Fachbereich 10: Biologie

Hochdurchsatz-DNS-Sequenzierung revolutioniert die genetische Diagnostik und Krankenversorgung

Dissertation

zur Erlangung des akademischen Grades
„Doktor der Naturwissenschaften“
am Fachbereich Biologie der
Johannes Gutenberg-Universität Mainz

Stefan Diederich
Mainz, 15. Dezember 2020

Durchgeführt am Institut für Humangenetik an der Universitätsmedizin der Johannes Gutenberg-Universität in Mainz.

Arbeit eingereicht am:

Arbeit angenommen am:

durch:

Mainz, den .....
(Unterschrift)

Tag der mündlichen Prüfung:

Inhaltsverzeichnis

Abkürzungsverzeichnis

μl: ..... Mikroliter

3'UTR: ..... 3' untranslatierter Bereich

A: ..... Adenin

ACMG: ..... American College of Medical Genetics and Genomics

ADHS: ..... Aufmerksamkeits-Defizit-Hyperaktivitäts-Syndrom

AMP: ..... Association for Molecular Pathology

ASCII: ..... American Standard Code for Information Interchange

ASD: ..... Autism Spectrum Disorder

BAC: ..... Bacterial Artificial Chromosome

BAF: ..... B-Allel-Frequenz

BAM: ..... Binary Alignment/Map

BED: ..... Browser-Extensible-Data

BMG: ..... Bundesministerium für Gesundheit

bp: ..... Basenpaar

BQSR: ..... Base Quality Score Recalibration

BWA: ..... Burrows-Wheeler Alignment

BWT: ..... Burrows-Wheeler-Transformation

C: ..... Cytosin

CNV: ..... Copy Number Variation

CPAN: ..... Comprehensive Perl Archive Network

dATP: ..... Desoxyadenintriphosphat

DB: ..... Datenbank

dbSNP: ..... Single Nucleotide Polymorphism Database

dCTP: ..... Desoxycytosintriphosphat

ddNTP: ..... Didesoxyribonukleosidtriphosphat

dGTP: ..... Desoxyguanintriphosphat

DMEM: ..... Dulbecco's Modified Eagle Medium

DNA: ..... Deoxyribonucleic Acid

DNB: ..... DNA-Nanobälle

dNTP: ..... Desoxyribonukleosidtriphosphat

DPBS: ..... Dulbecco's Phosphate-Buffered Saline

DRAGEN: ..... Dynamic Read Analysis for GENomics

dTTP: ..... Desoxythymidintriphosphat

EBI: ..... European Bioinformatics Institute

EBM: ..... Einheitlicher Bewertungsmaßstab

ExAC: ..... Exome Aggregation Consortium

EXO I: ..... Exonuklease I

FISH: ..... Fluoreszenz-In-Situ-Hybridisierung

FPGA: ..... Field-Programmable Gate Array

G: ..... Guanin

g: ..... Gramm

GATK: ..... Genome Analysis Toolkit

GB: ..... Gigabyte

Gbp: ..... Gigabasenpaare

gnomAD: ..... Genome Aggregation Database

GO: ..... Gene Ontology

GUI: ..... Graphical User Interface

H^3M^2: ..... Homozygosity Heterogeneous Hidden Markov Model

HDD: ..... Hard Drive Disk

HDS: ..... Hochdurchsatz-DNA-Sequenzierung

HGMD® : ..... Human Gene Mutation Database

HPG: ..... Humangenomprojekt

IDB: ..... Identity by Descent

IPS: ..... Induzierte Pluripotente Stammzellen

JDK: ..... Java Development Kit

JRE: ..... Java Runtime Environment

JVM: ..... Java Virtual Machine

kbp: ..... Kilobasenpaar

LB: ..... lysogeny broth

LINES: ..... Long Interspersed Nuclear Elements

LIPER: ..... Long-Insert Paired-End Read

LVM: ..... Logical Volume Manager

Mbp: ..... Megabasenpaar

MIM: ..... Mendelian Inheritance in Man

ml: ..... Milliliter

NCBI: ..... National Center for Biotechnology Information

NGS: ..... Next Generation Sequencing

NHS: ..... National Health Service

NIH: ..... National Institute of Health

nm: ..... Nanometer

NPC: ..... Neuronal Precursor Cell

Oak: ..... Object Application Kernel

OMIM: ..... Online Mendelian Inheritance in Man

PairHMM: ..... Pair Hidden Markow Model

PAR: ..... Pseudoautosomale Region

PCR: ..... Polymerase Chain Reaction

PE: ..... Paired-End

RAID: ..... Redundant Array of Independent Disks

RAM: ..... Read Access Memory

REST: ..... Representational State Transfer

RFLP: ..... Restriktionsfragmentlängenpolymorphismus

ROH: ..... Regions of Homozygosity

rpm: ..... Rounds per Minute

SAM: ..... Sequence Alignment/Map

SAP: ..... Shrimp Alkaline Phosphatase

SAS: ..... Serial Attached SCSI

SBS: ..... Sequencing by Synthesis

SE: ..... Single-End

SINES: ..... Short Interspersed Nuclear Elements

SIPER: ..... Short-Insert Paired-End Read

SNP: ..... Single Nucleotide Polymorphism

SNV: ..... Single Nucleotide Variation

SSD: ..... Solid-State-Drive

STR: ..... Short Tandem Repeat

SV: ..... Structual Variations

T: ..... Thymin

TB: ..... Terrabyte

TES: ..... Target Enrichment Sequencing

UCSC: ..... University of California, Santa Cruz

uORF: ..... Upstream Open Reading Frame

VCF: ..... Variant Call Format

VEP: ..... Variant Effect Predictor

WES: ..... Whole Exome Sequencing

WGS: ..... Whole Genome Sequencing

YAC: ..... Yeast Articial Chromosome

Zusammenfassung

Nach der Aufklärung der Grundstruktur des menschlichen Genoms durch das Humangenomprojekt rückte die Inventarisierung pathogener Sequenzvarianten in den Mittelpunkt der Genomforschung. Bis heute konnte erst ein winziger Teil der klinisch und funktionell relevanten Genomveränderungen identifiziert werden. Die sichere Unterscheidung pathogener und funktionell neutraler Veränderungen im menschlichen Genom ist jedoch ein noch weitgehend ungelöstes Problem. Genetisch bedingte Krankheiten, allen voran psychomotorische Entwicklungsstörungen, sind oft außerordentlich heterogen, und bei vielen Patienten kommen Hunderte verschiedener Gendefekte als Krankheitsursache infrage, die sich klinisch nicht eindeutig gegeneinander abgrenzen lassen.

Die Einführung der „Next Generation Sequencing"-Technologie (NGS), die eine gleichzeitige Untersuchung vieler infrage kommender Krankheitsgene erlaubt, hat die medizinische Genetik im vergangenen Jahrzehnt auf eine neue, viel breitere Basis gestellt und die genetische Diagnostik sowie die Krankenversorgung entscheidend verbessert und beschleunigt. Waren die ersten NGS-Geräte in der Lage, einige Dutzend ausgewählte Gene gleichzeitig zu sequenzieren, so ermöglichte die Kapazitätssteigerung dieser Geräte bald die Sequenzierung aller proteinkodierenden Bereiche (Whole Exome Sequencing, WES) des menschlichen Genoms bis hin zur gesamtgenomischen Sequenzierung (Whole Genome Sequencing, WGS). Aufgrund dieser Entwicklungen ist die NGS für die weltweit an akademischen Zentren konzentrierte genommedizinische Krankenversorgung und Forschung unverzichtbar geworden.

Bei dieser Dissertation steht die bioinformatische Verarbeitung und Interpretation der durch NGS erzeugten Sequenzdaten im Mittelpunkt, von der Assemblierung meist kurzer Sequenzfragmente zu Genen oder gar ganzen Genomen über den Aufbau einer bioinformatischen Pipeline zur Analyse und Filterung von Sequenzdaten und schließlich zur Identifikation pathogenetisch relevanter Sequenzvarianten sowie deren Abgleich mit der Literatur. Am Mainzer Institut für Humangenetik der Universitätsmedizin konnten erste Erfahrungen mit einem krankheitsgruppenspezifischen Genpanel (MPIMG1-Test) gewonnen werden. Mit der Entwicklung der WES und WGS stiegen die diagnostischen Möglichkeiten, aber auch die Ansprüche an die NGS-gestützte Diagnostik und die bioinformatische Analyse. Dies erforderte eine stetige Verbesserung der hausinternen bioinformatischen Pipeline und den Aufbau einer benutzerfreundlichen graphischen Oberfläche zur Interpretation der Ergebnisse. Zudem stellte sich in den letzten Jahren heraus, dass den Veränderungen in nicht kodierenden

Bereichen des Genoms eine große Rolle als Krankheitsursache zuzuteilen ist. Zur Aufspürung solcher Varianten wurde in dieser Arbeit ein Panel für regulatorische Bereiche (RegVar-Panel) entworfen, mit dem Ziel, die noch weitgehend unbekannten Varianten in solchen regulatorischen Bereichen zu identifizieren und deren Pathogenität abzuschätzen.

Die Entwicklung der der bioinformatischen Pipeline sowie die Einführung des MPIMG1-Tests zur Ergänzung der konventionellen Stufendiagnostik hatte bei Patienten mit psychomotorischen Entwicklungsstörungen einen hochsignifikanten Anstieg der Aufklärungsrate um 26% zur Folge. Nach der Ablösung des MPIMG1-Tests durch die Exomsequenzierung konnte diese Aufklärungsrate nochmals um 6% erhöht werden. Somit ist es möglich mit Hilfe dieser erweiterten Stufendiagnostik bei insgesamt 51% der Patienten eine Diagnose zu stellen. Die Anwendung des RegVar-Panels leistete keinen Beitrag zur Diagnose der untersuchten Patienten. Luziferase Assays für eine im RegVar-Panel detektierte Variante widerlegten deren *in silico* vorhergesagte Interaktionsveränderung mit verschiedenen mikro-RNAs.

Im Gegensatz zur WES erlaubt die WGS die Erkennung nahezu aller Veränderungen im Erbgut. Dadurch muss diese Untersuchung nur einmal im Leben erfolgen und eignet sich hervorragend als diagnostischer Eingangstest für Patienten mit Verdacht auf eine genetisch bedingte Erkrankung, wie durch eine in dieser Arbeit vorgestellte Fallstudie belegt und anschließend diskutiert.

Der Ausblick dieser Dissertation befasst sich unter anderem mit jüngsten Entwicklungen, die auf eine Einführung der WGS in die genetische Regelversorgung in Deutschland abzielen, und geht auch auf neue Sequenzierungstechniken ein, die eine *de novo*-Sequenzierung menschlicher Genome erlauben werden, ohne dabei auf Referenzgenome angewiesen zu sein. Diese Techniken erzeugen sehr lange zusammenhängende Sequenzfragmente, mit denen es möglich sein wird, auch die letzten Lücken im menschlichen Genom zu schließen und alle Sequenzvarianten zu erfassen, selbst in hochrepetitiven Genomabschnitten.

Abstract

After the basic structure of the human genome had been elucidated by the Human Genome Project, the inventory of pathogenic sequence variants became the focus of genome research. To date, only a tiny fraction of clinically and functionally relevant genome alterations have been identified. However, the reliable differentiation of pathogenic and functionally neutral changes in the human genome is still a largely unsolved problem. Genetically caused diseases, especially psychomotor developmental disorders, are often extremely heterogeneous, and in many patients hundreds of different genetic defects may be the cause of disease, which cannot be clinically distinguished from each other due to their variable manifestation.

The introduction of the „Next Generation Sequencing“ technology (NGS), which allows the simultaneous examination of many disease genes, has put medical genetics on a new, much broader basis over the past decade and has decisively improved and accelerated genetic diagnostics and patient care. While the first NGS devices were capable of sequencing a few dozen selected genes simultaneously, the increased capacity of these devices soon enabled the sequencing of all protein coding regions (Whole Exome Sequencing, WES) of the human genome up to whole genome sequencing (WGS). Due to these developments, NGS has become indispensable for the genomic medical care and research concentrated at academic centers worldwide.

This dissertation focuses on the bioinformatics processing and interpretation of sequence data generated by NGS, from the assembly of mostly short sequence fragments into genes or even entire genomes, through the construction of a bioinformatics pipeline for the analysis and filtering of sequence data, and finally to the identification of pathogenetically relevant sequence variants and their comparison with the literature. At the Institute of Human Genetics of University Medicine Mainz, first experiences with a disease-group specific gene panel (MPIMG1 test) were gained. With the development of WES and WGS, the diagnostic possibilities increased, but also the demands on NGS-supported diagnostics and bioinformatic analysis. This required a continuous improvement of the in-house bioinformatic pipeline and the development of a user-friendly graphical interface for the interpretation of results. In addition, in recent years it has become apparent that changes in non-coding regions of the genome play a major role as a cause of disease. In order to detect such variants, this thesis designed a panel for regulatory areas (RegVar-Panel) with the aim to identify the still largely unknown variants in such regulatory areas

and to assess their pathogenicity.

The development of the bioinformatics pipeline and the introduction of the MPIMG1 test to complement conventional stepwise diagnostics resulted in a highly significant 26% increase in the detection rate in patients with psychomotor developmental disorders. After the replacement of the MPIMG1 test by exome sequencing, this detection rate increased by a further 6%. Thus, it is possible to diagnose 51% of the patients with the help of this extended stepwise diagnostic procedure. The application of the RegVar panel did not contribute to the diagnosis of the examined patients. Luciferase assays for a variant detected in the RegVar panel disproved its *in silico* predicted interaction change with different microRNAs.

In contrast to WES, WGS allows the detection of almost all changes in the genome. Therefore, this examination has to be performed only once in a lifetime and is perfectly suited as an initial diagnostic test for patients suspected of having a genetic disease, as demonstrated by a case study presented in this thesis and discussed afterwards.

The outlook of this dissertation deals, among other things, with recent developments aimed at the introduction of WGS into regular genetic care in Germany and also discusses new sequencing techniques that will allow *de novo* sequencing of human genomes without having to rely on reference genomes. With the help of these techniques it will be possible to generate very long coherent sequence fragments, and thus it will be possible to close even the last gaps in the human genome and to capture all sequence variants, even in highly repetitive genome sections.

Kapitel 1

Einleitung

Der Begriff „Genetik“ wurde erstmals 1905 von William Bateson in einem Brief an seinen vorgesetzten Adam Sedgwick verwendet und ein Jahr später als Bezeichnung für das sich neu herausbildende wissenschaftliche Teilgebiet der Biologie vorgeschlagen [Bateson und Bateson, 1928]. Dieses befasst sich mit den Grundlagen zur Ausprägung erblicher Merkmale und der Weitergabe von Erbinformationen an die nachfolgenden Generationen. Im Jahr 1909 prägte der dänische Botaniker Wilhelm Johannsen darauf aufbauend die Begriffe „Gen“, „Erbgut“ und „Phänotyp“ [Johannsen, 1909].

1.1 Genetik, von Mendel bis heute

Bereits zum Ende des 19. Jahrhunderts gelang es bei mikroskopischen Untersuchungen von sich teilenden Zellen fadenartige Strukturen in den Zellkernen zu beobachten, die sich anfärben ließen und deshalb als Chromosomen bezeichnet wurden. Theodor Boveri und Walter Sutton entdeckten Gemeinsamkeiten zwischen der Gestalt bzw. dem Verhalten dieser Chromosomen bei der Zellteilung und den von Gregor Mendel 1866 beschriebenen Vererbungsmustern bestimmter genetischer Merkmale [Mendel, 1866]. So stellten sie ab 1903 die Chromosomentheorie der Vererbung auf, welche besagt, dass sich der materielle Träger des Erbguts im Zellkern befindet und Teile der Chromosomen den von Mendel postulierten Erbanlagen entsprechen [Boveri, 1902] [Sutton, 1903]. Erste Überlegungen über die Vererbung von Krankheiten stammten unter anderem von Archibald Edward Garrod. Er erkannte, dass einige bekannte Krankheiten wie die Alkaptonurie, Albinismus oder Cysteinurie nach den Regeln von Mendel vererbt werden und prägte 1908 den Begriff „angeborene Stoffwechselstörungen“ (engl.: inborn errors of metabolism) [Nuland, 2003].

Aus einem Experiment an R- und S-Pneumokokken-Stämmen folgerte Frederick Griffith 1928 erstmals, dass es eine transformierende Substanz geben muss, die eine Umwandlung von avirulenten in virulente Pneumokokken ermöglicht [Griffith, 1928]. 1944 gelang es Avery *et al.* die chemische Zusammensetzung dieser transformierenden Substanz durch Aufreinigung aufzuklären. Die Mengenanteile an Kohlenstoff, Waserstoff, Stickstoff und Phosphor entsprachen der von DNS (Desoxyribonukleinsäure; engl. Deoxyribonucleic Acid, DNA). Dies

war ein starkes Indiz dafür, dass die Erbinformation durch die DNA und nicht, wie bis dahin angenommen, durch Proteine kodiert wird [Avery *et al.*, 1944].

Etwa ein Jahrzehnt später, im Jahr 1952, veröffentlichte Erwin Chargaff die sogenannte Paritätsregel. Diese beschreibt, dass in DNA Molekülen auf beiden Strängen immer gleich viele Adenin- und Thymin- sowie Cytosin- und Guanin-Basen auftreten [Chargaff *et al.*, 1952]. Diese bahnbrechenden Erkenntnisse und die Ergebnisse von Rosalind Franklins Röntgenstrukturanalysen waren die Grundlage für die Aufklärung der Doppelhelixstruktur der DNA durch James Watson und Francis Crick im Jahre 1953 [Watson und Crick, 1953] [Franklin und Gosling, 1953].

Vernon Martin Ingram, ein deutsch-amerikanischer Biologe, wies 1956 durch Untersuchungen des Hämoglobin-Proteins bei der Sichelzellanämie erstmals nach, dass ein Austausch einer einzigen Aminosäure eine Erkrankung auslösen kann [Ingram, 1956]. Mithilfe elektrophoretischer und chromatographischer Methoden fand er heraus, dass der Glutaminsäurerest an Position 6 der β-Kette des normalen Hämoglobins im Sichelzell-Hämoglobin durch einen Valinrest ersetzt ist. Dieser Aminosäurenaustausch bewirkt eine geringfügig veränderte elektrische Ladung des Sichelzell-Hämoglobins bei neutralem pH-Wert, wodurch es bei Sauerstoffmangel zur Polymerisation und zur sichelförmigen Deformierung der Erythrozyten kommt. Die Folge sind Verschlüsse kleiner Arterien. Homozygote Träger des Gendefekts leiden an anfallsartigen und zum Teil lebensbedrohlichen Durchblutungsstörungen. Die Sichelzellanämie ist somit die erste molekulargenetisch aufgeklärte Krankheit [Lehninger, 2011].

Rund 25 Jahre nach der Aufklärung der DNA-Struktur stellten Frederick Sanger sowie Allan Maxam und Walter Gilbert 1975 Methoden zur DNA-Sequenzierung vor, mit der die Nukleotidabfolge eines Sequenzfragments von bis zu 500 Basenpaaren (bp) aufgeschlüsselt werden kann [Sanger *et al.*, 1977] [Maxam und Gilbert, 1977]. In den folgenden Jahren wurde mit der Polymerase-Kettenreaktion (engl. Polymerase Chain Reaction, PCR) ein einfaches Verfahren zur gezielten Vervielfältigung einzelner DNA-Abschnitte entwickelt [Mullis *et al.*, 1986]. Zuvor konnten solche DNA-Fragmente nur durch „Schrotschuss-Klonierung“ von Genomen und präziser Auslese der betreffenden Klone erzeugt werden.

Das 1990 gegründete internationale Humangenomprojekt (HGP) verfolgte das Ziel, das gesamte menschliche Erbgut (Genom) bis 2005 durch Sequenzierung zu entschlüsseln. Dies bildet die Grundlage zur Erforschung von Erbkrankheiten und soll zum besseren Verständnis von Krebserkrankungen auf molekularer Ebene dienen. Ihm ging die genetische und physikalische Kartierung einer Vielzahl von Genen und die Erstellung vollständiger genetischer Karten der menschlichen Chromosomen voraus [Donis-Keller *et al.*, 1987a] [Murray *et al.*, 1994]. Das mit öffentlichen Mitteln finanzierte HGP sah es vor, die Chromosomen auf der Grundlage überlappender und fast das vollständige Genom abdeckender Klone aus künstlichen Bakterien- (engl. Bacterial Arti-

ficial Chromosome, BAC) und Hefechromosomen (engl. Yeast Articial Chromosome, YAC) nacheinander zu sequenzieren. In der ersten Phase konzentrierte sich das HGP auf die Entwicklung diverser genetischer Marker, was ermöglichte mehr oder weniger detaillierte Genkarten zu erstellen und Kopplungsanalysen durchzuführen. Diese polymorphen DNA-Marker sind Einzelnukleotidaustausche (engl. Single Nucleotide Polymorphism, SNP), kleine sich wiederholende Sequenzen (engl. Short Tandem Repeats, STRs) und Restriktionsfragmentlängenpolymorphismen (RFLPs). Die STRs bestehen aus zwei bis vier sich wiederholenden Nukleotiden, wohingegen mit RFLPs Sequenzvarianten bezeichnet werden, die bei der Verdauung von DNA mithilfe sequenzspezifischer Restriktionsenzyme zu einem veränderten Muster der DNA-Fragmente führen und so erkannt werden können. Je geringer die Entfernung zweier syntänischer, also auf dem gleichen Chromosom lokalisierter Sequenzvarianten ist, desto wahrscheinlicher ist es, dass sie während der Keimzellbildung nicht durch ein Crossover zwischen den gepaarten Chromatiden getrennt, sondern gekoppelt an die Nachkommen weitergegeben werden.

Durch den Nachweis einer genetischen Kopplung mit bekannten, chromosomal lokalisierten DNA-Markern besteht die Möglichkeit in Familien mit sogenannten monogenen Krankheiten, die auf Defekte einzelner Gene zurückgehen, die ungefähre Lage der betreffenden Krankheitsgene zu ermitteln. Bereits bevor die Grundstruktur des menschlichen Genoms bekannt war, gelang es auf diese Weise, die für einige Dutzend monogener Krankheiten verantwortlichen Gendefekte im Genom zu lokalisieren und durch „Positional Cloning“ molekular aufzuklären. Dazu wurden für die durch flankierende genetische Marker begrenzten chromosomalen Abschnitte mittels „Chromosome Walking“ [Rédei, 2008] überlappende DNA-Klone hergestellt und diese sequenziert. Eine auf diese Weise klonierte Kandidatenregion umspannt häufig mehrere Millionen DNA-Bausteine und enthält meist eine Vielzahl von Genen. Das Herausfiltern des gesuchten Krankheitsgens stellte eine der größten Herausforderungen bei der Positionsklonierung dar [Collins, 1992].

Victor Almon McKusick, ein amerikanischer Humangenetiker, katalogisierte ab 1966 alle bekannten Merkmale des Menschen, die sich nach den Mendelschen Regeln vererben und gab zwischen den Jahren 1966 und 1998 insgesamt 12 gedruckte Ausgaben eines Verzeichnisses aller bekannten menschlichen Gene und ihren assoziierten Erkrankungen unter dem Titel „Mendelian Inheritance in Man“ (MIM) heraus. Ab dem Jahr 1998 wird diese Kollektion unter dem Namen „Online Mendelian Inheritance in Man“ (OMIM) digital als im Internet öffentlich zugängliche und fortlaufend aktualisierte Datenbank weitergeführt.

Im Jahre 1998 bekam das staatlich finanzierte Internationale Human Genome Projekt Konkurrenz von Craig Venter, dem Direktor von TIGR, einem privaten Institut für Genomforschung. Statt zuerst für alle menschlichen Chromosomen überlappende Raster von BAC-Klonen mit 100 bis 200 Kilobasenpaaren (kbp) langen menschlichen DNA-Fragmenten zu erstellen, verfolgte Craig Venter den Plan, unsortierte Genomfragmente im großen Stil zu sequenzieren und diese

mit massiver Computerunterstützung anhand ihrer überlappenden Enden zu immer längeren zusammenhängenden Sequenzen, den sogenannten „contigs“ zu verknüpfen.

Im Jahre 2001 verkündete das Humane Genomprojekt unter Leitung von Francis S. Collins gemeinsam mit Craig Venter, dem Leiter der privaten Firma TIGR, den Abschluss der Untersuchungen zur Grundstruktur des menschlichen Genoms in den Medien [Lander *et al.*, 2001] [Venter *et al.*, 2001]. Allerdings war die Sequenzierung zu diesem Zeitpunkt nur zu etwa 70% abgeschlossen und damit noch lückenhaft. Die Publikation einer „endgültigen“, aber seither mehrfach nachgebesserten Version des humanen Referenzgenoms erfolgte im Jahr 2003.

Als Begründung für die notwendige staatliche Finanzierung des HGP war anfangs aufgeführt worden, dass die Genomsequenzierung der Schlüssel für die Aufklärung, Diagnose und Therapie aller genetisch bedingter Krankheiten sei [National Human Genome Research Institute, 1990] [Collins und Fink, 1995]. Dabei wurde unterschlagen, dass dieses Vorhaben „nur“ auf die Aufklärung der Grundstruktur des menschlichen Genoms zielte, welches lediglich die Basis für die anschließende Suche nach krankheitsverursachenden Genomveränderungen darstellte. Zu Beginn des Humanen Genomprojekts schätzten Wissenschaftler die Anzahl der Gene eines menschlichen Organismus auf etwa 100.000 [Collins und Fink, 1995] [Keleher, 1993]. Diese Zahl variierte im Laufe der Jahre mehrfach. Nach Publikation des ersten Entwurfs der menschlichen Genomsequenz im Jahr 2001 korrigierten die Arbeitsgruppen um Lander *et al.*, Venter *et al.* und Claverie *et al.* die Anzahl auf etwa 30.000 bis 40.000 Gene [Lander *et al.*, 2001] [Venter *et al.*, 2001] [Claverie, 2001]. Nur 3 Jahre später, im Oktober 2004, bestätigte das Konsortium des humanen Genomprojekts die Existenz von 19.599 proteinkodierenden Genen [Abdellah *et al.*, 2004]. Die Zahl von 19.000 bis 20.000 kodierenden Genen im menschlichen Genom wird auch heute noch angenommen [Ezkurdia *et al.*, 2014].

Nach der Aufklärung der Grundstruktur des menschlichen Genoms und der Definition eines Referenzgenoms rückte die Inventarisierung pathogener Sequenzvarianten in den Mittelpunkt der Humangenomforschung. Aufgrund früherer Untersuchungen war bereits bekannt, dass die Genome nicht verwandter gesunder Menschen ca. 4 Millionen Unterschiede in ihrer DNA-Sequenz aufweisen, und erwartungsgemäß erwiesen sich die allermeisten dieser Varianten als funktionell neutral. Die systematische Suche nach den viel selteneren Sequenzvarianten, welche für die phänotypische Variabilität und insbesondere für genetisch bedingte Krankheiten verantwortlich sind, erlangte erst während des vergangenen Jahrzehnts zunehmende Popularität. Krankheitsauslösende Mutationen wurden bisher erst in ca. 4.200 der 20.000 proteinkodierenden Gene des Menschen gefunden (Stand 14.10.2020) [OMIM, 2020] und bis heute, mehr als 17 Jahre nach dem Abschluss des HGP, konnte erst ein winziger Teil der klinisch und funktionell relevanten Genomveränderungen identifiziert werden.

Die Einführung der „Next Generation Sequencing“-Technologie (NGS) zur massiv-parallelen Sequenzierung vieler verschiedener DNA-Fragmente hat diese Untersuchungen stark beschleunigt. Die Firma 454 Life Sciences (heute Roche) stellte 2005 mit dem 454 GS FLX Sequenziersystem das erste kommerziell vertriebene NGS-Gerät vor [Margulies *et al.*, 2005]. Kurze Zeit später folgten weitere Sequenziermaschinen anderer Hersteller, zum Beispiel der Genome Analyzer der Firma Solexa (heute Illumina) im Jahr 2006. Waren die ersten Geräte in der Lage, einige speziell ausgewählte und angereicherte Gene oder Genomabschnitte gleichzeitig zu sequenzieren (engl. Target Enrichment Sequencing, TES), so ermöglichte die Kapazitätssteigerung dieser Geräte bald die Sequenzierung aller proteinkodierenden Bereiche des menschlichen Genoms, des sogenannten Exoms (engl. Whole Exome Sequencing, WES). Mit der Einführung des Illumina Hiseq X10-Systems im Jahre 2014 wurde es möglich vollständige Genome einzelner Probanden innerhalb weniger Tage zu sequenzieren (engl. Whole Genome Sequencing, WGS) und gleichzeitig die Kosten für die WGS auf ca. 5000 Euro zu senken. Diese Entwicklungen haben die molekulare Aufklärung und Diagnose monogener Erbkrankheiten enorm beschleunigt und vereinfacht. So gelang es Najmabadi *et al.* bereits im Jahre 2011, durch Sequenzierung krankheitsrelevanter Genomabschnitte in 136 konsanguinen Familien 50 neue Krankheitsgene zu identifizieren [Najmabadi *et al.*, 2011].

Proteinkodierende Abschnitte umfassen etwa 1-2% des menschlichen Genoms [Ng *et al.*, 2009]. Weitere 3-4% des Genoms sind evolutionär hoch konserviert und dadurch offenbar funktionell wichtig, zum Beispiel für die spatiotemporale Steuerung der Genexpression während der Keimesentwicklung. Bis heute gelang es etwa 4.200 der 20.000 evolutionär konservierten Gene mit einer monogenen Erkrankung zu assoziieren [OMIM, 2020]. Das menschliche Genom enthält 3,2 Milliarden Basenpaare und besteht aus 4 verschiedenen Basen. Dadurch können maximal 9,6 Milliarden Austausche einzelner DNA-Bausteine auftreten. Bisher konnten etwa 150.000 Varianten als eindeutig pathogen identifiziert werden [The Monarch Initiative, 2020]. Dies entspricht 0,0016% des Genoms und 0,11% der kodierenden Bereiche. Monogene Krankheiten, die auf Defekten einzelner Gene beruhen, sind für sich genommen meist selten, jedoch als Krankheitsgruppe häufig. Ungefähr 20% der Todesfälle im Säuglingsalter und ca. 18% der pädiatrischen Krankenhausaufenthalte lassen sich auf seltene, rezessive Erbkrankheiten zurückführen [Kingsmore, 2012],und etwa 80% der seltenen Krankheiten haben monogene Ursachen [Ehrhart *et al.*, 2019] [Lee *et al.*, 2020]. Dennoch steht die Identifikation und Erforschung von pathogenen Varianten noch immer erst am Anfang.

Aufgrund der Verbesserung von Untersuchungs- und Analysemethoden besteht heute die Möglichkeit, mit Hilfe der WGS einen Großteil aller krankheitsverursachenden Varianten eines Genoms aufzudecken. Neben Einzelbasenaustauschen (engl. Single Nucleotide Variation, SNV) können unter anderem auch Deletionen und Insertionen, sowie größere chromosomale Rearrangements detektiert werden. Damit zeichnet sich ab, dass die WGS bei weiter sinkenden Kosten im Laufe der Zeit einen Großteil der heute in der genetischen Dia-

gnostik verwendeten Methoden, wie die klassische Zytogenetik, genomweite Mikroarrays und die Paneldiagnostik ersetzen könnte. Die sichere Unterscheidung pathogener und funktionell neutraler Veränderungen im menschlichen Genom ist jedoch ein noch weitgehend ungelöstes Problem.

Tabelle 1.1: *Meilensteine der genetischen Forschung*

Jahr	**Ereignis**	**Literatur**
1866	Mendelsche Regeln	[Mendel, 1866]
1903	Chromosomentheorie der Vererbung	[Boveri, 1902] [Sutton, 1903]
1908	Archibald Edward Garrod prägteden Begriff „angeborene Stoffwechselstörungen“	[Nuland, 2003]
1944	Erstes Indiz für DNA als Träger des Erbmaterials	[Avery *et al.*, 1944]
1952	Basenverhältnis folgt Muster; Paritätsregel (A=T, G=C)	[Chargaff *et al.*, 1952]
1953	Beschreibung der Struktur des DNA-Doppelstrangs	[Watson und Crick, 1953] [Franklin und Gosling, 1953]
1955	Das menschliche Genom besteht aus 46 Chromosomen	[Tjio und Levan, 1956]
1956	Aminosäureaustausch im Betaglobin-Protein verursacht Sichelzellanämie	[Ingram, 1956]
1959	Down-Syndrom entsteht durch Trisomie 21	[Lejeune *et al.*, 1959]
1961	Entdeckung der mRNA als Transmitter der Informationen aus Kern ins Zytoplasma	[Brenner *et al.*, 1961]
1966	Entschlüsselung des genetischen Codes	[Nirenberg *et al.*, 1966]
1966	Erste Auflage von MIM	[McKusick, 1966]
1969	Entwicklung erster	[Dayhoff, 1969]

Fortsetzung auf der nächsten Seite

Tabelle 1.1 – *Fortsetzung von vorheriger Seite*

Jahr	**Ereignis**	**Literatur**
bis 1970	Verfahren zur Sequenzanalyse	[Needleman und Wunsch, 1970]
1972	Erstes vollständig rekombinantes DNA-Molekül	[Jackson *et al.*, 1972]
1975	DNA Sequenzierung	[Sanger *et al.*, 1977] [Maxam und Gilbert, 1977]
1977	Erstbeschreibung von unterbrochenen Genen und Splicemechanismen	[Chow *et al.*, 1977]
1977	Klonierung des ersten menschlichen Gens	[Itakura *et al.*, 1977]
1980	Entdeckung der RFLPs	[Wai Kan und Dozy, 1978] [Botstein *et al.*, 1980]
1983	Erfindnung der PCR	[Mullis *et al.*, 1986]
1983	Das Huntington Gen liegt auf Chromosom 4	[Gusella *et al.*, 1983]
1985 bis 1991	Positionsklonierung der Gene für CGD, DMD, CF, CHM, Fra(X)	
1986	Entwicklung des ersten automatischen Sequenziergeräts	[Smith *et al.*, 1986]
1987	Erste umfassende Genkarte des menschlichen Genoms basierend auf RFLPs	[Donis-Keller *et al.*, 1987b]
1989	Entdeckung von Mikrosatellitenmarkern	[Kelly *et al.*, 1989]
1988 bis 1990	Entwicklung von Standardprogrammen zur Sequenzanalyse (FASTA, BLAST)	[Pearson und Lipman, 1988] [Altschul *et al.*, 1990]
1990 bis 1996	Entwickung von Genchips zur parallelen Messung des Transkriptionszustands ganzer Genome durch Hybridisierungsexperimente	[Carig *et al.*, 1990] [Lennon und Lehrach, 1991] [Schena *et al.*, 1995] [Lockhart *et al.*, 1996]
1990	Gründung des HGP	

Fortsetzung auf der nächsten Seite

Tabelle 1.1 – *Fortsetzung von vorheriger Seite*

Jahr	Ereignis	Literatur
1995	Beitritt Deutschlands zum HGP	
1999	Entwicklung der Shotgun-Methode zur Hochdurchsatz-sequenzierung ganzer Genome	[Weber und Myers, 1997]
1998	Veröffentlichung von OMIM	[Hamosh *et al.*, 2005]
1999	Sequenzierung des ersten menschlichen Chromosoms (Chromosom 22)	[Dunham *et al.*, 1999]
2001	Veröffentlichung einer ersten, noch lückenhaften Sequenz des menschlichen Genoms	[Lander *et al.*, 2001]
2003	Abschluss des HGP	
2005	454 Life Science stellt erstes NGS-Gerät vor	[Margulies *et al.*, 2005]
ab 2011	Serielle Aufklärung monogen bedingter Krankheiten des Menschen	

1.2 Gegenstand dieser Arbeit

Seit der Entwicklung der Hochdurchsatz-DNA-Sequenzierung (HDS) vor etwa 10 Jahren ist sie für die genetische Krankenversorgung und Forschung unentbehrlich geworden. Im Gegensatz zu England und anderen europäischen Nachbarländern fehlten in Deutschland jedoch bisher staatliche Initiativen zur koordinierten Einführung der HDS an qualifizierten akademischen Zentren. Zudem hielten die Regelungen zur Kostenerstattung durch die Krankenkassen derartiger Untersuchungen mit den Entwicklungen auf diesem Sektor nicht Schritt. Das Fehlen staatlicher Vorgaben zwang viele humangenetische Institute zum Aufbau der personellen und apparativen Infrastruktur für die NGS-gestützte Diagnostik in eigener Regie, so auch das Institut für Humangenetik der Universitätsmedizin Mainz.

Ein zentraler Aspekt dieser Aufgabe und gleichzeitig Thema dieser Dissertation ist die Entwicklung und Etablierung einer bioinformatischen Pipeline zur Analyse und Filterung von Sequenzdaten, der Identifikation pathogenetisch relevanter Sequenzvarianten und deren Abgleich mit klinischen Befunden. Erste

Erfahrungen konnten mit kleinen krankheitsgruppenspezifischen Genpanels gesammelt werden. Während des Berichtszeitraums kam es zu einer raschen Weiterentwicklung der betreffenden Methoden, Ressourcen und Konzepte. Parallel dazu stiegen die diagnostischen Möglichkeiten, aber auch die Ansprüche an die NGS-gestützte Diagnostik und die bioinformatische Analyse immer größerer und komplexerer Datensätze, besonders seit der Entwicklung der WES und WGS. Dies erforderte eine stetige Verbesserung der hausinternen bioinformatischen Pipeline und den Aufbau einer benutzerfreundlichen graphischen Oberfläche zur Interpretation der Ergebnisse.

Die vorliegende Arbeit geht (zum Teil chronologisch) auf diese Entwicklungen ein, die inzwischen auch in Mainz zur Einführung der diagnostischen Sequenzierung des gesamten Gesamtgenoms geführt haben.

Nicht in allen Fällen ist mit der Exomsequenzierung eine kausale Genveränderung zu finden. Dies lässt die Frage nach der Lage und Funktion bisher nicht als pathologisch definierter Mutationen aufkommen. In den letzten Jahren stellte sich heraus, dass den Veränderungen in nicht kodierenden Bereichen des Genoms eine große Rolle als Krankheitsursache zuzuteilen ist. Zur Detektion solcher Veränderungen wird versucht die zugrundeliegende Variante durch die Anwendung eines eigens designten TES Panels für regulatorische Bereiche (3'UTRs und mikro-RNAs) aufzuspüren. Dabei ist zu überlegen, wie eine Pathogenitätsabschätzung der noch weitgehend unbekannten Varianten in solchen regulatorischen Bereichen zu bewerkstelligen ist.

Anhand großer Patientenkollektive ist es möglich, die Aufklärungsraten zwischen Genpanel-Untersuchung und Exomsequenzierung zu vergleichen. In einem Fallbeispiel wird die gesamtgenomische Sequenzierung beschrieben und die Vor- und Nachteile der Einführung der WGS als universellen genetischen Test diskutiert.

Diese Arbeit schließt mit einem Ausblick auf neueste technische Entwicklungen auf diesem Gebiet, die schon bald in der genetischen Diagnostik Eingang finden könnten.

Kapitel 2

Verfahren zur Hochdurchsatz-Sequenzierung

Die Hochdzurchsatz-Sequenzierung von DNA oder RNA, welche auch als nächste Generation der Sequenzierung (engl. Next Generation Sequencing, NGS) bezeichnet wird, umfasst Methoden mit deren Hilfe bis zu mehrere Billionen von kurzen Sequenzbruchstücken (50 – 300 bp) parallel untersucht werden können.

2.1 Roche 454

Den ersten DNA-Sequenzierer der nächsten Generation (GS20) brachte das im Jahr 2007 von Roche aufgekaufte Unternehmen 454 Life Sciences 2005 auf den Markt. Bis zum Jahr 2009 verbesserte Roche/454 Life Sciences seine Sequenziergeräte, sodass diese in der Lage waren 400-600 Millionen Basenpaare pro 10-Stunden-Lauf zu sequenzieren [Voelkerding *et al.*, 2009]. Die 454-Technologie ist aus der Konvergenz von Pyrosequenzierung [Nyrén *et al.*, 1993] [Ronaghi *et al.*, 1996] [Ronaghi *et al.*, 1998] und Emulsions-PCR [Tawfik und Griffiths, 1998] entstanden.

