting of the Alzheimer's disease amyloid precursor protein mediated by the AP-4 complex. In: *Dev Cell* 18 (2010), Mar, Nr. 3, 425–436. http://dx.doi.org/10.1016/j.devcel.2010.01.015. – DOI 10.1016/j.devcel.2010.01.015

[12] Chen, Ken ; Wallis, John W. ; McLellan, Michael D. ; Larson, David E. ; Kalicki, Joelle M. ; Pohl, Craig S. ; McGrath, Sean D. ; Wendl, Michael C. ; Zhang, Qunyuan ; Locke, Devin P. ; Shi, Xiaoqi ; Fulton, Robert S. ; Ley, Timothy J. ; Wilson, Richard K. ; Ding, Li ; Mardis, Elaine R.: BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. In: *Nat Methods* 6 (2009), Sep, Nr. 9, 677–681. http://dx.doi.org/10.1038/nmeth.1363. – DOI 10.1038/nmeth.1363

[13] Choi, Murim ; Scholl, Ute I. ; Ji, Weizhen ; Liu, Tiewen ; Tikhonova, Irina R. ; Zumbo, Paul ; Nayir, Ahmet ; Bakkaloglu, Aycin ; Ozen, Seza ; Sanjad, Sami ; Nelson-Williams, Carol

; Farhi, Anita ; Mane, Shrikant ; Lifton, Richard P.: Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. In: *Proc Natl Acad Sci U S A* 106 (2009), Nov, Nr. 45, 19096–19101. http://dx.doi.org/10.1073/pnas.0910672106. – DOI 10.1073/pnas.0910672106

[14] Clark, Michael J. ; Chen, Rui ; Lam, Hugo Y K. ; Karczewski, Konrad J. ; Chen, Rong ; Euskirchen, Ghia ; Butte, Atul J. ; Snyder, Michael: Performance comparison of exome DNA sequencing technologies. In: *Nat Biotechnol* 29 (2011), Oct, Nr. 10, 908–914. http://dx.doi.org/10.1038/nbt.1975. – DOI 10.1038/nbt.1975

[15] Clarke, James ; Wu, Hai-Chen ; Jayasinghe, Lakmal ; Patel, Alpesh ; Reid, Stuart ; Bayley, Hagan: Continuous base identification for single-molecule nanopore DNA sequencing. In: *Nat Nanotechnol* 4 (2009), Apr, Nr. 4, 265–270. http://dx.doi.org/10.1038/nnano.2009.12. – DOI 10.1038/nnano.2009.12

[16] Clerget-Darpoux, F. ; Bonaiti-Pellie, C. ; Hochez, J.: Effects of misspecifying genetic parameters in lod score analysis. In: *Biometrics* 42 (1986), Jun, Nr. 2, S. 393–399

[17] Consortium, 1000 Genomes P.: A map of human genome variation from population-scale sequencing. In: *Nature* 467 (2010), Oct, Nr. 7319, 1061–1073. http://dx.doi.org/10.1038/nature09534. – DOI 10.1038/nature09534

[18] Consortium, International H.: The International HapMap Project. In: *Nature* 426 (2003), Dec, Nr. 6968, S. 789–796

[19] Consortium, International H.: A haplotype map of the human genome. In: *Nature* 437 (2005), Oct, Nr. 7063, S. 1299–1320

[20] Consortium, International H. ; Frazer, Kelly A. ; Ballinger, Dennis G. ; Cox, David R. ; Hinds, David A. ; Stuve, Laura L. ; Gibbs, Richard A. ; Belmont, John W. ; Boudreau, Andrew ; Hardenbol, Paul ; Leal, Suzanne M. ; Pasternak, Shiran ; Wheeler, David A. ; Willis, Thomas D. ; Yu, Fuli ; Yang, Huanming ; Zeng, Changqing ; Gao, Yang ; Hu, Haoran ; Hu, Weitao

; Li, Chaohua ; Lin, Wei ; Liu, Siqi ; Pan, Hao ; Tang, Xiaoli ; Wang, Jian ; Wang, Wei ; Yu, Jun ; Zhang, Bo ; Zhang, Qingrun ; Zhao, Hongbin ; Zhao, Hui ; Zhou, Jun ; Gabriel, Stacey B. ; Barry, Rachel ; Blumenstiel, Brendan ; Camargo, Amy ; Defelice, Matthew ; Faggart, Maura ; Goyette, Mary ; Gupta, Supriya ; Moore, Jamie ; Nguyen, Huy ; Onofrio, Robert C. ; Parkin, Melissa ; Roy, Jessica ; Stahl, Erich ; Winchester, Ellen ; Ziaugra, Liuda ; Altshuler, David ; Shen, Yan ; Yao, Zhijian ; Huang, Wei ; Chu, Xun ; He, Yungang ; Jin, Li ; Liu, Yangfan ; Shen, Yayun ; Sun, Weiwei ; Wang, Haifeng ; Wang, Yi ; Wang, Ying ; Xiong, Xiaoyan ; Xu, Liang ; Waye, Mary M Y. ; Tsui, Stephen K W. ; Xue, Hong ; Wong, J. Tze-Fei ; Galver, Luana M. ; Fan, Jian-Bing ; Gunderson, Kevin ; Murray, Sarah S. ; Oliphant, Arnold R. ; Chee, Mark S. ; Montpetit, Alexandre ; Chagnon, Fanny ; Ferretti, Vincent ; Leboeuf, Martin ; Olivier, Jean-Francois ; Phillips, Michael S. ; Roumy, Stephanie ; Sallie, Clementine ; Verner, Andrei ; Hudson, Thomas J. ; Kwok, Pui-Yan ; Cai, Dongmei ; Koboldt, Daniel C. ; Miller, Raymond D. ; Pawlikowska, Ludmila ; Taillon-Miller, Patricia ; Xiao, Ming ; Tsui, Lap-Chee ; Mak, William ; Song, You Q. ; Tam, Paul K H. ; Nakamura, Yusuke ; Kawaguchi, Takahisa ; Kitamoto, Takuya ; Morizono, Takashi ; Nagashima, Atsushi ; Ohnishi, Yozo ; Sekine, Akihiro ; Tanaka, Toshihiro ; Tsunoda, Tatsuhiko ; Deloukas, Panos ; Bird, Christine P. ; Delgado, Marcos ; Dermitzakis, Emmanouil T. ; Gwilliam, Rhian ; Hunt, Sarah ; Morrison, Jonathan ; Powell, Don ; Stranger, Barbara E. ; Whittaker, Pamela ; Bentley, David R. ; Daly, Mark J. ; Bakker, Paul I W. ; Barrett, Jeff ; Chretien, Yves R. ; Maller, Julian ; McCarroll, Steve ; Patterson, Nick ; Pe'er, Itsik ; Price, Alkes ; Purcell, Shaun ; Richter, Daniel J. ; Sabeti, Pardis ; Saxena, Richa ; Schaffner, Stephen F. ; Sham, Pak C. ; Varilly, Patrick ; Altshuler, David ; Stein, Lincoln D. ; Krishnan, Lalitha ; Smith, Albert V. ; Tello-Ruiz, Marcela K. ; Thorisson, Gudmundur A. ; Chakravarti, Aravinda ; Chen, Peter E. ; Cutler, David J. ; Kashuk, Carl S. ; Lin, Shin ; Abecasis, Goncalo R. ; Guan, Weihua ; Li, Yun ; Munro, Heather M. ; Qin, Zhaohui S. ; Thomas, Daryl J. ; McVean, Gilean ; Auton, Adam ; Bottolo, Leonardo ; Cardin, Niall ; Eyheramendy, Susana ; Freeman, Colin ; Marchini, Jonathan ; Myers,

Simon ; Spencer, Chris ; Stephens, Matthew ; Donnelly, Peter ; Cardon, Lon R. ; Clarke, Geraldine ; Evans, David M. ; Morris, Andrew P. ; Weir, Bruce S. ; Tsunoda, Tatsuhiko ; Mullikin, James C. ; Sherry, Stephen T. ; Feolo, Michael ; Skol, Andrew ; Zhang, Houcan ; Zeng, Changqing ; Zhao, Hui ; Matsuda, Ichiro ; Fukushima, Yoshimitsu ; Macer, Darryl R. ; Suda, Eiko ; Rotimi, Charles N. ; Adebamowo, Clement A. ; Ajayi, Ike ; Aniagwu, Toyin ; Marshall, Patricia A. ; Nkwodimmah, Chibuzor ; Royal, Charmaine D M. ; Leppert, Mark F. ; Dixon, Missy ; Peiffer, Andy ; Qiu, Renzong ; : A second generation human haplotype map of over 3.1 million SNPs. In: *Nature* 449 (2007), Oct, Nr. 7164, 851–861. http://dx.doi.org/10.1038/nature06258. – DOI 10.1038/nature06258