Zur Vorbereitung auf die Sequenzierung wird die genomische DNA in kleinere Sequenzen (300-800 bp) fragmentiert und an Adapter-Oligonukleotide ligiert. Anschließend erfolgt eine Verdünnung der Bibliothek auf Einzelmolekülkonzentration sowie eine Denaturierung und Hybridisierung an einzelne Beads. Diese werden wiederrum in Wasser-Öl-Mikrovesikel unterteilt, wo es während der Emulsions-PCR zu einer klonalen Amplifikation einzelner an die Beads gebundener DNA-Moleküle kommt. Nach der Amplifikation wird die Emulsion aufgetrennt und die Beads angereichert.

Zur Sequenzierung erfolgt erneut eine Grenzverdünnung der Beads, welche im weiteren Verlauf in einzelne Vertiefungen der Picotiterplatte deponiert und mit Sequenzierungsenzymen, wie DNA-Polymerase, ATP-Sulfurylase und Luciferase kombiniert werden. Im Sequenziergerät fungiert die Picotiterplatte als Fließzelle (engl. Flow Cell), in der die iterative Pyrosequenzierung durch sequenzielles Hinzufügen der vier Desoxyribonukleosidtriphosphate (dNTPs) stattfindet.

Eine Nukleotidinkorporation erzeugt eine Pyrophosphatfreisetzung und durch die Luciferase eine gut lokalisierte Lumineszenz, die sich durch die faseroptische Platte übertragen und mit einer Kamera aufzeichnen lässt. Nach jeder Zugabe eines dNTPs werden die Lichtsignale der Wells fotografiert, nach Qualitätskriterien gefiltert und anschließend algorithmisch in eine lineare Sequenzausgabe übersetzt [Voelkerding *et al.*, 2009].

Eine anerkannte Stärke der 454-Technologie ist die größere Leselänge (bis zu 700 bp), die eine *de novo*-Montage von Genomen erleichtert [Pearson *et al.*, 2007]. Ein schwerwiegendes Problem ist hingegen die genaue Bestimmung von Homopolymeren mit einer Länge von mehr als drei bis vier Basen. Ein Homopolymer mit sechs Basen sollte theoretisch die doppelte Lumineszenz eines Homopolymers mit drei Basen ergeben. In der Praxis variiert diese Lumineszenzausbeute. So sind Schätzungen der Homopolymerlänge mit zunehmender Länge weniger genau [Margulies *et al.*, 2005] [Huse *et al.*, 2007]. Die schlechte Qualität der sequenzierten Reads und der im Vergleich zu anderen neu aufkommenden Firmen wie Illumina niedrige Durchsatz der Sequenziergeräte von 454 führte zu einer stetigen Abnahme der Wettbewerbsfähigkeit und 2013 schließlich zur Einstellung der Sequenziersparte bei Roche.

2.2 Helicos

Bereits während der Hochphase des oben vorgestellten 454-Sequenzierers wurde 2008 das Heliscope-Sequenziergerät vorgestellt. Es ist als erstes kommerzielles NGS-Gerät in der Lage, einzelne DNA-Moleküle ohne vorherigen Amplifikationsschritt zu sequenzieren.

Bei der Vorbereitung zur Sequenzierung wird bei dieser Methode jedes DNA-Fragment mit einem mindestens 50 bp langen Poly-A-Schwanz am 3‘-Ende versehen. Zudem ist es notwendig, die 3’-Enden der zu sequenzierenden Moleküle zu blockieren, damit diese beim eigentlichen Sequenzierungsschritt nicht verlängert werden können. Dies geschieht typischerweise mit Hilfe eines Didesoxynukleotids.

Die Einzelmolekül-Fluoreszenzsequenzierung wird auf einer Flow Cell aus Glas mit 25 Kanälen durchgeführt. Diese ist mit 50 bp langen Oligo(dT) Sequenzen bestückt, an die später die DNA-Fragmente binden. Dabei kann jeder Kanal einzeln adressiert und wenn gewünscht mit unterschiedlichen Proben beladen werden. Es ist möglich das Sequenziergerät mit einer oder zwei Fließzellen gleichzeitig zu betreiben.

Im Allgemeinen werden die Proben für die Sequenzierung so vorbereitet, dass der Poly(A)-Schwanz länger ist als das Oligo(dT)50 auf der Oberfläche der Fließzelle. Um die Sequenzierung der ungepaarten A-Reste zu vermeiden, ist eine „Fill-and-Lock“-Behandlung erforderlich. Bei dieser werden nach der Hybridisierung dTTPs und fluoreszenzmarkierte Terminator-Nukleotide (dATP, dCTP und dGTP) [Bowers *et al.*, 2009] gemeinsam mit der DNA-Polymerase

hinzugefügt. Auf diese Weise lassen sich alle ungepaarten Adenosinbasen des Poly(A)-Schwanzes mit TTPs auffüllen. Das hybridisierte Molekül wird für die weitere Synthese geblockt, wenn die Polymerase auf den ersten Nicht-A-Rest trifft und das entsprechende Terminator-Nukleotid einfügt. Da durch die Terminator-Nukleotide nun an jedes DNA-Molekül ein Farbstoff gebunden ist, enthält ein entsprechend aufgenommenes Bild die Positionen aller sequenzierbarer Moleküle. Dieses initial aufgenommene Bild dient dem System zur Orientierung auf der Fließzelle und beinhaltet keine Sequenzinformation, da die Markierung jeder beliebigen Base entsprechen könnte. Daher beginnt die Sequenzierung der Moleküle mit der zweiten Base des ursprünglichen Moleküls [Glenn, 2011].

Um die hybridisierten DNAs zu sequenzieren, ist es zunächst notwendig, die fluoreszenzmarkierten Terminatornukleotide abzuspalten und wegzuwaschen. Anschließend wird neue Polymerase sowie nacheinander jeweils eines der vier verschiedenen Fluoreszenznukleotide hinzugefügt und nach Anregung des Fluoreszenzanteils durch einen Laser ein weiteres Bild aufgenommen. Bei einem Standard-Sequenzierungslauf wird dieser zyklische Prozess 120 Mal wiederholt.

Unter optimalen Bedingungen können für einen Standardlauf mit 120 Zyklen von jedem Kanal 12.000.000 bis 20.000.000 Reads mit einer Länge von 25 - 32 bp erwartet werden. Dies resultiert in insgesamt bis zu 1.000.000.000.000 Reads und 35 Gigabasenpaaren (Gbp) Sequenzinformation von jedem Lauf. Ein kompletter Durchlauf dauert allerdings bis zu 8 Tage.

Vorteile dieser Einzelmolekül-Sequenzierungsstrategie sind die Vereinfachung der DNA-Probenvorbereitung, die Vermeidung von PCR-induzierten Verzerrungen und Fehlern, eine zum Teil vereinfachte Datenanalyse und die Toleranz gegenüber degradierten Proben. Da der Sequenzierungsprozess allerdings nach jedem Verlängerungsschritt angehalten wird, ist der Zeitbedarf dieser Methode groß, und die realisierte Leselänge beträgt nur 32 Nukleotide. Zudem ist die Fehlerrate hoch. Dies kann mit repetitiver Sequenzierung überwunden werden, welches jedoch die Kosten pro sequenzierter Base erhöht. Vermutlich konnte sich diese Sequenziermethode aufgrund dieser Nachteile auf dem Markt nicht etablieren, was im November 2012 zur Insolvenz der Firma Helicos BioSciences führte [Robison, 2013].

2.3 Complete Genomics/BGI

Parallel zu Roche 454 und Helicos entwickelte die 2006 gegründete und im März 2013 von BGI-Shenzhen übernommene Firma Complete Genomics eine DNA-Sequenzierungsplattform auf der Basis von sogenannten DNA-Nanobällen (DNB).

Bei dieser Technologie wird die fragmentierte DNA durch spezielle Adapter zirkularisiert und anschließend durch die „Rolling Circle Replication“ [Ali *et al.*, 2014] vervielfältigt. Dabei bildet sich ein zusammenhängender

Strang aus sich wiederholenden Kopien, welcher sich zu einem DNB mit etwa 220-240 Nanometer (nm) Durchmesser verdichtet. Die so entstandenen DNBs sind negativ geladen und stoßen sich so auf natürliche Weise ab. Dies verhindert die Verbindung verschiedener Nanobälle untereinander. Zur Sequenzierung werden die DNA-Nanobälle auf eine Fließzelle gegeben. Auf dieser befinden sich als Array angeordnete positiv geladene Aminosilan-Spots mit einem Durchmesser von 220 nm, an welche die DNBs durch ihre negative Ladung selektiv binden.Die Sequenzierung erfolgt analog zum von Illumina bekannten Schema des „Sequencing by Synthesis“ (siehe Kapitel 2.4). Nach jedem Einbau fluoreszenzmarkierter DNA-Nukleotide werden die jeweiligen Fluorophore mithilfe eines Lasers angeregt und die Fluoreszenzemissionen aller an die Fließzelle gebundenen DNA-Nanobälle mit einer hochauflösenden CCD-Kamera erfasst. Aus der Abfolge der Bilder lässt sich dann die Intensität jedes einzelnen Punktes beurteilen und die Reihenfolge der eingebauten Basen bestimmen.

Die DNA-Nanoball-Sequenziertechnologie bietet einige Vorteile gegenüber anderen Sequenzierplattformen. Ein Vorteil ist die Eliminierung optischer Duplikate, da die DNA-Nanobälle auf dem geordneten Array der Fließzelle an Ort und Stelle bleiben und nicht mit benachbarten Nanobällen interagieren. Ein weiterer positiver Aspekt ist die Verwendung der High-Fidelity-DNA-Polymerase Phi 29, um eine genaue Amplifikation der kreisförmigen Matrize zu gewährleisten. Durch die „Rolling–Circle“-Replikation kommt es zudem nicht zur Verbreitung von durch die Polymerase verursachten Replikationsfehlern. Der Hauptnachteil der DNA-Nanoball-Sequenzierung ist, wie bei allen Next-Generation-Sequencing Methoden, die kurze Leselänge der mit dieser Methode erhaltenen DNA-Sequenzen. Kurze Reads repetitiver Genomabschnitte lassen sich zudem oft nicht einer einzigen Region des menschlichen Genoms zuordnen. Aufgrund der großen Präzision und des hohen Durchsatzes dieser Methode und den damit verbundenen geringen Kosten ist das BGI zu einem der großen, etablierten Anbieter für Sequenzierdienstleistungen und Sequenziersysteme geworden.

2.4 Solexa/Illumina

Die im Jahre 1998 gegründete, auf dem Gebiet von DNA-Mikroraster-Technologien spezialisierte amerikanische Firma Illumina, wurde nach der Übernahme der britischen Firma Solexa im Jahre 2007 zum größten Anbieter von Sequenziersystemen weltweit [Cimino, 2020]. Der heutige Marktanteil beträgt 90% (Stand 17.12.2018 [Scheid, 2018]).

Der MiSeq sowie der NextSeq500 sind zwei Beispiele für Sequenziersysteme der Firma Illumina, die in Mainz diagnostisch genutzt werden. Beide verwenden die Methode des „Sequencing by Synthesis“ (SBS) mithilfe von reversiblen Farbstoff-Terminatoren.

Zur Herstellung sog. Sequenzierbibliotheken („Sequencing Libraries“) wird die zu untersuchende DNA dabei zuerst fragmentiert, beidseitig mit Adaptern ver-

sehen und anschließend in Einzelstränge aufgespalten. Diese Fragmente sind je nach Anwendung etwa 500 - 5000 bp groß. Aufgrund verschiedener Arten von „Hintergrundrauschen“, das sich mit jedem Sequenzierungsschritt verstärkt, ist die SBS Technologie letztendlich in ihrer Leselänge begrenzt. Während der MiSeq optional 50 - 250 bp dieser Fragmente von einem Ende (engl. Single-End, SE) oder von beiden Enden (engl. Paired-End, PE) sequenzieren kann, liegt die Leseweite beim NextSeq500 jeweils zwischen 75 bp und 150 bp. Der MiSeq produziert bis zu 50 Millionen Reads bei einer PE Sequenzierung mit dem MiSeq Reagent Kit v3, wohingegen der NextSeq bis zu 800 Millionen Reads bei einer PE Sequenzierung mit einem High-Output Kit erzeugt [Illumina, 2017a] [Illumina, 2020b].

Bei einer PE-Sequenzierung wird im Gegensatz zur SE Sequenzierung ein zweiter SBS-Lesevorgang durchgeführt, indem das gegenüberliegende Ende jedes Bibliotheksfragments mit einem adapterspezifischen Oligonukleotid (Primer) vorbereitet und anschließend sequenziert wird [McCombie *et al.*, 2019]. Die typische Anwendung der PE-Sequenzierung ist jene mit Fragmentgrößen (Inserts) bis maximal 1000 bp, wobei die meisten Libraries Fragmente zwischen 200 bp und 800 bp beinhalten. Aufgrund der geringen Fragmentlänge werden die resultierenden Sequenzfragmente (Reads) solcher Libraries auch als „Short-Insert Paired-End Reads“ (SIPER) bezeichnet (siehe Abbildung 2.1 links). Sollen größere Bereiche zwischen den beiden sequenzierten Fragmenten liegen, muss das Protokoll für die Herstellung der Sequenzierbibliotheken angepaßt werden. Die DNA wird dabei initial in 2 - 5 kbp große Stücke fragmentiert und anschließend an beiden Enden biotinyliert. Die Folge ist eine Zirkularisierung der Fragmente. Im Anschluss werden diese DNA-Ringe wiederum zu 400 - 600 bp großen linearen DNA-Stücken fragmentiert und durch den Biotinanhang spezifisch angereichert. Die daraus resultierenden „Long-Insert Paired-End Reads“ (LIPERs) werden auch als Mate-Pair bezeichnet (siehe Abbildung 2.1 rechts) [ecSeq Bioinformatics, 2017].

Die Flow Cell besitzt auf ihrer Oberfläche Oligonukleotide, die zu den Adaptoren der DNA-Fragmente komplementär sind. Nach Applizieren der DNA auf die Flow Cell, binden die Adaptoren des DNA-Fragments an die Oligonukleotide. Anschließend erfolgt eine Amplifikation der DNA mit Hilfe der Brücken-PCR (Bridge Amplification). Dabei wird ein chemisches Milieu geschaffen, bei dem sich die DNA-Fragmente biegen und mit dem Adapter des freien Endes an ein Oligonukleotid der Flow Cell binden und sogenannte Brücken ausbilden. Dieses Oligonukleotid dient gleichzeitig als Primer (kurzes Oligonukleotid, das als Startpunkt für die DNA-Synthese dient) für die DNA-Synthese. Die komplementären Stränge werden entlang der Matrize synthetisiert. Durch einen Denaturierungsschritt (Trennung der komplementären DNA-Stränge) liegen die DNA-Fragmente wieder in Einzelsträngen vor. Nach einigen Wiederholungen der Brücken-PCR bilden sich Gruppen von gleichen DNA-Fragmenten, sogenannte Cluster, aus. Am Ende der Brückenamplifikation wird der zum originalen DNA-Fragment komplementäre DNA-Strang vom Oligonukleotid der Flow Cell getrennt und durch Waschen entfernt. So besteht ein Cluster nun

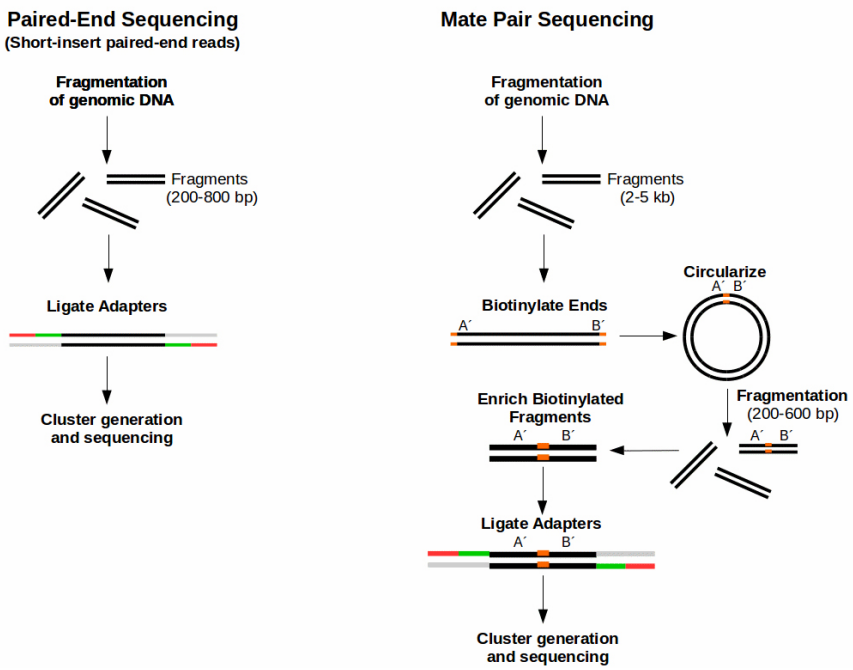


Abbildung 2.1: *Vergleich der Probenvorbereitung für die Illumina Paired-End- und Mate-Pair-Sequenzierung*
Bei der Probenaufbereitung der PE-Sequenzierung werden die Adapter direkt an die 200 - 800 bp langen DNA-Fragmente gebunden (links). Die Fragmente der Mate Pair-Sequenzierung sind mit 2 - 5 kbp länger. Durch die an beiden Enden angebrachte Biotinylierung (orange) zirkulieren die Fragmente. Nach der anschließenden Fragmentierung dieser zirkulären Struktur werden die Fragmente mit Bitinanhang spezifisch angereichert und mit Adaptern versehen (rechts). [ecSeq Bioinformatics, 2017]

aus mehreren gleichen Kopien des ursprünglichen DNA-Fragments. Der letzte Schritt vor der Sequenzierung besteht in der Hybridisierung eines universellen, zum Adapter komplementären Primers an jedes DNA-Fragment auf der Flow Cell (siehe Abbildung 2.2) [Illumina, 2013b].

Die Sequenzierung der Cluster erfolgt durch die Replikation (Synthese) des komplementären Strangs. Es werden farbstoffmarkierte (Fluorophor) Desoxynukleotide mit reversiblen Terminatorgruppen verwendet. Nach Zugabe dieses Nukleotidgemisches auf die Flow Cell hybridisiert das zur Matrize komplementäre Nukleotid an den DNA-Strang. Die Terminatorgruppe verhindert die Bindung eines weiteren Nukleotids und bricht die DNA-Synthese somit ab, wodurch der Strang pro Zyklus nur um ein Nukleotid erweitert werden kann [Illumina, 2013a]. Das Auswaschen entfernt überschüssige Nukleotide. Zur Identifizierung der eingebauten Base wird der Fluorophor durch einen Laser zum Leuchten angeregt. Eine Kamera nimmt dieses Fluoreszenzsignal auf und speichert es als Bild ab (siehe Abbildung 2.3). Um den nächsten Sequenzierungszyklus zu ermöglichen, wird der Fluorophor und die Terminatorgruppe von den jeweils eingebauten Nukleotiden abgespalten und aus der Flow Cell

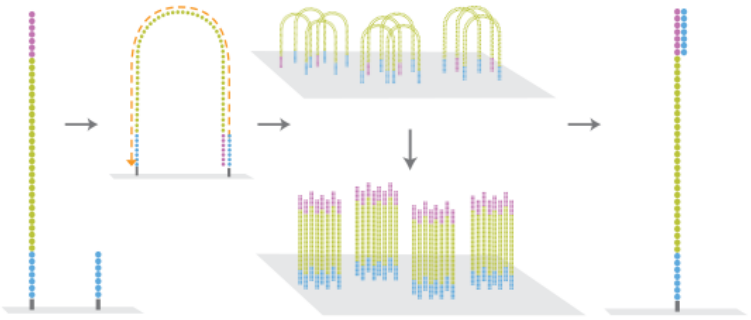


Abbildung 2.2: *Brücken-PCR und Cluster Generierung.*
Ein auf der Flow Cell immobilisierter DNA-Einzelstrang bindet mit seinem Adapter (lila) an einen Primer in der Umgebung (blau). Es entsteht eine brückenähnliche Form. Der komplementäre DNA-Strang wird ausgehend vom Primer (blau) auf der Flow Cell entlang der Matrize synthetisiert (orange gestrichelt). Nach dem Abspalten und Auswaschen der zum Original reversen DNA-Stränge sind Cluster gleicher DNA-Fragmente entstanden (Mitte unten). An jedes DNA-Fragment wird ein zum Adapter komplementärer Primer (blau) als Startpunkt der Sequenzierung hybridisiert. [Illumina, 2013b]

ausgewaschen [Illumina, 2013b].

Bei der 4-Kanal SBS Chemie trägt jedes der vier Nukleotide Adenin (A), Thymin (T), Guanin (G) und Cytosin (C) einen unterschiedlichen Fluorophor (farbliche Markierung). So sind für jeden Sequenzierungszyklus vier einzelne Schritte notwendig, bei denen die Fluorophore jeweils mit unterschiedlichen

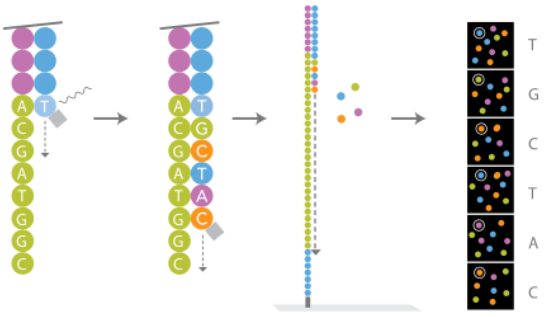


Abbildung 2.3: *Sequencing by Synthesis.*
Es wird ein Gemisch aus fluoreszenzmarkierten und mit einem Terminator versehenen Nukleotiden (A,T,C,G) auf die Flow Cell aufgetragen. Die Elongation erfolgt nur durch das zur Matrize komplementäre Nukleotid. Die Übrigen werden ausgewaschen. Anschließend regt ein Laser den Fluorophor zum Leuchten an. Die Kamera nimmt dieses Fluoreszenzsignal auf und speichert es als Bild ab. Nach der Abspaltung der Terminatorgruppe sowie des Fluorophors beginnt ein neuer Syntheseschritt. Durch die Abfolge der Bilder kann die Basensequenz bestimmt werden [Illumina, 2013b].

Wellenlängen zum Leuchten angeregt und einzelne Bilder aufgenommen werden (siehe Abblildung 2.4 links). Diese Technik kommt im MiSeq und HiSeq Gerät von Illumina zum Einsatz. Bei der 2-Kanal SBS Chemie, die beim MiniSeq, NextSeq und NovaSeq Anwendung findet [Illumina, 2018], werden hingegen nur zwei unterschiedliche Fluorophore (Rot und Grün) und somit nur zwei Bilder pro Sequenzierungszyklus verwendet, um alle vier Nukleotide unterscheiden zu können. Thymin-Basen sind durch ein grünes und Cytosine durch ein rotes markiert. Adenine hingegen tragen einen roten und einen grünen Farbstoff und Guanin-Basen besitzen kein angehängtes Fluorophor (siehe Abblildung 2.4 rechts) [Illumina, 2018].

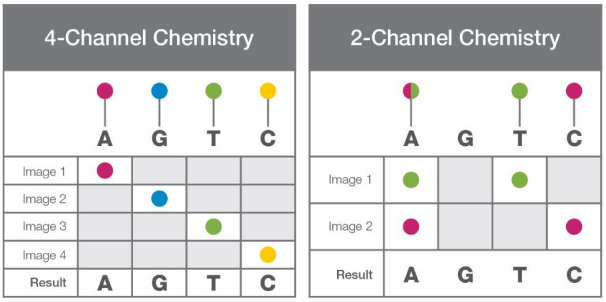


Abbildung 2.4: *2- und 4-Kanal SBS Chemie.*
Die 4-Kanal-Chemie verwendet eine Mischung von Nukleotiden, die mit vier verschiedenen Fluoreszenzfarbstoffen markiert sind (links). Bei der 2-Kanal-Chemie werden zwei verschiedene Fluoreszenzfarbstoffe verwendet (rechts). Die Bilder werden von einer Bildanalysesoftware verarbeitet, um die Nukleotididentität zu bestimmen [Illumina, 2018].

Aus der Reihenfolge der in jedem Zyklus entstandenen Bilder ist es möglich die Sequenz des originalen DNA-Fragments für jedes Cluster zu ermitteln (Basecalling). Die Basenabfolge wird zusammen mit einem errechneten Qualitätswert, dem „Quality Score“ als FASTQ-Datei (siehe Anhang II.II) abgespeichert.

Derzeit entwickeln die Firma Oxford Nanopore und Pacific Bioscience Sequenziermethoden der dritten Generation. Diese erlauben die Sequenzierung von DNA-Fragmenten mit einer Länge von bis zu 175 kbp [Eisenstein, 2012] [Mikheyev und Tin, 2014] (siehe Kapitel 6).

Kapitel 3

Bedeutung der NGS für die genetische Diagnostik

Grundsätzlich unterscheiden sich die Genome von zwei nicht verwandten Menschen je nach Population in etwa 0,1% - 0,6% ihrer DNA-Bausteine [Jorde und Wooding, 2004] [Tishkoff und Kidd, 2004] [Auton *et al.*, 2015]. Zusätzlich zur von Generation zu Generation weitergegebenen genetischen Information, wird jedes Individuum im Durchschnitt mit 44 – 82 neuen genetischen Veränderungen, sogenannten *de novo*-Mutationen, geboren [Michaelson *et al.*, 2012] [Kong *et al.*, 2012] [Gilissen *et al.*, 2014] [Francioli *et al.*, 2015] [Goldmann *et al.*, 2016]. Diese sind entweder während der Bildung der Keimzellen oder postzygotisch entstanden und meist funktionell neutral [Acuna-Hidalgo *et al.*, 2016]. Etwa ein bis zwei dieser *de novo*-Varianten betreffen einen proteinkodierenden Bereich im Genom.

Mit einer klassischen zytogenetischen Chromosomenanalyse ist es möglich numerische und strukturelle Chromosomenaberrationen in kultivierten Lymphozyten nachzuweisen. Diese Bänderungstechnik hat eine Auflösungsgrenze von etwa zehn Megabasenpaaren (Mbp) [Weise *et al.*, 2014]. Durch die molekularzytogenetische Fluoreszenz-In-Situ-Hybridisierungs (FISH)-Technik lassen sich mithilfe spezifischer DNA-Sonden auch einzelne kleinere strukturelle Veränderungen nachweisen [Weise *et al.*, 2014]. Mit der Einführung der Mikroarray-Technologie gelang es erstmals das Genom hochauflösend auf Verluste und Zugewinne (Kopienzahlveränderungen) zu analysieren [Müller und Röder, 2004]. Die Auflösungsgrenze solcher Arrays liegt heute bei etwa 10 kbp. Zudem war es mithilfe von Mikroarrays erstmals möglich, Homozygotieregionen (engl. Regions of Homozygosity, ROH) in verschiedenen Populationen umfassend nachzuweisen. Alle genetischen Varianten in solchen Regionen liegen homozygot vor, das heißt die Sequenzen dieser Genomabschnitte sind auf beiden Chromosomen identisch. Die Untersuchung solcher ROHs kann bei Patienten mit blutsverwandten Eltern ist eine erfolgreiche Strategie für die Aufklärung der molekularen Ursachen rezessiver Krankheiten. Zur Untersuchung kleiner Abschnitte des Genoms, zum Beispiel einzelner Gene auf Basenebene, galt jahrelang die in Kapitel 1.1 bereits vorgestellte Sanger-Sequenzierung als Goldstandard. Erst die Entwicklung von Methoden der Hochdurchsatz-Sequenzierung

ermöglichte die parallele Sequenzierung mehrerer Gene oder gar des gesamten menschlichen Genoms.

Da nur etwa 1% - 2% des Genoms in Proteine übersetzt wird [Ng *et al.*, 2009], fallen viele der insgesamt 4,1 – 5 Millionen Varianten eines Menschen in intergenische oder intronische Bereiche. Allerdings bewirkt auch nur ein Teil der Varianten in kodierenden Bereichen des Genoms eine funktionelle Veränderung des daraus entstehenden Proteins. Wie in Kapitel 1.1 bereits erwähnt, gelang es bis heute erst bei weniger als 25% aller Gene, Sequenzvarianten eindeutig mit genetisch bedingten Krankheiten in Verbindung zu bringen. Auch die Unterscheidung von funktionell neutralen und pathogenen Varianten in bekannten Krankheitsgenen ist alles Andere als trivial.

Bei der Beurteilung von Varianten kann auf Genotyp-Phänotyp-Datenbanken wie HGMD, ClinVar, OMIM oder Orphanet (siehe Anhang III.I, III.III, III.IV, III.V) zurückgegriffen werden. Diese Datenbanken enthalten jene Varianten, die bereits bei anderen Patienten gefunden und als ursächlich für eine Erkrankung beschrieben worden sind. Zusammengefasst machen diese allerdings nur etwa 150.000 der ca. 10 Milliarden möglichen Basenaustausche aus [The Monarch Initiative, 2020]. Zur pathogenetischen Beurteilung einer zuvor noch nicht mit einer Erkrankung assoziierten Variante haben sich zudem Datenbanken wie gnomAD, dbSNP oder ENSEMBL als nützlich erwiesen, denen man, gegebenenfalls auch aufgeschlüsselt nach bestimmten Subpopulationen, die Häufigkeit bestimmter Varianten bei klinisch gesunden Patienten entnehmen kann (siehe Anhang III.VI, III.II und III.VII).

Zusätzlich existieren unzählige Algorithmen, die eine Vorhersage bezüglich der Pathogenität von Varianten liefern. Zu den am häufigsten verwendeten Algorithmen gehören MutationTaster [Schwarz *et al.*, 2014], SIFT [Vaser *et al.*, 2016], PolyPhen2 [Adzhubei *et al.*, 2010] und CADD [Kircher *et al.*, 2014]. Allerdings sind die Ergebnisse solcher Algorithmen nur als richtungsweisend und nicht als Beweis der Pathogenität oder funktionellen Neutralität einer bisher unbekannten genetischen Variante zu werten. Für viele Varianten, vor allem jene in nicht kodierenden Genomabschnitten, gibt es immer noch kaum belastbare Informationen zu ihrer Pathogenität. Für definitive Aussagen zur Pathogenität derartiger Varianten kann auf funktionelle Studien nicht verzichtet werden, die aber in den allermeisten Fällen noch nicht existieren, zum Beispiel weil die Genfunktion nicht bekannt ist (siehe Kapitel 6).

In der Praxis der medizinisch-genetischen Diagnostik nimmt die Analyse des klinischen Bildes, des sogenannten Phänotyps, traditionell eine zentrale Rolle ein. Dies gilt insbesondere für Patienten mit Fehlbildungssyndromen, die auf einen oder wenige Gendefekte zurückgehen. Allerdings sind für viele derartige Störungen die molekularen Ursachen noch nicht bekannt. Insbesondere bei der zahlenmäßig wichtigsten Patientengruppe, den Probanden mit psychomotorischen Entwicklungsstörungen, kommen häufig Hunderte, wenn nicht gar

Tausende verschiedene Gendefekte als Ursache infrage, die sich aufgrund ihrer variablen Manifestation klinisch nicht gegeneinander abgrenzen lassen. Bis solche seltenen Erkrankungen richtig diagnostiziert werden, vergehen einer europäischen Studie zufolge [EURORDIS, 2020a] deshalb im Mittel 7 Jahre, und trotz langer Odysseen von Arzt zu Arzt und von Krankenhaus zu Krankenhaus wird bei vielen dieser Patienten und Familien die richtige Diagnose niemals gestellt.

Die Einführung von Verfahren, die eine gleichzeitige Untersuchung vieler infrage kommender Krankheitsgene (Genpanel), aller menschlichen Gene (Whole Exome Sequencing, WES) oder sogar des gesamten menschlichen Genoms erlaubt (WGS), hat die medizinische Genetik im vergangenen Jahrzehnt auf eine neue, viel breitere Basis gestellt, und die flächendeckende Einführung dieser Methoden verspricht die genetische Diagnostik und Krankenversorgung entscheidend zu beschleunigen und zu verbessern.

Die WES und die Genpaneluntersuchung werden unter der Bezeichnung „Target Enrichment Sequencing“ (TES) zusammengefasst und basieren auf der vorhergehenden, gezielten Anreicherung zu untersuchender Genomabschnitte. Die folgenden Unterkapitel beschreiben die Methode des TES sowie die gesamtgenomische Sequenzierung.

3.1 Target Enrichment Sequencing

In der Frühzeit der NGS-Diagnostik wurden Panel zur Anreicherung kodierender Abschnitte von Genen entwickelt, deren Defekte als Ursache von genetisch heterogenen monogenen Krankheiten (wie z.B. erbliche Netzhaut-Degeneration, familiäre Innenohr-Taubheit oder später psychomotorische Entwicklungsstörungen) bekannt waren. Ein in der Gendiagnostik am Institut für Humangenetik Mainz jahrelang verwendeter Test basiert auf einem 2012 beschriebenen Genpanel [Kingsmore, 2012]. Mit diesem auch MPIMG1 genannten Test lassen sich 1222 Gene, deren Mutationen ursächlich für seltene, rezessive pädiatrische Genleiden, geistige Behinderungen und verwandte Störungen sein können, untersuchen. Diese gezielte Analyse einer begrenzten Zahl von Genen ermöglicht eine kosteneffiziente Sequenzierung sowie aufgrund der dabei anfallenden geringeren Datenmenge eine schnellere Auswertung der Daten.

Andererseits sind diagnostische Genpanel grundsätzlich auf bereits bekannte Krankheitsgene beschränkt, wodurch sich bisher nicht beschriebene Gendefekte mit diesen Panels nicht identifizieren lassen. Mit der Entwicklung immer leistungsfähigerer Sequenziergeräte und der damit verbundenen Reduktion der Kosten pro sequenzierter Base wurde die gleichzeitige Untersuchung aller kodierenden Bereiche des Genoms im Rahmen der Krankenversorgung ökonomisch vertretbarer. In diesen kodierenden Bereichen liegen 85% der krankheitsverursachenden Varianten, sie stellen aber nur etwa 1% - 2% des gesamten Genoms dar [Ng *et al.*, 2009].

Da es sich bei der WES um eine Variante des „Target Enrichment Sequencing“ handelt, die (fast) alle menschlichen umfasst, sind die Arbeitsschritte zur Probenaufbereitung für beide Herangehensweisen gleich. Im Folgenden wird die Anreicherung (engl. Enrichment) der kodierenden Bereiche dieser Gene mithilfe des kommerziell erhältlichen SureSelectXT Kits der Firma Agilent erläutert.

Im ersten Schritt wird die doppelsträngige genomische DNA mittels Transposase enzymatisch fragmentiert, an Adaptoren gebunden und schließlich zu Einzelsträngen denaturiert. Für die Anreicherung der gewünschten Zielsequenzen kommen spezifische, diesen Regionen komplementäre und biotinylierte RNA-Sonden (engl. Baits) zum Einsatz. Im Falle einer WES gibt es bereits vorgefertigte, kommerziell vertriebene Kits mit entsprechenden Sonden (z.B. Agilent Human All Exon V7).

Für die Anreicherung von Sequenzen bestimmter Gene oder Genomabschnitte müssen anhand der genomischen Koordinaten der Zielregion spezifische RNA-Sonden entworfen werden. Hierzu dienen Online-Software Tools der jeweiligen Hersteller kommerzieller NGS-Kits. Diese RNA Baits hybridisieren durch Zugabe eines Hybridisierungspuffers an die DNA-Fragmente. Die im nächsten Schritt hinzugegebenen magnetischen Streptavidin-Beads binden an die biotinylierten Sonden und ziehen diese samt DNA-Fragment nach Anlegen eines magnetischen Feldes nach unten aus der Lösung. Ein Entfernen des Überstands verwirft nicht gewünschte DNA-Bereiche. Die DNA-Fragmente werden chemisch von den Baits getrennt und einer erneuten Anreicherungsreaktion unterzogen. Nach der Amplifizierung der angereicherten Regionen sind die DNA-Fragmente bereit zur Sequenzierung (siehe Abbildung 3.1).

Wie bereits erwähnt können mit Hilfe einer TES nur Varianten in angereicherten Regionen, bei einer WES hingegen in allen bekannten kodierenden Bereichen und deren flankierenden Regionen detektiert werden.

Nicht alle anscheinend hereditären Krankheitsbilder lassen sich auf Sequenzvarianten in kodierenden Genomabschnitten zurückführen, weshalb schon lange die Frage im Raum steht, wo die „fehlenden Mutationen“ zu finden sind. Bereits 1992 führte David N. Cooper aus, dass Mutationen in Promoterregionen die Genregulation beeinflussen und so Krankheiten auslösen können [Cooper, 1992]. Durch die Entdeckung und Aufklärung weiterer regulierender Elemente wie den miRNAs und deren Bindestellen innerhalb der untranslatierten Regionen am 3’ Ende des Gens (3’UTRs) liegt die Vermutung nahe, dass sich in diesen Genomregionen krankheitsassoziierte Varianten verbergen.

Um auch diese Varianten detektieren zu können wurde im Laufe der dieser Arbeit zugrundeliegenden Untersuchungen zudem ein TES-Genpanel erstellt, das die 3’UTR-Regionen aller bekannten, mit Krankheiten assoziierten Gene sowie alle bekannten miRNA-Gene enthält (siehe Kapitel 4.4.1). Dieses Panel wurde zur Nachuntersuchung von Patienten mit begründetem Verdacht auf

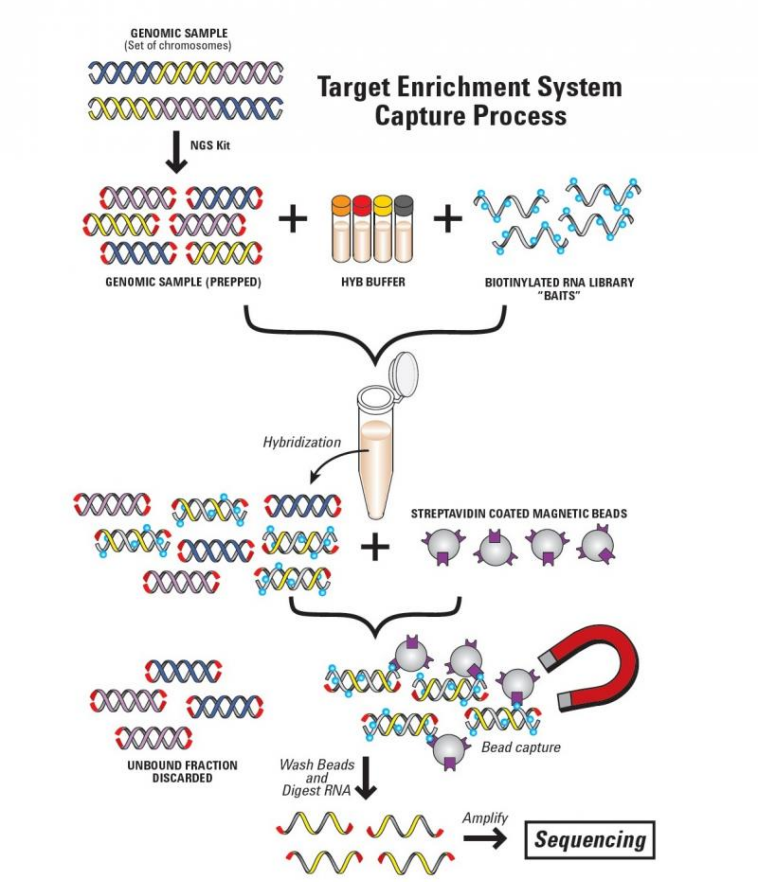


Abbildung 3.1: *Library Preparation für eine „Target Enrichment Sequenzierung“.*
Die doppelsträngige DNA wird enzymatisch fragmentiert und an Adaptoren gebunden. Nach der Denaturierung zu Einzelsträngen werden für die zu untersuchenden Regionen spezifische und komplementäre RNA-Sonden hinzugegeben, welche an die DNA-Fragmente binden. Durch Zugabe von magnetischen Streptavidin-Beads lassen sich so die gewünschten Fragmente isolieren und anreichern. Nach Ablösen der RNA-Sonden und Amplifizierung ist die Library bereit zur Sequenzierung [National Cancer Institute, 2014].

eine genetische Krankheitsursache verwendet, bei denen WES-Analysen nicht zu einer eindeutigen Diagnose geführt hatten.

3.2 Genomsequenzierung

Im Gegensatz zur TES, bei der nur spezifische zuvor angereicherte Gen- oder Genomabschnitte untersucht werden, erlaubt die sog. „Ganzgenomsequenzierung“ (Whole Genome Sequencing, WGS) im Prinzip die Identifizierung aller Varianten des 2x 3,2 Milliarden Basenpaare umfassenden Genoms menschlicher Körperzellen. In den ersten Jahren nach der Einführung der NGS-Techniken war die WGS aufgrund der damit verbundenen exorbitanten Kosten auf weni-

ge Forschungsprojekte beschränkt, jedoch hat sie aufgrund spektakulärer Kostensenkungen während des vergangenen Jahrzehnts zunehmend Eingang in die klinische Diagnostik gefunden [Gilissen *et al.*, 2014]. Inzwischen zeichnet sich ab, dass die WGS bereits in wenigen Jahren einen Großteil der heute in der genetischen Diagnostik verwendeten Methoden ersetzen könnte, zumal dieses Verfahren [Mooney, 2015], zumal dieses Verfahren sich grundsätzlich zur Erfassung aller Genomvarianten eignet (siehe Kapitel 5.3).

Die „Library Preparation“ zur Sequenzierung der genomischen DNA auf einem Illumina-Gerät (z.B. Nova-Seq) wird hier am Beispiel des TruSeq DNA PCR-Free Kits der Firma Illumina dargestellt.

Zu Beginn ist eine Fragmentierung der DNA mittels Ultraschall notwendig (siehe Abbildung 3.2 A). Überstehende Enden (Sticky Ends) der jeweiligen doppelsträngigen Fragmente werden durch Auffüllen mit Basen und Exonukleasereaktionen entfernt, sodass gerade Enden (Blunt Ends) entstehen (siehe Abbildung 3.2 B). Im Anschluss erfolgt eine Größenselektion der Fragmente durch Aufreinigungsbeads (siehe Abbildung 3.2 C). Zur Vorbereitung auf die Ligation der Sequenzier-Adapter wird an das 3'-Ende jedes Strangs ein Adeninmolekül angehängt (siehe Abbildung 3.2 D). Die Adapter besitzen einen

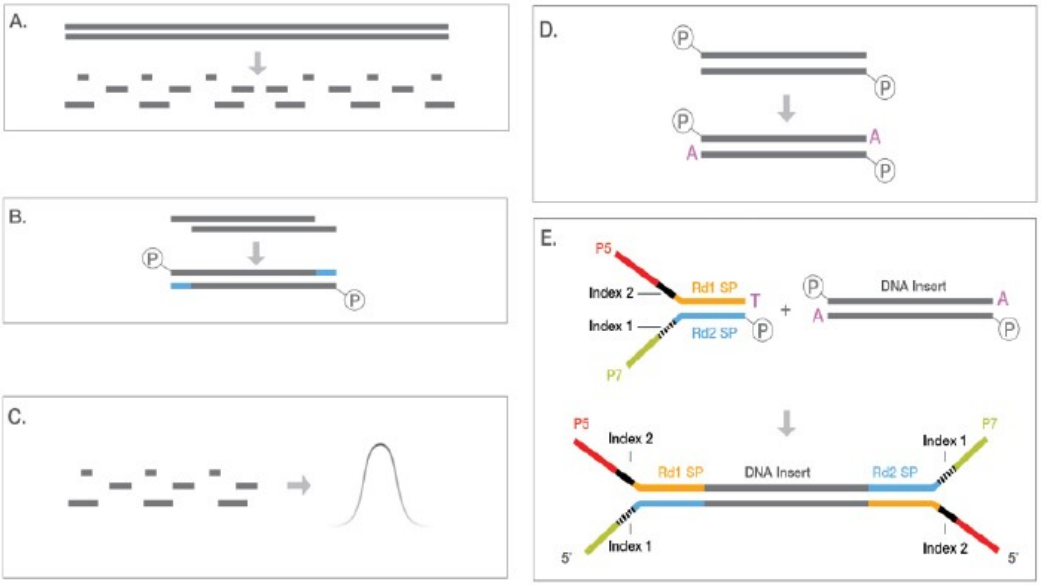


Abbildung 3.2: *„Library Preparation“ für eine WGS.*
Die fragmentierte DNA (A) erhält Blunt Ends (B) und wird im nächsten Schritt einer Größenselektion unterzogen (C). Durch die Ligation einer Adeninbase an das 3'-Ende (D) können die Sequenzierprimer mit Adaptoren und eindeutigen Patientenbarcodes angehangen werden (E) [Illumina, 2017c].

zum DNA-Fragment komplementären Thymin-Überhang an ihrem 3'-Ende, mit dem sie an das Fragment binden. Sie beinhalten Indizes zur Identifizierung der Probe sowie die zum Sequenzierprimer komplementäre Sequenz (sie-

he Abbildung 3.2 E). Da nun über die Indizes eine eindeutige Zuweisung jedes DNA-Bruchstückes zu einem Individuum möglich ist, werden die zuvor pro Individuum getrennt behandelten DNA-Fragmente gemischt (Poolen der Proben) [Illumina, 2017c]. Diese gepoolte Library ist nun fertig zur Sequenzierung.

3.3 Bioinformatische Analysemethoden

Nach Abschluss eines NGS-Laufs müssen die durch das Gerät während der einzelnen Sequenzierungszyklen aufgenommen Bilddateien in eine DNA-Sequenz umgewandelt werden. Bei Illuminas NextSeq500 Sequenziergerät erledigt diese Aufgabe der BCL2FASTQ-Agorithmus (siehe Anhang IV.V). Neben der Konvertierung ordnet der Algorithmus beim „Demultiplexing“ anhand eines vom Benutzer bereitgestellten Sample Sheets (siehe Anhang II.I) unter Verwendung der Nukleotidbarcodes die Sequenzfragmente wieder einem spezifischen Patienten zu.

Um aus diesen resultierenden Millionen kurzen Sequenzbruchstücken (Reads) Erkenntnisse über den Genotypen des Patienten zu erhalten, ist es notwendig diese wieder systematisch zu einer einzigen Sequenz zusammenzufügen. Bei einem Resequenzierungsprojekt besteht die Intention die Reads mit einem bereits vorhandenen Referenzgenom zu vergleichen und einer bestimmten Position in diesem zuzuordnen. Diese Aufgabe wird durch die genetische Variation der untersuchten Population, entstandene Sequenzierfehler, die geringe Länge der Reads und deren große Datenmenge erschwert.

Statistisch gesehen würde man erwarten, dass eine DNA-Sequenz mit einer Länge von mindestens 30 bp höchstens ein einziges Mal im menschlichen Referenzgenom vorkommt (Unique Hit), jedoch gilt das nicht für repetitive Sequenzen wie sogenannten Short- und Long Interspersed Nuclear Elements (SINES und LINES) [Singer, 1982], von denen viele tausend Kopien im Genom existieren, oder für evolutionär nahe verwandte Gene, die sich ab und an nur in wenigen DNA-Bausteinen unterscheiden [Mighell *et al.*, 2000]. In der Praxis kann es allerdings vorkommen, dass sich einige Reads aufgrund von DNA-Kontamination oder wegen Artefakten während der Sequenzierung keiner Region in der Referenz zuordnen lassen. Aufgrund der genetischen Variation innerhalb von oder zwischen Populationen ist es notwendig, effiziente Mapping-Strategien zu entwerfen, die eine begrenzte Anzahl an Diskrepanzen (Mismatches) und kleinen Insertionen oder Deletionen zulassen [Horner *et al.*, 2009].

Die DNA besteht aus antiparallelen Doppelsträngen, dem Forward- und dem Reverse-Strang. Ein bioinformatisches Referenzgenom hingegen beinhaltet nur einen dieser beiden Stränge, der definitionsgemäß als Forward-Strang bezeichnet wird. Reads, die vom gleichen Strang wie das Referenzgenom stammen, lassen sich direkt kartieren. Sequenzfragmente des zum Referenzgenom komplementären Strangs müssen in ihre revers-komplementäre Sequenz über-führt werden, um sie dem Referenzgenom zuordnen zu können. Die Information des zugehörigen Strangs wird in der Ausgabedatei des Mappings festgehalten.

Für das Auffinden von kurzen Reads mit gleichen beziehungsweise geringfügig abweichenden Sequenzen im Referenzgenom stehen bereits verschiedene effiziente Algorithmen und Strategien zur Verfügung. Eine dieser Strategien ist die Burrows-Wheeler-Transformation [Burrows und Wheeler, 1994] in Verbindung mit Suffix-Arrays, auf die Mapping-Algorithmen wie Bowtie [Langmead *et al.*, 2009], BWA [Li und Durbin, 2009] oder SOAP2 [Li *et al.*, 2009b] zurückgreifen (siehe Anhang IV.VI).

Der nächste Schritt, das sogenannte „Variant Calling“, dient dem Auffinden von Sequenzunterschieden zwischen der untersuchten DNA und der Sequenz des Referenzgenoms. Um durch die NGS-Technologie entstandene Sequenzierfehler von echten Varianten unterscheiden zu können, ist es notwendig jeden DNA-Abschnitt beziehungsweise jede Base mehrfach zu sequenzieren. Erst ab einer 20-fachen Abdeckung („Coverage“) einer Base lassen sich die allermeisten Sequenzier- oder Mappingfehler als solche erkennen [Nielsen *et al.*, 2011]. Im Allgemeinen wird auch die Bestimmung der Zygotie (Heterozygotie oder Homozygotie) einer Variante, das sogenannte „Genotype Calling“, in den Begriff des Variant Calling eingeschlossen. Zur Detektion einer Variante müssen die im Mapping-Schritt gefundenen Diskrepanzen zwischen dem Referenzgenom und den sequenzierten Reads betrachtet werden. Tritt eine Abweichung vom Referenzgenom in mehreren Reads an der gleichen Position auf, so lässt sich ein Sequenzierfehler ausschließen. Hierbei kann auch der Qualitätswert einer Base, zum Beispiel Q20 (QPhred=20) (siehe Kapitel II.II, als unterer Schwellenwert für echte Sequenzvarianten berücksichtigt werden, bevor diese in das anschließende Genotyp Calling einbezogen werden.

Eine einfache Methode zur Genotypisierung einer Variante ist die Analyse der absoluten Häufigkeiten der vier Basen A, T, C und G in den Reads an der variierten Position. Dabei dienen feste Grenzwerte zur Einteilung der Variante als homozygot oder heterozygot. Weist eine nicht der Referenz entsprechende Base zum Beispiel eine relative Häufigkeit von 20% bis 80% auf, so wird sie als heterozygote und andernfalls als homozygote Variante eingestuft. Diese Standardmethode arbeitet zuverlässig bei einer hohen Coverage und wird von kommerziellen Softwares, wie der CLC Genomics Workbench [QIAGEN, 2020] verwendet [Nielsen *et al.*, 2011].

In Sequenzierprojekten mit niedriger Coverage (<20x) führt die Bestimmung von Genotypen mit festen Schwellenwerten zur Fehlinterpretation von Varianten. Wahrscheinlichkeitsbasierte Methoden können in diesem Fall eine erhöhte Genauigkeit der Genotypisierung erzielen. Eine Genotyp-Wahrscheinlichkeit $p(X|G)$ für einen Genotypen G lässt sich mit der Bayes'schen Formel bestimmen. X beschreibt hier alle Basen, die für ein bestimmtes Individuum an einer bestimmten Position im Referenzgenom mappen. Der Genotyp mit der höchsten Wahrscheinlichkeit $p(X|G)$ wird gewählt. Das Verhältnis zwischen der höchsten und der zweithöchsten Wahrscheinlichkeit für den Genotyp einer bestimmten Variante kann als Vertrauensmaß dienen. Diese statistischen Mo-

delle lassen sich durch Einbeziehung weiterer Informationen wie zum Beispiel dem Quality Score der variierten Basen verbessern und finden zum Beispiel im GATK HaplotypeCaller Algorithmus (siehe Anhang IV.VIII.III) Anwendung.

Durch Abgleich der gefundenen Varianten mit Datenbanken wie der gnomAD-Datenbank (siehe Anhang III.VI) lässt sich rückschließen, ob es sich um eine in der Population gehäuft auftretende und daher wahrscheinlich nicht krankheitsverursachende Variante (Polymorphismus) handelt. Datenbanken wie zum Beispiel die Human Gene Mutation Database (HGMD) (siehe Anhang III.I) und Online Mendelian Inheritance in Man (siehe Anhang III.IV) können zur Identifizierung bereits bekannter pathologischer Veränderungen des Erbguts herangezogen werden.

3.4 NGS-Analyse Pipelines

Die einzelnen Softwareprogramme zum Kartieren, Auffinden von Varianten und zum Datenbankabgleich werden meist zu einer Daten-Analyse-Pipeline zusammengefügt und automatisiert ausgeführt. Das MiSeq-Sequenziersystem der Firma Illumina verfügt zum Beispiel über die MiSeq-Reporter-Software. Mit deren Hilfe ist es möglich, anwendungsspezifische Datenanalysen direkt nach einem Sequenzierungslauf durchführen zu lassen. So wird bei einem Resequenzierungsprojekt die Software BWA für das Mapping und das GATK zur Detektion von Varianten eingesetzt [Illumina, 2020a]. Der NextSeq500 bietet eine solche Möglichkeit der direkten Datenverarbeitung auf dem Sequenziergerät nicht an.

Zur Auswertung von Sequenzdaten können webbasierte Pipeline-Softwares, wie Galaxy [Afgan *et al.*, 2018] und WEP [D'Antonio *et al.*, 2013] herangezogen werden. Galaxy verfügt über eine Sammlung an bioinformatischen Tools, die vom Benutzer zu einem für eine Analyse spezifischen Workflow zusammengefügt werden können. WEP hingegen kombiniert eine feste Abfolge an ausgewählter Software zur Analyse von Exom-Datensätzen. Neben diesen vorgefertigten Pipeline-Tools besteht die Möglichkeit eine Pipeline aus eigens ausgewählten Software-Paketen zu erstellen. Diese Pipelines können flexibel gestaltet werden und besitzen den Vorteil alle Software-Parameter sowie Datenbankabgleiche genau definieren zu können. Zudem muss bei Eigenimplementationen von Pipelines und deren Ausführung auf lokalen Servern nicht auf ausreichend Rechenkapazität gewartet werden. Dies kann die Analyse der Daten beschleunigen. Eine weitere Möglichkeit einer performanteren Datenverarbeitung bietet die Illumina DRAGEN (Dynamic Read Analysis for GENomics) Bio-IT Plattform, welche die FPGA (engl. Field-Programmable Gate Array)-Technologie mit speziell implementierten Algorithmen verwendet, um einen gesamtgenomischen Datensatz in etwa 25 Minuten zu prozessieren [Illumina, 2019b]. Da nur speziell optimierte Algorithmen verwendet werden können, ist die Flexibilität bezüglich der Softwareauswahl und der individuellen Anpassung auf dieser Plattform reduziert.

Die Anforderungen an eine solche Auswertungspipeline für NGS Daten sind vielfältig. So sollen innerhalb kürzester Zeit Ergebnisse mit höchster Spezifität und Sensitivität vorliegen. Dies gelingt nur, wenn die Hardwareressourcen effizient ausgelastet und die verwendeten Softwarepakete stetig auf den aktuellen Stand gebracht werden. Um keine wertvolle Zeit zu verlieren ist es wünschenswert, die Analyse nach Abschluss des Sequenziervorgangs automatisiert zu starten. Es ist anzustreben alle Patienten parallel zu prozessieren, damit eine anschließende gleichzeitige Auswertung durch mehrere wissenschaftliche Mitarbeiter gewährleistet ist. Um den Fortschritt des Analyseprozesses verfolgen zu können, ist es sinnvoll, patientenbezogene Statusmeldungen in einer für Mitarbeiter zugänglichen Datenbank abzulegen. Neben diesen sollten auch alle gefundenen Varianten mit Bezug zum jeweiligen Patienten in dieser Datenbank gespeichert werden, um sie später für eine Kohortenanalyse und den Ausschluss von verfahrensspezifischen Sequenzartefakten nutzen zu können. Zur Erleichterung der Auswertung und Interpretation der Varianten ist eine reichhaltige Annotation mit stets aktuellen öffentlichen Datenbanken wie der dbSNP, OMIM und GnomAD sowie diverser *in silico* Pathogenitätsvorhersagetools wie CADD, MutationTaster und Polyphen2 von immenser Bedeutung. Eine maschinelle Filterung der Daten ermöglicht die Erstellung einer Liste von ca. 50 der vermutlich schädlichsten Varianten des Patienten, um die Suche nach der krankheitsverursachenden Mutation zu erleichtern und zu beschleunigen. Neben der Analyse von Nukleotidaustauschen und kleiner Insertionen und Deletionen sollen ebenfalls Kopienzahlveränderungen (engl. Copy Number Variations, CNVs) und große homozygote Regionen (engl. Regions of Homozygosity, ROH) detektiert werden. Die Bestätigung der gefundenen Kandidatenvarianten zu Segregations- und Validierungszwecken erfolgt mittels PCR und Sanger-Sequenzierung (siehe Anhang V.I). Ein automatisches Primerdesign für die „nasse“ Validierung dieser Varianten wäre hilfreich und zeitsparend. Nach Terminierung der Auswertepipeline müssen alle notwendigen Daten auf ein für Mitarbeiter zugängliches Netzlaufwerk abgelegt werden.

Diese bioinformatischen Pipelines produzieren, im Falle von WES und WGS, sehr große und unübersichtliche Textdateien als Ausgabe. So besteht die Notwendigkeit der Implementierung einer Benutzeroberfläche (engl. Graphical User Interface, GUI), die die resultierenden Ergebnisse wie die patientenbezogenen Varianten für die biologische und medizinische Auswertung visualisiert. Die in Zusammenarbeit mit Biologen und Ärzten erarbeiteten Anforderungen an diese GUI werden im Folgenden erläutert.

Im Allgemeinen soll die Oberfläche dem Benutzer die Möglichkeit geben, die Ergebnisse der Pipeline zu visualisieren. Es wird angestrebt, die Varianten eines Patienten mit deren Annotationen tabellarisch anzuzeigen. Dabei kann der Benutzer Spalten ein- und ausblenden sowie verschieben. Diverse Filter sollen es dem Nutzer erlauben die Varianten auf einen kleinen Datensatz mit möglichst pathologischen Varianten zu reduzieren. Hier ist das Filtern nach Allelfrequenzen aus dem gnomAD Browser, dem Grad der phylogenetischen Konservation, der Auswirkung auf Basen- sowie Proteinebene und bereits bekannten Asso-

ziationen mit Krankheiten in Datenbanken hilfreich. Zudem soll der Benutzer die Möglichkeit haben, die Variante auf Basenebene in den alignierten Rohdaten mit Hilfe eines Alignment-Viewers zu betrachten. Ebenso ist es erwünscht, Informationen über den Phänotyp des Patienten sowie über Krankheiten, die mit dem in der Tabelle ausgewählten Gen assoziiert werden, dem Nutzer direkt anzuzeigen. Eine Verlinkung von einzelnen Annotationsdaten, wie zum Beispiel dem dbSNP-Identifier, dem OMIM-Eintrag oder dem HGMD-Identifier zur zugehörigen Online-Datenbank, sowie eine graphische Darstellung der Ergebnisse von *in silico* Pathogenitätswerten (z.B.: in einem Spiderplot) ist erstrebenswert.

Kapitel 4

Aufbau einer Infrastruktur zur Auswertung von NGS-Daten

Mit der Anschaffung eines MiSeq-Sequenziergeräts der Firma Illumina im Jahr 2013 stieg das Institut für Humangenetik der Universitätsmedizin Mainz in die Next-Generation-Sequenzierung ein. In diesem Zusammenhang wurde der am Max-Planck-Institut für molekulare Genetik in Berlin entwickelte und in Kapitel 3.1 eingeführte TES Panel namens „MPIMG1-Test“ etabliert. Durch den Umstieg auf das NextSeq500 Sequenziergerät der Firma Illumina im Jahr 2016, welches mit bis zu 120 Gbp im Vergleich zum MiSeq mit 15 Gbp eine deutlich höhere Sequenzierleistung pro Lauf aufweist, war es ab diesem Zeitpunkt möglich, WES Analysen am Institut für Humangenetik durchzuführen. Da eine gesamtgenomische Sequenzierung mit dem NextSeq500 aus ökonomischer Sicht nicht sinnvoll ist, wurden zwischen 2018 und 2020 einzelne besonders betroffene Patienten auf Forschungsebene durch verschiedene Sequenzierungsdienstleister wie zum Beispiel dem Max-Delbrück-Centrum (MDC) in Berlin und dem Genomik Institut in Beijing (engl. Beijing Genomics Institute, BGI) einer WGS unterzogen.

Im Folgenden wird der Vollständigkeit halber kurz auf die Datenauswertungspipeline für den MPIMG1-Test eingegangen, da der Aufbau und die Etablierung dieser bereits ausführlich in meiner am Institut für Humangenetik der Universitätsmedizin-Mainz angefertigten Masterarbeit beschrieben wurde. Im Fokus steht in diesem Kapitel die Realisierung der in Kapitel 3.4 skizzierten automatisierten Analysepipeline für NGS-Daten und die Entwicklung einer zugehörigen benutzerfreundlichen graphischen Oberfläche zur Auswertung der Ergebnisse. Des Weiteren wird die Skalierbarkeit der Pipeline auf die Auswertung von WGS Daten dargelegt und die Ausarbeitung einer bioinformatischen Methode zur Pathogenitätsabschätzung von Varianten in miRNAs und regulatorischen Bereichen beschrieben.

4.1 Analysepipeline für den MPIMG1-Test

Der MPIMG1-Test stellt eine um viele Gene erweiterte Version des in den USA als Carrierscreening etablierten Kingsmore-Tests [Kingsmore, 2012] dar

und umfasst 1222 mit seltenen, pädiatrischen Erkrankungen, geistiger Behinderung sowie verwandten Störungen assoziierte Gene. Er wurde insbesondere zur Diagnose bei Kindern mit unklarem Krankheitsbild eingesetzt. Die parallele Sequenzierung von 12 Patienten erfolgte in zwei konsekutiven Paired-End Sequenzierläufen der gleichen Library auf einem Illumina MiSeq System mit einer Leseweite von jeweils 250 Basenpaaren. Die aus der zweifachen Sequenzierung resultierenden Datensätze für Forward und Reverse Reads wurden jeweils pro Patient kombiniert und anschließend mit Hilfe der früher beschriebenen Auswertungspipeline unter Verwendung des SOAP2-Aligners [Li *et al.*, 2009b] und des MERAP Softwarepakets [Hu *et al.*, 2014] (siehe Anhang IV.IV) mit den in Tabelle 4.1 angegebenen Parametern analysiert.

Tabelle 4.1: `MPIMG-1 Pipeline` *Parameter*

Parameter	**Wert**	**Bedeutung**
`Skip Alignment`	no	Die Pipeline wird mit Mapping-Schritt ausgeführt
`Bulid New Index`	no	Es wird ein bereits vorhandener Index verwendet
`Minimal Depth for an Indel`	10	Mind. 10 Reads müssen die Insertion/ Deletion an dieser Position tragen
`Minimal Percentage for an Indel`	0,3	Die Insertion/Deletion muss in mind. 30% der Reads vorkommen, die diese Position abdecken
`Minimal Depth of a Variant`	10	Mind. 10 Reads müssen die Variante an dieser Position tragen
`Minimal Percentage of a Variant`	0,3	Die varrierte Base muss in mind. 30% der Reads vorkommen, die diese Position abdecken

4.2 Aufbau einer Pipeline für WES Daten

Die Einführung der Whole-Exome-Sequenzierung im Institut für Humangenetik der Unimedizin Mainz und das stetig steigende Probenaufkommen fordern den Aufbau einer automatisierten Datenauswertungspipeline. Zur Speicherung und Verarbeitung der riesigen Datenmengen wurden zwei neue Server erworben und entsprechend konfiguriert (siehe Anhang I). Der NextSeq500 schreibt die während der Sequenzierung entstehenden Ergebnisse direkt auf das Netzlaufwerk `Rohdaten` des Primärservers. Sobald ein Lauf abgeschlossen ist, soll die Verarbeitung der Daten beginnen. Dazu prüft ein „Watcher“-Skript regelmäßig, ob es auf dem Laufwerk `Rohdaten` einen abgeschlossenen und noch nicht prozessierten Lauf gibt. Findet es einen solchen, konvertiert es die Bilddaten des Sequenzierers zu FASTQ-Dateien, legt diese auf dem `Berechnungen`

Laufwerk ab und ruft das „Organizer“-Skript auf. Dieses liest das beigefügte Samplesheet ein, erstellt die zur Auswertung benötigten Ordnerstrukturen und ruft das „Pipeline“-Skript mit den entsprechenden Parametern für jede Probe auf. Ist die Datenanalyse einer Probe abgeschlossen, werden die Ergebnisse auf das `Ergebnisse` Laufwerk verschoben (siehe Abbildung 4.1). Zur Veranschaulichung der Prozesse wird im Folgenden der fiktive Patient mit der Journalnummer 0001/01 exemplarisch einer Exomsequenzierung unterzogen.

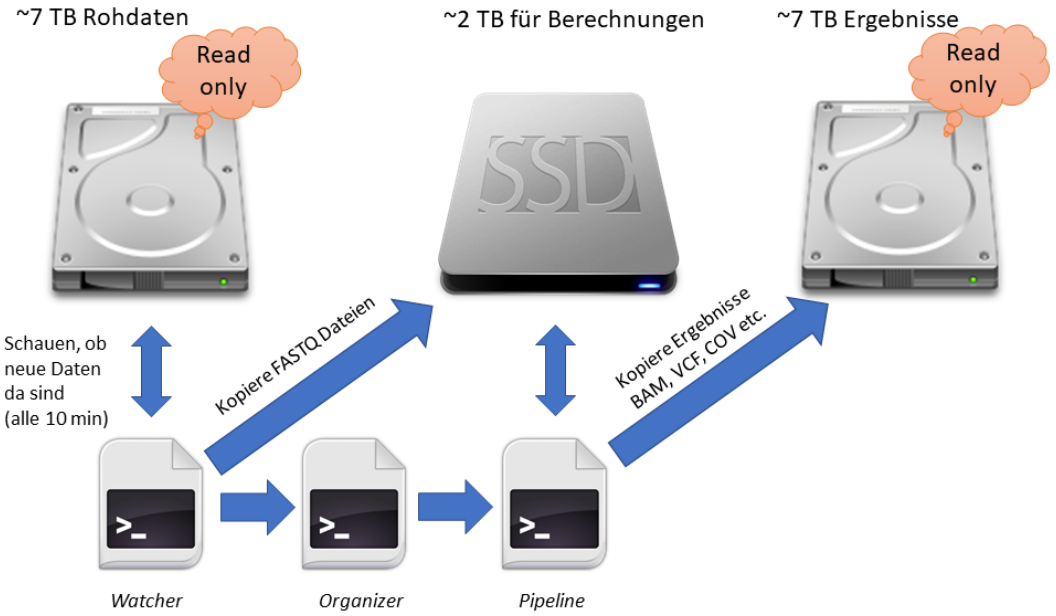


Abbildung 4.1: *Schematischer Aufbau der WES Pipeline.*
Das „Watcher“-Skript sucht auf dem `Rohdaten`*-Laufwerk regelmäßig nach abgeschlossenen Sequenzierläufen. Es konvertiert die Bilddaten zu FASTQ Dateien, legt diese auf dem* `Berechnungen`*-Laufwerk ab und ruft das „Organizer“-Skript auf, welches das Samplesheet einliest, die benötigte Ordnerstruktur erstellt und für jede Probe das „Pipeline“-Skript mit entsprechenden Parametern startet. Nach Abschluss der Datenanalyse werden die Ergebnisse auf das* `Ergebnisse`*-Laufwerk verschoben.*

4.2.1 Watcher-Skript

Das „Watcher“-Skript ist in der Skriptsprache Perl (siehe Anhang IV.I) implementiert und trägt den Namen `MaiWatcher.pl`. Zu Beginn der Ausführung liest es die zugehörige Konfigurationsdatei `MaiWatcher.config` ein. Diese Datei beinhaltet Informationen über den Ort des Rohdatenverzeichnis (`RAW_DATA`), den Pfad zum Verzeichnis in dem die Berechnungen später stattfinden sollen (`CALCULATE_DATA`), die bei einem Sequenzierungslauf vom Gerät als letztes geschriebene Datei (`LAST_WRITTEN_FILE_NEXTSEQ`) und den Pfad zum Ordner in dem das Samplesheet für den entsprechenden Lauf abgelegt ist (`SAMPLESHEETS`) als tabulatorgetrennte Schlüssel-Wert-Paare. Kommentare können in eine durch

```
##Version 1.2
##Release: 30.09.2016
########################
#Path to the folder where the sequencer writes the raw files to
RAW_DATA       /media/Rohdaten
#Path to the folder where the calculation should be done
CALCULATE_DATA          /media/Berechnungen
#Last file which is written from NextSeq
LAST_WRITTEN_FILE_NEXTSEQ RTAComplete.txt
#Path to SampleSheet folder
SAMPLESHEETS /media/Rohdaten/Samplesheets
```

Abbildung 4.2: *Beispiel einer MaiWatcher.config Datei.*
Sie beinhaltet Informationen über den Ort des Rohdatenverzeichnis, des Verzeichnisses für die Berechnungen, die vom Gerät als letztes geschriebene Datei und den Pfad zum Samplesheet-Verzeichnis als tabulatorgetrennte Schlüssel-Wert-Paare. Kommentare werden durch `#` *eingeleitet.*

`#` eingeleitete Zeile geschrieben werden (siehe Abbildung 4.2). Im nächsten Schritt durchläuft der Algorithmus alle Unterordner des Rohdatenverzeichnisses und schaut, ob einer der Ordner (Runfolder) einen fertigen Sequenzierungslauf enthält. Dazu wird überprüft, ob der Sequenzierer die Datei `RTAComplete.txt` bereits abgelegt hat. Ist dies der Fall, schreibt das Programm die Statusdatei `MAIWATCHER_START` in diesen Ordner. Über das Auslesen der FlowCell-Seriennummer aus der Datei `RunParameters.xml` kann das zum Lauf gehörige Samplesheet im in der Konfigurationsdatei angegebenen Verzeichnis der Samplesheets gefunden und in den Runfolder verschoben werden. Nach Anlegen eines gleichnamigen Ordners im Berechnungsverzeichnis (Arbeitsverzeichnis) startet die Konvertierung und das Demultiplexing der Rohdaten (BCL-Dateien) zu FASTQ-Dateien (siehe Anhang II.II) für jeden Patienten mit der durch Illumina bereitgestellten bcl2fastq Software (siehe Anhang IV.V) unter Angabe des Rohdatenverzeichnisses (`--runfolder-dir`), des Arbeitsverzeichnisses (`--output-dir`) und des Parameters `--no-lane-splitting`. Dieser verhindert die getrennte Ausgabe der Sequenzen nach FlowCell Lanes und lässt nur jeweils eine FASTQ-Datei pro Patient für die FWD und REV Sequenzen entstehen. Ist die Konvertierung abgeschlossen, kopiert `MaiWatcher.pl` das Samplesheet in das Ausgabeverzeichnis und startet das „Organizer“-Skript (`MaiOrganizer.pl`) unter Angabe des Pfads zum Samplesheet (`-sample_sheet`) und des Pfads zum entsprechenden Arbeitsverzeichnisses (`-run_folder`), welches nun die FASTQ-Dateien für jeden Patienten beinhaltet. Zuletzt schreibt das Skript eine weitere Statusdatei, `MAIWATCHER_END`, welche den Abschluss der Datenkonvertierung signalisiert.

Durch die beiden Statusdateien `MAIWATCHER_START` und `MAIWATCHER_END` erkennt der Algorithmus beim nächsten Traversieren des Rohdatenverzeichnisses, dass der entsprechende Lauf nicht nur fertig sequenziert, sondern die Prozessierung bereits begonnen bzw. abgeschlossen ist. Somit muss dieser Runfolder bzw. Sequenzierungslauf nicht mehr weiter beachtet werden.

Zur automatischen Suche nach fertigen und unprozessierten Läufen ist das `MaiWatcher.pl`-Skript in regelmäßigen Abständen auszuführen, was sich in Linux-Systemen durch Cron-Jobs realisieren lässt. Durch Eintragen einer entsprechenden Zeile (Zeile 18) in die CronTab Datei unter `/etc/crontab`, wird das Skript alle 10 Minuten nach Wechseln ins entsprechende Verzeichnis (`cd /Maipipe`) gestartet (`perl MaiWatcher.pl`) (siehe Abbildung 4.3).

```
1  # /etc/crontab: system-wide crontab
2  # Unlike any other crontab you don't have to run the `crontab'
3  # command to install the new version when you edit this file
4  # and files in /etc/cron.d. These files also have username fields,
5  # that none of the other crontabs do.
6
7  SHELL=/bin/sh
8  PATH=/usr/local/sbin:/usr/local/bin:/sbin:/bin:/usr/sbin:/usr/bin:/BioinfSoftware
9  MODULEPATH=/etc/environment-modules/modules:/usr/share/modules/versions:/usr/Modules/$MODULE_VERSION/
10 modulefiles:/usr/share/modules/modulefiles:/BioinfSoftware/modulefiles/Linux:/BioinfSoftware
11 /modulefiles/Core:/BioinfSoftware/lmod/lmod/modulefiles/Core
12
13 # m h dom mon dow user  command
14 17 *    * * *   root    cd / && run-parts --report /etc/cron.hourly
15 25 6    * * *   root    test -x /usr/sbin/anacron || ( cd / && run-parts --report /etc/cron.daily )
16 47 6    * * 7   root    test -x /usr/sbin/anacron || ( cd / && run-parts --report /etc/cron.weekly )
17 52 6    1 * *   root    test -x /usr/sbin/anacron || ( cd / && run-parts --report /etc/cron.monthly )
18 */10 *  * * *   die9s   umask 002;cd /Maipipe && perl MaiWatcher.pl
```

Abbildung 4.3: *Eintrag in der CronTab Datei.*
Zeile 19 zeigt den Cronjob, welcher das `MaiWatcher.pl`*-Skript alle 10 Minuten im Verzeichnis* `/Maipipe` *ausführt.*

4.2.2 Organizer-Skript

Das „Organizer“-Skript liest zu Beginn ebenfalls eine Konfigurationsdatei namens `MaiOrganizer.config` ein, welche den gleichen Aufbau wie die `MaiWatcher.config` Datei hat. In ihr ist der Pfad zum Referenzgenom (`genome`), der Pfad zum erstellten BWA-Index des Referenzgenoms (`BWA_index`), der Pfad zu Referenzdateien des GATK (`resource_dir`) sowie der Pfad zum Ordner in dem sich die Dateien mit Informationen über die angereicherten Regionen der unterschiedlichen Experimente befindet (`intervals_dir`), angegeben. Zudem lässt sich über `padding` eine beliebige Anzahl an Basen definieren, mit der die Intervalle up- und downstream zu erweitern sind. Durch `parallel` und `cores` wird bestimmt, wie viele Patientendatensätze parallel auf je welcher Anzahl Prozessorkernen analysiert werden. `STORE_DATA` gibt den Pfad zum Verzeichnis für die Ergebnisse an (siehe Abbildung 4.4). Durch Auswertung der `RunParameters.xml`-Datei und des Samplesheets können Informationen über den Lauf gewonnen und in der hausinternen Labormanagement Datenbank abgespeichert werden. Zu diesen Informationen gehört die Bezeichnung des Sequenzierers, die Geräte-ID, die fortlaufende Experiment-ID des Geräts, die für das Experiment verantwortliche Person, das Datum des Sequenzierlaufs, die Anzahl der sequenzierten Basen sowie die LOT Nummern der Kartusche und der FlowCell. Tabelle 4.2 zeigt einen Datenbankeintrag für die Tabelle `ngs.runs` für einen typischen Exomsequenzierungslauf auf dem NextSeq500 des Instituts für Humangenetik in Mainz.

Die im Daten-Bereich des Samplesheets befindlichen Informationen bezüglich der Probennummern, den verwendeten Indizes, der durchgeführten Untersu-

```
##Version 1.2
##Release: 30.09.2016
#################################
# Path to the reference genome in FASTA format (GATK style)
genome              /media/Berechnungen/Referenzgenom/HG19/HG19.karyo.fasta
# Path to the GATK index
BWA_index       /media/Berechnungen/Referenzgenom/HG19/BWA_index/HG19.karyo.fasta
# Path to the folder with all the neccessary database files (e.g. /Resources)
resource_dir     /media/Berechnungen/AnnotationDBs/Resources
# List of Intervals in GATK-Format (chr:start-end)
intervals_dir      /media/Berechnungen/AnnotationDBs/ROI
# Padding for intervals
padding              250
# Number of samples to process in parallel
parallel 8
# Number of cores per sample
cores     4
# Path to folder where result data should be stored
STORE_DATA    /media/Ergebnisse
```

Abbildung 4.4: *Beispiel einer MaiOrganizer.config Datei.*
Sie enthält den Pfad zum Referenzgenom, zum BWA-Index, zum Ergebnis Verzeichnis, zum Ordner, welcher Dateien zur Ausführung des GATK beinhaltet sowie den Pfad zur Intervall-Datei. Zudem kann angegeben werden, um wie viele Basen die jeweiligen Intervalle zu erweitern sind und wie viele Proben parallel auf welcher Anzahl an CPU Kernen prozessiert werden sollen. Die Parameter liegen als tabulator-getrennte Schlüssel-Wert-Paare vor. Kommentare werden durch `#` *eingeleitet.*

chung und des genutzten Anreicherungs-Kits werden mit Zuordnung zu einem Lauf in der Datenbank des Labormanagementsystems in der Tabelle `ngs.auf_lauf` abgelegt. Tabelle 4.3 zeigt exemplarisch den Eintrag für den Beispielpatienten 0001/01, welcher mit dem Agilent Human all Exon V7 Kit angereichert wurde und dessen Sequenzfragmente durch die Basenabfolge der Indizes Agi_1 und Agi_2 eindeutig identifizierbar sind.

Gleichzeitig wird mit Hilfe der zuvor gewonnenen Daten aus der Journalnummer, der durchzuführenden Untersuchung, der Geräte-ID und der Experiment-ID für jeden Patienten ein eindeutiger Verzeichnisname im Format `Journalnummer_Untersuchung_GeräteID_ExperimentID` generiert. Dieser Verzeichnisname lautet für den Patienten 0001/01 `0001-01_Exom_NB501654_0172`. Da das „/“-Zeichen in einem Dateipfad einen Unterordner einleitet, wird dieses in der Journalnummer der Patienten durch ein „-“ ersetzt. Zudem kann der Verzeichnisname Leerzeichen und Kommata enthalten. Einige Programme sind allerdings nicht in der Lage mit Kommata und Leerzeichen in Dateipfaden umzugehen, weshalb bei der Namensvergabe für den Patientenordner im Arbeitsverzeichnis Leerzeichen durch „=“ und Kommata durch „.“ ersetzt werden. Für jeden Patienten erfolgt die Anlage eines solchen Ordners im Arbeitsverzeichnis neben den FASTQ-Dateien. Bei der Erstellung des Patientenordners im Ergebnisverzeichnis werden diese Korrekturen nicht vorgenommen.