[21] Consortium, Mouse Genome S. ; Waterston, Robert H. ; Lindblad-Toh, Kerstin ; Birney, Ewan ; Rogers, Jane ; Abril, Josep F. ; Agarwal, Pankaj ; Agarwala, Richa ; Ainscough, Rachel ; Alexandersson, Marina ; An, Peter ; Antonarakis, Stylianos E. ; Attwood, John ; Baertsch, Robert ; Bailey, Jonathon ; Barlow, Karen ; Beck, Stephan ; Berry, Eric ; Birren, Bruce ; Bloom, Toby ; Bork, Peer ; Botcherby, Marc ; Bray, Nicolas ; Brent, Michael R. ; Brown, Daniel G. ; Brown, Stephen D. ; Bult, Carol ; Burton, John ; Butler, Jonathan ; Campbell, Robert D. ; Carninci, Piero ; Cawley, Simon ; Chiaromonte, Francesca ; Chinwalla, Asif T. ; Church, Deanna M. ; Clamp, Michele ; Clee, Christopher ; Collins, Francis S. ; Cook, Lisa L. ; Copley, Richard R. ; Coulson, Alan ; Couronne, Olivier ; Cuff, James ; Curwen, Val ; Cutts, Tim ; Daly, Mark ; David, Robert ; Davies, Joy ; Delehaunty, Kimberly D. ; Deri, Justin ; Dermitzakis, Emmanouil T. ; Dewey, Colin ; Dickens, Nicholas J. ; Diekhans, Mark ; Dodge, Sheila ; Dubchak, Inna ; Dunn, Diane M. ; Eddy, Sean R. ; Elnitski, Laura ; Emes, Richard D. ; Eswara, Pallavi ; Eyras, Eduardo ; Felsenfeld, Adam ; Fewell, Ginger A. ; Flicek, Paul ; Foley, Karen ; Frankel, Wayne N. ; Fulton, Lucinda A. ; Fulton, Robert S. ; Furey, Terrence S. ; Gage, Diane ; Gibbs, Richard A. ; Glusman, Gustavo ; Gnerre, Sante ; Goldman, Nick ; Goodstadt, Leo ; Grafham, Darren ; Graves, Tina A. ; Green, Eric D.

; Gregory, Simon ; Guige, Roderic ; Guyer, Mark ; Hardison, Ross C. ; Haussler, David ; Hayashizaki, Yoshihide ; Hillier, LaDeana W. ; Hinrichs, Angela ; Hlavina, Wratko ; Holzer, Timothy ; Hsu, Fan ; Hua, Axin ; Hubbard, Tim ; Hunt, Adrienne ; Jackson, Ian ; Jaffe, David B. ; Johnson, L. S. ; Jones, Matthew ; Jones, Thomas A. ; Joy, Ann ; Kamal, Michael ; Karlsson, Elinor K. ; Karolchik, Donna ; Kasprzyk, Arkadiusz ; Kawai, Jun ; Keibler, Evan ; Kells, Cristyn ; Kent, W. J. ; Kirby, Andrew ; Kolbe, Diana L. ; Korf, Ian ; Kucherlapati, Raju S. ; Kulbokas, Edward J. ; Kulp, David ; Landers, Tom ; Leger, J. P. ; Leonard, Steven ; Letunic, Ivica ; Levine, Rosie ; Li, Jia ; Li, Ming ; Lloyd, Christine ; Lucas, Susan ; Ma, Bin ; Maglott, Donna R. ; Mardis, Elaine R. ; Matthews, Lucy ; Mauceli, Evan ; Mayer, John H. ; McCarthy, Megan ; McCombie, W. R. ; McLaren, Stuart ; McLay, Kirsten ; McPherson, John D. ; Meldrim, Jim ; Meredith, Beverley ; Mesirov, Jill P. ; Miller, Webb ; Miner, Tracie L. ; Mongin, Emmanuel ; Montgomery, Kate T. ; Morgan, Michael ; Mott, Richard ; Mullikin, James C. ; Muzny, Donna M. ; Nash, William E. ; Nelson, Joanne O. ; Nhan, Michael N. ; Nicol, Robert ; Ning, Zemin ; Nusbaum, Chad ; O'Connor, Michael J. ; Okazaki, Yasushi ; Oliver, Karen ; Overton-Larty, Emma ; Pachter, Lior ; Parra, Genes ; Pepin, Kymberlie H. ; Peterson, Jane ; Pevzner, Pavel ; Plumb, Robert ; Pohl, Craig S. ; Poliakov, Alex ; Ponce, Tracy C. ; Ponting, Chris P. ; Potter, Simon ; Quail, Michael ; Reymond, Alexandre ; Roe, Bruce A. ; Roskin, Krishna M. ; Rubin, Edward M. ; Rust, Alistair G. ; Santos, Ralph ; Sapojnikov, Victor ; Schultz, Brian ; Schultz, Joerg ; Schwartz, Matthias S. ; Schwartz, Scott ; Scott, Carol ; Seaman, Steven ; Searle, Steve ; Sharpe, Ted ; Sheridan, Andrew ; Shownkeen, Ratna ; Sims, Sarah ; Singer, Jonathan B. ; Slater, Guy ; Smit, Arian ; : Initial sequencing and comparative analysis of the mouse genome. In: *Nature* 420 (2002), Dec, Nr. 6915, 520–562. http://dx.doi.org/10.1038/nature01262. – DOI 10.1038/nature01262

[22] Cooper, D. N. ; Ball, E. V. ; Krawczak, M.: The human gene mutation database. In: *Nucleic Acids Res* 26 (1998), Jan, Nr. 1, S. 285–287

[23] Cooper, David N. ; Stenson, Peter D. ; Chuzhanova, Nadia A.: The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms. In: *Curr Protoc Bioinformatics* Chapter 1 (2006), Jan, Unit 1.13. http://dx.doi.org/10.1002/0471250953.bi0113s12. – DOI 10.1002/0471250953.bi0113s12

[24] Danecek, Petr ; Auton, Adam ; Abecasis, Goncalo ; Albers, Cornelis A. ; Banks, Eric ; DePristo, Mark A. ; Handsaker, Robert E. ; Lunter, Gerton ; Marth, Gabor T. ; Sherry, Stephen T. ; McVean, Gilean ; Durbin, Richard ; Group, 1000 Genomes Project A.: The variant call format and VCFtools. In: *Bioinformatics* 27 (2011), Aug, Nr. 15, 2156–2158. http://dx.doi.org/10.1093/bioinformatics/btr330. – DOI 10.1093/bioinformatics/btr330

[25] Dayyani, Farshid ; Wang, Jianfeng ; Yeh, Jing-Ruey J. ; Ahn, Eun-Young ; Tobey, Erica ; Zhang, Dong-Er ; Bernstein, Irwin D. ; Peterson, Randall T. ; Sweetser, David A.: Loss of TLE1 and TLE4 from the del(9q) commonly deleted region in AML cooperates with AML1-ETO to affect myeloid cell proliferation and survival. In: *Blood* 111 (2008), Apr, Nr. 8, 4338–4347. http://dx.doi.org/10.1182/blood-2007-07-103291. – DOI 10.1182/blood–2007–07–103291

[26] Derrington, Ian M. ; Butler, Tom Z. ; Collins, Marcus D. ; Manrao, Elizabeth ; Pavlenok, Mikhail ; Niederweis, Michael ; Gundlach, Jens H.: Nanopore DNA sequencing with MspA. In: *Proc Natl Acad Sci U S A* 107 (2010), Sep, Nr. 37, 16060–16065. http://dx.doi.org/10.1073/pnas.1001831107. – DOI 10.1073/pnas.1001831107

[27] Dohm, Juliane C. ; Lottaz, Claudio ; Borodina, Tatiana ; Himmelbauer, Heinz: Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. In: *Nucleic Acids Res* 36 (2008), Sep, Nr. 16, e105. http://dx.doi.org/10.1093/nar/gkn425. – DOI 10.1093/nar/gkn425

[28] Drmanac, Radoje ; Sparks, Andrew B. ; Callow, Matthew J. ; Halpern, Aaron L. ; Burns, Norman L. ; Kermani, Bahram G. ;