Mit Hilfe des Perl Moduls `File::Find::Rule` von Richard Clamp aus dem CPAN[1] werden die zum entsprechenden Patienten gehörigen FASTQ-Dateien

[1] `https://metacpan.org/pod/File::Find::Rule`

Tabelle 4.2: *Datenbankeintrag Sequenzierungslauf*

Spalte	**Wert**
ID *(fortlaufend)*	353
Gerät	NextSeq
Geräte ID	NB501654
Experiment-ID	172
Investigator	Forscher X
Experiment-Name	Exom Run 72
Datum	02.04.2020
Zyklen	152
LOT Kartusche	20421603
LOT FlowCell	20407922

Tabelle 4.3: *Datenbankeintrag Patienten auf Lauf*

Spalte	**Wert**
ID *(fortlaufend)*	3702
Journalnummer	0001/01
Run ID	353
Untersuchung	Exom
Index 1 ID	Agi_1
Index 1 NT	TAAGGCGA
Index 2 ID	Agi_14
Index 2 NT	ATAGAGAG
Description	Exom
Kit	Agilent Human All Exon V7

im Arbeitsverzeichnis ausfindig gemacht, der Befehl zum Aufruf der im Kapitel 4.2.3 beschriebenen Analysepipeline `MaiPipeline4.0.pl` mit den entsprechenden Parametern erstellt und in einer Arraystruktur gespeichert. Für den fiktiven Patienten 0001/01 ergibt sich daraus der folgende Befehl zum Start der Analyse:

```
perl MaiPipeline4.0.pl
   -fq1 /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_R1_S1.fastq.gz
   -fq2 /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_R2_S1.fastq.gz
   -sample_name 0001-01
   -output /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172
   -genome /media/Berechnungen/Referenzgenom/HG19/
      HG19.karyo.fasta
   -BWA_index /media/Berechnungen/Referenzgenom/HG19/BWA_index/
```

```
   HG19.karyo.fasta
-resource_dir /media/Berechnungen/AnnotationDBs/Resources
-intervals /media/Berechnungen/AnnotationDBs/ROI/
   Exom.intervals
-padding 250
-cov_file /media/Berechnungen/AnnotationDBs/ROI/Exom.cov.bed
-t 4
```

Neben dem Kommando zum Starten der Analysepipeline wird ein Befehl für das Kopieren der resultierenden Daten aus dem Patientenordner des Arbeitsverzeichnisses in den Patientenordner des Ergebnisvereichnisses kreiert und in einer weiteren Arraystruktur abgelegt.

Anschließend durchläuft der Algorithmus die Arraystruktur und startet die pro Patient erstellten Befehle mit Hilfe des `Parallel::ForkManager`-Moduls[2] von Yanick Champoux nebeneinander. Die Anzahl der parallel zu analysierenden Datensätze ist durch den Wert für `parallel` in der Konfigurationsdatei, in diesem Fall acht, festgelegt. Befinden sich mehr als acht Elemente und damit Befehlsaufrufe in der Arraystruktur, so wird zum Beispiel der neunte Befehl erst ausgeführt, sobald ein anderer abgeschlossen ist. Nach Abschluss eines Pipelinebefehls verschiebt der Kopierbefehl die Daten in den entsprechenden Patientenordner des Ergebnisverzeichnisses.

Sind alle Datensätze verarbeitet und die Ergebnisse in den entsprechenden Ordnern auf dem Ergebnisverzeichnis kopiert, so wird das Arbeitsverzeichnis auf der Berechnungspartition gelöscht.

4.2.3 Pipeline-Skript

Im „Pipeline“-Skript `MaiPipeline4.0.pl` werden die NGS Datensätze durch eine Verkettung von verschiedenen, in den folgenden Unterkapiteln aufgeführten Programmen ausgewertet. Um dies vollautomatisch durchführen zu können, ist es obligatorisch, die in Tabelle 4.4 beschriebenen Parameter beim Skriptaufruf anzugeben.

Vor dem Alignment durchlaufen die Daten des Patienten eine Qualitätskontrolle. In den folgenden Schritten werden duplizierte Reads entfernt und die Basenqualitätswerte neu kalibriert. Im Abschnitt zum Auffinden der Varianten schließen sich die CNV-Analyse, die Berechnung der Coverage, das Auffinden von ROHs und das Variant Calling an. Anschließend werden die resultierenden Varianten und CNVs mittels Intervar bezüglich ihrer Pathogenität bewertet, annotiert und gefiltert. Zur Vermeidung von Probenvertauschungen erfolgt ein Abgleich des aus den Daten ermittelten Geschlechts mit dem in der Datenbank festgehaltenen Geschlecht sowie der Vergleich mit 22 mittels Pyrosequenzierung genotypisierter SNPs. Alle Varianten werden in der hausinternen Datenbank abgelegt. Um die Validierung und Segregationsanalyse einer

[2] `https://metacpan.org/pod/Parallel::ForkManager`

Tabelle 4.4: `MaiPipeline4.0.pl` *Parameter*

Parameter	Erläuterung
`-fq1`	Pfad zur FASTQ Datei mit FWD Reads
`-fq2`	Pfad zur FASTQ Datei mit REV Reads
`-sample_name`	Journalnummer der Probe
`-output`	Pfad zum Patientenordner im Arbeitsverzeichnis
`-genome`	Pfad zum Referenzgenom
`-BWA_index`	Pfad zum BWA Index des Referenzgenoms
`-resource_dir`	Pfad zum Verzeichnis mit Daten für das GATK
`-intervals`	Pfad zur Datei mit Informationen zu angereicherten Regionen für diese Untersuchung
`-padding`	Anzahl der Basen zur Erweiterung der Intervale
`-cov_file`	Pfad zur Datei mit Informationen zur Berechnung der Coverage
`-t`	Anzahl der zu verwendenden Prozessorkerne

möglicherweise pathologischen Variante mittels Sanger-Sequenzierung effizient und schnell durchführen zu können, wird für jede Variante automatisch ein Primerpaar designed. Abbildung 4.5 zeigt diesen Workflow, dessen einzelne Schritte und Programmaufrufe mit den entsprechenden Parametern anhand des Patienten 0001/01 näher erläutert werden.

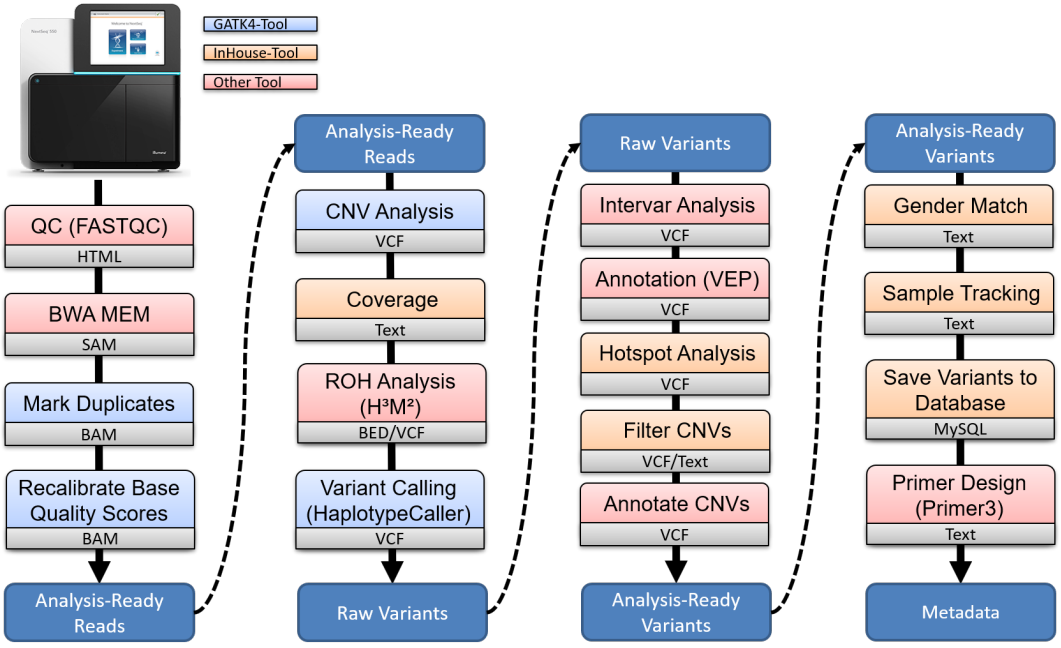


Abbildung 4.5: *Schematische Darstellung des Pipeline Workflows.*

4.2.3.1 Qualitätskontrolle

Zur Qualitätskontrolle der Rohdaten findet das in Anhang IV.VII beschriebene Programm `fastqc` Anwendung. Es liest die patientenspezifischen FASTQ-

Dateien ein und erstellt eine HTML-Datei pro FASTQ-Datei im angegebenen Patientenordner. Durch Verteilung der Berechnungen auf zwei CPU Kerne (`-t 2`) werden beide FASTQ-Dateien gleichzeitig analysiert. Der entsprechende Programmaufruf lautet:

```
fastqc
   -t 2
   -outdir /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172
   /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_R1_S1.fastq.gz
   /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_R2_S1.fastq.gz
```

4.2.3.2 Alignment

Im Alignment-Schritt wird für jedes Sequenzfragment des Patienten eine entsprechende Position im Referenzgenom gesucht. Aus der Anordnung der einzelnen Fragmente, dem Alignment, ist es möglich die Konsensussequenz der Patienten-DNA abzuleiten. Um diese Arbeit verrichten zu können, benötigt `bwa` (siehe Anhang IV.VI) einen Index des Referenzgenoms. Dieser wird durch

```
bwa index
   -p /media/Berechnungen/Referenzgenom/HG19/BWA_Index/HG19.karyo.
      fasta
   /media/Berechnungen/Referenzgenom/HG19/HG19.karyo.fasta
```

erstellt. Der Parameter `-p` beschreibt hierbei den Präfix des BWA-Indexes. Das Alignment wird anschließend, festgelegt durch die Angaben der MaiOrganizer Konfigurationsdatei, auf vier Kernen unter Angabe der Read-Gruppe (`-R`) mit folgenden Parametern ausgeführt:

```
bwa mem
   -R '@RG\tID:WES\tSM:0001-01\tPL:illumina\tLB:Agilent\tPU:
      NB501654'
   -t 4
   /media/Berechnungen/Referenzgenom/HG19/BWA_Index/HG19.karyo.
      fasta
   /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_R1_S1.fastq.gz
   /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_R2_S1.fastq.gz
   > /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.sam
```

4.2.3.3 Entfernen von Duplikaten

Die Entfernung von duplizierten Reads innerhalb des Alignments erfolgt mit dem Tool `MarkDuplicatesSpark` des GATK Pakets (siehe Anhang IV.VIII.I) durch den Befehl:

```
gatk MarkDuplicatesSpark
   -I /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.sam
   -O /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.dedup.bam
   --spark-master local[4]
   --tmp-dir /media/Ergebnisse/picardtmp
   -OBI
```

Dabei liest das Programm eine SAM-Datei ein und erstellt eine sortierte und indizierte (`-OBI`) BAM-Datei als Ausgabe in welcher die duplizierten Reads mit einem Flag markiert sind. Der Befehl `--spark-master local[4]` sorgt für die Ausführung des Kommandos auf vier Prozessorkernen des lokalen Servers.

4.2.3.4 Rekalibrieren der Basenqualitäten

Zur Rekalibrierung der Basenqualitätswerte sind zwei Einzelschritte notwendig (siehe Anhang IV.VIII.II). Das Programm `BaseRecalibrator` erzeugt mit dem Aufruf

```
gatk BaseRecalibrator
   -R /media/Berechnungen/Referenzgenom/HG19/HG19.karyo.fasta
   -I /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.dedup.bam
   -known-sites /media/Berechnungen/AnnotationDBs/Resources/
      All_20180423.vcf
   -known-sites /media/Berechnungen/AnnotationDBs/Resources/
      Mills_and_1000G_gold_standard.indels.hg19.vcf
   -L /media/Berechnungen/AnnotationDBs/ROI/Exom.intervals
   --interval-padding 250
   -O /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.before_recal_data.gpr
```

unter Berücksichtigung der mit `-known-sites` übergebenen polymorphen Stellen eine Ausgabedatei mit Tabellen (`-O`), die im folgenden Schritt vom `ApplyBQSR`-Tool verwendet wird, um die Rekalibrierung durchzuführen. Die Berechnung wird hierbei auf die mit `-L` angegebenen und durch 250 bp up- und downstream erweiterten (`--interval-passing`) ROIs eingeschränkt.

Im zweiten Schritt erfolgt die Übergabe dieser Ausgabedatei über den Parameter `-bqsr` an das `ApplyBQSR`-Programm welches von diesem für die Rekalibrierung der Basenqualitäten verwendet wird. Der Befehl

```
gatk ApplyBQSR
   -R /media/Berechnungen/Referenzgenom/HG19/HG19.karyo.fasta
   -I /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.dedup.bam
   -bqsr /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.before_recal_data.gpr
   -O /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.recal.bam
   --tmp-dir /media/Ergebnisse/picardtmp
   -OBI
```

erzeugt dabei eine indizierte (`-OBI`) BAM-Datei mit rekalibrierten Basenqualitätswerten.

4.2.3.5 Detektion von Kopienzahlveränderungen

Für die Detektion von CNVs ist es notwendig vor der eigentlichen Analyse entsprechende Modelle anzulegen. Im ersten Abschnitt werden daher die Schritte zur Generierung eines solchen Modells für die Tools `DetermineGermlineContigPloidy` und `GermlineCNVCaller` (siehe Anhang IV.VIII.IV) beschrieben. Im Anschluss folgt die Anwendung dieser Modelle auf einzelne Proben im Ablauf der Pipeline.

Erstellung der Modelle Die entsprechenden Modelle werden aus 346 einzelnen Proben berechnet, die im Zeitraum von Juli 2018 bis Januar 2020 einer WES mit dem Agilent Human All Exon V7 Kit unterzogen wurden. Zuerst erfolgt bei der Präprozessierung der Intervall-Datei (`-L`) die Erweiterung der ROIs um die mit dem Parameter `--padding` angegebene Anzahl an Basen up- und downstream. Daraus resultierende überlappende Regionen fasst der Algorithmus zu einer ROI zusammen. Eine Unterteilung der Regionen in sogenannte „bins“ ist für eine WES nicht ratsam und wird durch Angabe von `-bin-length 0` unterbunden. Die aus dem Befehl

```
gatk PreprocessIntervals
   -R /media/Berechnungen/Referenzgenom/HG19/HG19.karyo.fasta
   -L /media/Berechnungen/AnnotationDBs/ROI/Exom.intervals
   --padding 250
   -imr OVERLAPPING_ONLY
   --bin-length 0
   -O /media/Berechnungen/AnnotationDBs/CNV/Exom/
      Exom.preprocessed.interval_list
```

resultierende Datei mit präprozessierten ROIs (`-O`) wird durch das `Annotate-Intervals`-Tool mit Mappability-Informationen und dem GC-Gehalt für jede Region annotiert. Dazu ist die Mappability-Datei (`k100.umap.bed`) vom HofmanLab herunter zu laden. Überlappende Regionen in dieser Mappability-Datei werden mit dem Befehl

```
bedtools merge
   -i /media/Berechnungen/AnnotationDBs/CNV/k100.umap.bed
   > /media/Berechnungen/AnnotationDBs/CNV/k100.umap.merged.bed
```

verbunden. Anschließend erfolgt die Indizierung der Datei `k100.umap.merged.bed` durch

```
gatk IndexFeatureFile
   -I /media/Berechnungen/AnnotationDBs/CNV/k100.umap.merged.bed
```

Die Annotation der präprozessierten Regionen mit Mappability-Informationen (`--mappability-track`) durch den Befehl

```
gatk AnnotateIntervals
   -L /media/Berechnungen/AnnotationDBs/CNV/Exom/
      Exom.preprocessed.interval_list
   -R /media/Berechnungen/Referenzgenom/HG19/HG19.karyo.fasta
   -imr OVERLAPPING_ONLY
   --mappability-track /media/Berechnungen/AnnotationDBs/CNV/
      k100.umap.merged.bed
   -O /media/Berechnungen/AnnotationDBs/CNV/Exom/
      Exom.annotated.tsv
```

erzeugt eine tabulator-getrennte Datei mit den entsprechenden ROIs und deren Annotationen, die später zur Filterung der Regionen verwendet werden kann.

Für alle 346 Probendatensätze werden mit

```
gatk CollectReadCounts
   -L /media/Berechnungen/AnnotationDBs/CNV/Exom/
      Exom.preprocessed.interval_list
   -R /media/Berechnungen/Referenzgenom/HG19/HG19.karyo.fasta
   -imr OVERLAPPING_ONLY
   -I <<Pfad zur *.dedup.bam Datei der jeweiligen Probe>>
   -O /media/Berechnungen/AnnotationDBs/CNV/Exom/HDF5/
      <<ProbenID>>.hdf5
```

die Read Counts für jede ROI (`-L`) ermittelt und in einer HDF5-Datei (`-O`)

gespeichert. Mit Hilfe aller 346 HDF5-Dateien sowie der annotierten Intervalle lassen sich nun schlecht abgedeckte oder schwierig zu sequenzierende ROIs durch

```
gatk FilterIntervals
   -L /media/Berechnungen/AnnotationDBs/CNV/Exom/
      Exom.preprocessed.interval_list
   --annotated-intervals /media/Berechnungen/AnnotationDBs/CNV/
      Exom/Exom.annotated.tsv
   -I <<Pfad zur *.dedup.bam Datei der jeweiligen Probe>>
      <<Probe 1>>.hdf5
   -I <<Pfad zur *.dedup.bam Datei der jeweiligen Probe>>
      <<Probe 2>>.hdf5
   -I <<Pfad zur *.dedup.bam Datei der jeweiligen Probe>>
      <<Probe ...>>.hdf5
   -I <<Pfad zur *.dedup.bam Datei der jeweiligen Probe>>
      <<Probe 346>>.hdf5
   -imr OVERLAPPING_ONLY
   -O /media/Berechnungen/AnnotationDBs/CNV/Exom/
      Exom.filtered.interval_list
```

herausfiltern. Die resultierende Datei `Exom.filtered.interval_list` dient im weiteren Verlauf als ROI Datei (`-L`) für die Etablierung der Modelle. Das Modell für die spätere Fall-Analyse des `DetermineGermlineContigPloidy`-Tools wird mit folgendem Befehl unter Berücksichtigung aller 346 Read Count-Dateien erstellt:

```
gatk DetermineGermlineContigPloidy
   -L /media/Berechnungen/AnnotationDBs/CNV/Exom/
      Exom.filtered.interval_list
   -I /media/Berechnungen/AnnotationDBs/CNV/Exom/Probe1.hdf5
   -I /media/Berechnungen/AnnotationDBs/CNV/Exom/Probe2.hdf5
   -I /media/Berechnungen/AnnotationDBs/CNV/Exom/ProbeX.hdf5
   -I /media/Berechnungen/AnnotationDBs/CNV/Exom/Probe346.hdf5
   -imr OVERLAPPING_ONLY
   --contig-ploidy-priors /media/Berechnungen/AnnotationDBs/CNV/
      contig_ploidy_prior.txt
   --output /media/Berechnungen/AnnotationDBs/CNV/Exom/
   --output-prefix ploidy
```

Es resultieren die zwei Ausgabeordner `ploidy-model` und `ploidy-calls` im Verzeichnis `/media/Berechnungen/AnnotationDBs/CNV/Exom/`. Der erstgenannte Ordner enthält die Daten des später in der Pipeline verwendeten Modells. Der `ploidy-calls` Ordner enthält Unterverzeichnisse für jede Probe. In diesem ist die Ploidie jedes Chromosoms der entsprechenden Probe angegeben. Das Modell für die spätere CNV Analyse auf ROI Ebene wird beruhend

auf den im vorherigen Schritt berechneten Ploidien (`--contig-ploidy-calls`) unter Berücksichtigung aller 346 Read Count-Datensätze (`-I`), der gefilterten ROIs (`-L`) sowie der Annotationen für die entsprechenden ROIs (`--annotated-intervals`) erstellt. Dazu dient der Aufruf des Progamms `GermlineCNVCaller` im Kohortenmodus wie folgt:

```
gatk GermlineCNVCaller
   --run-mode COHORT
   -L /media/Berechnungen/AnnotationDBs/CNV/Exom/
      Exom.filtered.interval_list
   -I /media/Berechnungen/AnnotationDBs/CNV/Exom/Probe1.hdf5
   -I /media/Berechnungen/AnnotationDBs/CNV/Exom/Probe2.hdf5
   -I /media/Berechnungen/AnnotationDBs/CNV/Exom/ProbeX.hdf5
   -I /media/Berechnungen/AnnotationDBs/CNV/Exom/Probe346.hdf5
   --contig-ploidy-calls /media/Berechnungen/AnnotationDBs/CNV/
      Exom/ploidy-calls
   --annotated-intervals /media/Berechnungen/AnnotationDBs/CNV/
      Exom/Exom.annotated.tsv
   -imr OVERLAPPING_ONLY
   --output /media/Berechnungen/AnnotationDBs/CNV/Exom/
   --output-prefix CNV
   --tmp-dir /media/Ergebnisse/picardtmp
```

Der Ordner `CNV-model` enthält auch in diesem Fall das Modell für die spätere Fallanalyse und im `CNV-calls`-Verzeichnis befinden sich die CNV Analysen für alle 346 Datensätze in entsprechenden Unterordnern. Die Berechnung der Modelle ist sehr rechenaufwändig und dauert auf dem in Anhang I beschriebenen Primärserver etwa 2 Wochen.

Verwendung der Modelle in der Pipeline Zur Detektion von Kopienzahlveränderungen einer einzelnen Probe müssen zunächst ihre Read Counts durch

```
gatk CollectReadCounts
   -L /media/Berechnungen/AnnotationDBs/CNV/Exom/
      Exom.preprocessed.interval_list
   -imr OVERLAPPING_ONLY
   -I /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.dedup.bam
   -O /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.hdf5
```

ermittelt werden. Anschließend erfolgt die Bestimmung der Ploidie für jedes Chromosom. Dabei dient die HDF5-Datei (`-I`) mit den Read Counts für jede einzelne ROI aus dem vorherigen Schritt als Eingabedatei, welche mit dem zuvor erstellten Ploidie-Modell verglichen wird. Durch die Angabe des Pfads

zu einem solchen Modell (`--model`) startet das Tool im Fall-Modus. Der Befehl

```
gatk DetermineGermlineContigPloidy
   --model /media/Berechnungen/AnnotationDBs/CNV/Exom/ploidy-model
   -imr OVERLAPPING_ONLY
   -I /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.hdf5
   --output /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/
   --output-prefix 0001-01_DGCP
```

erzeugt die Datei `contig_ploidy.tsv` im Unterverzeichnis `SAMPLE_0` des Ausgabeordners `0001-01_DGCP-calls`. Diese enthält neben einer durch ein `@` Zeichen eingeleiteten Überschrift mit dem Probennamen drei tabulator-getrennte Spalten mit Informationen über die für ein entsprechendes Chromosom festgestellte Ploidie und die Qualität dieser Vorhersage als nummerischen Wert (siehe Abbildung 4.6). Die Ploidie-Datei wird unter dem Namen `0001-01.ploidy` ins

```
1  @RG ID:GATKCopyNumber   SM:0001-01
2  CONTIG PLOIDY PLOIDY_GQ
3  chr1    2   126.46313192399798
4  chr2    2   125.55616667963986
5  chr3    2   126.29542683546178
6  chr4    2   123.88380782707704
7  chr5    2   126.01991815063992
8  chr6    2   126.58083642001087
9  chr7    2   127.17365305116807
10 chr8    2   126.98844003044249
11 chr9    2   128.00623190070851
12 chr10   2   127.06458820860553
13 chr11   2   127.98303315493101
14 chr12   2   126.87714848390824
15 chr13   2   125.83906775220991
16 chr14   2   127.41227233407929
17 chr15   2   127.4462034833345
18 chr16   2   129.54502964339827
19 chr17   2   128.71419125517167
20 chr18   2   125.72255866296544
21 chr19   2   129.8214939514437
22 chr20   2   129.27607993137252
23 chr21   2   129.20156243402468
24 chr22   2   130.19959713100764
25 chrX    2   122.49723520481145
26 chrY    0   37.670968579722022
```

Abbildung 4.6: *Beispiel einer Ploidy-Datei.*
Sie enthält durch ein `@`-Zeichen eingeleitete Angaben zur Probenidentifikation sowie drei tabulator-getrennte Spalten mit Informationen über die festgestellte Ploidie für jedes Chromosom mit zugehöriger Qualität dieser Vorhersage.

Patientenverzeichnis kopiert. Die Ausgabe des `DetermineGermlineContigPloidy`-Tools dient im nächsten Schritt gemeinsam mit dem CNV-Modell als Grundlage zur Berechnung der Kopienzahlveränderungen der einzelnen ROIs. Mit dem folgenden Befehl wird das `GermlineCNVCaller`-Tool im Fallmodus

(`--run-mode CASE`) ausgeführt.

```
gatk GermlineCNVCaller
   --run-mode CASE
   --contig-ploidy-calls /media/Berechnungen/200402_NB501654_0172_
      AHTYGGHBGX/0001-01_Exom_NB501654_0172/0001-01_DGCP-calls
   --model /media/Berechnungen/AnnotationDBs/CNV/Exom/CNV-model
   --input /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.hdf5
   --output /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/
   --output-prefix 0001-01_GCNV
```

Die Ergebnisse der CNV-Analyse werden unter Berücksichtigung der Ploidie für jedes Chromosom (`--contig-ploidy-calls`) und der Angabe der allosomalen Chromosomen (`--allosomal-contig`) sowie der allgemeinen Normploidie für Autosomen (`--autosomal-ref-copy-number 2`) durch das Programm `PostprocessGermlineCNVCalls` in zwei VCF-Dateien zusammengefasst. Das Kommando

```
gatk PostprocessGermlineCNVCaller
   --calls-shard-path /media/Berechnungen/200402_NB501654_0172_
      AHTYGGHBGX/0001-01_Exom_NB501654_0172/0001-01_GCNV-calls
   --model-shard-path /media/Berechnungen/AnnotationDBs/CNV/
      Exom/CNV-model
   --contig-ploidy-calls /media/Berechnungen/200402_NB501654_0172_
      AHTYGGHBGX/0001-01_Exom_NB501654_0172/0001-01_DGCP-calls
   --autosomal-ref-copy-number 2
   --allosomal-contig chrX
   --allosomal-contig chrY
   --output-genotyped-intervals /media/Berechnungen/200402_
      NB501654_0172_AHTYGGHBGX/0001-01_Exom_NB501654_0172/
      0001-01.genotyped_intervals.vcf
   --output-genotyped-segments /media/Berechnungen/200402_
      NB501654_0172_AHTYGGHBGX/0001-01_Exom_NB501654_0172/
      0001-01.genotyped_segments.vcf
```

gibt in der Datei `0001-01.genotyped_intervals.vcf` die ermittelte Kopienzahl für jede untersuchte Region aus. Da Kopienzahlveränderungen meist mehrere aneinander grenzende Regionen (bei der WES Exons) überspannt, werden in der `0001-01.genotyped_segments.vcf` nebeneinanderliegende Intervalle mit gleicher Kopienzahl zu einem Segment zusammengefasst ausgegeben.

4.2.3.6 Coverage berechnen

Zur Bestimmung der Effizienz und Qualität der WES ist es möglich, die prozentuale Abdeckung (engl. Coverage) aller ROIs sowie deren Sequenziertiefe (engl. Depth) heranzuziehen. Um diese beiden Parameter für jedes Exon eines jeden Transkripts zu bestimmen, wird das Skript `calculateCoverage3.pl` implementiert. Dieses Skript ermittelt durch den Befehl

```
gatk CollectAllelicCounts
   -I /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.recal.bam
   -R /media/Berechnungen/Referenzgenom/HG19/HG19.karyo.fasta
   -L /media/Berechnungen/AnnotationDBs/ROI/Exom.intervals
   -O /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.collectAllelicCounts
```

die Sequenziertiefe jedes detektierten Allels jeder in der Intervall-Datei (`-L`) angegebenen Basenposition. Durch Einlesen und Addition der Sequenziertiefen jedes einzelnen Allels einer Position wird die Gesamtsequenziertiefe aller Basen innerhalb jedes Intervalls in einer Hash-Tabelle gespeichert. Eine mit Hilfe des ENSEMBL Biomarts[3] erstellte `Exom.cov.bed`-Datei beinhaltet die Start- und Endposition für jedes Exon aller Transkripte eines Gens von Interesse. Die Datei besteht aus den in Tabelle 4.5 gelisteten tabulator-getrennten Spalten.

Tabelle 4.5: *Spalten der Exom.cov.bed-Datei*

Spalte	Erläuterung
1	Angabe des Chromosoms
2	Startposition des Exons
3	Endposition des Exons
4	ENSEMBL Exon Identifier (ENSE)
5	Position des Exons im Transkript
6	ENSEMBL Transkrip Identifier (ENST)
7	ENSEMBL Gen Identifier (ENSG)
8	ENSEMBL Genname

Durch die Kombination der Daten aus der Hash-Tabelle und der Start- und Endpositionen jedes Exons aus der `Exom.cov.bed` Datei wird eine neue Datei erzeugt (`0001-01.bedtoolsCoverage`), die für jede Base jedes angegebenen Exons die Sequenziertiefe beinhaltet. Im nächsten Schritt wird die soeben erstellte Datei wieder durch das Skript eingelesen und für jedes Exon einer Gen-Transkript-Exon Kombination die Länge der Region, die Summe der Sequenziertiefe aller beinhalteten Basen, die maximale und minimale Sequenziertiefe sowie die Anzahl an Basen mit einer Sequenziertiefe ≥ 1, ≥ 10 und ≥ 20 in einer Hash-Tabelle festgehalten. Mit Hilfe der in dieser Hash-Tabelle gespei-

[3] `http://grch37.ensembl.org/biomart/martview/`

cherten Daten, kann nun die mittlere Sequenziertiefe ($\frac{Summe\ der\ Sequenziertiefe}{Laenge\ der\ Region}$), der prozentuale Anteil des Exons mit einer minimalen Sequenziertiefe von einem Read ($\frac{Anzahl\ an\ Basen\ mit\ einer\ Sequenziertiefe >= 1}{Laenge\ der\ Region}$), der prozentuale Anteil des Exons mit einer minimalen Sequenziertiefe von zehn Reads und der prozentuale Anteil des Exons mit einer minimalen Sequenziertiefe von 20 Reads für jedes Exon berechnet werden. Durch Summierung der Werte aller zu einem Transkript gehörender Exons ist es möglich die Werte für ein bestimmtes Transkript zu ermitteln. Aus der Summierung aller zu einem Gen gehöriger Transkripte ergeben sich die Werte für das entsprechende Gen.

Um die Performance des gesamten Versuchs zu bestimmen, findet das `CollectHsMetrics`-Tool des GATK Pakets wie folgt Anwendung:

```
gatk CollectHsMetrics
   -BI /media/Berechnungen/AnnotationDBs/ROI/Exom.picard.intervals
   -TI /media/Berechnungen/AnnotationDBs/ROI/Exom.picard.intervals
   -I /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.recal.bam
   -O /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.picard_metrics
```

Aus der resultierenden Datei wird die Genomgröße, die Gesamtanzahl der Basen in ROIs, die Anzahl an Reads, die mittlere und mediane Coverage der angereicherten Regionen sowie die prozentuale Angabe von Regionen mit einer Sequenziertiefe von 1, 2, 10, 20, 30, 40, 50 und 100 Reads extrahiert.

Die Zusammenfassung der Coverage-Ergebnisse auf Ebene des gesamten Experiments, der Gene, der Transkripte und der Exons in eine Datei erlaubt später eine detaillierte Analyse von nicht abgedeckten Regionen zur weiteren Untersuchung oder Sequenzierung dieser „Lücken“ mittels Sanger-Sequenzierung.

Zur Bestimmung der Coverage wird das `CalculateCoverage3.pl`-Skript folgendermaßen aus der Pipeline aufgerufen:

```
perl CalculateCoverage3.pl
   -bam /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.recal.bam
   -investigation Exom
   -genome /media/Berechnungen/Referenzgenom/HG19/HG19.karyo.fasta
   -sample_name 0001-01
   -output /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/
```

4.2.3.7 Bestimmung von Homozygotieregionen (ROH)

Die Analyse auf Homozygotie-Regionen wird mit dem H^3M^2-Programm (siehe Anhang IV.XI) durchgeführt. Im ersten Schritt ermittelt das Kommando

```
H3M2BamParsing.sh
   /media/Berechnungen/AnnotationDBs/H3M2Tool/
   /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/
   0001-01.recal.bam
   /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/
   ROH
   0001-01
   /media/Berechnungen/Referenzgenom/HG19/HG19.karyo.fasta
   /media/Berechnungen/AnnotationDBs/H3M2Tool/
      SNP1000GP.HGb37_Exome.mod.bed
```

die B-Allel-Frequenz der im `SNP1000GP.HGb37_Exome.mod.bed` Datensatz angegebenen SNP-Positionen auf dessen Grundlage im zweiten Schritt mit dem Befehl

```
H3M2Analyse.sh
   /media/Berechnungen/AnnotationDBs/H3M2Tool/
   /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/
   ROH
   0001-01
   /media/Berechnungen/AnnotationDBs/H3M2Tool/
      SNP1000GP.HGb37_Exome.mod.bed
   100000
   0.1
   0.1
   5
```

die ROHs berechnet werden. Die Ausgabedatei `0001-01_1e+05_0.1_0.1_HomozygosityTableCall.bed` beinhaltet fünf tabulator-getrennte Spalten, in denen das entsprechende Chromosom, die Startposition, die Endposition, die Wahrscheinlichkeit für eine ROH und die Anzahl der SNP-Marker innerhalb dieser ROH abgelegt sind. Jede Zeile beschreibt eine ROH Region. Das eigens implementierte `MaiROHparser.pl`-Skript liest die Ausgabedatei ein (`-ROH_bed`) und gibt unter Ausschluss der Gonosomen eine prozentuale Angabe der Regionen mit ROH >1 Mbp, >3 Mbp, >5 Mbp und >8 Mbp in die Datei `0001-01.roh_stats` aus. Zudem filtert das Skript die ROHs und schreibt nur jene größer des mit `-min_size` angegebenen Schwellenwertes in die Datei `0001-01.roh` im BED-Format. Diese wird anschließend mit `bgzip` kompri-

miert und mit `tabix` indiziert, um sie im Annotationsschritt zu nutzen. Der Aufruf des Skripts aus der Pipeline ist wie folgt:

```
perl MaiROHparser.pl
   -ROH_bed /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01_1e+05_0.1_0.1_
      HomozygosityTableCall.bed
   -sample_name 0001-01
   -output /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/
   -min_Size 1000000
```

Zur Visualisierung und späteren Einbindung in die Auswertungssoftware werden die einzelnen ROH Bereiche jedes Chromosoms als Box dargestellt. Ein Strich spiegelt einen biallelischen Bereich des Chromosoms wider. Abbildung 4.7 zeigt diese Plots exemplarisch für die Chromosomen eins bis fünf.

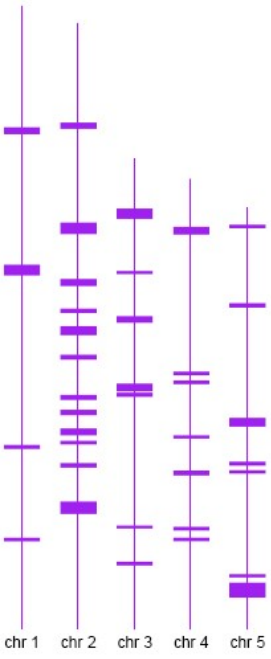


Abbildung 4.7: *Beispielhafte ROH Plots für Chromosom eins bis fünf. Ein Strich spiegelt einen biallelischen Bereich des Chromosoms wider. Regionen mit ROH sind als Box dargestellt.*

Diese Bilder werden mit dem R-Script `MaiRohPic.R` unter Verwendung der `ggplot2`-Library von Hadley Wickham *et al.* aus dem CRAN-Archiv[4] erstellt. Dazu liest das Script die gefilterte ROH Datei (`0001-01.roh`) sowie eine Datei mit Angaben über die Länge der Chromosomen (`*.genomefile`) ein. Folgender Befehl startet die Analyse.

[4] `https://cran.r-project.org/web/packages/ggplot2/index.html`

```
Rscript MaiRohPic.R
   /media/Berechnungen/Referenzgenom/HG19/
      HG19.karyo.fasta.genomefile
   /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.roh
   /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/
```

4.2.3.8 Detektion von Varianten

Nukleotidaustausche sowie kleine Insertionen und Deletionen im Patientendatensatz werden, wie in Kapitel 3.3 beschrieben, beim Variant Calling ermittelt. Dazu vergleicht der `HaplotypeCaller` (siehe Anhang IV.VIII.III) jede Base des Alignments (`-I`) des Patienten mit dem Referenzgenom (`-R`). Der Befehl

```
gatk HaplotypeCaller
   -R /media/Berechnungen/Referenzgenom/HG19/HG19.karyo.fasta
   -I /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.recal.bam
   -L /media/Berechnungen/AnnotationDBs/ROI/Exom.intervals
   --interval-padding 250
   -O /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.vcf
```

sucht nach Abweichungen innerhalb der ROIs (`-L`). Alle gefundenen Varianten werden in einer VCF-Datei (siehe Anhang II.IV) ausgegeben.