Carnevali, Paolo ; Nazarenko, Igor ; Nilsen, Geoffrey B. ; Yeung, George ; Dahl, Fredrik ; Fernandez, Andres ; Staker, Bryan ; Pant, Krishna P. ; Baccash, Jonathan ; Borcherding, Adam P. ; Brownley, Anushka ; Cedeno, Ryan ; Chen, Linsu ; Chernikoff, Dan ; Cheung, Alex ; Chirita, Razvan ; Curson, Benjamin ; Ebert, Jessica C. ; Hacker, Coleen R. ; Hartlage, Robert ; Hauser, Brian ; Huang, Steve ; Jiang, Yuan ; Karpinchyk, Vitali ; Koenig, Mark ; Kong, Calvin ; Landers, Tom ; Le, Catherine ; Liu, Jia ; McBride, Celeste E. ; Morenzoni, Matt ; Morey, Robert E. ; Mutch, Karl ; Perazich, Helena ; Perry, Kimberly ; Peters, Brock A. ; Peterson, Joe ; Pethiyagoda, Charit L. ; Pothuraju, Kaliprasad ; Richter, Claudia ; Rosenbaum, Abraham M. ; Roy, Shaunak ; Shafto, Jay ; Sharanhovich, Uladzislau ; Shannon, Karen W. ; Sheppy, Conrad G. ; Sun, Michel ; Thakuria, Joseph V. ; Tran, Anne ; Vu, Dylan ; Zaranek, Alexander W. ; Wu, Xiaodi ; Drmanac, Snezana ; Oliphant, Arnold R. ; Banyai, William C. ; Martin, Bruce ; Ballinger, Dennis G. ; Church, George M. ; Reid, Clifford A.: Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. In: *Science* 327 (2010), Jan, Nr. 5961, 78–81. http://dx.doi.org/10.1126/science.1181498. – DOI 10.1126/science.1181498

[29] Eid, John ; Fehr, Adrian ; Gray, Jeremy ; Luong, Khai ; Lyle, John ; Otto, Geoff ; Peluso, Paul ; Rank, David ; Baybayan, Primo ; Bettman, Brad ; Bibillo, Arkadiusz ; Bjornson, Keith ; Chaudhuri, Bidhan ; Christians, Frederick ; Cicero, Ronald ; Clark, Sonya ; Dalal, Ravindra ; Dewinter, Alex ; Dixon, John ; Foquet, Mathieu ; Gaertner, Alfred ; Hardenbol, Paul ; Heiner, Cheryl ; Hester, Kevin ; Holden, David ; Kearns, Gregory ; Kong, Xiangxu ; Kuse, Ronald ; Lacroix, Yves ; Lin, Steven ; Lundquist, Paul ; Ma, Congcong ; Marks, Patrick ; Maxham, Mark ; Murphy, Devon ; Park, Insil ; Pham, Thang ; Phillips, Michael ; Roy, Joy ; Sebra, Robert ; Shen, Gene ; Sorenson, Jon ; Tomaney, Austin ; Travers, Kevin ; Trulson, Mark ; Vieceli, John ; Wegener, Jeffrey ; Wu, Dawn ; Yang, Alicia ; Zaccarin, Denis ; Zhao, Peter ; Zhong, Frank ; Korlach, Jonas ; Turner, Stephen: Real-time DNA sequencing from single polymerase molecules. In: *Science* 323 (2009), Jan, Nr. 5910, 133–138.

http://dx.doi.org/10.1126/science.1162986. – DOI 10.1126/science.1162986

[30] Ewing, B. ; Green, P.: Base-calling of automated sequencer traces using phred. II. Error probabilities. In: *Genome Res* 8 (1998), Mar, Nr. 3, S. 186–194

[31] Ewing, B. ; Hillier, L. ; Wendl, M. C. ; Green, P.: Base-calling of automated sequencer traces using phred. I. Accuracy assessment. In: *Genome Res* 8 (1998), Mar, Nr. 3, S. 175–185

[32] Gilissen, Christian ; Arts, Heleen H. ; Hoischen, Alexander ; Spruijt, Liesbeth ; Mans, Dorus A. ; Arts, Peer ; Lier, Bart van ; Steehouwer, Marloes ; Reeuwijk, Jeroen van ; Kant, Sarina G. ; Roepman, Ronald ; Knoers, Nine V A M. ; Veltman, Joris A. ; Brunner, Han G.: Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. In: *Am J Hum Genet* 87 (2010), Sep, Nr. 3, 418–423. http://dx.doi.org/10.1016/j.ajhg.2010.08.004. – DOI 10.1016/j.ajhg.2010.08.004

[33] Gilissen, Christian ; Hoischen, Alexander ; Brunner, Han G. ; Veltman, Joris A.: Disease gene identification strategies for exome sequencing. In: *Eur J Hum Genet* 20 (2012), May, Nr. 5, 490–497. http://dx.doi.org/10.1038/ejhg.2011.258. – DOI 10.1038/ejhg.2011.258

[34] Girard, Simon L. ; Gauthier, Julie ; Noreau, Anne ; Xiong, Lan ; Zhou, Sirui ; Jouan, Loubna ; Dionne-Laporte, Alexandre ; Spiegelman, Dan ; Henrion, Edouard ; Diallo, Ousmane ; Thibodeau, Pascale ; Bachand, Isabelle ; Bao, Jessie Y J. ; Tong, Amy Hin Y. ; Lin, Chi-Ho ; Millet, Bruno ; Jaafari, Nematollah ; Joober, Ridha ; Dion, Patrick A. ; Lok, Si ; Krebs, Marie-Odile ; Rouleau, Guy A.: Increased exonic de novo mutation rate in individuals with schizophrenia. In: *Nat Genet* 43 (2011), Nr. 9, 860–863. http://dx.doi.org/10.1038/ng.886. – DOI 10.1038/ng.886

[35] Glenn, Travis C.: Field guide to next-generation DNA sequencers. In: *Mol Ecol Resour* 11 (2011), Sep, Nr. 5, 759–769.

http://dx.doi.org/10.1111/j.1755-0998.2011.03024.x. – DOI 10.1111/j.1755–0998.2011.03024.x

[36] Gnirke, Andreas ; Melnikov, Alexandre ; Maguire, Jared ; Rogov, Peter ; LeProust, Emily M. ; Brockman, William ; Fennell, Timothy ; Giannoukos, Georgia ; Fisher, Sheila ; Russ, Carsten ; Gabriel, Stacey ; Jaffe, David B. ; Lander, Eric S. ; Nusbaum, Chad: Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. In: *Nat Biotechnol* 27 (2009), Feb, Nr. 2, 182–189. http://dx.doi.org/10.1038/nbt.1523. – DOI 10.1038/nbt.1523

[37] Gotoh, O.: An improved algorithm for matching biological sequences. In: *J Mol Biol* 162 (1982), Dec, Nr. 3, S. 705–708

[38] Greif, P. A. ; Eck, S. H. ; Konstandin, N. P. ; Benet-Pages, A. ; Ksienzyk, B. ; Dufour, A. ; Vetter, A. T. ; Popp, H. D. ; Lorenz-Depiereux, B. ; Meitinger, T. ; Bohlander, S. K. ; Strom, T. M.: Identification of recurring tumor-specific somatic mutations in acute myeloid leukemia by transcriptome sequencing. In: *Leukemia* 25 (2011), May, Nr. 5, 821–827. http://dx.doi.org/10.1038/leu.2011.19. – DOI 10.1038/leu.2011.19

[39] Hierro, Aitor ; Rojas, Adriana L. ; Rojas, Raul ; Murthy, Namita ; Effantin, Gregory ; Kajava, Andrey V. ; Steven, Alasdair C. ; Bonifacino, Juan S. ; Hurley, James H.: Functional architecture of the retromer cargo-recognition complex. In: *Nature* 449 (2007), Oct, Nr. 7165, 1063–1067. http://dx.doi.org/10.1038/nature06216. – DOI 10.1038/nature06216

[40] Hoischen, Alexander ; Bon, Bregje W M. ; Gilissen, Christian ; Arts, Peer ; Lier, Bart van ; Steehouwer, Marloes ; Vries, Petra de ; Reuver, Rick de ; Wieskamp, Nienke ; Mortier, Geert ; Devriendt, Koen ; Amorim, Marta Z. ; Revencu, Nicole ; Kidd, Alexa ; Barbosa, Mafalda ; Turner, Anne ; Smith, Janine ; Oley, Christina ; Henderson, Alex ; Hayes, Ian M. ; Thompson, Elizabeth M. ; Brunner, Han G. ; Vries, Bert B A.