4.2.3.9 Intervar-Analyse

Um die Varianten bezüglich ihrer Pathogenität besser einschätzen zu können, nutzt die `Intervar`-Software (siehe Anhang IV.X) die ACMG Empfehlungen zur Einstufung einer Variante in eine der fünf Pathogenitätsstufen. Das Modul zur Anwendung des `Intervar`-Programms und zur Verarbeitung der Ergebnisse wurde von Sri Dewi (Institut für Humangenetik, Unimedizin Mainz) als R-Skript `InterVar_pipeline.R` implementiert. Es nutzt den Patientenordner sowie die Patienten-ID als Eingabe, wendet den `Intervar`-Algorithmus auf die im vorhergehenden Schritt ermittelten Varianten an, verarbeitet die resultierende Datei und speichert die Varianten mit deren vorhergesagter Pathogenitätsstufe in einer VCF-Datei ab. Zur Einbindung dieser Vorhersage in die Annotation wird die VCF-Datei komprimiert und mittels `tabix` indiziert. Der Aufruf des R-Skripts erfolgt durch folgenden Befehl:

```
Rscript InterVar_pipeline.R
   /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/
   0001-01
```

4.2.3.10 Annotation der Varianten

Die Annotation der im Patientendatensatz gefundenen Varianten wird mit Hilfe des ENSEMBL Variant Effect Predictors (siehe Anhang IV.XII) durchgeführt. Neben den durch den VEP bereitgestellten Datenbanken werden die Plugins CADD, LoFtool und dbNSFP verwendet (`--plugin`). Zudem fließen externe Datenbanken und Pathogenitätswerte (PhyloP, OMIM, ClinPred, HGMD, Deafness Variation Database (DVD), SpliceAI, PrimateAI) sowie die Populationsdaten der hausinternen Datenbank (IHDB), die zuvor berechneten ROH Bereiche und Pathogenitätsstufen von Intervar in die Annotation der Varianten ein (`--custom`). Der Programmaufruf

```
vep
   --offline
   --dir /media/Berechnungen/AnnotationDBs/vep
   --fasta /media/Berechnungen/Referenzgenom/HG19/
      HG19.karyo.fasta
   --everything
   --assembly GRCh37
   -i /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.vcf
   -o /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.annot.vcf
   --plugin CADD,/media/Berechnungen/AnnotationDBs/cadd/20160314/
      whole_genome_SNVs.tsv.gz,/media/Berechnungen/AnnotationDBs/
      cadd/20160314/cadd_InDels.tsv.gz
   --plugin LoFtool
   --plugin dbNSFP,/media/Berechnungen/AnnotationDBs/dbNSFP/
      20171009/dbNSFP2.9.3.gz,SIFT_score,SIFT_converted_rankscore,
      SIFT_pred,Polyphen2_HVAR_score,Polyphen2_HVAR_rankscore,
      Polyphen2_HVAR_pred,LRT_score,LRT_converted_rankscore,
      LRT_pred,MutationTaster_score,MutationTaster_converted_
      rankscore,MutationTaster_pred,FATHMM_score,FATHMM_
      rankscore,FATHMM_pred,MetaSVM_score,MetaSVM_rankscore,
      MetaSVM_pred,MetaLR_score,MetaLR_rankscore,MetaLR_pred,
      Reliability_index,PROVEAN_score,PROVEAN_converted_rankscore,
      PROVEAN_pred,M-CAP_score,M-CAP_rankscore,M-CAP_pred,CADD_raw,
      CADD_raw_rankscore,CADD_phred,REVEL_score,REVEL_rankscore,
      VEST3_score,VEST3_rankscore
   -custom /media/Berechnungen/AnnotationDBs/phylop/20160314/
      hg19.100way.phyloP100way.sorted.bedgraph.gz,PhyloP,bed,
      exact,0
   -custom /media/Berechnungen/AnnotationDBs/ihdb/current/
      ihdb_af.vcf.gz,IHDB_AF,vcf,exact,0
   -custom /media/Berechnungen/AnnotationDBs/ihdb/current/
      ihdb_count.vcf.gz,IHDB_Count,vcf,exact,0
   -custom /media/Berechnungen/AnnotationDBs/omim/20200210/
```

```
   OmimInheritanceHG19.sorted.bed.gz,OMIM,bed,overlap,0
-custom /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
   0001-01_Exom_NB501654_0172/0001-01.intervar.vcf.gz,
   InterVar,vcf,exact,0
-custom /media/Berechnungen/AnnotationDBs/ClinPred/
   ClinPred.vcf.gz,ClinPred,vcf,exact,0
-custom /media/Berechnungen/AnnotationDBs/HGMD/2018.2/
   hgmd_pro_2018.2_hg19.vcf.gz,HGMD_PRO,vcf,exact,0,CLASS,PHEN
-custom /media/Berechnungen/AnnotationDBs/DVD/Version8.2.1/
   DVD_v8.2.1_pathogenicity.vcf.gz,DVD_Pathogenicity,vcf,
   exact,0
-custom /media/Berechnungen/AnnotationDBs/DVD/Version8.2.1/
   DVD_v8.2.1_disease.vcf.gz,DVD_Disease,vcf,exact,0
-custom /media/Berechnungen/AnnotationDBs/DVD/Version8.2.1/
   DVD_v8.2.1_pubmed.vcf.gz,DVD_Pubmed,vcf,exact,0
-custom /media/Berechnungen/AnnotationDBs/gnomad/
   gnomad.vcf.gz,gnomAD_GENOMES,vcf,exact,0,AF
-custom /media/Berechnungen/AnnotationDBs/SpliceAI/
   spliceai_scores.masked.snv.hg19.vcf.gz,SpliceAI,
   vcf,exact,0,DS_AG,DS_AL,DS_DG,DS_DL,DP_AG,
   DP_AL,DP_DG,DP_DL
-custom /media/Berechnungen/AnnotationDBs/PrimateAI/
   AIScores_sorted.vcf.gz,PrimateAI,vcf,exact,0
-custom /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
   0001-01_Exom_NB501654_0172/0001-01.roh.gz,ROH,bed,overlap,0
--buffer_size 10000
--vcf
--pick
--pick_order canonical,rank,appris,tsl,biotype,ccds,length
--force_overwrite
--fork 4
```

produziert eine VCF-Datei, welche für jede Variante die Annotationen im `CSQ`-Feld der `INFO`-Spalte trägt. Die einzelnen Annotationen sind dabei durch ein „|“-Zeichen getrennt.

4.2.3.11 Priorisierung von Varianten

Basierend auf diesen Annotationen findet mit dem eigens entwickelten und in Perl implementierten Skript `MaiHotspot.pl` eine Priorisierung bezüglich der Pathogenität einer Variante statt. Nach Einlesen der annotierten VCF-Datei werden nur jene Varianten weiter betrachtet, die eine Qualität ≥ 300 und eine Allelfrequenz von unter 1% in der gnomAD-Population (siehe Anhang III.VI) und der Population der internen Datenbank aufweisen. So lässt sich die Analyse auf seltene Varianten beschränken, bei denen es sich nicht um Artefakte handelt. Ist eine dieser Varianten in der HGMD oder ClinVar Datenbank (siehe Anhang III.I und III.III) als wahrscheinlich pathologisch oder pathologisch

beschrieben, erfolgt die Abspeicherung in einer temporären Tabelle. Proteintrunkierende Varianten sowie andere Arten von Varianten in Verbindung mit einem CADD-Pathogenitätswert ≥ 15 werden ebenfalls dieser temporären Tabelle hinzugefügt.

Im zweiten Schritt wird die temporäre Tabelle durchlaufen und überprüft, ob der Genotyp mit dem für das betroffene Gen bekannten Erbgang theoretisch eine Erkrankung auslösen kann. Befinden sich zwei verschiedene Varianten im gleichen Gen in dieser Tabelle, so werden die beiden Varianten ausgegeben, um auch compound heterozygote Fälle bei autosomal rezessiven Erkrankungen abdecken zu können. Die Ausgabe homozygot vorliegender Varianten bei einem autosomal rezessiven Erbgang ist obligatorisch. Folgt das entsprechende Gen einem autosomal dominanten Erbgang, wird die zugehörige Variante nur ausgegeben, wenn sie im heterozygoten Zustand vorliegt. Für einen X-gekoppelt rezessiven oder dominanten Erbgang verläuft das Verfahren analog. Gespeichert werden die so gefilterten annotierten Varianten wieder im VCF-Format.

Durch diese Priorisierung lassen sich die etwa 120.000 Varianten pro Patient auf etwa 50 Varianten mit der höchsten Pathogenitätswahrscheinlichkeit reduzieren. Der aus der Pipeline ausgeführte Befehl lautet:

```
perl MaiHotspot.pl
   -vcf /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.annot.vcf
   -out_vcf /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.hotspots.vcf
```

4.2.3.12 Filtern und Annotieren der Kopienzahlveränderungen

Da die Ausgabedatei der CNV-Analyse `0001-01.genotyped_segments.vcf` auch Segmente mit Kopienzahl zwei beinhaltet, ist es sinnvoll diese Daten zu filtern. Eine Kopienzahl von zwei stellt bezogen auf die Autosomen die wildtypische Kopienzahl eines diploiden Organismus dar. Neben der Extraktion von aberranten CNV-Segmenten verwendet das implementierte Perl-Skript `MaiCNVFilter.pl` die im Patientendatensatz ermittelten Varianten, um die Spezifität und Sensitivität einer Kopienzahlveränderung zu ermitteln. So sollten sich zum Beispiel im Bereich einer vorhergesagten homozygoten Deletion (Kopienzahl 0) keine Varianten befinden. Eine heterozygote Deletion hingegen (Kopienzahl 1) sollte nur homozygote Varianten beinhalten. Bei einer Duplikation mit der Kopienzahl drei verschiebt sich das Allelverhältnis eines nicht homozygoten SNPs theoretisch auf 66%. Bei Einbeziehung der Gonosomen besteht bei männlichen Individuen nur jeweils eine Kopie des Chromosoms X und Y. Somit stellt hier bereits die Kopienzahl von zwei eine Duplikation dar, welches bei der Beurteilung der Spezifität und Sensitivität der CNVs durch Betrachtung der Varianten berücksichtigt werden muss.

Zu Beginn liest das Skript die VCF-Datei mit SNP- und InDel-Varianten

(`0001-01.vcf`) ein und speichert die Sequenziertiefe pro Allel für jede Variantenposition zwischen. Die Ploidie-Datei (`0001-01.ploidy`) wird zur Ermittlung des Geschlechts verwendet. Anschließend durchläuft das Programm jedes in der Datei `0001-01.genotyped_segments.vcf` enthaltene Segment, ermittelt die Anzahl der heterozygoten und homozygoten Varianten in diesem Bereich und berechnet das Allelverhältnis der heterozygoten SNPs. Im zweiten Schritt ist es möglich diese Werte zu verwenden, um die CNVs, wie im vorherigen Abschnitt beschrieben, als falsch-positiv oder richtig-positiv zu klassifizieren. Die Ausgabe der Klassifizierung erfolgt in einer tabulator-getrennten Datei `0001-01.cnv` unter Angabe der Position, der Anzahl an unterstützenden SNPs und deren Allelverhältnis. Alle vom wildtyp abweichenden CNV Ereignisse werden zudem in eine VCF-Datei geschrieben (`0001-01.cnv.vcf`).

Die durch den Programmaufruf

```
perl MaiCNVFilter.pl
   -genotyped_segments /media/Berechnungen/200402_NB501654_0172_
      AHTYGGHBGX/0001-01_Exom_NB501654_0172/
      0001-01.genotyped_segments.vcf
   -vcf /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.annot.vcf
   -ploidy /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.ploidy
   -out_vcf /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.cnv.vcf
```

gefilterten CNVs unterlaufen im nachfolgenden Schritt einer Annotation durch den Variant Effect Predictor (siehe Anhang IV.XII), um unter anderem die betroffenen Gennamen mit den zugehörigen Exons und Introns zu erhalten. Umfasst eine Kopienzahlveränderung mehrere Gene, so sind die Annotationsblöcke durch ein Komma getrennt aufgelistet. Der Befehl

```
vep
   --offline
   --dir /media/Berechnungen/AnnotationDBs/vep
   --fasta /media/Berechnungen/Referenzgenom/HG19/
      HG19.karyo.fasta
   --assembly GRCh37
   -i /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.cnv.vcf
   -o /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.cnv.annot.vcf
   --vcf
   --per_gene
   --pick_order canonical,rank,appris,tsl,biotype,ccds,length
   --force_overwrite
```

führt die Annotation der CNVs durch. Anschließend erstellt das R-Skript `MaiCNVPics.R` durch

```
Rscript MaiCNVPic.R
   /media/Berechnungen/Referenzgenom/HG19/
      HG19.karyo.fasta.genomefile
   /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.cnv.vcf
   /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/
```

Plots der Kopienzahlveränderungen analog zu den Plots der ROH Regionen (siehe Kapitel 4.2.3.7).

4.2.3.13 Gender-Match

Ein einfacher Ansatz zur Aufdeckung vermeintlicher Probenvertauschungen ist der Abgleich des in der Patientendatenbank abgelegten mit dem aus den Variantendaten bestimmten Geschlechts. Zur Berechnung des Geschlechts wird im `MaiGender.pl`-Skript die Anzahl heterozygoter und homozygoter SNPs auf dem X-Chromosom mit einem Qualitätswert >300 und einer minimalen Sequenziertiefe von 10 Reads gezählt. Anschließend erfolgt die Berechnung des Verhältnisses aus der Anzahl heterozygot vorliegender SNPs zur Anzahl aller SNPs des X-Chromosoms. Durch die pseudoautosomalen Regionen (PAR) der Chromosomen X und Y sind auch bei männlichen Individuen heterozygote SNPs auf dem X-Chromosom detektierbar. Liegt das Heterozygotieverhältnis unter 0,25 (25%), so ist die Probe als männlich zu klassifizieren. Andernfalls liegt ein weiblicher Datensatz vor. Das Ergebnis des Vergleichs zwischen dem ermittelten und dem in der Datenbank festgehaltenen Geschlechts wird ebenso wie die zur Berechnung verwendeten Parameter in der Ausgabedatei `0001-01.gender` festgehalten (siehe Abbildung 4.8). Auf diese Weise ist es allerdings nicht möglich Probenvertauschungen gleichgeschlechtiger Patienten aufzudecken. Der Programmaufruf aus dem Pipeline-Skript lautet folgendermaßen:

```
perl MaiGender.pl
   -dbGender w
   -vcf /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.annot.vcf
   -output /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.gender
```

```
# Parameter
input = /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/0001-01_Exom_NB501654_0172/0001-01.annot.vcf
output = /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/0001-01_Exom_NB501654_0172/0001-010219-20.gender
quality_threshold = 500
depth_threshold = 10
qd_threshold = 10
# Results
analyzed_SNPs = 814
hom_SNPs = 790
het_SNPs = 24
het_ratio = 0.0294840294840295
estimated_sex = male
database_sex = m
gender_match = yes
```

Abbildung 4.8: *Beispiel einer Ergebnisdatei des GenderMatches. Neben den Pfaden zur Ein- und Ausgabedatei sind die verwendeten Schwellenwerte zur Auswahl der Varianten im „Parameter“ Abschnitt angegeben. Im „Ergebnis“ Bereich wird die Anzahl der einbezogenen SNPs, der Anteil an heterozyoten und homozygoten SNPs sowie die Heterozygotierate gelistet. Das daraus ermittelte Geschlecht und der Abgleich mit der Patientendatenbank sind ebenfalls aufgeführt.*

4.2.3.14 Probennachverfolgung

Um eine Verwechslung von Proben komplett ausschließen zu können, ist eine vorherige unabhängige Genotypisierung von mehreren SNPs mittels Pyrosequenzierung notwendig. Die Kombination dieser polymorphen SNP-Marker ist für jedes Individuum einzigartig. Im Institut für Humangenetik werden die Genotypen der in Tabelle 4.6 gelisteten SNPs aus jeder Patientenprobe bestimmt und die Ergebnisse anschließend in der internen Patientendatenbank gespeichert.

Das eigens entwickelte Perl Skript `MaiSampleTracker.pl` liest die aus den NGS-Daten ermittelten Varianten ein und vergleicht die entsprechenden Positionen mit den in der Datenbank festgehaltenen Ergebnissen der Pyrosequenzierung dieses Patienten. Die Ausgabedatei enthält neben der Proben-ID die Anzahl untersuchter SNPs, die Anzahl der übereinstimmenden Genotypen zwischen NGS und Pyrosequenzierung, die Prozentzahl der Übereinstimmung, und eine detaillierte Auflistung aller untersuchter SNPs mit ihren ermittelten Genotypen der Pyro- sowie Next-Generation-Sequenzierung und deren zugehörigen Qualitätswerten (`0001-01.trk`). Zudem wird das Ergebnis des SampleTrackings in der Patientendatenbank festgehalten. Der Programmaufruf lautet:

```
perl MaiSampleTracking.pl
   -sample_name 0001-01
   -vcf /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.vcf
   -output /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/
```

Tabelle 4.6: *SNPs fürs Sample Tracking*

Chromosom	Position	Referenzallel	dbSNP Identifier
1	179520506	G	rs1410592
2	169789016	T	rs497692
3	4403767	A	rs2819561
4	5749904	T	rs4688963
5	82834630	T	rs309557
6	146755140	G	rs2942
7	48450157	T	rs17548783
8	94935937	T	rs4735258
9	100190780	A	rs1381532
10	100219314	G	rs10883099
11	16133413	A	rs4617548
12	993930	C	rs7300444
13	39433606	A	rs9532292
14	50769717	G	rs2297995
15	34528948	G	rs4577050
16	70303580	G	rs2070203
17	71197748	G	rs1037256
18	21413869	T	rs9962023
19	10267077	T	rs2228611
20	6100088	A	rs10373
21	44323590	T	rs4148973
22	21141300	T	rs4675

4.2.3.15 Varianten in Datenbank eintragen

Alle beim Patienten gefundenen Varianten werden mit Bezug zur Patienten-ID in der internen Datenbank abgelegt. Die wichtigsten Annotationen, wie zum Beispiel der Genname, die Gen- und Transkript-ID der ENSEMBL Datenbank, das betroffene Exon bzw. Intron, die Auswirkung auf Basen- und Proteinebene sowie Ergebnisse von Pathogenitätsvorhersagetools schreibt das Perl-Skript `MaiVar2DB.pl` ebenfalls in diese Datenbank und verknüpft sie mit der entsprechenden Variante. Das Schema der Varianten-Datenbank ist in Abbildung 4.9 dargestellt. Jede Variante kann dabei nur einmal in der `varianten`-Tabelle vorkommen. Eine Variante kann über die `traegt_varianten`-Tabelle mehreren Patienten zugeordnet sein. Umgekehrt besteht die Möglichkeit, dass ein Patient auf diese Weise auch mehrere Varianten besitzt. Durch die Zuordnung einer Variante zu einem Patienten ist es möglich in dieser Tabelle aus der Sequenzierung resultierende spezifische Merkmale wie den Name der Untersuchung, die Zygotie, die Qualität und die Sequnziertiefe einer jeden Variante des Patientendatensatzes festzuhalten. Jede Variante kann durch einen Eintrag der `annotationen`-Tabelle beschrieben werden. Zudem ist es über die Auswertesoftware (siehe Kapitel 4.5) möglich, jeder Variante Klassifizierungseinträge und Pathogenitätsstufen zuzuordnen. Diese werden später durch die Auswertungssoftware in der `klassifizierung`-Tabelle abgelegt.

Das Perl-Skript zum Eintragen der Varianten in die Datenbank wird durch den Befehl

```
perl MaiVar2DB.pl
   -vcf /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.annot.vcf
   -sample-name 0001-01
   -investigation Exom
```

aufgerufen, durchläuft die annotierte VCF-Datei, extrahiert die entsprechenden Daten und trägt sie in die Tabellen der Datenbank ein.

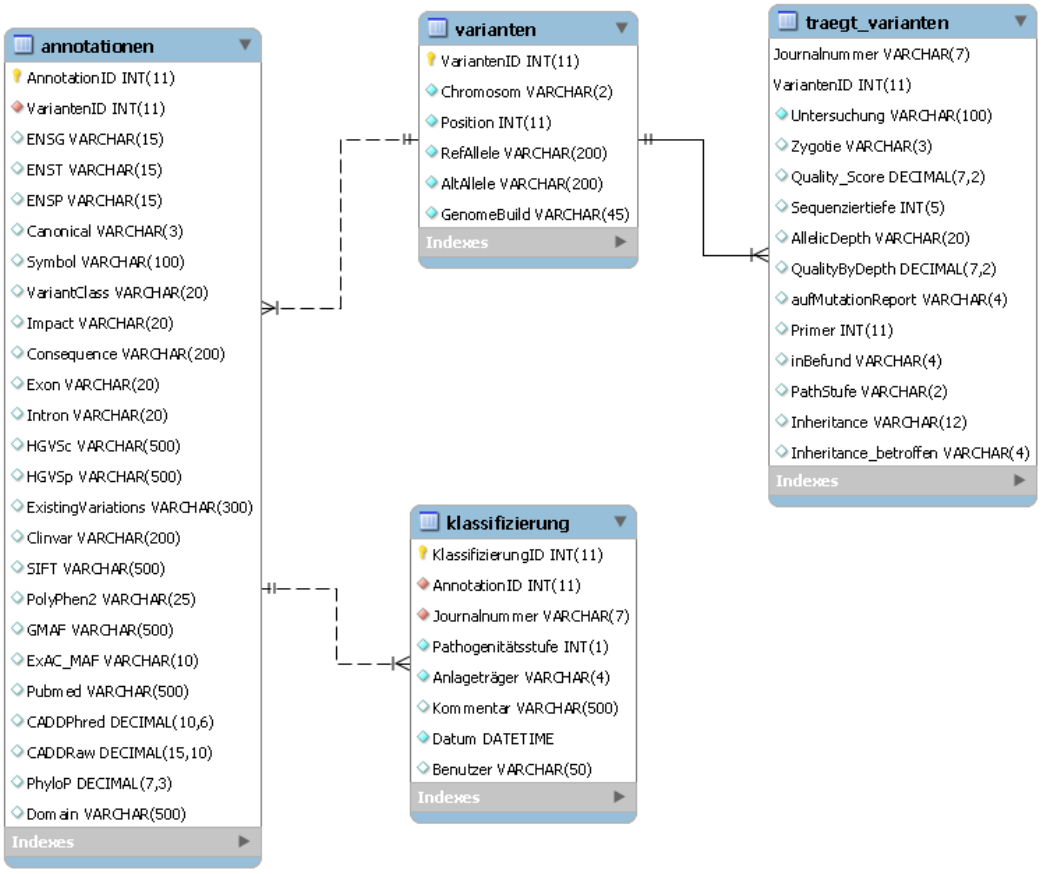


Abbildung 4.9: *Schema der Varianten-Datenbank.*
Ein Patient steht über die `traegt_varianten`*-Tabelle in einer many-to-many Relation zur* `varainten`*-Tabelle. Dabei kann einer Variante genau eine Annotation (*`annotationen`*), und einer Annotation eine Klassifizierung (*`klassifizierung`*) zugeordnet werden.*

4.2.3.16 Primer erstellen

Der letzte Schritt der Pipeline besteht aus dem Design von Primerpaaren für alle Varianten eines Patienten. Diese werden zur Validierung und Segre-

gationsanalyse von putativ pathogenen Varianten mittels PCR und Sanger-Sequenzierung (siehe Anhang V.I) benötigt. Das Auswertungs-Personal kann so einen benötigten Primer direkt über die Auswertungssoftware zur Bestellung freigeben.

Um Primer für jede Variante designen zu können, durchläuft das Perl-Skript `MaiPrimer.pl` die VCF-Datei und extrahiert mit Hilfe der `samtools faidx`-Software die Sequenzabfolge 300 bp up- und downstream jeder Variante aus dem Referenzgenom. So entstehen 601 bp lange Referenzsequenz-Bruchstücke, welche als Wert für den Schlüssel `SEQUENCE_TEMPLATE` gemeinsam mit der Angabe eines Idenifiers (`SEQUENCE_ID`) in einer Datei (`0001-01.samplefile`) gespeichert werden. Zudem ist der durch die PCR zu vervielfältigende Bereich mit Hilfe des Schlüssels `SEQUENCE_TARGET=200,201` zu definieren. Das Primerpaar soll ein Amplikon erstellen, welches die DNA-Sequenz 100 bp vor und nach der Variante einschließt. Da die Variante an Position 301 des extrahierten Referenzsequenzbruchstücks liegt, startet dieser Bereich bei Base 200 und ist 201 bp lang. Ein alleinstehendes =-Zeichen trennt die einzelnen Datensätze (siehe Abbildung 4.10).

```
1  SEQUENCE_ID=chr3:58132509-58133109
2  SEQUENCE_TEMPLATE=TCTCCCTAACACCCCTGCATCCCTGTTCCCACTTG...
3  SEQUENCE_TARGET=200,201
4  =
5  SEQUENCE_ID=chr3:58144963-58145563
6  SEQUENCE_TEMPLATE=GGTTTTCTGGCCTGCATGGGCATTATTTTGGATGC...
7  SEQUENCE_TARGET=200,201
8  =
```

Abbildung 4.10: *Aufbau der* `Samplefile` *für den Primer3 Algorithmus. In ihr wird das Sequenzfragment (*`SEQUENCE_TEMPLATE`*) sowie der zu vervielfältigende Bereich* `SEQUENCE_TARGET` *mit einer eindeutigen ID (*`SEQUENCE_ID`*) für jede Variante angegeben.*

Das Programm `primer3_core` liest die `0001-01.samplefile` ein und erstellt wenn möglich ein Primerpaar für jede Variante. Im Folgenden extrahiert das Perl-Skript die Länge des Amplikons, die Sequenz des Forward- und Reverse-Primers sowie deren GC-Gehalt und Annealing-Temperatur für jede Variante aus der resultierenden Datei `0001-01.primer3`. Die so gewonnenen Informationen werden in der tabulator-getrennten Datei `0001-01.primer` abgelegt. Das Skript zum Primerdesign wird mit dem folgenden Programmaufruf gestartet:

```
perl MaiPrimer.pl
   -vcf /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.annot.vcf
   -sample_name 0001-01
   -output /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/
```

4.3 Anwendung der Exom-Pipeline auf Genomdaten

Die im Kapitel 4.2 beschriebene Pipeline zur Analyse von WES-Daten ist so ausgelegt, dass sie ebenso auf Datensätze einer gesamtgenomischen Sequenzierung anwendbar ist. Auch wenn die meisten „Library Preparation"-Kits für eine WGS ohne eine PCR-Amplifikation arbeiten, ist es sinnvoll den Schritt zur Detektion von Duplikaten (siehe Kapitel 4.2.3.3 und Anhang IV.VIII.I) durchzuführen, um die gegebenenfalls bei der Sequenzierung entstandenen optischen Duplikate zu entfernen.

Eine Analyse auf Kopienzahlveränderungen im Patientendatensatz der WGS bedarf für den in der Pipeline verwendeten Algorithmus aus dem GATK-Paket (siehe Kapitel 4.2.3.5 und Anhang IV.VIII.IV) der Erstellung eines Modells aus mehreren auf gleiche Art und Weise sequenzierten Proben. Da das Institut für Humangenetik der Universitätsmedizin Mainz bisher keine ausreichende Sequenzierkapazität zur eigenständigen Durchführung einer WGS besitzt, ist es notwendig die Genomsequenzierung von externen Firmen, wie dem BGI oder MDC für vereinzelte Patientenproben auf wissenschaftlicher Basis durchführen zu lassen. Durch den Einkauf externer Sequenzierleistungen bei verschiedenen Firmen entsteht ein unterschiedliches Spektrum an verwendeten Probenaufbereitungs- und Sequenziermethoden, auf dessen Basis kein Modell für die CNV Analyse etabliert werden kann. Somit ist zum Zeitpunkt des Anfertigens dieser Dissertationsschrift in der Universitätsmedizin Mainz keine Bestimmung der Kopienzahlveränderung bei WGS-Datensätzen möglich. Die entsprechenden Schritte zum Filtern und Annotieren der CNVs (siehe Kapitel 4.2.3.12) können somit ebenfalls nicht angewendet werden.

Alle restlichen Auswertungsschritte sind wie in Kapitel 4.2 mit den beschriebenen Algorithmen und Parametern zu verwenden.

4.4 Die Suche nach regulatorischen Varianten

In durch WES molekulargenetisch nicht diagnostizierten Fällen mit anzunehmendem genetischen Hintergrund sind Krankheitsvarianten in regulatorischen Regionen zu vermuten. Da eine Genomsequenzierung aller nicht diagnostizierter Patienten zu kostenintensiv ist, wird ein Panel entwickelt, der wichtige, in der WES nicht enthaltene regulatorische Regionen abdeckt.

4.4.1 Erstellung eines TES-Panels

Zur Untersuchung regulatorischer Regionen wurde ein TES-Panel (RegVar-Panel) auf DNA-Ebene entworfen. Dieser umfasst 1881 humane primäre miRNA-Transkripte der miRBase-Datenbank (siehe Anhang III.VIII) sowie die letzten 0,5 kbp der 3'UTRs von 2600 krankheitsassoziierten Genen (Stand 2016). Da die WES bei den meisten Genen einen großen Teil des 5'-Endes des

3'UTRs eines Gens abdeckt, wurde nur das 3' Ende des 3'UTRs berücksichtigt. Zudem umfasst das Panel 345 von Irima *et al.* publizierte Mikroexons, die bei der Neurogenese eine Rolle spielen [Irimia *et al.*, 2014]. Alle in das Panel integrierten Regionen wurden durch 50 bp stromaufwärts und stromabwärts ergänzt. Aufgrund von Kombinationsmöglichkeiten und Synergieeffekten mit anderen im Institut für Humangenetik der Universitätsmedizin Mainz angewendeten und etablierten Gen-Panel fiel die Wahl des „Library Preparation"-Kit zur Anreicherung auf das SureSelect XT Custom Enrichment Kit der Firma Agilent. Mit Hilfe der SureDesign Online-Software[5] wurden die RNA-Sonden für den TES-Panel mit einer Gesamtgröße von 2,876 Mb erstellt. Die Sequenzierung erfolgt auf dem institutseigenen Illumina NextSeq500-System (2x126 bp PE).

4.4.2 Pipeline zur Bewertung regulatorischer Varianten

Die Datenanalyse des RegVar-Panels folgt dem Schema der Auswertung von WES-Daten. Da bisher keine ausreichende Anzahl an Proben untersucht wurde, um ein verlässliches CNV-Modell erstellen zu können, ist die Analyse und die damit verbundene Filterung und Annotation von Kopienzahlveränderungen nicht möglich. Außerdem ist das H^3M^2 Softwarepaket nur für WES- und WGS-Daten ausgelegt, wodurch seine Anwendung auf Daten des RegVar-Panels aufgrund des Fehlens vieler SNP-Marker Informationen nicht aussagekräftig ist. Da es sich bei den Ergebnissen hauptsächlich um regulatorische Varianten in nicht kodierenden Bereichen handelt, ist eine automatisierte Kategorisierung der Varianten in die fünf Pathogenitätsstufen durch die Intervar-Software aufgrund der unzureichenden und oftmals unzuverlässigen Annotationsdaten nicht möglich.

Die einzelsträngigen pre-miRNAs bilden nach ihrer Translation durch Wasserstoffbrücken mit komplementären Teilbereichen von sich selbst eine sogenannte Sekundärstruktur (2D-Struktur). Liegen Varianten in einem für die Ausbildung der 2D-Struktur wichtigen Bereich, ist ein Einfluss auf die korrekte 2D-Faltung möglich. Dies könnte die Bindung von Proteinen, die für weitere Prozessierungsschritte und den Transport der pre-miRNAs zuständig sind, beeinflussen. Möglicherweise wird so die Reifung der pre-miRNA zur mature-miRNA verhindert.

Unter dieser Hypothese wurde im Rahmen dieser Dissertation ein Skript (`MaiMirnaStructureSimilarity.pl`) entwickelt, welches die Sekundärstruktur von mutierter und wildtypischer miRNA vorhersagt und miteinander vergleicht. Aus diesem Vergleich resultiert ein relativer Ähnlichkeitswert (Similarity-Score). Ein Wert von 1 beschreibt absolut identische Sekundärstrukturen. Je niedriger ein Similarity-Score ist, desto mehr weichen die beiden 2D-Strukturen voneinander ab. Dieser Similarity-Score fließt im Anschluss in die Annotation der Varianten ein.

[5] `https://earray.chem.agilent.com/suredesign/`

Das `MaiMirnaStructureSimilarity.pl`-Skript zur Berechnung der Ähnlichkeitswerte zwischen mutierter und wildtypischer miRNA wird mit dem Programmaufruf

```
perl MaiMirnaStructureSimilarity.pl
   -dat /media/Berechnungen/AnnotationDBs/miRNAstructure/
      miRNA.dat
   -pos /media/Berechnungen/AnnotationDBs/miRNAstructure/
      miRNA_HG19.bed
   -vcf /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/0001-01.vcf
   -out_folder /media/Berechnungen/200402_NB501654_0172_AHTYGGHBGX/
      0001-01_Exom_NB501654_0172/
   -cutoff 1.1
```

gestartet. Die Datei `miRNA.dat` steht im Downloadbereich der miRBase Datenbank[6] zum Heruntergeladen zur Verfügung. Diese enthält detaillierte Informationen zu allen in der miRBase gespeicherten miRNAs im EMBL-Format. Des Weiteren benötigt das Skript die chromosomale Start- und Endposition, den Strang auf dem die miRNA liegt (+ oder -) und den Namen der entsprechenden miRNA im `BED`-Dateiformat. Abbildung 4.11 beschreibt den Ablauf des Algorithmus.

Zu Beginn der Analyse liest das Skript die `miRNA.dat`-Datei ein, extrahiert die Sequenz jeder miRNA und speichert diese unter Zuordnung zum miRNA Namen in einem Hash zwischen. Im nächsten Schritt durchläuft der Algorithmus die VCF-Datei des untersuchten Patientendatensatzes und überprüft anhand der `miRNA_HG19.bed`-Datei, ob die entsprechende Variante innerhalb einer miRNA liegt. Ist dies der Fall, so erstellt das Skript mit Hilfe der im Hash zwischengespeicherten wildtypischen miRNA-Sequenz, der chromosomalen Positionsangaben und der mutierten Base aus der VCF Datei eine mutierte miRNA-Sequenz. Enthält die Variante eine Thymin-Base (T) wird sie in ihr RNA-äquivalent Uracil (U) transformiert. Falls die miRNA auf dem reversen (-) Strang liegt, ist die Variante in ihre komplementäre Base zu übersetzen. Die wildtypische und die mutierte miRNA-Sequenz werden anschließend in jeweils eigene FASTA-Dateien mit einem für eine Variante eindeutigen Namen bestehend aus Chomosom, Position, Referenzbase und Muatation im Unterordner `miRNA_predictions` des Ausgabeverzeichnisses (`-out_folder`) abgelegt. Die Vorhersage der Sekundärstruktur beider miRNA-Sequenzen übernimmt das `RNAstructure`-Softwarepaket (siehe Anhang IV.XIV) mit folgenden Befehlen:

```
Fold
   <<Chromosom_Position_Referenz-Mutation_WT>>.fa
   <<Chromosom_Position_Referenz-Mutation_WT>>.ct
   -m 1
```

[6] `http://www.mirbase.org/ftp.shtml`

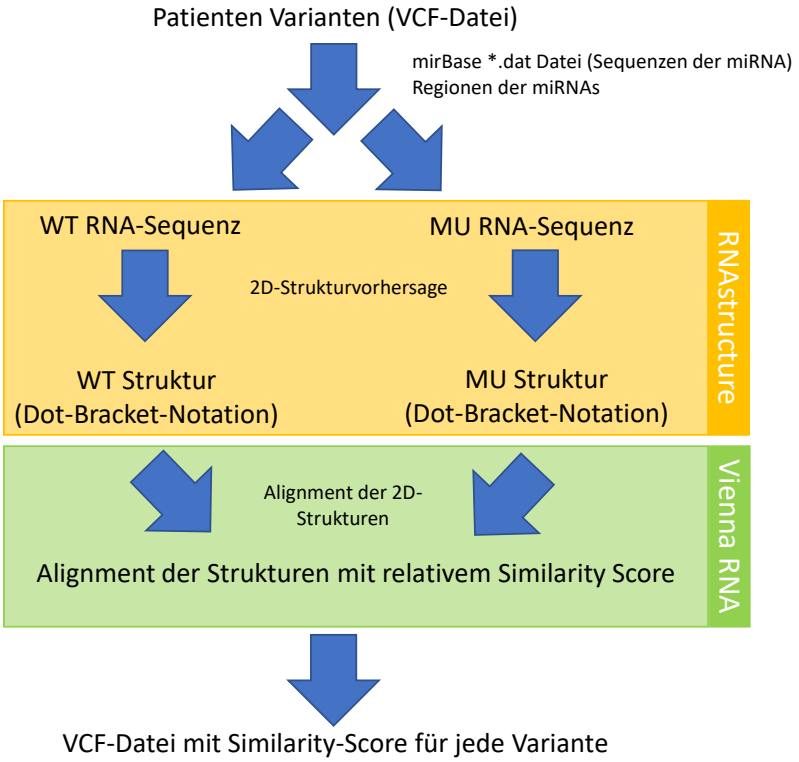


Abbildung 4.11: *Workflow des* `MaiMirnaStructureSimilarity.pl`*-Skripts. Für jede Variante innerhalb einer miRNA wird jeweils für das wildtypische (WT) und mutierte (MU) Allel eine Sekundärstruktur berechnet. Anschließend erfolgt der Vergleich der beiden 2D-Strukturen durch ein Alignment. Aus diesem resultiert ein relativer Ähnlichkeitswert (Similarity-Score), welcher mit der zugehörigen Variante in einer VCF-Datei abgespeichert wird.*

```
Fold
   <<Chromosom_Position_Referenz-Mutation_MU>>.fa
   <<Chromosom_Position_Referenz-Mutation_MU>>.ct
   -m 1
```

Die hieraus resultierenden Strukturen im CT-Dateiformat müssen für die weitere Analyse durch Aufruf von

```
ct2dot
   <<Chromosom_Position_Referenz-Mutation_WT>>.ct
   -1
   <<Chromosom_Position_Referenz-Mutation_WT>>.db
   -f 1
```

```
ct2dot
   <<Chromosom_Position_Referenz-Mutation_MU>>.ct
   -1
   <<Chromosom_Position_Referenz-Mutation_MU>>.db
   -f 1
```

aus dem `RNAstructure`-Softwarepaket in die Dot-Bracket-Notation überführt und mittels

```
cat
   <<Chromosom_Position_Referenz-Mutation_WT>>.db
   <<Chromosom_Position_Referenz-Mutation_MU>>.db
   > <<Chromosom_Position_Referenz-Mutation_COMP>>.db
```

zu einer Datei kombiniert werden. Abbildung 4.12 zeigt exemplarisch eine solche Datei. Mit Hilfe des `RNAforester`-Programms aus dem `Vienna RNA`-Paket

```
1  >chr1_12251808T-G_WT
2  GAGGGCAGCGUGGGUGUGGCGGAGGCAGGCGUGACCGUUUGCCGCCCUCUCGCUGCUCUAG
3  .(((((((((.(((.((((((((((....)).)).)))))))).))).)))))))..
4  >chr1_12251808T-G_MU
5  GAGGGCAGCGUGGGUGUGGCGGAGGCAGGCGUGACCGUGUGCCGCCCUCUCGCUGCUCUAG
6  .(((((((((.(((.((((((..((((....)).)))...))))))).))).)))))))..
```

Abbildung 4.12: *Dot-Bracket-Datei für Alignment der 2D-Struktur. Sie enthält die Sequenzen sowie die 2D-Strukturen des wildtypischen sowie des mutierten Allels in der Dot-Bracket-Notation*

(siehe Anhang IV.XV) wird ein Alignment der in der `<<Chromosom_Position_Referenz-Mutation_COMP>>.db`-Datei enthaltenen wildtypischen und mutierten 2D-Strukturen erstellt. Der verwendete Programmaufruf lautet:

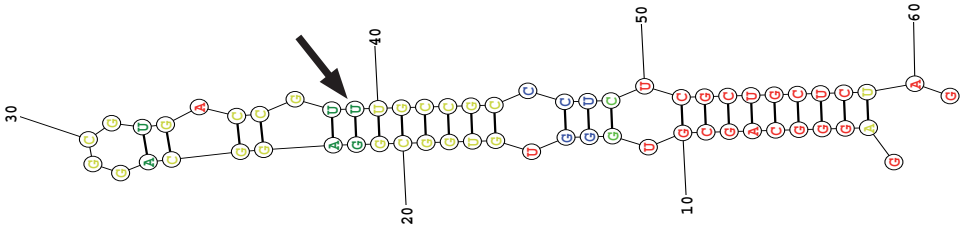
```
RNAforester
   < <<Chromosom_Position_Referenz-Mutation_COMP>>.db
   -r
   > <<Chromosom_Position_Referenz-Mutation_COMP>>.aln
```

Die Datei `<<Chromosom_Position_Referenz-Mutation_COMP>>.aln` trägt Angaben über die verwendeten Parameter (blau), führt das Alignment der miRNA-Sequenz (grün) sowie das Alignment der 2D-Struktur in Dot-Bracket-Notation (gelb) und den relativen Similarity-Score (rot) (siehe Abbildung 4.13 a). Die exemplarisch gezeigten Sekundärstrukturen der wildtypischen und durch die Variante g.chr1:12251808T>G betroffenen Sequenz der miRNA hsa-mir-4632 sind in Abbildung 4.13 b bzw. c dargestellt. Im letzten Schritt extrahiert das `MaiMirnaStructureSimilarity.pl`-Skript den relativen Similarity-Score und speichert ihn gemeinsam mit den Angaben der zugehörigen Variante in einer VCF-Datei ab. Diese wird mit `bgzip` komprimiert, durch `tabix` indiziert und fließt so über den `--custom`-Parameter des VEP-Programms als Datenbank in die Annotation der Varianten ein.

a)

```
*** Scoring parameters ***

Scoring type: global similarity
Scoring parameters:
pair match:       10
pair indel:       -5
base match:       1
base replacement: 0
base indel:       -10


Input string (upper or lower case); & to end for multiple alignments, @ to quit
m....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,....8
global optimal score: 226
0.922449
chr1_12251808T-G_WT      GAGGGCAGCGUGGGUGUGGCGGAGGCAGGCGUGACCGUUUGCCGCCCUCUCGCUG
chr1_12251808T-G_MU      GAGGGCAGCGUGGGUGUGGCGGAGGCAGGCGUGACCGUGUGCCGCCCUCUCGCUG
                         ************************************ ****************

chr1_12251808T-G_WT      CUCUAG
chr1_12251808T-G_MU      CUCUAG
                         ******

chr1_12251808T-G_WT      .((((((((((.(((.((((((((((((....)).)).)))))))).))))).)))))
chr1_12251808T-G_MU      .((((((((((.(((.((((((..((((....)).)).)...)))))).))))).)))))
                         ********************  **************  ****************

chr1_12251808T-G_WT      ))))..
chr1_12251808T-G_MU      ))))..
                         ******
```

b) Wildtypische 2D-Struktur



c) Mutierte 2D-Struktur

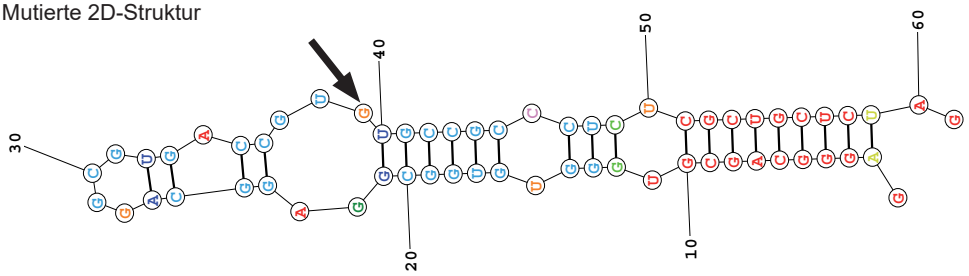


Abbildung 4.13: *Alignment der Sekundärstruktur.*
Dargestellt ist das Sekundärstrukturalignment am Beispiel der Variante g.chr1:12251808T>G betroffenen Sequenz der miRNA hsa-mir-4632. (a) Die Datei trägt Angaben über die verwendeten Parameter (blau), führt das Alignment der miRNA-Sequenz (grün) sowie der 2D-Struktur (gelb) und den relativen Similarity-Score (rot). (b) zeigt die Sekundärstruktur des wildtypischen und (c) des mutierten Allels.

4.5 Graphische Benutzeroberfläche

Da die beschriebenen Pipelines keinerlei graphische Benutzeroberfläche anbieten und die Ergebnisse als Textdatei unter anderem im VCF-Format ausgeben, ist eine Visualisierung dieser resultierenden Textdateien wünschenswert.

Im Folgenden wird die graphische Umsetzung sowie die Implementierung der in Kapitel 3.4 beschriebenen Funktionsanforderungen dieser graphischen Benutzeroberfläche beschrieben.

Die `MaiVarView`-Software ist in Java (siehe Anhang IV.II) programmiert und als ausführbare JAR-Datei (`MaiVarViewer.jar`) exportiert. Diese Datei kann mit dem Befehl

```
java
   -Xmx3000m
   -jar
   MaiVarViewer.jar
```

aus dem Terminal gestartet werden. Um das Programm unter Windows mit einem Doppelklick ausführen zu können, ist der oben beschriebene Programmaufruf in einer Batch-Datei (`*.bat`) abzuspeichern. Nach Öffnen der Software muss sich der Benutzer mit seinem Account für die interne Datenbank anmelden. In dieser Datenbank ist die benutzerspezifische Anordnung der Spalten in der Variantentabelle abgelegt. Zudem ermöglichen die Zugangsdaten das Laden und Anzeigen von patientenspezifischen Daten, wie zum Beispiel Name, Geburtsdatum, Geschlecht und eventuell vorhandene Phänotypinformationen.

Über das Menü `File -> Open VCF File` ist es möglich, die entsprechende annotierte VCF-Datei (`*.annot.vcf`) eines Patientendatensatzes zu laden. Im Hintergrund lädt die Software weitere zur Visualisierung und Auswertung nützliche Dateien des Patientendatensatzes. Dazu zählt das Alignment (`*.recal.bam`), die VCF-Datei mit annotierten CNVs (`*.cnv.annot.vcf`), die Informationen des Gender-Matches (`*.gender`), die Ergebnisse des Sample-Trackings (`*.trk`), die Coverage-Informationen (`*.cov`), die vorgefilterte Liste mit höchst pathologischen Varianten (`*.hotspots.vcf`), die Primer für jede Variante (`*.primer`), die Angabe zur Ploidie jedes Chromosoms (`*.ploidy`) sowie die Informationen über die Verteilung der ROH Regionen des Patienten (`*.roh_stat`). Nach Öffnen des Datensatzes stehen dem Benutzer 216 sogenannte „virtuelle Subpanel“ aus 25 Krankheitskategorien zur Auswahl (siehe Abbildung 4.14). Diese Subpanel bestehen jeweils aus mit der entsprechenden Erkrankung assoziierten Gensets. Durch diese grenzt die Software die Anzahl der auszuwertenden Varianten auf für die Erkrankung relevante Veränderungen ein. Die Gensets werden halbjährlich auf ihre Vollständigkeit überprüft und aktualisiert.

Das sich anschließend öffnende Fenster zeigt im oberen Bereich (siehe Abbildung 4.15 rot) die beim Patienten gefundenen Varianten und deren zugehörigen Annotationen tabellarisch an. Über die Tabs an der linken Seite ist es möglich zwischen allen Varianten, den vorgefilterten Hotspot-Varianten und den CNVs zu wechseln. Unten links werden die aus der Datenbank abgerufenen Patientendaten (Name, Geburtsdatum, Geschlecht, Phänotypinformationen) sowie die

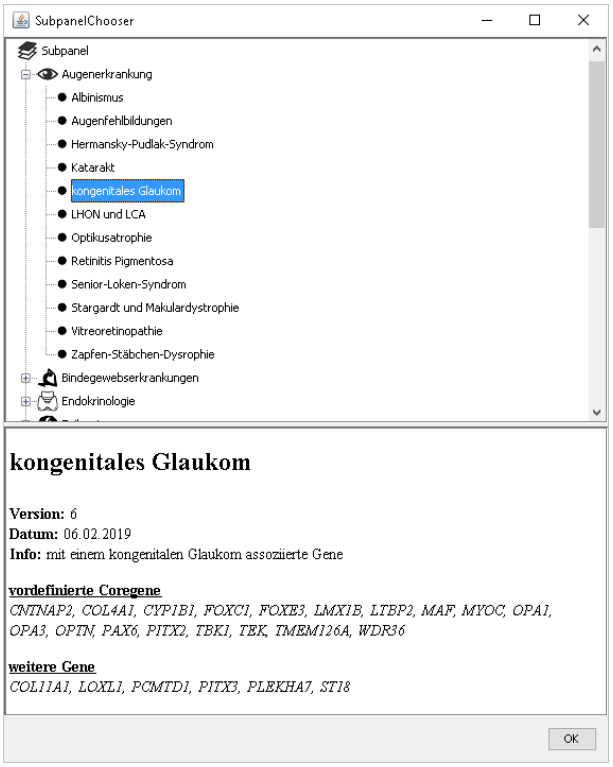


Abbildung 4.14: *Auswahl der Subpanel.*
Die Subpanel bestehen jeweils aus einem mit der entsprechenden Erkrankung assoziierten Genset. Durch dieses kann die Anzahl der Varianten auf für die Erkrankung relevante Veränderungen eingegrenzt werden.

Übereinstimmung des Gender-Matches und des Sample-Trackings aufgeführt (siehe Abbildung 4.15 gelb). Rechts daneben zeigt eine Tabelle die mit dem ausgewählten Gen assoziierten OMIM Einträge falls vorhanden (siehe Abbildung 4.15 blau). Auf der rechten Seite (siehe Abbildung 4.15 grün) befindet sich neben den Filteroptionen für die Variantentabelle ein Button zur Änderung der Subpanelauswahl. Zudem besteht die Möglichkeit die ausgewählte Variante in den alignierten Rohdaten mit Hilfe des IGV anzuzeigen und bezüglich ihrer Pathogenität zu bewerten bzw. zu kommentieren.

Bei einer WES ist es nicht ungewöhnlich, ca. 150.000 Sequenzvarianten pro Patient zu erhalten. Darunter muss die krankheitsverursachende Variante gefunden werden. Dazu stehen in der Auswertesoftware die folgenden Filter zur Verfügung:

IHDB (<5%, <1%, <0,1%) Bei der Auswahl dieses Filters werden alle Varianten, welche häufiger als 5% (bzw. 1% bzw. 0,1%) in der internen Kohorte (alle bis dahin sequenzierten Patienten) des Instituts für Humangenetik Mainz vorkommen, verworfen. Dadurch können neben den häufig in der Population vorkommenden SNPs auch systematische Fehler der Sequenzierung und „Library Preparation“ herausgefiltert werden.

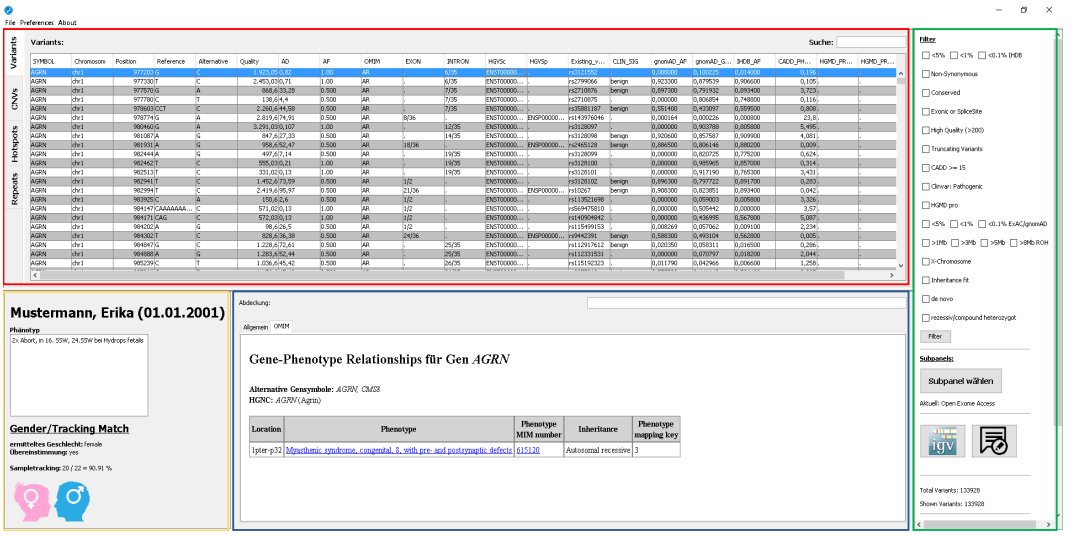


Abbildung 4.15: *Hauptfenster der MaiVarView Software.*
Die beim Patienten gefundenen Varianten und deren Annotationen werden tabellarisch angezeigt (rot). Die Reiter an der linken Seite ermöglichen einen Wechsel zwischen allen Varianten, den vorgefilterten Hotspot-Varianten und den CNVs. Die patientenbezogenen Daten sowie die Ergebnisse des Gender-Matches und des Sample-Trackings werden aufgeführt (gelb). Rechts daneben zeigt eine Tabelle die mit dem ausgewählten Gen assoziierten OMIM Einträge (falls vorhanden) (blau). Neben den Filteroptionen für die Variantentabelle findet sich ein Button zur Änderung der Subpanelauswahl sowie zur Betrachtung der ausgewählte Varianten im IGV (grün).

Non-Synonymous Da die meisten krankheitsverursachenden Varianten die Aminosäurenabfolge des zugehörigen Proteins beeinflussen, kann dieser Filter genutzt werden, um nur solche Varianten anzeigen zu lassen.

Conserved Viele pathologische Varianten befinden sich in hoch konservierten Bereichen des Genoms bzw. eines Gens, da hier meist katalytische Zentren oder Bindungsstellen liegen, welche bei einer Sequenzveränderung zerstört oder zumindest in ihrer Funktion eingeschränkt würden. Mit dieser Filteroption werden nur Varianten angezeigt, die einen PhyloP-Wert >2,5 aufweisen und damit als konserviert gelten.

Exonic or SpliceSite Da es hin und wieder auch krankheitsverursachende Varianten gibt, die keinen unmittelbaren Effekt auf der Proteinebene auslösen, werden durch diesen Filter alle exonischen und SpliceSite-Varianten angezeigt.

High Quality (>200) Während des Variant Callings wird den Varianten ein Qualitätswert zugewiesen. Ab einem Qualitätswert von 200 kann von einer richtig-positiven Variante ausgegangen werden. Somit reduziert dieser Filter die falsch-positiven Varianten.

Truncating Variants Proteintrunkierende Varianten haben meist weitreichendere funktionelle Folgen als Aminosäureaustausche. Dieser Filter zeigt nur trunkierende Varianten an.

CADD >= 15 Dieser Filter zeigt nur Varianten mit einem *in silico* CADD-Pathogenitätswert von größer gleich 15 an. Diese Varianten gelten laut dem im Institut für Humangenetik verwendeten Schwellenwert für den CADD-Vorhersagealgorithmus als pathologisch.

ClinVar Pathogenic Durch Anwahl dieser Filteroption werden nur Varianten angezeigt, die in der ClinVar-Datenbank (siehe Anhang III.III) bereits als „likely pathogenic“ oder „pathogenic“ eingetragen sind.

HGMD Pro Bei Auswahl dieses Filters werden nur die Varianten angezeigt, die in der HGMD-Datenbank (siehe Anhang III.I) als krankheitsverursachend (DM = Disease Mutation) oder wahrscheinlich krankheitsverursachend (DM? = Possible Disease Mutation) gelistet sind.

ExAC/gnomAD (<5%, <1%, <0,1%) Dieser Filter zeigt nur Varianten mit einer Allelfrequenz kleiner 5% (bzw. 1% bzw. 0,1%) der gnomAD Datenbank (siehe Anhang III.VI) an. Hierbei liegen bei älteren prozessierten Fällen die Allelfrequenzen der ExAC und bei den neueren Fällen die der gnomAD Datenbank zugrunde.

X-Chromosome Legt die Stammbaumanalyse eine X-chromosomale Vererbung nahe, kann die Anzeige mit diesem Filter auf Varianten des X-Chromosoms reduziert werden.

Inheritance Fit Die OMIM-Datenbank (siehe Anhang III.IV) beinhaltet Informationen über die Vererbung eines Gens bzw. einer Krankheit. Mit diesem Filter wird das Vererbungsmodell mit der Zygotie der Variante verglichen. Bei Vorliegen eines rezessiven Vererbungsmodells und einer heterozygoten Variante lässt sich diese also herausfiltern. Dabei ist zu beachten, dass durch diesen Filter krankheitsverursachende, compound heterozygote Varianten und Varianten in Genen, welche in OMIM keinem Vererbungsmodell zugeteilt wurden, verloren gehen können.

De novo Bei einer Trioanalyse werden Mutter, Vater und das betroffene Kind gleichzeitig per WES untersucht. In diesem Fall ist es möglich die drei Variantendatensätze miteinander zu vergleichen und die Vererbung (maternal, paternal oder *de novo*) auf das Kind zu bestimmen. Mit diesem Filter würden im Falle einer Trioanalyse nur die *de novo* Varianten angezeigt.

Rezessiv/compound heterozygot Im Falle einer Trioanalyse können mit diesem Filter compound heterozygote und homozygote (Kind homozygot, Mutter und Vater jeweils heterozygot) Varianten angezeigt werden.

Alle Filter sind dynamisch und es besteht die Möglichkeit diese beliebig miteinander zu kombinieren. Sie sind mit einem logischen UND verknüpft. Dies bedeutet, dass die resultierende Variantenmenge die Schnittmenge aller angewandten Filter darstellt.

Durch Doppelklick auf eine Zeile in der Variantentabelle öffnet sich ein Fenster mit zusammenfassenden Informationen zur *in silico* Pathogenitätsvorhersage. Links sind die Werte der Tools gelistet, welche im darunterliegenden „Spiderplot“ visualisiert sind. Je größer der rote Bereich innerhalb dieses Netzes ausfällt, desto pathologischer wird die Auswirkung der Variante vorhergesagt. Auf der rechten Seite sind die für die entsprechende Variante verfügbaren Angaben aus den Datenbanken dbSNP, ClinVar, HGMD, gnomAD, 1000Genomes und der internen Datenbank aufgelistet (siehe Abbildung 4.16). Durch einen

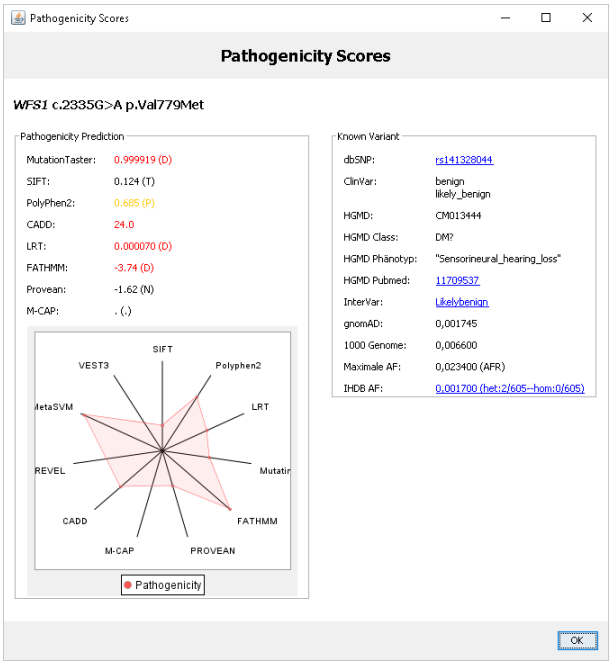


Abbildung 4.16: *Detailansicht der Pathogenitätswerte.*
Im linken Bereich werden die Ergebnisse der einzelnen Pathogenitätsvorhersage-Tools aufgelistet und in einem Spider-Plot visualisiert. Rechts finden sich Informationen aus der dbSNP, ClinVar, HGMD, gnomAD, 1000Genomes und der hausinternen Datenbank zur ausgewählten Variante.

Klick auf das Mausrad kann die überprüfte Variante in der Tabelle rot markiert werden. Dies funktioniert sowohl in der Varianten-, als auch in der Hotspot-Tabelle.

Eine oder mehrere vermeintlich pathologische Varianten können über einen

Rechtsklick auf die Variantenzeile in der Tabelle und anschließender Auswahl von `Add to mutation report` aus dem Kontextmenü dem Mutationsbericht hinzugefügt werden. Über das Menü `Preferences -> Mutation Report` öffnet sich ein neues Fenster (siehe Abbildung 4.17), in dem es möglich ist

Abbildung 4.17: *Bearbeiten des Mutation-Reports.*
Ausgewählte Varianten können mit Notizen versehen oder gelöscht werden. Außerdem besteht die Möglichkeit vorberechnete Primer für die ausgewählte Variante zu bestellen. Dem Report kann eine allgemeine Bemerkung hinzugefügt werden. Zur weiteren Probenbearbeitung lässt sich der Mutation Report als PDF speichern oder drucken.

eine Notiz zu jeder Variante (`Notiz Mutation`) oder zum Report allgemein (`Bemerkung hinzufügen`) einzufügen. Über den `Löschen`-Button lässt sich die angewählte Variante vom Bericht entfernen. Mit Hilfe der während der Prozessierung für jede Variante vorberechneten Primer kann das zur Resequenzierung und Segregationsanalyse mittels Sanger-Sequenzierung benötigte Primerpaar direkt aus der Software zur Bestellung freigegeben werden (siehe Abbildung

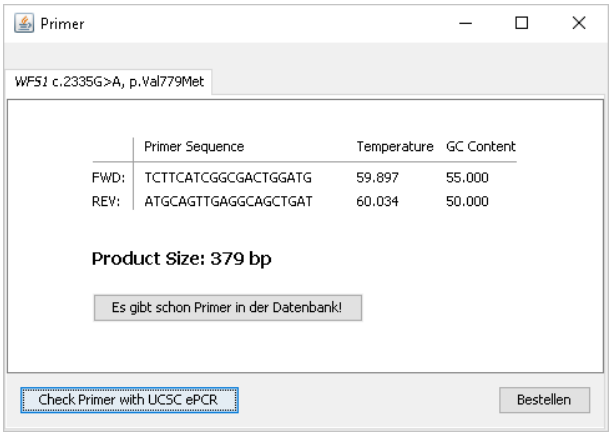


Abbildung 4.18: *Bestellung der Primer.*
Bevor ein Primerpaar für eine Variante bestellt werden kann, ist dieser mit Hilfe des UCSC ePCR Programms auf das Vorkommen häufiger SNPs zu überprüfen.

4.18). Vor der Freigabe ist das entsprechende Primerpaar mit Hilfe des ePCR-Tools des UCSC-Browsers zu überprüfen. Abbildung 4.19 zeigt einen gespeicherten (`Speichern`) bzw. ausgedruckten (`Drucken`) Mutationsbericht.

Universitätsmedizin Mainz Institut für Humangenetik	Molekulargenetisches Labor	Version 01
	Kapitel 5.5.1	Erstelldatum 15.05.2017
	Titel: Auswertung Exom/Masterpanel	Seite 1

Name:	Mustermann, Erika	Journalnummer:	1280/19
Mutter:	N/A	Vater:	N/A
Geburtsdatum:	01.01.2001	Lauf (Illumina ID):	140
10x Coverage:	96 %	Lauf Name:	Exom Run 60
Fingerprint:	90.91 %	Run Datum:	18.09.2019
Untersuchung:	Exom	Report Datum:	06.05.2020
Bearbeiter:	die9s		
Phänotyp:	2x Abort, in 16. SSW, 24.SSW bei Hydrops fetalis		

Allgemeine Bemerkung:

Hier steht eine allgemeine Bemerkung

WFS1 c.2335G>A, p.Val779Met (heterozygot)

g.chr4:6303857G>A **Sequenziertiefe:** 134,97

Transkript: ENST00000226760.1 **Protein:** ENSP00000226760.1

Exon: 8/8 **Intron:** .

Known Variant: rs141328044 CM013444
ClinVar: benign likely_benign
HGMD: CM013444 (DM?) "Sensorineural_hearing_loss"

1k Genome:	0.0066	**SIFT:**	tolerated(0.15)
Max AF:	0.0234 (AFR)	**PolyPhen2:**	possibly_damaging(0.685)
IHDB:	0.0017 het:2/605--hom:0/605	**CADD:**	24.0
InterVar:	Likelybenign	**MutationTaster:**	D (0.999919)

Subpanel: Open Exome Access

Filter: N/A

Notiz:

Hier steht eine Notiz zur Variante

Erstellt durch: S. Diederich	Geprüft durch: S. Springer	Freigegeben durch: Dr. J. Winter
Datum: 15.05.2017 gez. S. Diederich	Datum: 15.05.2017 gez. S. Springer	Datum: 15.05.2017 gez. Dr. J. Winter
Datenquelle: K:\Qualitätsmanagement\Neufassung des QMH\Dokumente zu QMH-Kapitel\5_Technische Anforderungen\5.5 Untersuchungsverfahren\Mole\Auswertebögen\Auswertung Exom 15.05.2017.pdf		

Abbildung 4.19: *Beispiel eines Mutation-Reports. Der Mutation-Report besteht aus einem Kopfbereich, der patientenspezifische Daten enthält. Anschließend steht die allgemeine Bemerkung (falls angegeben). Jede Variante wird mit Angaben über die Position im Genom und die Auswirkung auf das entsprechende Gen aufgelistet. Zudem sind Informationen der Populations und Variantendatenbanken sowie der Pathogenitätsvorhersage-Tools aufgelistet. Die variantenspezifische Notiz steht unterhalb der Variantendaten.*

Die CNVs werden einerseits tabellarisch mit Start- und Endposition im Genom, der entsprechenden Länge der Kopienzahlveränderung in Basenpaaren sowie den betroffenen Genen und Exons bzw. Introns angezeigt. Zudem beinhaltet die Tabelle Informationen über die vorliegende Kopienzahl, die Anzahl der in diesem Bereich liegenden Exons und die Qualität der CNV-Berechnung. Graphisch ist im oberen Bereich der Anzeige jede CNV als roter Balken neben dem entsprechenden Chromosomenpaar dargestellt. Liegen in der `*.ploidy`-Datei Angaben über eine Aneuploidie eines oder mehrerer Chromosomen vor, so wird dies durch ein entsprechend fehlendes oder zusätzliches Chromosomenpiktogramm dargestellt. Die lilafarbenen Balken zeigen die Bereiche mit ROH im Patientengenom an. Rechts neben dieser karyotypischen Darstellung finden sich Informationen über den prozentualen Anteil an ROHs mit Bereichen größer 8Mbp, 5Mbp, 3Mbp und 1Mbp (siehe Abbildung 4.20).

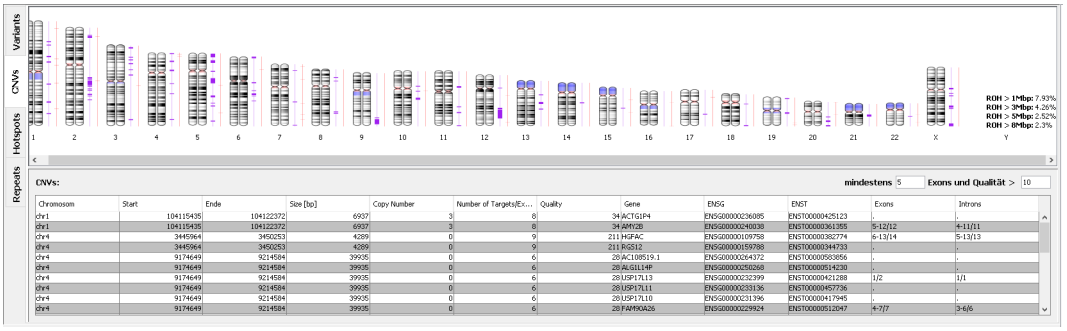


Abbildung 4.20: *Darstellung der CNVs, ROHs und Aneuploidien. Informationen über die vorliegende Kopienzahl und den davon betroffenen Bereich sind tabellarisch dargestellt. Im oberen Bereich ist jede CNV graphisch als roter Balken neben dem entsprechenden Chromosomenpaar angezeigt. Liegen Angaben über eine Aneuploidie eines oder mehrerer Chromosomen vor, so wird dies durch ein entsprechend fehlendes oder zusätzliches Chromosomenpiktogramm visualisiert. Lilafarbene Balken zeigen die Bereiche mit ROH im Patientengenom an. Rechts neben der karyotypischen Darstellung finden sich Informationen über den prozentualen Anteil an ROHs.*

Kapitel 5

Ergebnisse

5.1 Target Enrichment Sequencing

Zur Abklärung möglicher genetischer Ursachen von seltenen pädiatrischen Erbkrankheiten wird zunächst eine Chromosomenanalyse durchgeführt (Aufklärungsrate 3%). Anschließend folgt je nach klinischem Befund eine Untersuchung auf das Vorliegen eines Fragilen-X-, Angelman- oder Rett-Syndroms (Aufklärungsrate 1%). Darüber hinaus wird mithilfe der SNP-Array-Technologie nach submikroskopischen Deletionen und Duplikationen im Genom gesucht (Aufklärungsrate 15%). Zusammengenommen führen diese Untersuchungen bei 19% der Fälle zu einer diagnostischen Klärung.

5.1.1 MPIMG1-Test

Der MPIMG1-Test ist speziell zur Diagnostik von seltenen Erbkrankheiten erstellt worden. Er umfasst 1222 mit geistiger Behinderung und verwandten Störungen assoziierte Gendefekte. Die mit diesem TES-Panel untersuchten Patienten weisen eine geistige Behinderung (IQ <70), eine motorische Entwicklungsverzögerung oder einen regressiven neurodegenerativen Verlauf auf. Zudem eignet sich dieser Test zur molekularen Diagnose genetisch bedingter Störungen, die mit Autismus (engl. autism spectrum disorders, ASD), dem Aufmerksamkeits-Defizit-Hyperaktivitäts-Syndrom (ADHS) und anderen Verhaltensauffälligkeiten einhergehen.

Zwischen November 2013 und Dezember 2016 wurde der oben beschriebene MPIMG1-Test am Institut für Humangenetik der Universitätsmedizin Mainz als ergänzende Untersuchung im Rahmen der Stufendiagnostik eingesetzt. In diesem Zeitraum erfolgte die Befundung von insgesamt 257 Patienten. Bei 146 Patienten (56,81%) konnte keine pathogene Sequenzvariante identifiziert werden. Bei 15 Patienten (5,84%) lag eine wahrscheinlich gutartige Variante (Pathogenitätsstufe 2) vor, bei 29 Patienten (11,28%) fanden sich Varianten unklarer Signifikanz (VUS, Pathogenitätsstufe 3), für die anhand der verfügbaren Literatur keine eindeutige Klassifizierung möglich war. 48 der Patienten (18,68%) wiesen wahrscheinlich pathogene Varianten (Pathogenitätsstufe 4) auf. Dabei handelt es sich um Veränderungen, die bisher in der Literatur nicht beschrieben

wurden, welche aber aufgrund ihrer Beschaffenheit zu einem Verlust der Proteinfunktion führen. 19 Patienten (7,39%) trugen bereits bekannte, sicher als pathogen eingeschätzte Sequenzvarianten (Pathogenitätsstufe 5), die bereits früher bei Patienten beobachtete wurden. Abbildung 5.1 zeigt die Verteilung der Pathogenitätsstufen aller 257 mit dem MPIMG1-Test untersuchten Patienten.

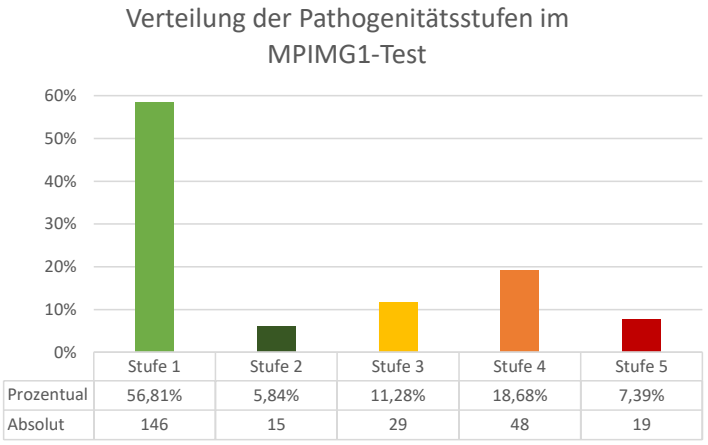


Abbildung 5.1: *Verteilung der Pathogenitätsstufen im MPIMG1-Test*

Auch wenn mit 63% für fast zwei Drittel (Stufe 1 und 2) der Patienten keine für die vorliegende Symptomatik ursächliche Variante detektiert werden konnte, so war es möglich, die allgemeine Aufklärungsrate der Stufendiagnostik durch die Etablierung des MPIMG1-Tests um 26% (Stufe 4 und 5) auf 45% zu erhöhen (siehe Abbildung 5.2).

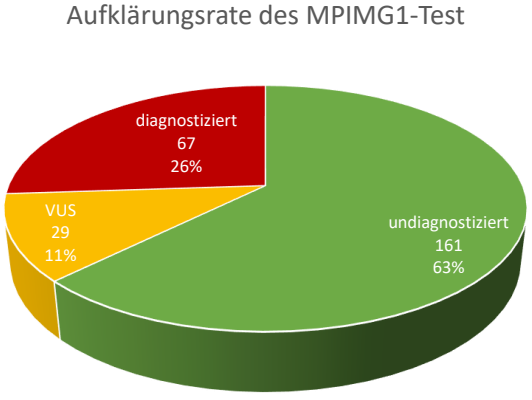


Abbildung 5.2: *Aufklärungsrate des MPIMG1-Tests*

5.1.2 Exom-Sequenzierung

Die Einführung der Whole-Exome-Sequenzierung am Institut für Humangenetik im Oktober 2016 löste die Diagnostik mittels MPIMG1-Test als letzte

Stufe der Diagnostik ab. Zudem ist es durch die Sequenzierung aller bekannten Gene des menschlichen Genoms möglich, das Spektrum der zu untersuchenden Erkrankungen im Vergleich zum MPIMG1-Test zu erweitern. Bis Ende April 2020 wurden 673 Patienten einer WES unterzogen. Da für 21 Patienten noch kein abschließender Bericht vorliegt, beschränkt sich die folgende Auswertung auf die 652 befundeten Patienten (siehe Abbildung 5.3). In 342 Patienten

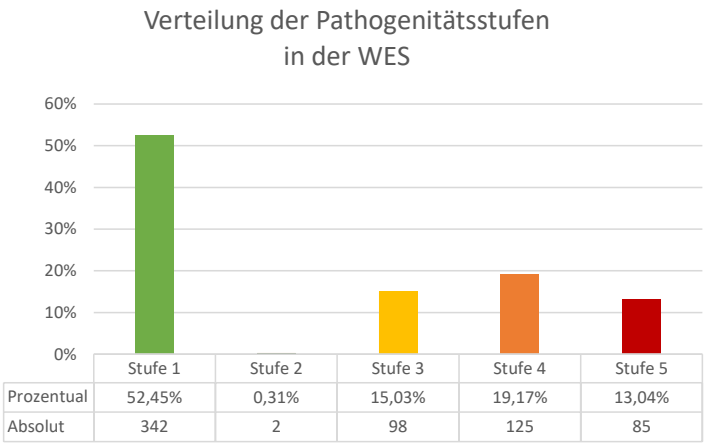


Abbildung 5.3: *Verteilung der Pathogenitätsstufen in der WES*

(52,42%) konnte keine ursächliche Variante (Pathogenitätsstufe 1) gefunden werden. Zwei Patienten (0,31%) erhielten ein wahrscheinlich unauffälliges Ergebnis (Pathogenitätsstufe 2). Bei 98 Patienten (15,03%) musste die gefundene Variante aufgrund der Datenlage als VUS (Pathogenitätsstufe 3) eingestuft werden. Eine wahrscheinlich pathologische Variante (Pathogenitätsstufe 4) lag in 19,17% der Fälle (125 Patienten) vor. Die in 85 Patienten (13,04%) gefundenen kausalen Varianten sind in der Literatur bereits als krankheitsverursachend beschrieben (Pathogenitätsstufe 5).

Die Zusammenfassung der Pathogenitätsstufen 4 und 5 ergibt eine Aufklärungsrate von 32% für die Whole-Exome-Sequenzierung. In 53% der Fälle konnte keine krankheitsverursachende Variante gefunden werden (Pathogenitätsstufen 1 und 2). Der Anteil der aufgrund von fehlender Literatur nicht einschätzbaren Varianten (Pathogenitätsstufe 3) liegt bei 15% (siehe Abbildung 5.4).

Der Vergleich des MPIMG1-Tests mit der Exom-Sequenzierung zeigt eine um 6% erhöhte Aufklärungsrate durch die WES. Die Anzahl der undiagnostizierten Fälle sinkt um etwa 10% von 62,65% beim MPIMG1-Test auf 52,76% in der WES. Die Differenz ist durch einen etwa 4%igen Anstieg der Varianten unklarer Signifikanz in der Whole-Exome-Sequenzierung zu erklären (siehe Abbildung 5.5). Beim Betrachten der Werte für die einzelnen Pathogenitätsstufen fällt auf, dass Stufe 2 in der Exom-Sequenzierung mit 0,31% im Gegensatz zu 5,84% im MPIMG1-Tests kaum präsent ist. Ebenso prägnant ist der Anstieg der als sicher pathologisch eingestuften Varianten von 7,39% im MPIMG1-Test auf 13,04% in der WES. In der Exom-Sequenzierung sind zudem vermehrt (et-

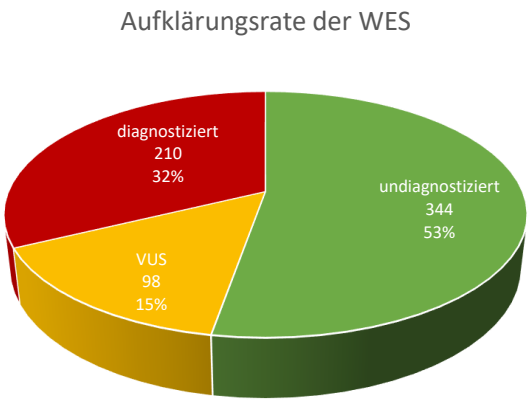


Abbildung 5.4: *Aufklärungsrate der WES*

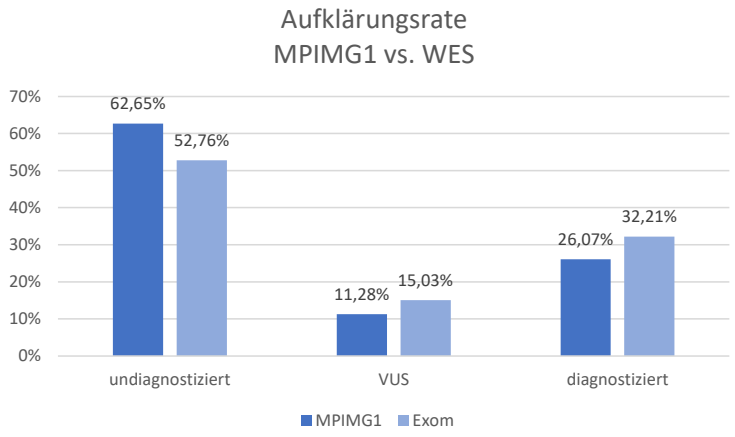
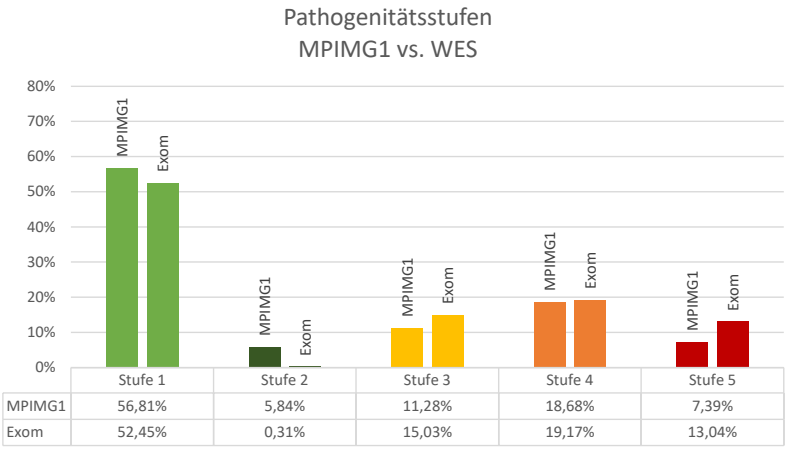


Abbildung 5.5: *Aufklärungsrate des MPIMG1-Tests im Vergleich zur WES*

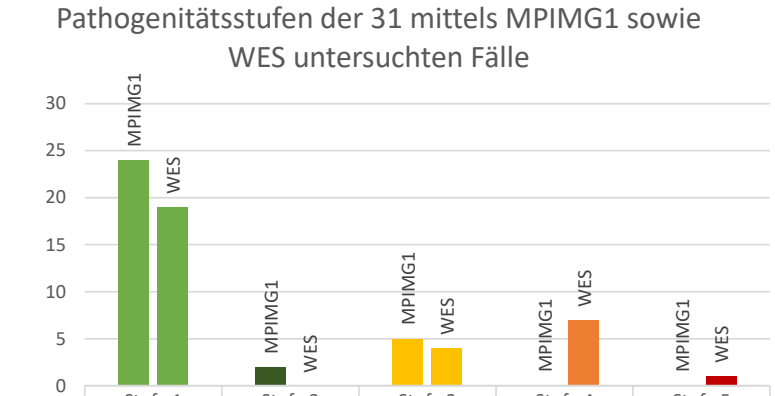
wa 4%) Varianten unklarer Signifikanz detektiert worden, wobei der Anteil von wahrscheinlich pathologischen Varianten mit 18,68% und 19,17% für den MPIMG1-Test beziehungsweise die WES nahezu gleich bleibt. Abbildung 5.6 zeigt vergleichend die relative Anzahl an Varianten der jeweiligen Pathogenitätsstufe.