; Veltman, Joris A.: De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. In: *Nat Genet* 42 (2010), Jun, Nr. 6, 483–485. http://dx.doi.org/10.1038/ng.581. – DOI 10.1038/ng.581

[41] Ikram, M. K. ; Sim, Xueling ; Xueling, Sim ; Jensen, Richard A. ; Cotch, Mary F. ; Hewitt, Alex W. ; Ikram, M. A. ; Wang, Jie J. ; Klein, Ronald ; Klein, Barbara E K. ; Breteler, Monique M B. ; Cheung, Ning ; Liew, Gerald ; Mitchell, Paul ; Uitterlinden, Andre G. ; Rivadeneira, Fernando ; Hofman, Albert ; Jong, Paulus T V M. ; Duijn, Cornelia M. ; Kao, Linda ; Cheng, Ching-Yu ; Smith, Albert V. ; Glazer, Nicole L. ; Lumley, Thomas ; McKnight, Barbara ; Psaty, Bruce M. ; Jonasson, Fridbert ; Eiriksdottir, Gudny ; Aspelund, Thor ; Consortium, Global B. ; Harris, Tamara B. ; Launer, Lenore J. ; Taylor, Kent D. ; Li, Xiaohui ; Iyengar, Sudha K. ; Xi, Quansheng ; Sivakumaran, Theru A. ; Mackey, David A. ; Macgregor, Stuart ; Martin, Nicholas G. ; Young, Terri L. ; Bis, Josh C. ; Wiggins, Kerri L. ; Heckbert, Susan R. ; Hammond, Christopher J. ; Andrew, Toby ; Fahy, Samantha ; Attia, John ; Holliday, Elizabeth G. ; Scott, Rodney J. ; Islam, F. M A. ; Rotter, Jerome I. ; McAuley, Annie K. ; Boerwinkle, Eric ; Tai, E. S. ; Gudnason, Vilmundur ; Siscovick, David S. ; Vingerling, Johannes R. ; Wong, Tien Y.: Four novel Loci (19q13, 6q24, 12q24, and 5q14) influence the microcirculation in vivo. In: *PLoS Genet* 6 (2010), Oct, Nr. 10, e1001184. http://dx.doi.org/10.1371/journal.pgen.1001184. – DOI 10.1371/journal.pgen.1001184

[42] Jamra, Rami A. ; Philippe, Orianne ; Raas-Rothschild, Annick ; Eck, Sebastian H. ; Graf, Elisabeth ; Buchert, Rebecca ; Borck, Guntram ; Ekici, Arif ; Brockschmidt, Felix F. ; Noethen, Markus M. ; Munnich, Arnold ; Strom, Tim M. ; Reis, Andre ; Colleaux, Laurence: Adaptor protein complex 4 deficiency causes severe autosomal-recessive intellectual disability, progressive spastic paraplegia, shy character, and short stature. In: *Am J Hum Genet* 88 (2011), Jun, Nr. 6, 788–795. http://dx.doi.org/10.1016/j.ajhg.2011.04.019. – DOI 10.1016/j.ajhg.2011.04.019

[43] Kent, W. J.: BLAT–the BLAST-like alignment tool. In: *Genome Res* 12 (2002), Apr, Nr. 4, 656–664. http://dx.doi.org/10.1101/gr.229202. Article published online before March 2002. – DOI 10.1101/gr.229202. Article published online before March 2002

[44] Korlach, Jonas ; Marks, Patrick J. ; Cicero, Ronald L. ; Gray, Jeremy J. ; Murphy, Devon L. ; Roitman, Daniel B. ; Pham, Thang T. ; Otto, Geoff A. ; Foquet, Mathieu ; Turner, Stephen W.: Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. In: *Proc Natl Acad Sci U S A* 105 (2008), Jan, Nr. 4, 1176–1181. http://dx.doi.org/10.1073/pnas.0710982105. – DOI 10.1073/pnas.0710982105

[45] Krawczak, M. ; Ball, E. V. ; Fenton, I. ; Stenson, P. D. ; Abeysinghe, S. ; Thomas, N. ; Cooper, D. N.: Human gene mutation database-a biomedical information and research resource. In: *Hum Mutat* 15 (2000), Nr. 1, 45–51. http://dx.doi.org/3.0.CO;2-T. – DOI 3.0.CO;2–T

[46] Krivanek, Ondrej L. ; Chisholm, Matthew F. ; Nicolosi, Valeria ; Pennycook, Timothy J. ; Corbin, George J. ; Dellby, Niklas ; Murfitt, Matthew F. ; Own, Christopher S. ; Szilagyi, Zoltan S. ; Oxley, Mark P. ; Pantelides, Sokrates T. ; Pennycook, Stephen J.: Atom-by-atom structural and chemical analysis by annular dark-field electron microscopy. In: *Nature* 464 (2010), Mar, Nr. 7288, 571–574. http://dx.doi.org/10.1038/nature08879. – DOI 10.1038/nature08879

[47] Krivov, Georgii G. ; Shapovalov, Maxim V. ; Dunbrack, Roland L.: Improved prediction of protein side-chain conformations with SCWRL4. In: *Proteins* 77 (2009), Dec, Nr. 4, 778–795. http://dx.doi.org/10.1002/prot.22488. – DOI 10.1002/prot.22488

[48] Kucho, Ken ichi ; Yoneda, Hidekatsu ; Harada, Manabu ; Ishiura, Masahiro: Determinants of sensitivity and specificity in spotted DNA microarrays with unmodified oligonucleotides. In: *Genes Genet Syst* 79 (2004), Aug, Nr. 4, S. 189–197

[49] Langmead, Ben ; Trapnell, Cole ; Pop, Mihai ; Salzberg, Steven L.: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. In: *Genome Biol* 10 (2009), Nr. 3, R25. http://dx.doi.org/10.1186/gb-2009-10-3-r25. – DOI 10.1186/gb–2009–10–3–r25

[50] Lee, Sang H. ; Wray, Naomi R. ; Goddard, Michael E. ; Visscher, Peter M.: Estimating missing heritability for disease from genome-wide association studies. In: *Am J Hum Genet* 88 (2011), Mar, Nr. 3, 294–305. http://dx.doi.org/10.1016/j.ajhg.2011.02.002. – DOI 10.1016/j.ajhg.2011.02.002

[51] Levene, M. J. ; Korlach, J. ; Turner, S. W. ; Foquet, M. ; Craighead, H. G. ; Webb, W. W.: Zero-mode waveguides for single-molecule analysis at high concentrations. In: *Science* 299 (2003), Jan, Nr. 5607, 682–686. http://dx.doi.org/10.1126/science.1079700. – DOI 10.1126/science.1079700

[52] Ley, Timothy J. ; Mardis, Elaine R. ; Ding, Li ; Fulton, Bob ; McLellan, Michael D. ; Chen, Ken ; Dooling, David ; Dunford-Shore, Brian H. ; McGrath, Sean ; Hickenbotham, Matthew ; Cook, Lisa ; Abbott, Rachel ; Larson, David E. ; Koboldt, Dan C. ; Pohl, Craig ; Smith, Scott ; Hawkins, Amy ; Abbott, Scott ; Locke, Devin ; Hillier, Ladeana W. ; Miner, Tracie ; Fulton, Lucinda ; Magrini, Vincent ; Wylie, Todd ; Glasscock, Jarret ; Conyers, Joshua ; Sander, Nathan ; Shi, Xiaoqi ; Osborne, John R. ; Minx, Patrick ; Gordon, David ; Chinwalla, Asif ; Zhao, Yu ; Ries, Rhonda E. ; Payton, Jacqueline E. ; Westervelt, Peter ; Tomasson, Michael H. ; Watson, Mark ; Baty, Jack ; Ivanovich, Jennifer ; Heath, Sharon ; Shannon, William D. ; Nagarajan, Rakesh ; Walter, Matthew J. ; Link, Daniel C. ; Graubert, Timothy A. ; DiPersio, John F. ; Wilson, Richard K.: DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. In: *Nature* 456 (2008), Nov, Nr. 7218, 66–72. http://dx.doi.org/10.1038/nature07485. – DOI 10.1038/nature07485

[53] Li, Heng ; Durbin, Richard: Fast and accurate short read alignment with Burrows-Wheeler transform.

In: *Bioinformatics* 25 (2009), Jul, Nr. 14, 1754–1760. http://dx.doi.org/10.1093/bioinformatics/btp324. – DOI 10.1093/bioinformatics/btp324

[54] Li, Heng ; Handsaker, Bob ; Wysoker, Alec ; Fennell, Tim ; Ruan, Jue ; Homer, Nils ; Marth, Gabor ; Abecasis, Goncalo ; Durbin, Richard ; Subgroup, 1000 Genome Project Data P.: The Sequence Alignment/Map format and SAMtools. In: *Bioinformatics* 25 (2009), Aug, Nr. 16, S. 2078–2079

[55] Li, Heng ; Homer, Nils: A survey of sequence alignment algorithms for next-generation sequencing. In: *Brief Bioinform* 11 (2010), Sep, Nr. 5, 473–483. http://dx.doi.org/10.1093/bib/bbq015. – DOI 10.1093/bib/bbq015

[56] Li, Heng ; Ruan, Jue ; Durbin, Richard: Mapping short DNA sequencing reads and calling variants using mapping quality scores. In: *Genome Res* 18 (2008), Nov, Nr. 11, 1851–1858. http://dx.doi.org/10.1101/gr.078212.108. – DOI 10.1101/gr.078212.108

[57] Li, Ruiqiang ; Yu, Chang ; Li, Yingrui ; Lam, Tak-Wah ; Yiu, Siu-Ming ; Kristiansen, Karsten ; Wang, Jun: SOAP2: an improved ultrafast tool for short read alignment. In: *Bioinformatics* 25 (2009), Aug, Nr. 15, 1966–1967. http://dx.doi.org/10.1093/bioinformatics/btp336. – DOI 10.1093/bioinformatics/btp336