31 Patienten mit negativem (Stufe 1 und 2) oder unklarem Ergebnis (Stufe 3) des MPIMG1-Tests wurden außerdem mittels Exom-Sequenzierung untersucht. Bei rund einem Viertel (8) war es so zusätzlich möglich, mithilfe der WES die für die Symptomatik des jeweiligen Patienten verantwortliche Veränderung aufzuklären (Pathogenitätsstufe 4 und 5). Eine dieser Varianten ist in der Literatur bereits als krankheitsverursachend beschrieben (Stufe 5), die restlichen 7 sind aufgrund analoger publizierter Befunde als wahrscheinlich pathogen (Stufe 4) anzusehen. Zudem konnte eine Variante mit unklarer Signifikanz sowie zwei vermutlich gutartige Varianten auf der Grundlage aktueller Literatur und Datenbanken neu bewertet und als nicht pathogen (Stufe 1) eingestuft werden (siehe Abbildung 5.7).



	Stufe 1	Stufe 2	Stufe 3	Stufe 4	Stufe 5
MPIMG1	56,81%	5,84%	11,28%	18,68%	7,39%
Exom	52,45%	0,31%	15,03%	19,17%	13,04%

Abbildung 5.6: *Verteilung der Pathogenitätsstufen im Vergleich zwischen MPIMG1-Test und WES*



	Stufe 1	Stufe 2	Stufe 3	Stufe 4	Stufe 5
MPIMG1	24	2	5	0	0
WES	19	0	4	7	1

Abbildung 5.7: *Verteilung der Pathogenitätsstufen der 31 mittels MPIMG1 sowie WES untersuchten Fälle*

5.1.3 NGS: Eine Revolution der klinischen Diagnostik

Mit der Etablierung des MPIMG1-Tests zur Ergänzung der konventionellen Stufendiagnostik wurde die NGS-Technologie am Institut für Humangenetik der Universitätsmedizin Mainz eingeführt. Bei Patienten mit psychomotorischen Entwicklungsstörungen hatte dies einen hochsignifikanten Anstieg der Aufklärungsrate um 26% zur Folge. Somit konnte mit Hilfe dieser erweiterten Stufendiagnostik bei insgesamt 45% der Patienten eine Diagnose gestellt werden. Ein Vergleich der Detektionsrate des MPIMG1-Tests mit anderen publizierten Panel-Tests ist jedoch schwierig, da Target Enrichment-Panels meist für die Diagnose unterschiedlicher genetisch bedingter Krankheiten konzipiert sind und daher in aller Regel verschiedene Gene abdecken.

Nach einer kürzlich erschienenen Untersuchung betrug die durchschnittliche Aufklärungsrate von im Jahre 2019 publizierten WES-gestützten diagnosti-

schen Untersuchungen 30% [Tzur *et al.*, 2020]. Dies ist vergleichbar mit der Aufklärungsrate von 32% durch die WES am Institut für Humangenetik Mainz.

Der Rückgang der als wahrscheinlich nicht pathogen eingestuften Varianten (Stufe 2) von 5,84% im MPIMG1-Test auf 0,31% in der WES lässt darauf schließen, dass die Gen- und Genomanalyse bei immer größeren Kohorten von Probanden und gesunden Kontrollen (z.B. gnomAD) zu einer Verbesserung der Diskrimination gutartiger und pathogener Sequenzvarianten geführt hat. So gelang es, unter anderem durch stetig wachsende Datenbanken mit Genotyp-Phänotyp-Informationen (z.B. OMIM, ClinVar, HGMD oder Orphanet), in der WES mit 13,04% bei fast doppelt so vielen Fällen eine sicher pathogene Variante zu identifizieren (Stufe 5) als im MPIMG1-Test mit 7,39%. Dieser Anstieg der durch WES gelösten Fälle könnte zudem auf die größere Anzahl an untersuchten Genen zurückzuführen sein. Die mit 15,03% recht hohe Rate an Varianten unklarer Signifikanz lässt sich durch die Vielzahl von Genen erklären, deren Krankheitsbezug bisher noch ungeklärt ist.

Unter Einbeziehung der Pathogenitätsstufen 4 und 5 führt die WES (32%) zu einer um 6% höheren Rate an gelösten Fällen als der MPIMG1-Test (26%). In den 31 Patienten, die sowohl mit dem MPIMG1-Test als auch durch WES untersucht wurden, liegt die Rate der durch die WES zusätzlich aufgeklärten Patienten bei 25%. Dies läßt sich zum Teil durch die Analyse zusätzlicher Gene erklären, jedoch haben andere Studien gezeigt, dass schon eine systematische Reanalyse von WES-Daten ein bis zwei Jahre nach der initialen Untersuchung die Diagnoserate um durchschnittlich 10% (5-22%) erhöht [Baker *et al.*, 2019] [Berkovic *et al.*, 2019] [Bruel *et al.*, 2019] [Liu *et al.*, 2019b] [Salfati *et al.*, 2019]. Diese Steigerung der diagnostischen Erfolgsrate beruht größtenteils auf der bereits erwähnten laufenden Aktualisierung von Referenzdatenbanken und auf verbesserten Algorithmen zur Detektion, Priorisierung und Pathogenitätseinstufung von Varianten. Die noch höhere Aufklärungsrate der hier „reanalysierten“ Fälle lässt sich vor Allem durch die große Zahl der bei der WES zusätzlich analysierten Gene erklären.

5.2 RegVar Panel

Da es auch mithilfe der Exom-Sequenzierung nicht gelingt, mehr als einen Teil der monogenen Krankheiten aufzuklären, wurde ein Target-Enrichment-Genpanel zur Detektion von Varianten in 3'UTRs und mikroRNAs entwickelt (siehe Kapitel 4.4.1). Bei der Analyse der detektierten Varianten steht die Veränderung eines regulatorischen Bereichs, wie zum Beispiel einer mikroRNA-Bindestelle im 3'UTR eines krankheitsassoziierten Gens sowie die Veränderung auf Seiten der mikroRNAs im Vordergrund.

Alle insgesamt 57 mit diesem TES Panel untersuchten Patienten hatten unauffällige Befunde in der Stufendiagnostik bis hin zur Exom-Sequenzierung und weisen eine Intelligenzminderung (syndromal: 27; nicht syndromal 10), Entwicklungsverzögerung (12), Autismus-Spektrum-Störung (4), Epilepsie (3)

oder Sprachentwicklungsstörung (1) auf (siehe Abbildung 5.8).

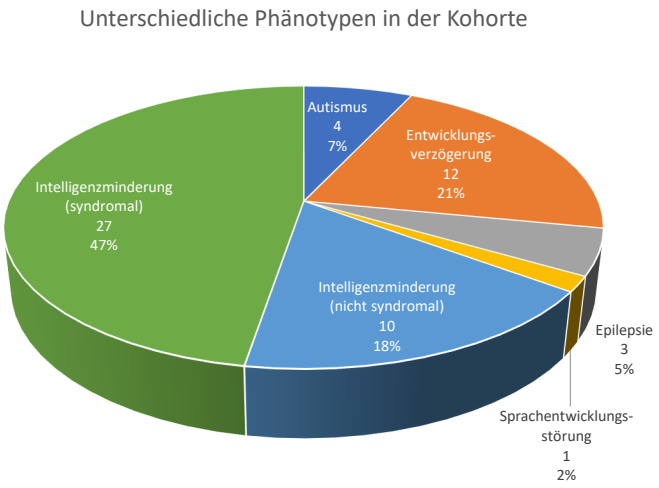


Abbildung 5.8: *Phänotypen der Kohorte des 3'UTR/miRNA Panels*

Die Auswertung der Daten erfolgte mit der in Kapitel 4.4.2 dargelegten Pipeline und der in Kapitel 4.5 beschriebenen graphischen Benutzeroberfläche. Patient 5 (m, 13 Jahre) stellte sich mit einer Autismus-Spektrum-Störung, ADHS, Wutanfällen sowie einer Lese- und Rechtschreibschwäche in der Genetischen Beratungsstelle des Instituts für Humangenetik Mainz vor. Bei ihm konnte die Variante c.*651C>T im *IL1RAPL1*-Gen auf dem X-Chromosom in hemizygoter Form festgestellt werden. Eine Segregationsanalyse mittels PCR und Sanger-Sequenzierung (siehe Anhang V.I) zeigte, dass seine bis auf ein ADHS unauffällige Mutter diese Variante heterozygot trägt. Deletionen und trunkierende Varianten des *IL1RAPL1*-Gens sind in der Literatur als ursächlich für die X-chromosomale Mentale Retardierung 21/34 beschrieben (OMIM #300143) [Kozak *et al.*, 1993] [Piton *et al.*, 2008] [Franek *et al.*, 2011]. Diese weist ein variables Erscheinungsbild auf und äußert sich unter anderem durch autistische Merkmale, ein ADHS, eine Intelligenzminderung und leichte Gesichtsdysmorphien.

Der detektierte Austausch einer Cytosin zu einer Thymin Base liegt im 3'UTR des *IL1RAPL1*-Gens, 651 Basen hinter dem Stop-Codon in der Bindungsstelle für die miRNAs miR-6810-3p und miR-6801-3p. Zur Überprüfung der Auswirkung dieser Variante wurden alle bekannten miRNAs der miRBase (siehe Anhang III.VIII) mit den *in silico* Target Prediction Tools miRanda [John *et al.*, 2004] und RNAhybrid [Krüger und Rehmsmeier, 2006] jeweils gegen die wildtypische als auch gegen die mutierte Sequenz getestet. miRanda beschreibt einen Verlust der Bindestelle für die miRNAs miR-6801-3p und miR-6810-3p durch den Basenaustausch. Allerings entsteht laut miRanda-Software durch diese Variante auch eine neue Bindestelle für die miRNAs miR-4672 und miR-5195-5p. RNAhybrid erkennt lediglich für miR-4672 und miR-6801-3p eine leicht reduzierte Bindungsaffinität an den mutierten 3'UTR des *IL1RAPL1*-Gens. Damit stimmen die Vorhersagen der beiden Target Prediction Algo-

rithmen miRanda und RNAhybrid nur für die miRNA miR-6801-3p überein. Tabelle 5.1 fasst die Ergebnisse der beiden Target Prediction-Tools zusammen.

Tabelle 5.1: *Target Prediction von miRanda und RNAhybrid*

miRNA	**3'UTR Kondition**	**miRanda**	**RNAhybrid**
miR-6801-3p	wildtypisch	bindet	Affinität: -28,1 kcal/mol
miR-6801-3p	mutiert	bindet nicht	Affinität reduziert (-26,3 kcal/mol)
miR-6810-3p	wildtypisch	bindet	Affinität: -31,3 kcal/mol
miR-6810-3p	mutiert	bindet nicht	Affinität gleich (-31,3 kcal/mol)
miR-4672	wildtypisch	bindet nicht	Affinität: -24,9 kcal/mol
miR-4672	mutiert	bindet	Affinität reduziert (-23,2 kcal/mol)
miR-5195-5p	wildtypisch	bindet nicht	Affinität: -26,3 kcal/mol
miR-5195-5p	mutiert	bindet	Affintät gleich (-26,3 kcal/mol)

Zur Überprüfung der *in silico* vorhergesagten Bindungsaffinitäten der vier mikroRNAs an die wildtypische und mutierte 3'UTR Sequenz des *IL1RAPL1*-Gens dienen Luziferase-Assays (siehe Anhang V.II). Es erfolgte die Durchführung von zwei Batches á drei Versuche mit jeweils drei technischen Replikaten. Im ersten Batch wurden alle 4 der in Tabelle 5.1 gelisteten miRNAs einbezogen. Batch 2 enthielt nur die miRNAs miR-6801-3p und miR-5195-5p, da diese in Batch 1 eine auffällige beziehungsweise teilweise leicht signifikante Tendenz aufwiesen. Abbildung 5.9 zeigt die Mittelwerte und die zugehörigen Standardabweichungen der technischen Replikate über die 3 Versuche von Batch 1 (oben) und Batch 2 (unten). In Batch 1 liegt im Vergleich zum wildtypischen 3'UTR Konstrukt eine signifikant ($p = 0{,}049$) niedrigere Luziferaseaktivität des Konstrukts mit mutierter 3'UTR Sequenz für die miRNA-6801-3p vor. Bei miRNA-5195-5p ist eine auffällige aber nicht signifikante Tendenz ersichtlich. Hier ist die Luziferaseaktivität bei mutiertem 3'UTR bedeutend niedriger als im Konstrukt mit dem wildtypischem 3'UTR. Diese Signifikanz bei miRNA-6801-3p sowie die deutliche Differenz bei miRNA-5195-5p konnte durch Batch 2 nicht bestätigt werden. Die Unterschiede korrelieren zwar mit den Ergebnissen aus Batch 1, sind aber weniger ausgebildet und nicht signifikant.

Wie in Abbildung 5.9 ersichtlich liegen die Luziferaseaktivitätswerte in Batch 2 im Allgemeinen deutlich niedriger als in Batch 1. Um diesen Batch-Effekt zu beseitigen und eine Vergleichbarkeit herzustellen, wird jeweils ein Wert der Kontrolle pro Kondition in jedem Versuch auf 1 gesetzt und die entsprechenden Luziferaseaktivitätswerte auf diese Weise normalisiert (siehe Abbildung 5.10). Nach dieser Normalisierung fallen sehr hohe Standardabweichungen der gemessenen Konditionen auf. Auch die Unterschiede zwischen der mutierten und wildtypischen Kondition bei miR-6801-3p und miR-5195-5p sind kaum noch sichtbar.

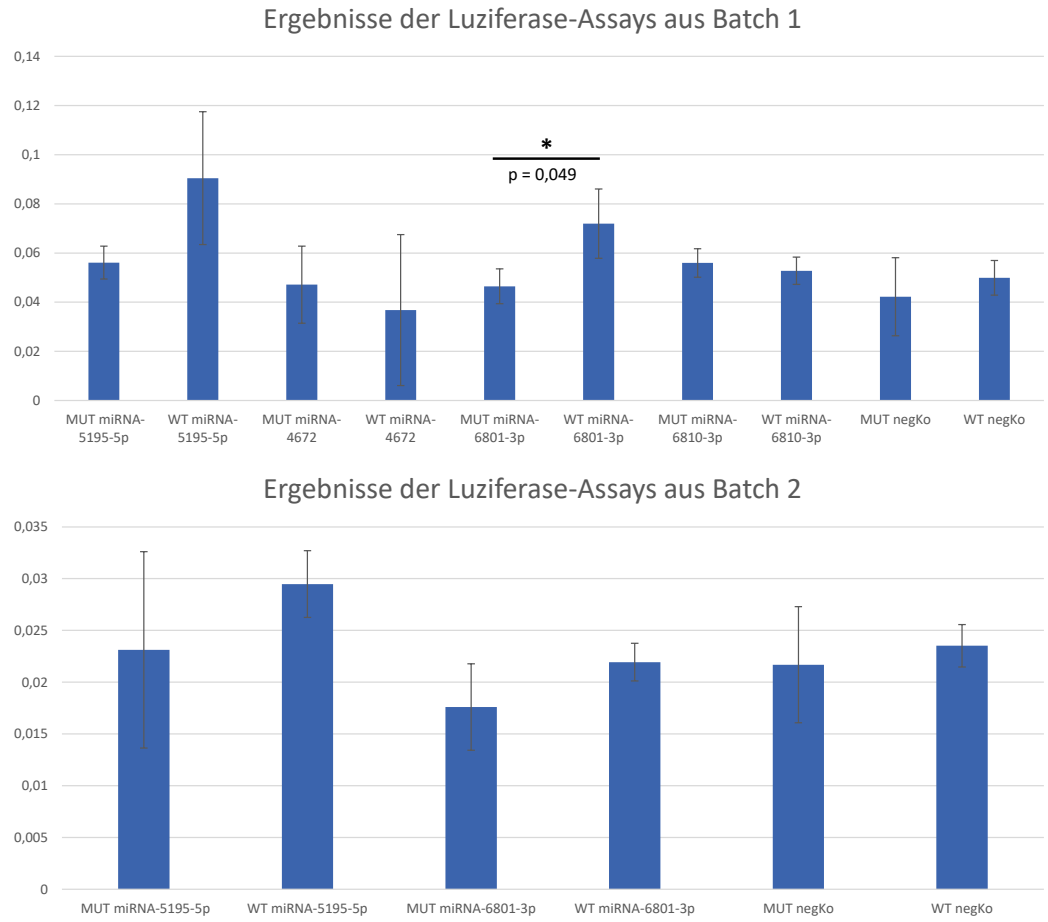


Abbildung 5.9: *Ergebnisse der Luziferase-Assays*

Zusammenfassend zeigt miRNA miR-5195-5p zwar *in vitro* durch eine reduzierte Luzifereaseaktivität des mutierten Konstrukts eine ähnliche Tendenz wie von der Software miRanda *in silico* vorhergesagt. Im Falle von miR-6801-3p widersprechen sich *in silico* und *in vitro* Ergebnisse hingegen, da im mutierten Konstrukt eine reduzierte Luziferaseaktivität vorliegt. Diese deutet allerdings auf eine erhöhte Bindungsaffinität der miRNA mir-6801-3p an die mutierte 3'UTR Sequenz des *IL1RAPL1*-Gens hin.

Zur Analyse der Varianten in miRNAs wurden alle 249.856 Varianten der 57 untersuchten Patienten in einer Datei kombiniert und mit dem `MaiMirnaStructureSimilarity.pl`-Skript prozessiert (siehe Kapitel 4.4.2). Durch die Annotation der knapp 250.000 Varianten mit den im vorherigen Schritt erhaltenen Similarity-Scores und der gnomAD-Allelfrequenz ist es nun möglich diese nach Relevanz zu filtern. 11.865 Varianten liegen in miRNA-Genen, kommen aber teilweise häufig in der untersuchten Kohorte vor. Nach Verwerfen der Varianten mit einer gnomAD-Allelfrequenz von mehr als 1%, einem Vorkommen in mehr als zwei Patienten der Kohorte sowie einer Eingrenzung des Similarity-Scores von 0% bis 70% Ähnlichkeit zwischen mutierter und wildtypischer Sekundärstruktur reduziert sich die Anzahl auf 23 Varianten in 17 verschiedenen betroffenen miRNAs. Abbildung 5.11 zeigt die 23 Varianten

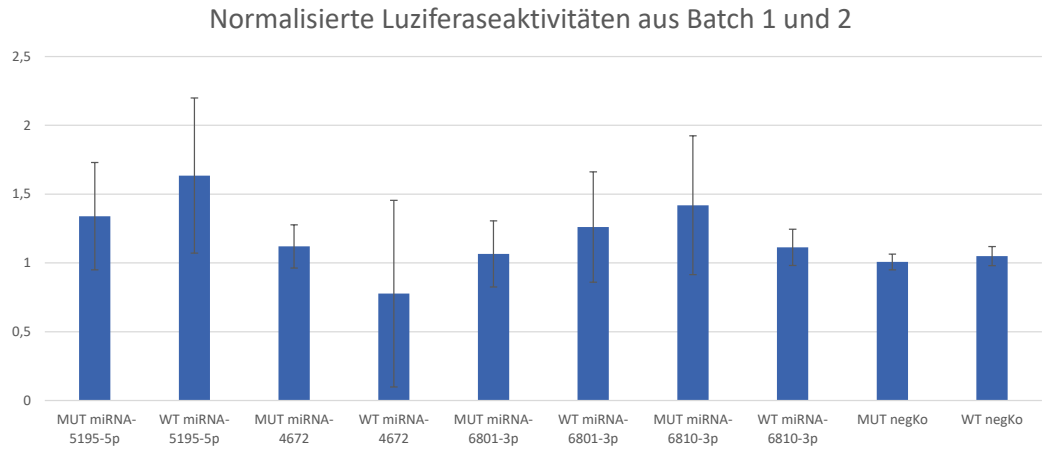


Abbildung 5.10: *Normalisierte Ergebnisse der Luziferase-Assays*

mit ihrer Häufigkeit in der Kohorte (rot = 1x; blau = 2x), dem zugehörigen Similarity-Score (X-Achse) und der entsprechenden gnomAD-Allelfrequenz (Y-Achse).

Tabelle 5.2 listet die 17 betroffenen miRNAs mit Informationen zu ihrer entsprechenden Expression im Blut und Gehirn [Petryszak *et al.*, 2016] sowie der durch das Online-Tool Target Scan [Agarwal *et al.*, 2015] vorhergesagten Anzahl an Zielgenen auf. Bis auf die beiden miRNAs mir-200a und mir-3666 beschreibt Target Scan die restlichen miRNAs als nicht vertrauenswürdig und die vorhergesagten Zielgene hauptsächlich als falsch-positive Ergebnisse. Aufgrund dessen werden im Folgenden nur mir-200a und mir-3666 näher betrachtet.

Mit Hilfe der Gene Ontology (GO) Anreicherungsanalyse [Ashburner *et al.*, 2000] der jeweiligen Zielgene der beiden miRNAs mir-200a und mir-3666 lässt sich evaluieren, welche biologischen Prozesse überrepresentativ beeinflusst sind. Die signifikanten (Bonferroni p-Wert <= 0,05) Ergebnisse der funktionellen Annotation durch das Online-Tool DAVID [Huang *et al.*, 2009] sind in Abbildung 5.12 und 5.13 für die jeweilige miRNA graphisch dargestellt.

Die prognostizierten Zielgene von mir-200a scheinen eine Rolle in der Neurogenese zu spielen. Mit jeweils 26 Genen sind die biologischen Prozesse der Neuronen-Projektionsmorphogenese (2,65-fach), der an der Neuronendifferenzierung beteiligten Zellmorphogenese (2,7-fach) und der Axogenese (2,93-fach) verstärkt von putativen Veränderungen der miRNA mir-200a betroffen. 29 der Zielgene wirken in der Entwicklung der Neuronenprojektion (2,46-fach) mit und 47 stehen im Zusammenhang mit der neuronalen Differenzierung (2,33-fach). Die weiteren signifikant angereicherten biologischen Prozesse hängen unter anderem mit der Transkription der DNA sowie der allgemeinen Morphogenese und Differenzierung von Zellen zusammen (siehe Abbildung 5.12). Bei der entsprechenden Variante handelt es sich um einen heterozygoten Austausch der Base Cytosin nach Thymin an Position 42 von mir-200a. Eine Segregati-

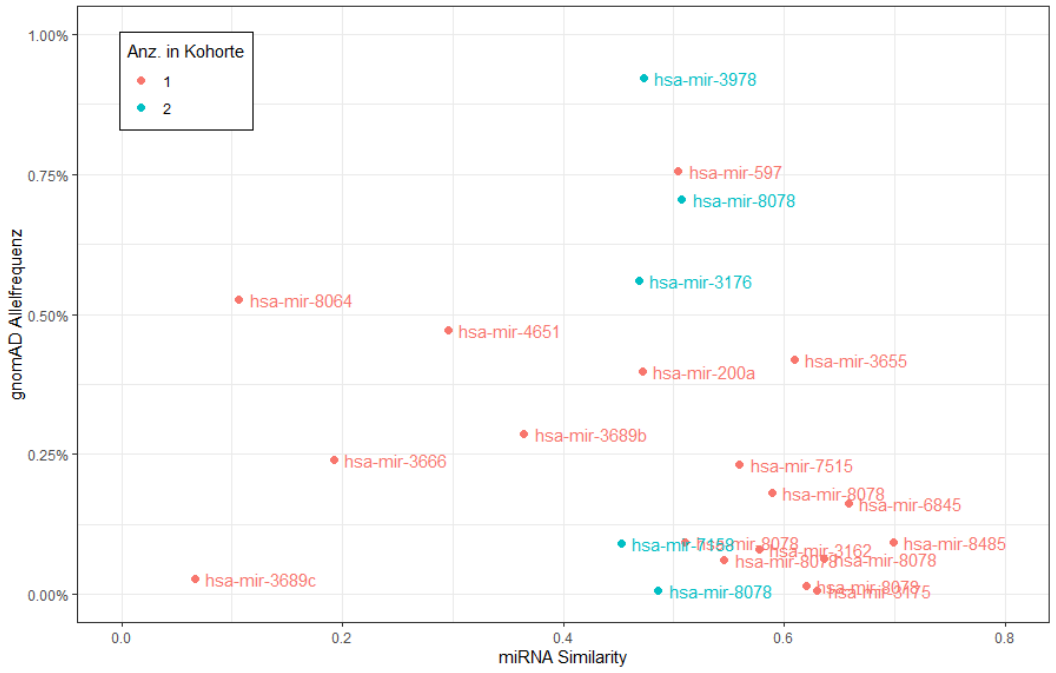


Abbildung 5.11: *miRNA Similarity vs. gnomAD Allelfrequenz. Dargestellt ist die Häufigkeit in der Kohorte (rot = 1x; blau = 2x), der zugehörige Similarity-Score (X-Achse) und die entsprechende gnomAD-Allelfrequenz (Y-Achse) für jede der gefilterten Varianten.*

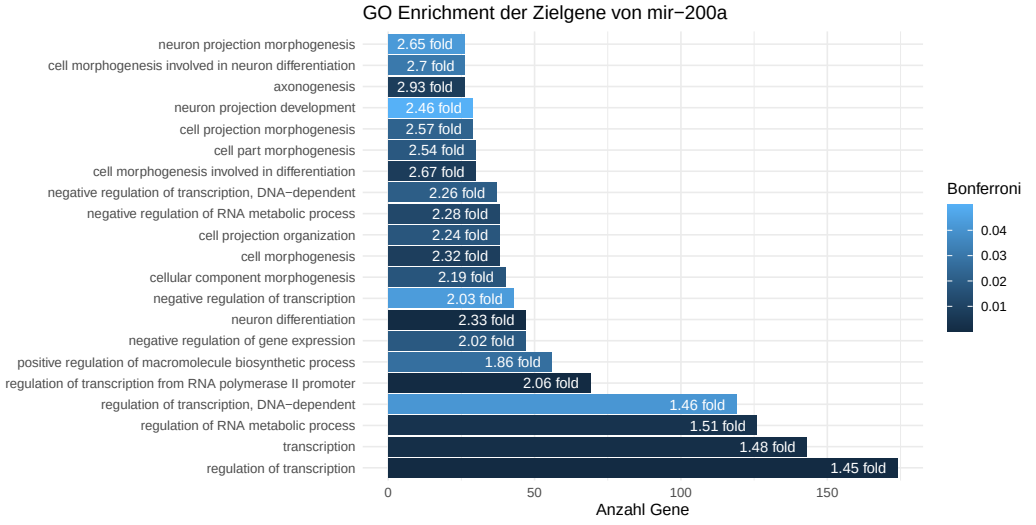


Abbildung 5.12: *GO Enrichment-Analyse der Zielgene von mir-200a. Dargestellt sind die signifikant (Bonferroni p-Value <0,05) angereicherten Ontologien.*

onsanalyse der Eltern zeigte eine paternale Vererbung der Variante. Zudem listet gnomAD-Exoms die Variante 653-mal heterozygot und 9-mal homozygot in 105.224 Individuen. gnomAD-Genomes listet sie 126-mal heterozygot sowie 1-mal homozygot in 15.691 Kontrollpersonen.

Tabelle 5.2: *miRNA Expression und vorhergesagte Targets*

miRNA	**Expression Blut**	**Expression Gehirn**	**# Targets (Target Scan)**
mir-200a	<Schwellenwert	niedrig	905
mir-3162	NA	niedrig	2435
mir-3175	NA	NA	5554
mir-3176	niedrig	mittel	4948
mir-3655	NA	NA	2151
mir-3666	<Schwellenwert	niedrig	1029
mir-3689b	<Schwellenwert	<Schwellenwert	6958
mir-3689c	NA	niedrig	6958
mir-3978	NA	niedrig	6002
mir-4651	NA	NA	5292
mir-597	NA	niedrig	3118
mir-6845	NA	NA	5333
mir-7158	<Schwellenwert	niedrig	3089
mir-7515	NA	NA	4985
mir-8064	<Schwellenwert	niedrig	3717
mir-8078	<Schwellenwert	<Schwellenwert	1986
mir-8485	NA	NA	5912

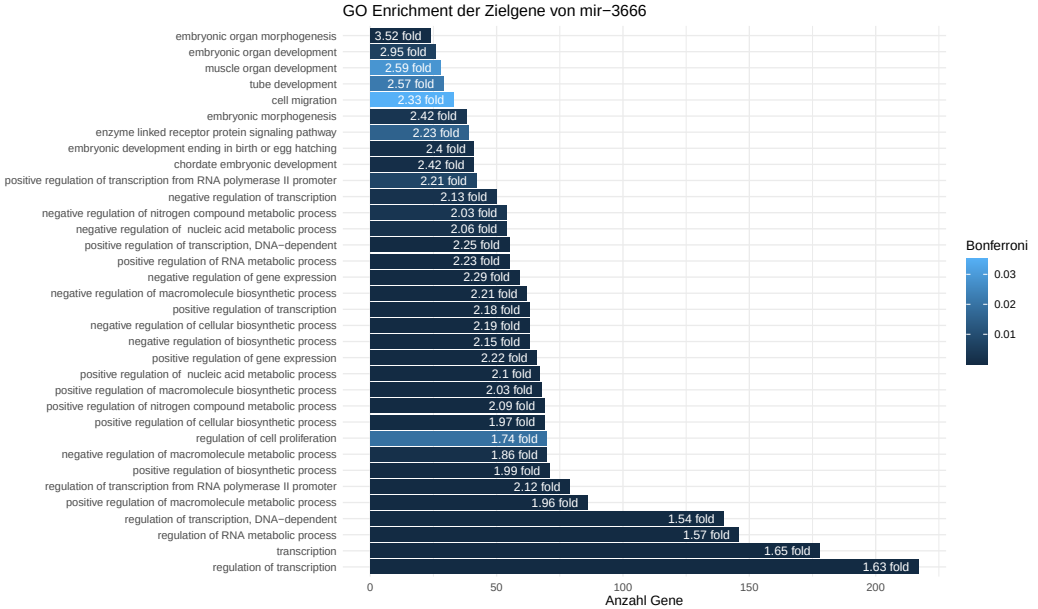


Abbildung 5.13: *GO Enrichment Analyse der Zielgene von mir-3666. Dargestellt sind die signifikant (Bonferroni p-Value <0,05) angereicherten Ontologien.*

Die GO-Anreicherungsanalyse von mir-3666 weist keinen direkten Zusammenhang zu neurologischen Prozessen auf. Vielmehr kann eine signifikante Beteiligung der vorhergesagten Zielgene bei der embryonalen Entwicklung und bei allgemeinen biologischen Prozessen wie der Transkription und biosynthetischen

sowie metabolischen Aufgaben der Zelle verzeichnet werden (siehe Abbildung 5.13). Somit scheidet eine Veränderung der 2D-Struktur von mir-3666 als Ursache für eine neurologische Erkrankung aus.

5.2.1 Varianten im Bereich der 3'UTRs

Werkzeuge zur Vorhersage von miRNA-mRNA-Interaktionen sind darauf ausgelegt, möglichst genaue Ergebnisse zu erstellen und die Anzahl an falsch-positiven Vorhersagen zu reduzieren. Dennoch bleibt die effektive Vorhersage dieser Interaktionen durch die sehr komplexen und teilweise noch nicht vollständig entschlüsselten biologischen Prozesse eine große Herausforderung [Riffo-Campos *et al.*, 2016]. Nur etwa 25% aller vorhergesagten miRNA-mRNA Bindungen können biologisch validiert werden [Mullany *et al.*, 2015]. Um die Anzahl der falsch-positiv vorhergesagten miRNA-mRNA-Interaktionen so gering wie möglich zu halten und die Auswahl der experimentell validierbaren Kandidaten zu vergrößern, wird die Verwendung von mehr als einem Vorhersageprogramm empfohlen.

Eine Vielzahl der entwickelten Tools zur Vorhersage von Interaktionen zwischen miRNAs und ihren Zielgenen sind frei verfügbar. Alle beruhen auf unterschiedlichen Algorithmen, welche auf verschiedene biologische Eigenschaften der miRNA-mRNA Bindung aufbauen. So wählt der miRanda-Algorithmus die Zielgene anhand der Sequenzkomplementarität unter Verwendung eines positionsgewichteten lokalen Alignment-Algorithmus, der freien Energie von RNA-RNA-Duplexen (Vienna RNA fold Paket) und der Konservierung von Zielgenen aus. RNAhybrid findet hingegen mit Hilfe von statistischen Modellen die energetisch günstigsten Hybridisierungsstellen zwischen miRNAs und ihren Ziel-mRNAs. Diese differierenden Herangehensweisen der beiden in dieser Arbeit verwendeten miRNA-mRNA-Interaktions-Vorhersagetools erklärt die in Kapitel 5.2 beschriebenen abweichenden Vorhersagen für die Bindungsaffinitäten der mikroRNAs miR-6810-3p, miR-4672 und miR-5195-5p an den wildtypischen bzw. mutierten 3'UTR des *IL1RAPL1*-Gens. Nur für miR-6801-3p berichten beide Algorithmen eine reduzierte Bindung der mikroRNA an den mutierten 3'UTR des *IL1RAPL1*-Gens (siehe Tabelle 5.1 in Kapitel 5.2). Diese Diskrepanzen werden durch die Verwendung unterschiedlicher mRNA und miRNA Datenbanken der beiden Programme zusätzlich verstärkt.

Eine Validierung der *in silico* durch miRanda und RNAhybrid vorhergesagten Interaktionen von miRNAs mit dem 3'UTR des *IL1RAPL1*-Gens durch Luziferaseassays zeigte größtenteils abweichende Ergebnisse. Die von beiden verwendeten Tools für die miRNA miR-6801-3p prognostizierte niedrigere Bindungsaffinität an den mutierten 3'UTR erwies sich *in vitro* als signifikant stärkere Bindung als an den wildtypischen 3'UTR (Batch 1). Die für miRNA miR-5195-5p im Luziferaseassay ersichtliche Tendenz für eine stärkere Bindung an den mutierten 3'UTR wird ebenfalls durch den Algorithmus von miRanda vorhergesagt, RNAhybrid beschreibt die Bindungsaffinitäten für miR-5195-5p an den mutierten und wildtypischen 3'UTR als gleich stark. Für

die miRNAs miR-4672 und mir-6810-3p sind *in vitro* keine unterschiedlichen Expressionswerte feststellbar. Batch 2 zeigt im Allgemeinen tendenziell gleiche wenn auch schwächer ausgeprägte Ergebnisse als Batch 1 für die Bindung von miR-5195-5p und miR-6801-3p an den mutierten bzw. wildtypischen 3'UTR. Auffällig sind die in Batch 2 allgemein niedriger gemessenen Luziferase-Aktivitätswerte. Dies kann durch eine erhöhte Renilla-Transfektionsrate erklärt werden. Möglicherweise weist der in Batch 2 neu angesetzte Renilla Vektor (pRL) eine höhere Transfektionsrate auf, als der in Batch 1 verwendete. Das Transfektionsreagenz Lipofektamin2000 kann als Ursache ausgeschlossen werden, da dieses die Transfetionsraten von Renilla-Vektor und Vektor-Insert-Konstrukt gleichermaßen beeinflussen würde. Da zwischen der Durchführung der beiden Batches eine Pause von etwa einem Monat lag und das Luminometer und damit ebenfalls die Firefly- und Renilla-Puffer auch von anderen Mitgliedern der Arbeitsgruppe verwendet wurden, ist es möglich, dass diese beiden Puffer innerhalb der Batches aus verschiedenen Chargen stammen. Insbesondere beim Firefly Puffer können hier individuelle Unterschiede entstehen, da der Puffer im Labor selbst hergestellt wird.

Zusammenfassend lässt sich sagen, dass trotz der Verwendung von zwei unterschiedlichen miRNA-mRNA-Interaktions-Vorhersagetools keine zuverlässige und biologisch validierbare Aussage getroffen werden kann. Selbst im Falle von miR-6801-3p, bei der beide Algorithmen das gleiche Ergebnis zur Bindung der miRNA an den mutierten bzw. wildtypischen 3'UTR prognostizierten, wurde diese Vorhersage durch einen Luziferase-Assay experimentell widerlegt. Zur Reduktion der Rate falsch-positiv vorhergesagter Interaktionen zwischen miRNA und mRNA ist es vermutlich notwendig, zusätzliche Algorithmen hinzuzuziehen und die Schnittmenge übereinstimmender Vorhersagen für die weiteren Analysen zu betrachten.

5.2.2 Varianten in mikroRNAs

Um den Effekt der mit Hilfe des RegVar-Panels detektierten Varianten in miRNA-Genen besser beurteilen zu können, wurde die Auswirkung dieser Varianten auf die Sekundärstruktur der miRNAs untersucht (siehe Kapitel 4.4.2). 11.865 aller etwa 250.000 bei der Analyse von 57 Patienten mit dem RegVar-Panel detektierten Varianten liegen in mikroRNA-Genen (siehe Kapitel 5.2). Die Schwellenwerte für die angewandten Filter wurden empirisch gewählt. Varianten, welche eine Allelfrequenz von mehr als 1% im gnomAD Datensatz aufweisen, kommen zu häufig in der Bevölkerung vor, als dass sie ursächlich für eine in der untersuchten Kohorte vorliegende früh manifeste Erkrankung sein können. Eine strengere Filterung, zum Beispiel mit einer Allelfrequenz von höchstens 0,1%, würde die Anzahl der Varianten weiter eingrenzen, hätte aber auch dazu führen können, Varianten für eine autosomal rezessive Erkrankung mit hoher Heterozygotenfrequenz zu verlieren. Auch wenn die Patienten der untersuchten Kohorte bezüglich ihrer phänotypischen Erscheinungsbilder relativ homogen erscheinen, ist es unwahrscheinlich, dass bei 57 untersuchten Patienten mehr als zwei Träger der gleichen Variante vorkommen. Eine höhere

Anzahl an Mutationsträgern in der Kohorte spricht für eine in der Population häufig auftretende und nicht krankheitsverursachende Mutation oder einen systematischen Sequenzierfehler, der durch die Probenaufbereitung oder den Sequenziervorgang entstanden sein kann. Somit wurden nur Varianten betrachtet, die in ein oder zwei Patienten der Kohorte detektierbar sind. Da anzunehmen ist, dass eine Veränderung der 2D-Struktur der pre-mikroRNA um mehr als 30% zu einer Beeinträchtigung bei der weiteren Verarbeitung der miRNA führen könnte, werden nur Varianten mit einem Similarity-Score von 0 - 70% betrachtet. Auch hier galt es den Schwellenwert lieber etwas großzügiger zu setzen und ein paar zusätzliche falsch-positive Varianten zu erhalten, anstatt richtig positive zu verlieren.