[58] Li, Yingrui ; Vinckenbosch, Nicolas ; Tian, Geng ; Huerta-Sanchez, Emilia ; Jiang, Tao ; Jiang, Hui ; Albrechtsen, Anders ; Andersen, Gitte ; Cao, Hongzhi ; Korneliussen, Thorfinn ; Grarup, Niels ; Guo, Yiran ; Hellman, Ines ; Jin, Xin ; Li, Qibin ; Liu, Jiangtao ; Liu, Xiao ; Sparso, Thomas ; Tang, Meifang ; Wu, Honglong ; Wu, Renhua ; Yu, Chang ; Zheng, Hancheng ; Astrup, Arne ; Bolund, Lars ; Holmkvist, Johan ; Joergensen, Torben ; Kristiansen, Karsten ; Schmitz, Ole ; Schwartz, Thue W. ; Zhang, Xiuqing ; Li, Ruiqiang ; Yang, Huanming ; Wang, Jian ; Hansen, Torben ; Pedersen, Oluf ; Nielsen, Rasmus ; Wang, Jun: Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. In: *Nat Genet* 42 (2010),

Nov, Nr. 11, 969–972. http://dx.doi.org/10.1038/ng.680. – DOI 10.1038/ng.680

[59] Liu, Jian-Ping ; Liu, Nan-Song ; Yuan, Han-Ying ; Guo, Qian ; Lu, Hong ; Li, Yu-Yang: Human homologue of SETA binding protein 1 interacts with cathepsin B and participates in TNF-Induced apoptosis in ovarian cancer cells. In: *Mol Cell Biochem* 292 (2006), Nov, Nr. 1-2, 189–195. http://dx.doi.org/10.1007/s11010-006-9214-7. – DOI 10.1007/s11010–006–9214–7

[60] Ma, Bin ; Tromp, John ; Li, Ming: PatternHunter: faster and more sensitive homology search. In: *Bioinformatics* 18 (2002), Mar, Nr. 3, S. 440–445

[61] Manolio, Teri A.: Genomewide association studies and assessment of the risk of disease. In: *N Engl J Med* 363 (2010), Jul, Nr. 2, 166–176. http://dx.doi.org/10.1056/NEJMra0905980. – DOI 10.1056/NEJMra0905980

[62] Manolio, Teri A. ; Collins, Francis S. ; Cox, Nancy J. ; Goldstein, David B. ; Hindorff, Lucia A. ; Hunter, David J. ; McCarthy, Mark I. ; Ramos, Erin M. ; Cardon, Lon R. ; Chakravarti, Aravinda ; Cho, Judy H. ; Guttmacher, Alan E. ; Kong, Augustine ; Kruglyak, Leonid ; Mardis, Elaine ; Rotimi, Charles N. ; Slatkin, Montgomery ; Valle, David ; Whittemore, Alice S. ; Boehnke, Michael ; Clark, Andrew G. ; Eichler, Evan E. ; Gibson, Greg ; Haines, Jonathan L. ; Mackay, Trudy F C. ; McCarroll, Steven A. ; Visscher, Peter M.: Finding the missing heritability of complex diseases. In: *Nature* 461 (2009), Oct, Nr. 7265, 747–753. http://dx.doi.org/10.1038/nature08494. – DOI 10.1038/nature08494

[63] Mardis, Elaine R. ; Ding, Li ; Dooling, David J. ; Larson, David E. ; McLellan, Michael D. ; Chen, Ken ; Koboldt, Daniel C. ; Fulton, Robert S. ; Delehaunty, Kim D. ; McGrath, Sean D. ; Fulton, Lucinda A. ; Locke, Devin P. ; Magrini, Vincent J. ; Abbott, Rachel M. ; Vickery, Tammi L. ; Reed, Jerry S. ; Robinson, Jody S. ; Wylie, Todd ; Smith, Scott M. ; Carmichael, Lynn ; Eldred, James M. ; Harris, Christopher C. ; Walker, Jason ; Peck,

Joshua B. ; Du, Feiyu ; Dukes, Adam F. ; Sanderson, Gabriel E. ; Brummett, Anthony M. ; Clark, Eric ; McMichael, Joshua F. ; Meyer, Rick J. ; Schindler, Jonathan K. ; Pohl, Craig S. ; Wallis, John W. ; Shi, Xiaoqi ; Lin, Ling ; Schmidt, Heather ; Tang, Yuzhu ; Haipek, Carrie ; Wiechert, Madeline E. ; Ivy, Jolynda V. ; Kalicki, Joelle ; Elliott, Glendoria ; Ries, Rhonda E. ; Payton, Jacqueline E. ; Westervelt, Peter ; Tomasson, Michael H. ; Watson, Mark A. ; Baty, Jack ; Heath, Sharon ; Shannon, William D. ; Nagarajan, Rakesh ; Link, Daniel C. ; Walter, Matthew J. ; Graubert, Timothy A. ; DiPersio, John F. ; Wilson, Richard K. ; Ley, Timothy J.: Recurring mutations found by sequencing an acute myeloid leukemia genome. In: *N Engl J Med* 361 (2009), Sep, Nr. 11, 1058–1066. http://dx.doi.org/10.1056/NEJMoa0903840. – DOI 10.1056/NEJMoa0903840

[64] Maxam, A. M. ; Gilbert, W.: A new method for sequencing DNA. In: *Proc Natl Acad Sci U S A* 74 (1977), Feb, Nr. 2, S. 560–564

[65] McKernan, Kevin J. ; Peckham, Heather E. ; Costa, Gina L. ; McLaughlin, Stephen F. ; Fu, Yutao ; Tsung, Eric F. ; Clouser, Christopher R. ; Duncan, Cisyla ; Ichikawa, Jeffrey K. ; Lee, Clarence C. ; Zhang, Zheng ; Ranade, Swati S. ; Dimalanta, Eileen T. ; Hyland, Fiona C. ; Sokolsky, Tanya D. ; Zhang, Lei ; Sheridan, Andrew ; Fu, Haoning ; Hendrickson, Cynthia L. ; Li, Bin ; Kotler, Lev ; Stuart, Jeremy R. ; Malek, Joel A. ; Manning, Jonathan M. ; Antipova, Alena A. ; Perez, Damon S. ; Moore, Michael P. ; Hayashibara, Kathleen C. ; Lyons, Michael R. ; Beaudoin, Robert E. ; Coleman, Brittany E. ; Laptewicz, Michael W. ; Sannicandro, Adam E. ; Rhodes, Michael D. ; Gottimukkala, Rajesh K. ; Yang, Shan ; Bafna, Vineet ; Bashir, Ali ; MacBride, Andrew ; Alkan, Can ; Kidd, Jeffrey M. ; Eichler, Evan E. ; Reese, Martin G. ; Vega, Francisco M De L. ; Blanchard, Alan P.: Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. In: *Genome Res* 19 (2009), Sep, Nr. 9, 1527–1541. http://dx.doi.org/10.1101/gr.091868.109. – DOI 10.1101/gr.091868.109

[66] McNally, Ben ; Singer, Alon ; Yu, Zhiliang ; Sun, Yingjie ; Weng, Zhiping ; Meller, Amit: Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays. In: *Nano Lett* 10 (2010), Jun, Nr. 6, 2237–2244. http://dx.doi.org/10.1021/nl1012147. – DOI 10.1021/nl1012147

[67] Moran-Mirabal, Jose M. ; Craighead, Harold G.: Zero-mode waveguides: sub-wavelength nanostructures for single molecule studies at high concentrations. In: *Methods* 46 (2008), Sep, Nr. 1, 11–17. http://dx.doi.org/10.1016/j.ymeth.2008.05.010. – DOI 10.1016/j.ymeth.2008.05.010

[68] Mrozek, Krzysztof ; Marcucci, Guido ; Paschka, Peter ; Whitman, Susan P. ; Bloomfield, Clara D.: Clinical relevance of mutations and gene-expression changes in adult acute myeloid leukemia with normal cytogenetics: are we ready for a prognostically prioritized molecular classification? In: *Blood* 109 (2007), Jan, Nr. 2, 431–448. http://dx.doi.org/10.1182/blood-2006-06-001149. – DOI 10.1182/blood–2006–06–001149

[69] Need, Anna C. ; Shashi, Vandana ; Hitomi, Yuki ; Schoch, Kelly ; Shianna, Kevin V. ; McDonald, Marie T. ; Meisler, Miriam H. ; Goldstein, David B.: Clinical application of exome sequencing in undiagnosed genetic conditions. In: *J Med Genet* 49 (2012), Jun, Nr. 6, 353–361. http://dx.doi.org/10.1136/jmedgenet-2012-100819. – DOI 10.1136/jmedgenet–2012–100819

[70] Ng, Pauline C. ; Levy, Samuel ; Huang, Jiaqi ; Stockwell, Timothy B. ; Walenz, Brian P. ; Li, Kelvin ; Axelrod, Nelson ; Busam, Dana A. ; Strausberg, Robert L. ; Venter, J. C.: Genetic variation in an individual human exome. In: *PLoS Genet* 4 (2008), Nr. 8, e1000160. http://dx.doi.org/10.1371/journal.pgen.1000160. – DOI 10.1371/journal.pgen.1000160

[71] Ng, Sarah B. ; Bigham, Abigail W. ; Buckingham, Kati J. ; Hannibal, Mark C. ; McMillin, Margaret J. ; Gildersleeve, Heidi I. ; Beck, Anita E. ; Tabor, Holly K. ; Cooper, Gregory M. ; Mefford, Heather C. ; Lee, Choli ; Turner, Emily H. ; Smith, Joshua D. ; Rieder, Mark J. ; Yoshiura, Koh-Ichiro ; Matsumoto,

Naomichi ; Ohta, Tohru ; Niikawa, Norio ; Nickerson, Deborah A. ; Bamshad, Michael J. ; Shendure, Jay: Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. In: *Nat Genet* 42 (2010), Sep, Nr. 9, 790–793. http://dx.doi.org/10.1038/ng.646. – DOI 10.1038/ng.646

[72] Ng, Sarah B. ; Buckingham, Kati J. ; Lee, Choli ; Bigham, Abigail W. ; Tabor, Holly K. ; Dent, Karin M. ; Huff, Chad D. ; Shannon, Paul T. ; Jabs, Ethylin W. ; Nickerson, Deborah A. ; Shendure, Jay ; Bamshad, Michael J.: Exome sequencing identifies the cause of a mendelian disorder. In: *Nat Genet* 42 (2010), Jan, Nr. 1, 30–35. http://dx.doi.org/10.1038/ng.499. – DOI 10.1038/ng.499

[73] Ng, Sarah B. ; Turner, Emily H. ; Robertson, Peggy D. ; Flygare, Steven D. ; Bigham, Abigail W. ; Lee, Choli ; Shaffer, Tristan ; Wong, Michelle ; Bhattacharjee, Arindam ; Eichler, Evan E. ; Bamshad, Michael ; Nickerson, Deborah A. ; Shendure, Jay: Targeted capture and massively parallel sequencing of 12 human exomes. In: *Nature* 461 (2009), Sep, Nr. 7261, 272–276. http://dx.doi.org/10.1038/nature08250. – DOI 10.1038/nature08250

[74] Osato, Motomi: Point mutations in the RUNX1/AML1 gene: another actor in RUNX leukemia. In: *Oncogene* 23 (2004), May, Nr. 24, 4284–4296. http://dx.doi.org/10.1038/sj.onc.1207779. – DOI 10.1038/sj.onc.1207779

[75] Ott, J. ; Falk, C. T.: Epistatic association and linkage analysis in human families. In: *Hum Genet* 62 (1982), Nr. 4, S. 296–300

[76] Pennisi, Elizabeth: Genomics. Semiconductors inspire new sequencing technologies. In: *Science* 327 (2010), Mar, Nr. 5970, 1190. http://dx.doi.org/10.1126/science.327.5970.1190. – DOI 10.1126/science.327.5970.1190

[77] Peterson, Luke F. ; Zhang, Dong-Er: The 8;21 translocation in leukemogenesis. In: *Oncogene* 23 (2004), May, Nr. 24, 4255–4262. http://dx.doi.org/10.1038/sj.onc.1207727. – DOI 10.1038/sj.onc.1207727

[78] Porreca, Gregory J.: Genome sequencing on nanoballs. In: *Nat Biotechnol* 28 (2010), Jan, Nr. 1, 43–44. http://dx.doi.org/10.1038/nbt0110-43. – DOI 10.1038/nbt0110–43

[79] Ramensky, Vasily ; Bork, Peer ; Sunyaev, Shamil: Human non-synonymous SNPs: server and survey. In: *Nucleic Acids Res* 30 (2002), Sep, Nr. 17, S. 3894–3900

[80] Reich, D. E. ; Lander, E. S.: On the allelic spectrum of human disease. In: *Trends Genet* 17 (2001), Sep, Nr. 9, S. 502–510

[81] Reindl, Carola ; Quentmeier, Hilmar ; Petropoulos, Konstantin ; Greif, Philipp A. ; Benthaus, Tobias ; Argiropoulos, Bob ; Mellert, Gudrun ; Vempati, Sridhar ; Duyster, Justus ; Buske, Christian ; Bohlander, Stefan K. ; Humphries, Keith R. ; Hiddemann, Wolfgang ; Spiekermann, Karsten: CBL exon 8/9 mutants activate the FLT3 pathway and cluster in core binding factor/11q deletion acute myeloid leukemia/myelodysplastic syndrome subtypes. In: *Clin Cancer Res* 15 (2009), Apr, Nr. 7, 2238–2247. http://dx.doi.org/10.1158/1078-0432.CCR-08-1325. – DOI 10.1158/1078–0432.CCR–08–1325

[82] Risch, N.: Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes. In: *Am J Hum Genet* 36 (1984), Mar, Nr. 2, S. 363–386

[83] Rogaeva, Ekaterina ; Meng, Yan ; Lee, Joseph H. ; Gu, Yongjun ; Kawarai, Toshitaka ; Zou, Fanggeng ; Katayama, Taiichi ; Baldwin, Clinton T. ; Cheng, Rong ; Hasegawa, Hiroshi ; Chen, Fusheng ; Shibata, Nobuto ; Lunetta, Kathryn L. ; Pardossi-Piquard, Raphaelle ; Bohm, Christopher ; Wakutani, Yosuke ; Cupples, L. A. ; Cuenco, Karen T. ; Green, Robert C. ; Pinessi, Lorenzo ; Rainero, Innocenzo ; Sorbi, Sandro ; Bruni, Amalia ; Duara, Ranjan ; Friedland, Robert P. ; Inzelberg, Rivka ; Hampe, Wolfgang ; Bujo, Hideaki ; Song, You-Qiang ; Andersen, Olav M. ; Willnow, Thomas E. ; Graff-Radford, Neill ; Petersen, Ronald C. ; Dickson, Dennis ; Der, Sandy D. ; Fraser, Paul E. ; Schmitt-Ulms, Gerold ; Younkin, Steven ; Mayeux, Richard ; Farrer, Lindsay A.

; George-Hyslop, Peter S.: The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. In: *Nat Genet* 39 (2007), Feb, Nr. 2, 168–177. http://dx.doi.org/10.1038/ng1943. – DOI 10.1038/ng1943

[84] Rojas, Raul ; Kametaka, Satoshi ; Haft, Carol R. ; Bonifacino, Juan S.: Interchangeable but essential functions of SNX1 and SNX2 in the association of retromer with endosomes and the trafficking of mannose 6-phosphate receptors. In: *Mol Cell Biol* 27 (2007), Feb, Nr. 3, 1112–1124. http://dx.doi.org/10.1128/MCB.00156-06. – DOI 10.1128/MCB.00156–06

[85] Rothberg, Jonathan M. ; Leamon, John H.: The development and impact of 454 sequencing. In: *Nat Biotechnol* 26 (2008), Oct, Nr. 10, 1117–1124. http://dx.doi.org/10.1038/nbt1485. – DOI 10.1038/nbt1485

[86] Samani, Nilesh J. ; Erdmann, Jeanette ; Hall, Alistair S. ; Hengstenberg, Christian ; Mangino, Massimo ; Mayer, Bjoern ; Dixon, Richard J. ; Meitinger, Thomas ; Braund, Peter ; Wichmann, H-Erich ; Barrett, Jennifer H. ; Koenig, Inke R. ; Stevens, Suzanne E. ; Szymczak, Silke ; Tregouet, David-Alexandre ; Iles, Mark M. ; Pahlke, Friedrich ; Pollard, Helen ; Lieb, Wolfgang ; Cambien, Francois ; Fischer, Marcus ; Ouwehand, Willem ; Blankenberg, Stefan ; Balmforth, Anthony J. ; Baessler, Andrea ; Ball, Stephen G. ; Strom, Tim M. ; Braenne, Ingrid ; Gieger, Christian ; Deloukas, Panos ; Tobin, Martin D. ; Ziegler, Andreas ; Thompson, John R. ; Schunkert, Heribert ; C., W. T. C. C. ; Cardiogenics Consortium the: Genomewide association analysis of coronary artery disease. In: *N Engl J Med* 357 (2007), Aug, Nr. 5, S. 443–453

[87] Sanger, F. ; Coulson, A. R.: A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. In: *J Mol Biol* 94 (1975), May, Nr. 3, S. 441–448

[88] Schadt, Eric E. ; Turner, Steve ; Kasarskis, Andrew: A window into third-generation sequencing. In: *Hum Mol Genet* 19 (2010), Oct, Nr. R2, R227–R240. http://dx.doi.org/10.1093/hmg/ddq416. – DOI 10.1093/hmg/ddq416

[89] Schunkert, Heribert ; Koenig, Inke R. ; Kathiresan, Sekar ; Reilly, Muredach P. ; Assimes, Themistocles L. ; Holm, Hilma ; Preuss, Michael ; Stewart, Alexandre F R. ; Barbalic, Maja ; Gieger, Christian ; Absher, Devin ; Aherrahrou, Zouhair ; Allayee, Hooman ; Altshuler, David ; Anand, Sonia S. ; Andersen, Karl ; Anderson, Jeffrey L. ; Ardissino, Diego ; Ball, Stephen G. ; Balmforth, Anthony J. ; Barnes, Timothy A. ; Becker, Diane M. ; Becker, Lewis C. ; Berger, Klaus ; Bis, Joshua C. ; Boekholdt, S. M. ; Boerwinkle, Eric ; Braund, Peter S. ; Brown, Morris J. ; Burnett, Mary S. ; Buysschaert, Ian ; Cardiogenics ; Carlquist, John F. ; Chen, Li ; Cichon, Sven ; Codd, Veryan ; Davies, Robert W. ; Dedoussis, George ; Dehghan, Abbas ; Demissie, Serkalem ; Devaney, Joseph M. ; Diemert, Patrick ; Do, Ron ; Doering, Angela ; Eifert, Sandra ; Mokhtari, Nour Eddine E. ; Ellis, Stephen G. ; Elosua, Roberto ; Engert, James C. ; Epstein, Stephen E. ; Faire, Ulf de ; Fischer, Marcus ; Folsom, Aaron R. ; Freyer, Jennifer ; Gigante, Bruna ; Girelli, Domenico ; Gretarsdottir, Solveig ; Gudnason, Vilmundur ; Gulcher, Jeffrey R. ; Halperin, Eran ; Hammond, Naomi ; Hazen, Stanley L. ; Hofman, Albert ; Horne, Benjamin D. ; Illig, Thomas ; Iribarren, Carlos ; Jones, Gregory T. ; Jukema, J. W. ; Kaiser, Michael A. ; Kaplan, Lee M. ; Kastelein, John J P. ; Khaw, Kay-Tee ; Knowles, Joshua W. ; Kolovou, Genovefa ; Kong, Augustine ; Laaksonen, Reijo ; Lambrechts, Diether ; Leander, Karin ; Lettre, Guillaume ; Li, Mingyao ; Lieb, Wolfgang ; Loley, Christina ; Lotery, Andrew J. ; Mannucci, Pier M. ; Maouche, Seraya ; Martinelli, Nicola ; McKeown, Pascal P. ; Meisinger, Christa ; Meitinger, Thomas ; Melander, Olle ; Merlini, Pier A. ; Mooser, Vincent ; Morgan, Thomas ; Muehleisen, Thomas W. ; Muhlestein, Joseph B. ; Muenzel, Thomas ; Musunuru, Kiran ; Nahrstaedt, Janja ; Nelson, Christopher P. ; Noethen, Markus M. ; Olivieri, Oliviero ; Patel, Riyaz S. ; Patterson, Chris C. ; Peters, Annette ; Peyvandi, Flora ; Qu, Liming ; Quyyumi, Arshed A. ; Rader, Daniel J. ; Rallidis, Loukianos S. ; Rice, Catherine ; Rosendaal, Frits R. ; Rubin, Diana ; Salomaa, Veikko ; Sampietro, M. L. ; Sandhu, Manj S. ; Schadt, Eric ; Schaefer, Arne ; Schillert, Arne ; Schreiber, Stefan ; Schrezenmeir, Juergen ; Schwartz,

Stephen M. ; Siscovick, David S. ; Sivananthan, Mohan ; Sivapalaratnam, Suthesh ; Smith, Albert ; Smith, Tamara B. ; Snoep, Jaapjan D. ; Soranzo, Nicole ; Spertus, John A. ; Stark, Klaus ; Stirrups, Kathy ; Stoll, Monika ; Tang, W. H W. ; Tennstedt, Stephanie ; Thorgeirsson, Gudmundur ; Thorleifsson, Gudmar ; Tomaszewski, Maciej ; Uitterlinden, Andre G. ; Rij, Andre M. ; Voight, Benjamin F. ; Wareham, Nick J. ; Wells, George A. ; Wichmann, H-Erich ; Wild, Philipp S. ; Willenborg, Christina ; Witteman, Jaqueline C M. ; Wright, Benjamin J. ; Ye, Shu ; Zeller, Tanja ; Ziegler, Andreas ; Cambien, Francois ; Goodall, Alison H. ; Cupples, L. A. ; Quertermous, Thomas ; Maerz, Winfried ; Hengstenberg, Christian ; Blankenberg, Stefan ; Ouwehand, Willem H. ; Hall, Alistair S. ; Deloukas, Panos ; Thompson, John R. ; Stefansson, Kari ; Roberts, Robert ; Thorsteinsdottir, Unnur ; O'Donnell, Christopher J. ; McPherson, Ruth ; Erdmann, Jeanette ; Consortium, C. A. R. D. IoG. R. A. M. ; Samani, Nilesh J.: Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. In: *Nat Genet* 43 (2011), Apr, Nr. 4, 333–338. http://dx.doi.org/10.1038/ng.784. – DOI 10.1038/ng.784

[90] Seelow, Dominik ; Schuelke, Markus ; Hildebrandt, Friedhelm ; Nuernberg, Peter: HomozygosityMapper–an interactive approach to homozygosity mapping. In: *Nucleic Acids Res* 37 (2009), Jul, Nr. Web Server issue, W593–W599. http://dx.doi.org/10.1093/nar/gkp369. – DOI 10.1093/nar/gkp369

[91] Shendure, Jay ; Ji, Hanlee: Next-generation DNA sequencing. In: *Nat Biotechnol* 26 (2008), Oct, Nr. 10, 1135–1145. http://dx.doi.org/10.1038/nbt1486. – DOI 10.1038/nbt1486

[92] Smith, L. M. ; Sanders, J. Z. ; Kaiser, R. J. ; Hughes, P. ; Dodd, C. ; Connell, C. R. ; Heiner, C. ; Kent, S. B. ; Hood, L. E.: Fluorescence detection in automated DNA sequence analysis. In: *Nature* 321 (1986), Nr. 6071, 674–679. http://dx.doi.org/10.1038/321674a0. – DOI 10.1038/321674a0

[93] Smith, T. F. ; Waterman, M. S.: Identification of common molecular subsequences. In: *J Mol Biol* 147 (1981), Mar, Nr. 1, S. 195–197

[94] Stoddart, David ; Heron, Andrew J. ; Mikhailova, Ellina ; Maglia, Giovanni ; Bayley, Hagan: Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. In: *Proc Natl Acad Sci U S A* 106 (2009), May, Nr. 19, 7702–7707. http://dx.doi.org/10.1073/pnas.0901054106. – DOI 10.1073/pnas.0901054106

[95] Tarpey, Patrick S. ; Stevens, Claire ; Teague, Jon ; Edkins, Sarah ; O'Meara, Sarah ; Avis, Tim ; Barthorpe, Syd ; Buck, Gemma ; Butler, Adam ; Cole, Jennifer ; Dicks, Ed ; Gray, Kristian ; Halliday, Kelly ; Harrison, Rachel ; Hills, Katy ; Hinton, Jonathon ; Jones, David ; Menzies, Andrew ; Mironenko, Tatiana ; Perry, Janet ; Raine, Keiran ; Richardson, David ; Shepherd, Rebecca ; Small, Alexandra ; Tofts, Calli ; Varian, Jennifer ; West, Sofie ; Widaa, Sara ; Yates, Andy ; Catford, Rachael ; Butler, Julia ; Mallya, Uma ; Moon, Jenny ; Luo, Ying ; Dorkins, Huw ; Thompson, Deborah ; Easton, Douglas F. ; Wooster, Richard ; Bobrow, Martin ; Carpenter, Nancy ; Simensen, Richard J. ; Schwartz, Charles E. ; Stevenson, Roger E. ; Turner, Gillian ; Partington, Michael ; Gecz, Jozef ; Stratton, Michael R. ; Futreal, P. A. ; Raymond, F. L.: Mutations in the gene encoding the Sigma 2 subunit of the adaptor protein 1 complex, AP1S2, cause X-linked mental retardation. In: *Am J Hum Genet* 79 (2006), Dec, Nr. 6, 1119–1124. http://dx.doi.org/10.1086/510137. – DOI 10.1086/510137

[96] Vilarino-Gueell, Carles ; Wider, Christian ; Ross, Owen A. ; Dachsel, Justus C. ; Kachergus, Jennifer M. ; Lincoln, Sarah J. ; Soto-Ortolaza, Alexandra I. ; Cobb, Stephanie A. ; Wilhoite, Greggory J. ; Bacon, Justin A. ; Behrouz, Bahareh ; Melrose, Heather L. ; Hentati, Emna ; Puschmann, Andreas ; Evans, Daniel M. ; Conibear, Elizabeth ; Wasserman, Wyeth W. ; Aasly, Jan O. ; Burkhard, Pierre R. ; Djaldetti, Ruth ; Ghika, Joseph ; Hentati, Faycal ; Krygowska-Wajs, Anna ; Lynch, Tim ; Melamed, Eldad ; Rajput, Alex ; Rajput, Ali H. ; Solida, Alessandra ; Wu, Ruey-Meei ; Uitti, Ryan J. ; Wszolek, Zbigniew K. ; Vingerhoets, Francois ; Farrer, Matthew J.: VPS35 mutations

in Parkinson disease. In: *Am J Hum Genet* 89 (2011), Jul, Nr. 1, 162–167. http://dx.doi.org/10.1016/j.ajhg.2011.06.001. – DOI 10.1016/j.ajhg.2011.06.001

[97] Visscher, Peter M. ; Brown, Matthew A. ; McCarthy, Mark I. ; Yang, Jian: Five years of GWAS discovery. In: *Am J Hum Genet* 90 (2012), Jan, Nr. 1, 7–24. http://dx.doi.org/10.1016/j.ajhg.2011.11.029. – DOI 10.1016/j.ajhg.2011.11.029

[98] Vissers, Lisenka E L M. ; Ligt, Joep de ; Gilissen, Christian ; Janssen, Irene ; Steehouwer, Marloes ; Vries, Petra de ; Lier, Bart van ; Arts, Peer ; Wieskamp, Nienke ; Rosario, Marisol del ; Bon, Bregje W M. ; Hoischen, Alexander ; Vries, Bert B A. ; Brunner, Han G. ; Veltman, Joris A.: A de novo paradigm for mental retardation. In: *Nat Genet* 42 (2010), Dec, Nr. 12, 1109–1112. http://dx.doi.org/10.1038/ng.712. – DOI 10.1038/ng.712

[99] Walsh, Tom ; Shahin, Hashem ; Elkan-Miller, Tal ; Lee, Ming K. ; Thornton, Anne M. ; Roeb, Wendy ; Rayyan, Amal A. ; Loulus, Suheir ; Avraham, Karen B. ; King, Mary-Claire ; Kanaan, Moien: Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. In: *Am J Hum Genet* 87 (2010), Jul, Nr. 1, 90–94. http://dx.doi.org/10.1016/j.ajhg.2010.05.010. – DOI 10.1016/j.ajhg.2010.05.010

[100] WATSON, J. D. ; CRICK, F. H.: Genetical implications of the structure of deoxyribonucleic acid. In: *Nature* 171 (1953), May, Nr. 4361, S. 964–967

[101] WATSON, J. D. ; CRICK, F. H.: Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. In: *Nature* 171 (1953), Apr, Nr. 4356, S. 737–738

[102] Wheeler, David A. ; Srinivasan, Maithreyan ; Egholm, Michael ; Shen, Yufeng ; Chen, Lei ; McGuire, Amy ; He, Wen ; Chen, Yi-Ju ; Makhijani, Vinod ; Roth, G. T. ; Gomes, Xavier ; Tartaro, Karrie ; Niazi, Faheem ; Turcotte, Cynthia L. ; Irzyk, Gerard P. ; Lupski, James R. ; Chinault, Craig ; Song, Xing zhi ; Liu, Yue

; Yuan, Ye ; Nazareth, Lynne ; Qin, Xiang ; Muzny, Donna M. ; Margulies, Marcel ; Weinstock, George M. ; Gibbs, Richard A. ; Rothberg, Jonathan M.: The complete genome of an individual by massively parallel DNA sequencing. In: *Nature* 452 (2008), Apr, Nr. 7189, 872–876. http://dx.doi.org/10.1038/nature06884. – DOI 10.1038/nature06884

[103] Wichmann, H-E. ; Gieger, C. ; Illig, T. ; Group, M. O. N. I. C. A/K. O. R. A. S.: KORA-gen–resource for population genetics, controls and a broad spectrum of disease phenotypes. In: *Gesundheitswesen* 67 Suppl 1 (2005), Aug, S. S26–S30

[104] Willnow, Thomas E. ; Petersen, Claus M. ; Nykjaer, Anders: VPS10P-domain receptors - regulators of neuronal viability and function. In: *Nat Rev Neurosci* 9 (2008), Dec, Nr. 12, 899–909. http://dx.doi.org/10.1038/nrn2516. – DOI 10.1038/nrn2516

[105] Yi, Xin ; Liang, Yu ; Huerta-Sanchez, Emilia ; Jin, Xin ; Cuo, Zha Xi P. ; Pool, John E. ; Xu, Xun ; Jiang, Hui ; Vinckenbosch, Nicolas ; Korneliussen, Thorfinn S. ; Zheng, Hancheng ; Liu, Tao ; He, Weiming ; Li, Kui ; Luo, Ruibang ; Nie, Xifang ; Wu, Honglong ; Zhao, Meiru ; Cao, Hongzhi ; Zou, Jing ; Shan, Ying ; Li, Shuzheng ; Yang, Qi ; Asan ; Ni, Peixiang ; Tian, Geng ; Xu, Junming ; Liu, Xiao ; Jiang, Tao ; Wu, Renhua ; Zhou, Guangyu ; Tang, Meifang ; Qin, Junjie ; Wang, Tong ; Feng, Shuijian ; Li, Guohong ; Huasang ; Luosang, Jiangbai ; Wang, Wei ; Chen, Fang ; Wang, Yading ; Zheng, Xiaoguang ; Li, Zhuo ; Bianba, Zhuoma ; Yang, Ge ; Wang, Xinping ; Tang, Shuhui ; Gao, Guoyi ; Chen, Yong ; Luo, Zhen ; Gusang, Lamu ; Cao, Zheng ; Zhang, Qinghui ; Ouyang, Weihan ; Ren, Xiaoli ; Liang, Huiqing ; Zheng, Huisong ; Huang, Yebo ; Li, Jingxiang ; Bolund, Lars ; Kristiansen, Karsten ; Li, Yingrui ; Zhang, Yong ; Zhang, Xiuqing ; Li, Ruiqiang ; Li, Songgang ; Yang, Huanming ; Nielsen, Rasmus ; Wang, Jun ; Wang, Jian: Sequencing of 50 human exomes reveals adaptation to high altitude. In: *Science* 329 (2010), Jul, Nr. 5987, 75–78. http://dx.doi.org/10.1126/science.1190371. – DOI 10.1126/science.1190371

[106] Zimprich, Alexander ; Benet-Pages, Anna ; Struhal, Walter ; Graf, Elisabeth ; Eck, Sebastian H. ; Offman, Marc N. ; Haubenberger, Dietrich ; Spielberger, Sabine ; Schulte, Eva C. ; Lichtner, Peter ; Rossle, Shaila C. ; Klopp, Norman ; Wolf, Elisabeth ; Seppi, Klaus ; Pirker, Walter ; Presslauer, Stefan ; Mollenhauer, Brit ; Katzenschlager, Regina ; Foki, Thomas ; Hotzy, Christoph ; Reinthaler, Eva ; Harutyunyan, Ashot ; Kralovics, Robert ; Peters, Annette ; Zimprich, Fritz ; Bruecke, Thomas ; Poewe, Werner ; Auff, Eduard ; Trenkwalder, Claudia ; Rost, Burkhard ; Ransmayr, Gerhard ; Winkelmann, Juliane ; Meitinger, Thomas ; Strom, Tim M.: A Mutation in VPS35, Encoding a Subunit of the Retromer Complex, Causes Late-Onset Parkinson Disease. In: *Am J Hum Genet* 89 (2011), Jul, Nr. 1, 168–175. http://dx.doi.org/10.1016/j.ajhg.2011.06.008. – DOI 10.1016/j.ajhg.2011.06.008

Curriculum Vitae

Personal

Name	Sebastian H. Eck
Birth	11.07.1981
Marital Status	Unmarried
Adress	Appenzellerstrasse 81
	81475 Munich
Phone	+49 170 2929016
Email	sebastian.eck@medizinische-genetik.de

Education

1988 - 1992	Elementary School An der Oselstrasse, Munich
1992 - 2001	Gymnasium Max-Planck Gymnasium, Munich, Abitur

Civil Service

2001 - 2002	Mathser Disability Transportation Service

University Education

2002 - 2008	Student of Bioinformatics Ludwig-Maximilians-University, Munich, Germany Technical University, Munich, Germany Diploma (M.Sc.)
2008 - 2012	PhD Student, Bioinformatics Departement of Human Genetics, Helmholtz Research Center for Environmental Health Munich, Germany

Practical Experience

2007 - 2008	Research Assistant, Departement of Bioinformatics Helmholtz Research Center for Environmental Health, Munich, Germany
2008	Research assistant, Departement of Evolutionary Biology Ludwig-Maximilians-University, Munich, Germany
2011	Tutor Synbreed Summer School, Next Generation Sequence Analysis: Practice and Departure to New Frontiers

Programming / Computer Skills

Perl
Java
mySQL
R statical computing
Unix / Linux
Latex

Languages

German	Native Language
English	Excellent
French	Basic

Interests

Indoor Climbing
English Novels
Beachvolleyball
Biking