Für die nach der Filterung übrig gebliebenen 17 Varianten in miRNA-Genen erfolgte eine Vorhersage der Zielgene mittels TargetScan. 15 der 17 untersuchten miRNAs werden von TargetScan als nicht sicher identifizierte miRNAs beschrieben. Alles *et al.* zeigte 2019 in einer Studie, dass sich lediglich 18,3% der neuen Kandidaten-miRNAs experimentell validieren lassen [Alles *et al.*, 2019]. Die Zielgen-Vorhersagen für solche miRNAs sind daher sehr ungenau und wahrscheinlich funktionell irrelevant. Für die beiden übrig gebliebenen mikroRNAs miR-200a und miR-3666 ergab nur die GO-Anreicherungsanalyse für miR-200a eine signifikante Beteiligung der Zielgene an neurologischen Prozessen. Der entsprechende heterozygote Basenaustausch von Cytosin nach Thymin an Position 42 des *mir-200a*-Gens bewirkt eine Änderung der 2D-Struktur um 53% (siehe Abbildung 5.14). Diese Transformation ist von Base 31 bis 58 erkennbar. Der in der wildtypischen Struktur ersichtliche große Bulge zwischen Base 47 und 57 verschwindet und bildet nun zwei kleinere Loops, die durch eine Helix verbunden sind. Zudem wird die Haarnadelstruktur der 2D Struktur des mutierten miR-200a größer. Zu beachten ist allerdings, dass sich die Sekundärstruktur des von der Veränderung betroffenen Bereichs der wildtypischen miR-200a in der Regel nur mit einer Wahrscheinlichkeit von 50 - 90% vorhersagen lässt. Somit ist ungewiss, ob die Variante C>T an Position 42 die 2D-Struktur tatsächlich so stark verändert oder ob die wildtypische Sekundärstruktur *in vivo* anders aussieht als die *in silico* prognostizierte. Somit würde der strukturelle Unterschied zwischen wildtypischer und mutierter miRNA nur marginal ausfallen. Die Allelfrequenz dieser Variante liegt mit 0,31% (gnomAD Exomes) und 0,4% (gnomAD Genomes) in einem sehr niedrigen Bereich. Ein tieferer Blick in die Daten zeigt jedoch, dass der Basenaustausch in 662 der 105.224 Individuen des gnomAD Exom-Datensatzes (653 heterozygot; 9 homozygot) vorkommt. Unter den 15.691 Kontrollpersonen des gnomAD Genom-Datensatzes tragen 126 Individuen diese Variante heterozygot und 1 Proband homozygot. Diese hohen absoluten Allelhäufigkeiten in einem Datensatz von vermeintlich gesunden Probanden, insbesondere in homozygoter Form, spricht gegen die Variante als alleiniger Auslöser eines früh manifesten phänotypischen Erscheinungsbildes. Aufgrund dieser Erkenntnisse wurde der Basenaustausch von Cytosin nach Thymin an Position 42 des miR-200a-Gens nicht weiter als krankheitsverursachende Variante für eine monogene Erkrankung betrachtet. Eine Beteiligung der Variante in einem Krankheitsbild mit di- bzw. oligogenem Vererbungsmus-

ter oder einer multifaktoriellen Erkrankung kann allerdings nicht ausgeschlossen werden.

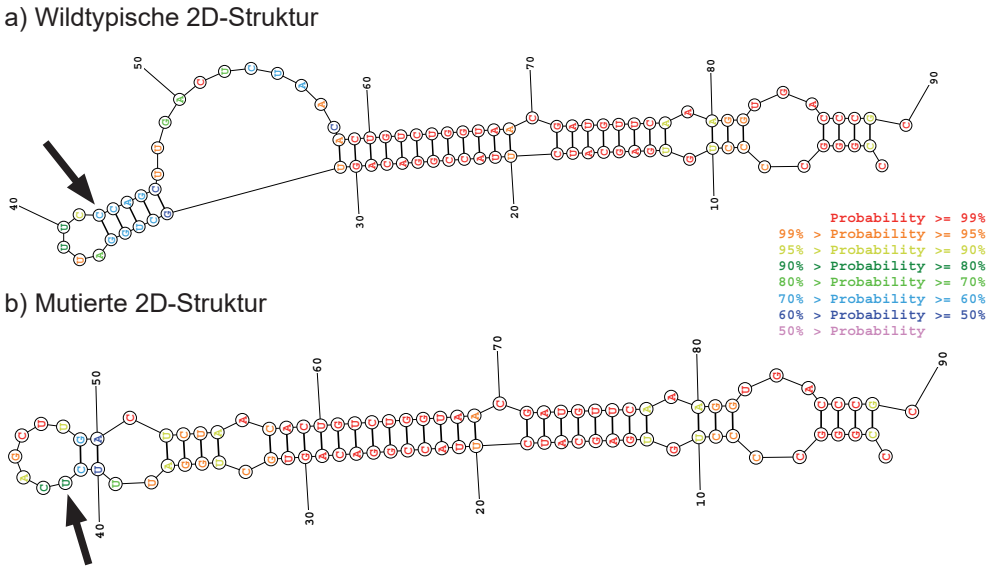


Abbildung 5.14: *Wildtypische und mutierte 2D-Struktur von miR-200a*

5.3 Genom-Sequenzierung beendet eine lange diagnostische Odyssee

Anfang des Jahres 2016 stellte eine Familie ihre Tochter (Patient 4, w, 14 Jahre) in der Genetischen Beratungsstelle des Instituts für Humangenetik Mainz vor. Das klinische Erscheinungsbild wies mit fehlender Sprache, Epilepsie (erster Anfall 12/2015), einer globalen Entwicklungsretardierung der Grob- und Feinmotorik, einer Intelligenzminderung, stereotypischen Waschbewegungen der Hände und einem ataktischen Gangbild auf ein atypisches Rett-Syndrom hin. Vor der Vorstellung in Mainz wurde an der Universitätsmedizin Göttingen eine Sanger-Sequenzierung und MLPA aller kodierender Exons des *MECP*2-Gens (07/2010) sowie der Gene *FOXG1* und *CDKL5* (01/2012) durchgeführt. Die Analysen waren ohne pathologischen Befund. Eine anschließend (2012) veranlasste Exom-Sequenzierung mit dem Agilent SureSelect XT all exon V4-Kit am Max-Delbrück-Centrum (MDC) in Berlin zeigte 37.080 Varianten. Im Durchschnitt wurden 96,5% der angereicherten Regionen ausreichend abgedeckt. Im Exomdatensatz allgemein, sowie in der Detailauswertung der Gene *MECP2*, *FOXG1* und *CDKL5* durch das MDC war keine krankheitsverursachende Variante festzustellen.

Am Institut für Humangenetik Mainz wurde dann zusätzlich eine genomweite SNP-Array-Analyse sowie Untersuchungen zum Ausschluss einer uniparentalen Disomie des Chromosoms 14, eines Angelman- und eines Temple-Syndroms veranlasst. Alle Untersuchungsergebnisse waren ohne pathologischen Befund.

Der durch ein HUMARA-Assay [Allen *et al.*, 1992] getestete Inaktivierungsstatus des X-Chromosoms wies keine Verschiebung auf. Abschließend wurden 2018 die Proben von Mutter, Vater und Kind wissenschaftlich als Trio am MDC Berlin einer gesamtgenomischen Sequenzierung unterzogen. Die resultierenden 451 Millionen Paired-End Reads mit einer Länge von jeweils 150 bp deckten das Genom mit einer durchschnittlichen Tiefe von 38 Reads ab. 97,5% der kodierenden Sequenzen aller bekannten Gene waren mit einer Sequenziertiefe von mindestens 10 Reads abgedeckt. Insgesamt konnten 5.088.367 Varianten im Genom von Patient 4 detektiert werden.

Bei der Auswertung der Sequenzierungsdaten ließ sich bei Patient 4 im kodierenden Exon 3 des *MECP2*-Gens (OMIM: *300005, Ensembl: ENST00000453960) eine heterozygote 23 bp Deletion c.1200_1222del nachweisen. Diese führt zu einem frühzeitigen Abbruch der Proteinbiosynthese an Position 409 des MECP2-Proteins (p.(Pro401ArgfsTer8)) und war in der Literatur bereits bei einer Patientin mit Rett-ähnlichem phänotypischen Erscheinungsbild als pathologisch beschrieben worden [Rauch *et al.*, 2012]. Das bioinformatische Pathogenitäts-Vorhersagetool CADD (Score: 35,0) stuft diese Variante als krankheitsverursachend ein. In der Populationsdatenbank gnomAD (Exomes/Genomes) ist sie nicht gelistet. Die Datensätze der Eltern weisen die Variante c.1200_1222del (p.(Pro401ArgfsTer8)) des *MECP2*-Gens nicht auf (siehe Abbildung 5.15). Somit besteht der Verdacht, dass diese bei Patientin 4 neu entstanden (*de novo*) ist.

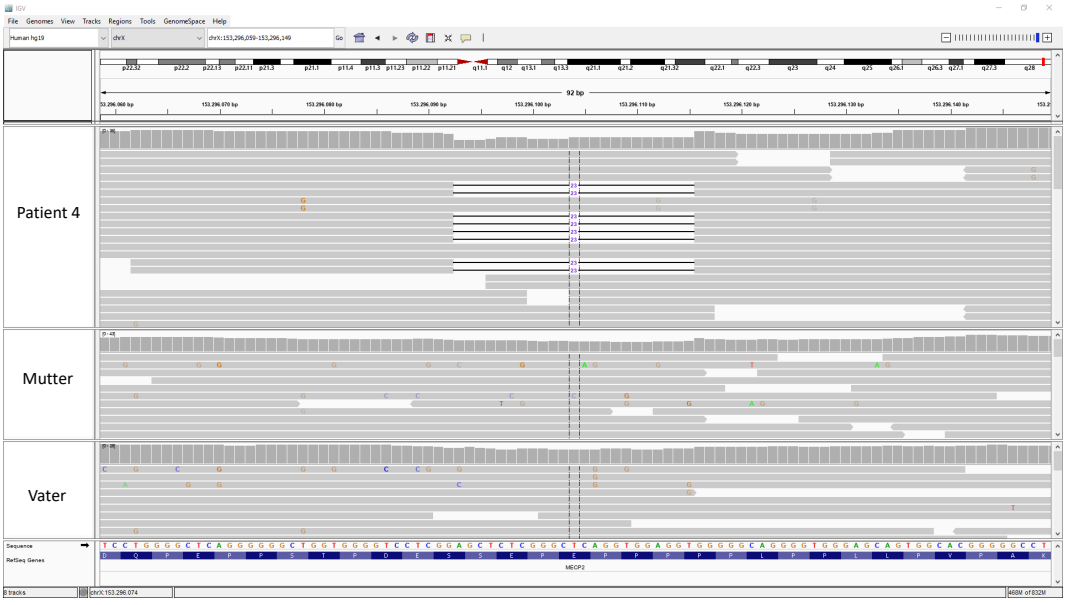


Abbildung 5.15: *23 bp de novo Deletion im MECP2-Gen.*
Die 23 bp Deletion ist im Patientendatensatz durch schwarze Striche erkennbar (oben). Die Datensätze von Mutter (Mitte) und Vater (unten) zeigen diese Deletion nicht.

Die Variante c.1200_1222del ließ sich aus unbekannten Gründen nicht mittels Sanger-Sequenzierung bestätigen. Daher wurde die von Rauch *et al.* erwähnte

Melting-Curve Analyse mittels Real-Time PCR zur Validierung der 23 bp Deletion etabliert [Rauch *et al.*, 2012]. Durch PCR-Amplifikation der mutierten Genomregion ensteht für das deletierte Allel ein um 23 bp kleineres Produkt als für das wildtypische Allel. Aufgrund der Größendifferenz unterscheiden sich die Schmelztemperaturen der beiden Amplikons. Dies kann mit Hilfe der Schmelzkurvenanalyse detektiert werden. Abbildung 5.16 stellt die unterschiedlichen

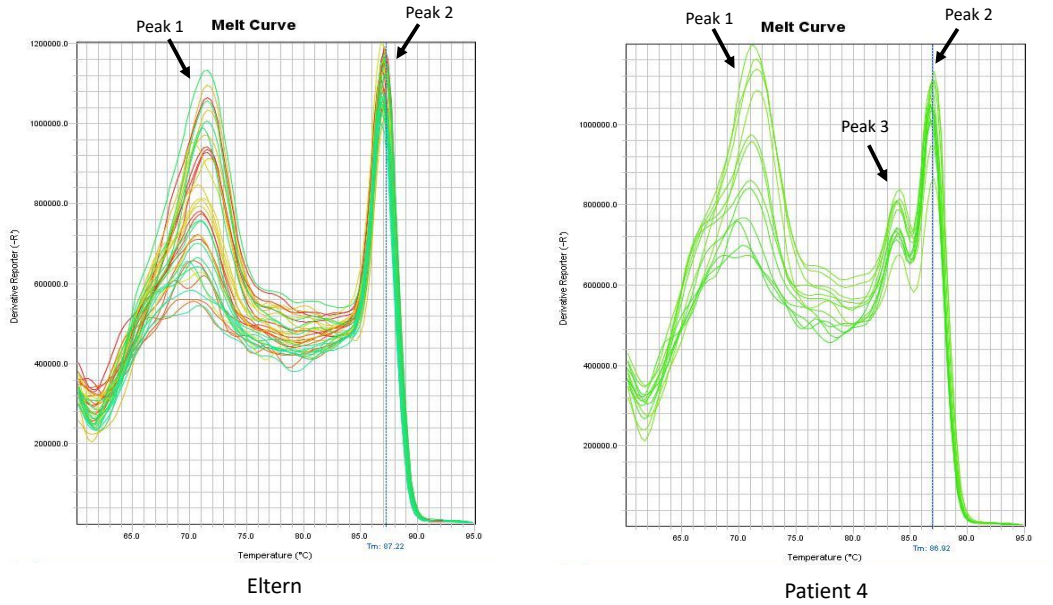


Abbildung 5.16: *Schmelzkurvenanalyse der 23 bp MECP2-Deletion. Der bei etwa 71°C liegende Peak 1 deutet auf Primer Dimere hin. Bei etwa 87°C (Peak 2) liegt die Schmelztemperatur des längeren wildtypischen Allels. Der Peak bei etwa 84°C (Peak 3 rechts) weist das um 23 bp verkürzte Amplikon des mutierten Allels nach. Dieses ist nur in der Probe des Patienten zu erkennen.*

Schmelzkurven in der DNA der Eltern und des Patienten 4 dar. Der Peak 1 liegt bei etwa 71°C und deutet auf ein kurzes Nebenprodukt (Primer Dimere) hin. Der Anstieg bei etwa 87°C (Peak 2) beschreibt das längere wildtypische Allel. Peak 1 und 2 liegen sowohl bei Patient 4 als auch bei den Eltern vor. Die Schmelzkurvenanalyse von Patient 4 zeigt einen zusätzlichen Peak bei etwa 84°C (Peak 3). Dieser weist das um 23 bp verkürzte Amplikon des mutierten Allels beim Patienten nach. Da der Peak 3 bei den Eltern nicht detektierbar ist, konnte das Ergebnis der WGS somit bestätigt werden.

Warum die 23 bp Deletion im *MECP2*-Gen nicht mittels Sanger- und Exomsequenzierung entdeckt werden konnte, lässt sich nicht abschließend klären. Vermutlich entgeht das mutierte Allel einer Amplifikation durch die PCR, da es eine niedrigere Schmelztemperatur aufweist. Durch diesen Allel-Dropout kann in der anschließenden Sanger-Sequenzierung nur das wildtypische Allel abgelesen werden. Die Exomsequenzierung basiert auf einer Anreicherung der

Zielregionen. Es ist möglich, dass die dazu eingesetzten RNA-Sonden aufgrund der 23 bp Deletion nicht an das veränderte Allel binden und somit ebenfalls einen Allel-Dropout verursachen. Zudem beinhaltet die Probenaufbereitung einer Exomsequenzierung mehrere PCR-Schritte, die wie zuvor erwähnt, ebenfalls zur selektiven Amplifikation des wildtypischen Allels beitragen könnten. Der Arbeitsgruppe um Rauch *et al.* gelang es allerdings, die 23 bp Deletion im *MECP2*-Gen in einer anderen Patientin mittels Exomsequenzierung nachzuweisen [Rauch *et al.*, 2012]. Da die Exomsequenzierung von Patientin 4 in einem externen Labor durchgeführt wurde, standen die Daten nicht für eine Reevaluation zur Verfügung. Es ist denkbar, dass die Deletion im Exomdatensatz der Patientin 4 nachweisbar war, dann aber als Artefakt eingestuft wurde, da sie mit Hilfe der Sanger-Sequenzierung nicht validiert werden konnte. Da die Exomsequenzierung der in dieser Arbeit beschriebenen Patientin im Jahr 2012 erfolgte und die Arbeit von Rauch *et al.* im November 2012 erschienen ist, war es den Kollegen des externen Labors vermutlich nicht möglich, bei der Auswertung der Daten auf diese Literatur zuzugreifen und die dort beschriebene Schmelzkurvenanalyse zur Validierung heranzuziehen. Bei der gesamtgenomischen Sequenzierung werden die genomischen DNA-Bruchstücke typischerweise ohne vorherigen Amplifikations- und Anreicherungsschritt sequenziert. Dadurch konnte die 23 bp Deletion im *MECP2*-Gen eindeutig nachgewiesen und mit Hilfe der in der Literatur erwähnten Schmelzkurvenanalyse bestätigt werden.

Kapitel 6

Diskussion und Ausblick

Dass heute in mehr als 4.000 der 20.000 proteinkodierenden menschlichen Gene krankheitsverursachende Mutationen bekannt sind und jedes Jahr etwa 200 weitere genetische Krankheitsursachen identifiziert werden, ist hauptsächlich auf die neue Technologie der Hochdurchsatz-DNA-Sequenzierung zurückzuführen. Die Etablierung der NGS-Technologie und der damit verbundene Aufbau einer bioinformatischen Pipeline am Institut für Humangenetik in Mainz ermöglichte in den letzten Jahren eine Steigerung der Aufklärungsrate der konventionellen Stufendiagnostik in der genetischen Diagnostik und Krankenversorgung um bis zu 32%.

6.1 Grenzen der WES

Dennoch ließen sich in 53% der mit einer Exomsequenzierung untersuchten Patienten keine ursächlichen Varianten finden. In den nicht aufgeklärten Fällen liegen vermutlich Varianten in Bereichen vor, die mit den verwendeten Sequenzierungsmethoden nicht ausreichend detektiert oder mit den derzeitigen Algorithmen und dem aktuellen Wissensstand nicht genügend beurteilt werden können. Außerdem lassen sich somatische, auf nicht zugängliche Gewebe (z.B. das Hirn) beschränkte Mutationen und Mosaike bei der Sequenzierung von DNA aus Blutzellen nicht nachweisen. Zudem kommen epigenetische Veränderungen in Frage, die durch herkömmliche Sequenzierungsmethoden nicht diagnostizierbar sind. Komplexe digene, polygene und multifaktorielle Ursachen sind ebenfalls schwer ausfindig zu machen, ebenso wie Varianten mit geringer Penetranz, die nur unwesentlich häufiger bei Patienten als bei Gesunden vorkommen und daher meist nicht als Krankheitsursache erkannt werden.

Sequenzveränderungen innerhalb der kanonischen Spleißstellen (2 bp vor und hinter dem Exon) lassen sich durch Algorithmen meist zuverlässig erkennen. Varianten an weiter entfernten Positionen oder mit ungenauer definierten Spleiß-Konsensussequenzen sind hingegen schlechter charakterisierbar [Grozeva *et al.*, 2015]. Zudem sind tief intronische sowie im 5'- und 3'-UTR liegende regulatorische Varianten durch die Exomsequenzierung nicht abgedeckt und dadurch nicht erfasst. Durch eine gesamtgenomische Sequenzierung können diese Varianten zwar sichtbar gemacht werden, die Beurteilung ihrer

Pathogenität ist jedoch bisher sehr unzuverlässig.

Wie jüngste Untersuchungen belegen [Murdock *et al.*, 2020] eignet sich die Sequenzierung des Transkriptoms von Blutzellen und Fibroblasten (d.h. die Sequenzierung der mRNA dieser Zellen) als flankierender Ansatz zur Identifizierung von Mutationen, welche die Expression von Genen beeinträchtigen oder eine Störung des Spleißens primärer Transkripte zur Folge haben. Bei 17% aller Probanden, deren Krankheitsursache mithilfe konventioneller Methoden (Mikroarray-Analyse und anschließende WES/WGS) nicht aufgeklärt werden konnte, ließen sich durch Sequenzierung des Transkriptoms regulatorische und das Spleißen verändernde, pathogene Varianten identifizieren [Murdock *et al.*, 2020]. Allerdings erfordert dieser Ansatz die Anzucht von Fibroblasten oder die Gewinnung anderer, als Untersuchungsmaterial geeigneter Gewebe. Für die Detektion von Gendefekten, die nur im Hirn nachweisbar sind, scheidet dieser Ansatz von vornherein aus, es sei denn es soll der Weg über die noch aufwendigere Herstellung und Differenzierung von induzierten pluripotenten Stammzellen (IPS) zu neuronalen Vorläuferzellen (engl. Neuronal Precursor Cells, NPC) [Adelaja, 2017] gegangen werden. Zudem spricht Einiges dafür, dass die Entwicklung neuer Algorithmen zur Erkennung pathogener Varianten in nicht kodierenden Genomabschnitten schon bald an die Stelle der ergänzenden Transkriptomsequenzierung treten wird [Yepez *et al.*, 2020].

Derzeit werden zunehmend bioinformatische Verfahren beschrieben, um mithilfe künstlicher Intelligenz und „deep-learning"-Methoden [Zou *et al.*, 2019] nicht-kodierende Varianten als Krankheitsursache zu identifizieren und ihren Effekt auf die transkriptionelle und posttranskriptionelle Regulation und das alternative Spleißen einzuschätzen [Louadi *et al.*, 2019] [Liu *et al.*, 2019a] [Varma *et al.*, 2019] [Jaganathan *et al.*, 2019] [Caron *et al.*, 2019]. Den raschen Fortschritt auf diesem Gebiet dokumentiert auch das kürzlich entwickelte Plugin „UTRannotator" [Zhang *et al.*, 2020] für das Annotations-Tool Variant Effect Predictor (siehe Kapitel IV.XII). Damit wird es einfacher, die Auswirkung von Varianten im 5'UTR auf die Expression von „upstream open reading frames" (uORFs) zu beurteilen. Diese und andere in Entwicklung befindlichen Methoden versprechen die Aufklärungsrate der NGS-gestützten genetischen Diagnostik rasch weiter zu steigern.

6.2 Somatische Mosaike und Kopienzahlveränderungen

Mutationen sind jedoch nicht auf Zellen der Keimbahn begrenzt, sondern können auch postzygotisch auftreten (z.B. während der Keimesentwicklung oder sogar im Erwachsenenalter). Sie sind dann nur in einem Teil der Körperzellen oder Gewebe nachweisbar [Gajecka, 2016]. Es ist zu vermuten, dass die Häufigkeit derartiger somatischer Mutationen und der dadurch entstehenden Mosaike unterschätzt wird, weil sie oft nicht zu erkennen sind. So zeigen Ballif *et al.*, dass sich 8% der pathologischen Mikroarraybefunde auf chromosomale

Mosaike zurückführen lassen [Ballif *et al.*, 2006]. Allerdings liegen somatische Mosaike, wie oben bereits diskutiert, nicht immer in diagnostisch zugänglichen Zellen und Geweben vor. Der hohe Durchsatz der NGS-Technologie und deren Sensitivität ermöglicht eine sehr tiefe Sequenzierung der DNA-Fragmente und den Nachweis geringer Mengen mutierter Varianten unter den Wildtyp-Allelen, welche oft als Hintergrundrauschen interpretiert und bei der Sanger-Sequenzierung übersehen werden. Mutationen in weniger als 10-15% der untersuchten Zellen sind erst bei einer durchschnittlichen Sequenziertiefe von mehr als 200 Reads zuverlässig nachweisbar. Die relativ hohen Kosten eines NGS-Experiments mit enormer Sequenziertiefe können jedoch eine große Einschränkung für die großflächige Anwendung dieser Methode im diagnostischen Labor darstellen [Gajecka, 2016]. Exomsequenzierungen mit durchschnittlich 95 Reads sowie Genomsequenzierungen mit durchschnittlich 30 Reads lassen einen sicheren Nachweis klinisch relevanter Mosaike meist nicht zu, obwohl dieser für die Einschätzung des Wiederholungsrisikos in Familien von großer Bedeutung wäre. So ließ sich bei ca. 13% der Eltern von Epilepsie-Patienten mit pathogenen *de novo* Mutationen ein parentales Mosaik aus mutierten und nicht mutierten Zellen nachweisen [Stosser *et al.*, 2018].

Etwa 15% der Patienten, die für klinische Gentests überwiesen werden, weisen krankheitsassoziierte Kopienzahlveränderungen (CNVs) auf [Miller *et al.*, 2010]. Für den Nachweis von CNVs anhand von WES-Daten existieren verschiedene bioinformatische Analyse-Programme. Typischerweise verwenden die Algorithmen verschiedene statistische Verteilungen, um die aggregierte Lesetiefe der Exons zu modellieren und die unterschiedlichen Lesetiefen benachbarter Exons zur Identifikation von Duplikationen oder Deletionen zu nutzen [Fromer *et al.*, 2012] [Krumm *et al.*, 2012] [Backenroth *et al.*, 2014] [Jiang *et al.*, 2015] [Packer *et al.*, 2016]. Trotz der Verwendung mehrerer hundert Referenzproben ist die Detektionsqualität von CNVs aus Exom-Sequenzierungsdaten durch hohe Falsch-Positiv-Raten begrenzt. Die Strategie, verschiedene Algorithmen parallel zur Erkennung von CNVs zu verwenden und nur diejenigen Veränderungen als relevant zu betrachten, die von allen Programmen angezeigt werden, hat demgegenüber den Nachteil, dass dabei richtig-positive Ergebnisse verloren gehen. Hier kommen immer mehr Algorithmen aus dem Bereich des maschinellen Lernens zum Einsatz, um die Ergebnisse mehrerer verschiedener CNV-Tools effizient zu kombinieren und möglichst viele falsch-positiv detektierte Kopienzahlveränderungen zu eliminieren [Pounraja *et al.*, 2019].

6.3 Vorteile der gesamtgenomischen Sequenzierung

Im Gegensatz zur WES, die auf die Erfassung von Sequenzvarianten in proteinkodierenden Abschnitten von Genen beschränkt ist, erlaubt die WGS die Erkennung nahezu aller Veränderungen im Erbgut. Dadurch muss diese Untersuchung nur einmal im Leben erfolgen und eignet sich hervorragend als diagnos-

tischer Eingangstest für Patienten mit Verdacht auf eine genetisch bedingte Erkrankung. Wie in Kapitel 5.3 gezeigt, kann eine gesamtgenomische Sequenzierung Varianten detektieren, die mit herkömmlichen Methoden wie der MLPA, Sanger- und Exomsequenzierung nicht nachweisbar sind. Dieses Fallbeispiel unterstreicht den Nutzen der Genomsequenzierung als primären molekulargenetischen Test in der klinischen Genetik. Durch seine amplifikations- und anreicherungsfreie Probenaufbereitung spiegelt die WGS mit ihrer gleichmäßigen Abdeckung die tatsächliche Genomstruktur eines Individuums genauer wider als eine Exomsequenzierung. So lassen sich neben Einzelnukleotidaustauschen, InDels und Kopienzahlveränderungen auch strukturelle Chromosomenabberationen [Kosugi *et al.*, 2019] und Repeatexpansionen [Dolzhenko *et al.*, 2017] nachweisen. Die Verarbeitung der großen Datenmengen, kann durch die stetige Weiterentwicklung im IT-Sektor zum Beispiel durch spezialisierte Hardware wie die Illumina DRAGEN Bio-IT-Plattform oder einen auf Grafikkarten basierenden Server beschleunigt werden. Immer effizientere Algorithmen und reichhaltigere Datenbanken helfen zudem, die Vielzahl an Varianten automatisiert zu kategorisieren.

Die derzeitigen Entwicklungen sprechen dafür, dass die Kosten einer gesamtgenomischen Sequenzierung weiter fallen und auch der für eine WGS-gestützte molekulargenetische Diagnose erforderliche Zeitbedarf drastisch sinken wird [Farnaes *et al.*, 2018]. Monogene Krankheiten sind für etwa 80% der seltenen Erkrankungen verantwortlich, von denen etwa 4% der Bevölkerung betroffen sind [EURORDIS, 2020b]. Oft handelt es sich hierbei um chronische und lebensbedrohliche Erkrankungen. Aus diesen Gründen hat England bereits im Jahre 2018 die WGS als Standardverfahren in die genetische Routinediagnostik eingeführt, und andere Länder sind diesem Beispiel inzwischen gefolgt.

6.4 Nationale Genommedizin-Progamme

Die englische Regierung begann im Jahr 2012 mit der Sequenzierung von 100.000 menschlichen Genomen im Rahmen des „Genomics England Project“, welches sich auf seltene Erkrankungen und - im geringeren Umfang - auf Krebs konzentriert. Die ursprünglich ebenfalls vorgesehene Untersuchung pathogener Mikroorganismen bei Patienten mit Infektionskrankheiten wurde schon bald ausgegliedert und auf andere Weise gefördert [Department of Health and Social Care, 2013]. Zur methodischen Standardisierung konzertriert sich die eigentliche Genomsequenzierung an einem einzigen nationalen WGS-Zentrum. Gleichzeitig wurden auch für die Phänotypisierung, das heißt für die der Sequenzierung vorausgehende klinische Untersuchung und Erfassung auffälliger Befunde, einheitliche Standards entwickelt und an 13 über das Land verteilten Genommedizin-Zentren etabliert. Die Genomdaten wurden mit Informationen über den Gesundheitszustand der Probanden und deren Krankengeschichte verknüpft und in pseudonymisierter Form in einer zentralen Datenbank abgelegt. Sämtliche resultierenden Datensätze sind für die Aufklärung der Ursachen von Krankheiten und die Suche nach neuen Behandlungsmöglichkeiten nutzbar. Bereits im Dezember 2018 wurden 100.000

Genome von insgesamt 70.000 Familien sequenziert und damit das Ziel des Projekts erreicht [Genomics England, 2018]. Aufgrund der dabei erzielten hohen diagnostischen Aufklärungsquote beschloss der National Health Service (NHS), die Genomsequenzierung in die genetische Routinediagnostik zu überführen. Seither nimmt England auf dem Gebiet der Genommedizin eine international führende Rolle ein und diente anderen Ländern wie Frankreich und Schweden als Vorbild für deren eigene Genomprojekte.

Auch in Deutschland gibt es seit geraumer Zeit Bestrebungen, die Ganzgenomsequenzierung im Rahmen eines nationalen Genommedizinprogramms in die Regelversorgung zu überführen. Im August 2019 hat der Gesundheitsminister diese Anregungen aufgenommen und ein nationales Genommedizin-Programm („genomDE“) ins Leben gerufen, welches mit einjähriger, der Corona-Pandemie geschuldeter Verzögerung im November 2020 der Fachwelt vorgestellt wurde. Als Nachzügler kann Deutschland bei der Realisierung dieses Vorhabens auf die einschlägigen Erfahrungen Englands, aber auch verschiedener anderer europäischer Nachbarländer zurückgreifen. Aufgrund der vom Bundesministerium für Gesundheit (BMG) vorgegebenen Zielsetzung, Genomdaten international auszutauschen, wird Deutschland sich an den inzwischen für die Genomsequenzierung und die Phänotypisierung, aber auch die Datenanalyse und –speicherung etablierten Standards und Formaten orientieren können und müssen. Dies dürfte auch den erforderlichen Aufbau einheitlicher Strukturen im Bereich der genetischen Krankenversorgung erleichtern.

6.5 WGS - Finanzierung und Datenschutz

Deutschland gehört bisher zu einem der wenigen Ländern, in denen mit der Genpanel-Untersuchung bereits die Kosten einer Art der Hochdurchsatzsequenzierung durch die Krankenkassen übernommen wird. Aufgrund jüngster Vereinbarungen ist mit einer auf qualifizierte akademische Zentren begrenzte Kassenfinanzierung der WES zu rechnen. Mit der genomDE-Initiative des BMG könnte der Grundstein für eine kontrollierte Einführung der WGS in Deutschland gelegt sein. Eine Realisierung dieser, seit kurzem auch von der EU unterstützten Initiative, würde nicht nur die genetische Krankenversorgung auf eine neue, breitere Basis stellen, sondern über die Etablierung einer auch für die Forschung nutzbaren zentralen Genomdatenbank die Chancen Deutschlands entscheidend verbessern, im Bereich der Genomforschung international Anschluss zu gewinnen.

Gelegentlich geäußerte Befürchtungen, dass in der zentralen Datenbank abgelegte Genomdaten zu forensischen Zwecken missbraucht oder an Versicherungen sowie Arbeitgeber weitergegeben werden könnten, entbehren jedoch jeder Grundlage, weil Deutschland ein sehr strenges Datenschutz- und Gendiagnostikgesetz hat, welches eine Weitergabe und missbräuchliche Nutzung dieser Daten unter Strafe stellt (Abs. 5 §19 GenDG). §18 Abs. 4 des Gendiagnostikgesetzes verbietet es Versicherungsgesellschaften überdies, vom Versicherten genetische Untersuchungen jeglicher Art zu verlangen. Die mehr as 90%

der deutschen Bevölkerung, die Mitglied in einer gesetzlichen Krankenkasse sind, sind ohnehin davon nicht betroffen, da gesetzliche Krankenkassen in ihrem Tarifsystem keine individuellen Gesundheitsrisiken berücksichtigen. Ebenso scheinen die Bedenken bezüglich des Datenschutzes für viele Ratsuchende nur eine sekundäre Rolle zu spielen. Wie das Fallbeispiel aus Kapitel 5.3 zeigt, überwiegt bei den Betroffenen die Erleichterung, nach jahrelanger Odyssee von Arzt zu Arzt endlich eine korrekte Diagnose und eine Antwort auf die Fragen zum weiteren Verlauf der Erkrankung, einer möglichen Therapie und dem Wiederholungsrisiko bei der weiteren Familienplanung zu erhalten.

Das Genom eines Menschen stimmt jeweils zur Hälfte mit den Genomen von Vater und Mutter überein und wird zur Hälfte an die Kinder weitergegeben. Nicht nur innerhalb von Familien, sondern auch in ganzen Populationen lassen sich gemeinsame genetische Merkmale finden, die entfernte Verwandtschaftsbeziehungen widerspiegeln. Durch Typisierung weniger Tausend genomischer Marker lässt sich die geographische Herkunft einzelner Menschen sehr genau bestimmen. Unser Genom gehört uns nicht alleine, die eigenen Genomdaten sind nicht unser Privatbesitz, den es zu verteidigen gilt, sondern wir teilen sie uns mit einer Vielzahl anderer Menschen. Nur durch die Bereitschaft vieler, ihre eigenen Genomdaten und klinische Befunde mit anderen zu teilen, wird es möglich sein, die Funktion einzelner Gene und schließlich des ganzen Genoms zu verstehen.

Auf der anderen Seite sollte sich niemand das Recht auf Nicht-Wissen nehmen lassen. Allerdings ist das Risiko gesunder Erwachsener sehr gering, durch eine Genomsequenzierung zu erfahren, eine Veranlagung für eine Erkrankung zu tragen, die in der Familie früher noch nicht vorgekommen war und für die es keine Therapie gibt. Schließlich ist die Nutzung pseudonymisierter Genomdaten und damit verknüpfter klinischer Daten, wie sie bei der diagnostischen Genomsequenzierung anfallen, eine Voraussetzung für die Aufklärung von bisher unbekannten genetischen Krankheitsursachen, deren Pathomechanismen und die Entwicklung neuer Medikamente oder Behandlungsmethoden, von der die gesamte Bevölkerung profitieren wird.

Im Bereich der Genetik sind Krankenversorgung und Forschung untrennbar miteinander verbunden. Die Genomsequenzierung wird Millionen von seltenen, noch nicht beschriebenen genetischen Veränderungen identifizieren, deren pathogenetische Relevanz mithilfe bioinformatischer Algorithmen, aber vor Allem durch spezifische funktionelle Studien überprüft werden muss. Wie in Kapitel 5.2 am Beispiel von Luziferase-Assays gezeigt, reichen allgemeine Tests oft nicht aus um *in silico* vorhergesagte funktionelle Störungen zu verifizieren. Daraus folgt die Notwendigkeit, funktionelle Tests für eine Vielzahl „neuer“ Krankheitsgene oder Genfamilien zu entwickeln, unter Einschluss spezifischer biochemischer Assays, zellulärer Modellsysteme, Tiermodelle oder sogar patientenspezifischer induzierter pluripotenter Stammzelllinien [Rodenburg, 2018]. Ohne Aufklärung der Funktion „neuer“ Krankheitsgene und ohne Einblick in die pathophysiologischen Konsequenzen krankheitsverursachender Mutationen

wird es keine erfolgversprechenden Therapieansätze und keine neuen Medikamente geben.

Somit ist die Next-Generation-Sequenzierung ein unverzichtbares Instrument zur Verbesserung der klinisch-genetischen Krankenversorgung geworden. Bereits heute können über die Hälfte der Ratsuchenden mit monogenen Krankheiten damit rechnen, mithilfe der Hochdurchsatzsequenzierung eine eindeutige Diagnose zu erhalten. Mit wachsenden Fallzahlen, verbesserten Algorithmen und stetig aktualisierten Datenbanken sind in Zukunft weiter steigende Aufklärungsraten zu erwarten.

6.6 EBM-Ziffern für die WES und WGS

Seit der Novellierung des Tarifsystems der Krankenkassen (Einheitlicher Bewertungsmaßstab, EBM) im 3. Quartal 2016 gibt es in Deutschland eine einheitliche Regelung für die Durchführung der NGS-gestützten genetischen Diagnostik. Danach werden solche Untersuchungen von den Kostenträgern vergütet, wenn die Gesamtlänge der dabei sequenzierten Genomabschnitte 25 kbp nicht übersteigt, was im Durchschnitt etwa der kodierenden Länge von 18 Genen entspricht. Viele genetisch bedingte Erkrankungen sind allerdings weitaus heterogener, weshalb Mutationen in Hunderten oder gar Tausenden von Genen als Ursache infrage kommen. Wie bereits erwähnt fallen z.B. psychomotorische Entwicklungsstörungen in diese Kategorie, und es läge daher nahe, die Mutationssuche bei Patienten mit solchen Krankheiten nicht auf wenige Krankheitsgene zu beschränken.

In Deutschland ist die Finanzierung der WES-gestützten genetischen Diagnostik zur Zeit noch auf wenige akademische Zentren beschränkt, die an dem jüngst beendeten T-NAMSE-Pilotprojekt des G-BA-Innovationsfonds beteiligt waren. Nach jüngsten Informationen soll diese Finanzierung durch Selektivverträge mit den Krankenkassen verstetigt werden, wobei noch offen ist, ob und in welchem Umfang auch andere Zentren für Humangenetik und die genetische Krankenversorgung allgemein davon profitieren werden.

Allerdings gibt es berechtigte Einwände gegen die Einführung von EBM-Ziffern für die Sequenzierung aller protein-kodierenden Genomabschnitte mithilfe der WES oder gar für die Sequenzierung des gesamten menschlichen Genoms. Zur Qualitätssicherung sollten derartige Untersuchungen auf Zentren beschränkt sein, die über eine ausreichende personelle und apparative Infrastruktur für die standardisierte Erhebung klinischer Daten verfügen. Diese müssen sich für die Indikationsstellung zur WES oder WGS an einheitliche, allgemein akzeptierte Standards halten und in der Lage sein, die durch WES oder WGS erzeugten Sequenzdaten analysieren und im Hinblick auf das klinische Krankheitsbild relevante Varianten identifizieren und priorisieren zu können. Da selbst unter diesen Voraussetzungen bisher erst ca. die Hälfte aller Fälle mit begründetem Verdacht auf monogene Krankheiten aufgeklärt werden können, sollten diese Zentren zusätzlich über Forschungskapazitäten zur Nachuntersuchung nicht

aufgeklärter Fälle verfügen.

Die von der genomDE-Initiative des BMG geplante Einführung der WGS-gestützten Diagnostik, welche als essentiellen Bestandteil die Etablierung einer nationalen Genomdatenbank vorsieht, erfordert eine noch stringentere Standardisierung der Phäno- und Genotypisierung, besonders im Hinblick auf den beabsichtigten Austausch der Daten im Rahmen der Kooperation mit anderen nationalen Genommedizin-Programmen, wie bereits ausgeführt. Dies spricht dafür, auch die Methoden und die apparative Infrastruktur für die Genomsequenzierung zu vereinheitlichen, was sich am einfachsten durch eine Zentralisierung der WGS an einem oder ganz wenigen Standorten erreichen lässt.

Andererseits ist jedoch die Zahl der akademischen Institute, die apparativ und personell zur Durchführung der WES und der sachgerechten, internationalen Standards genügenden Interpretation der Befunde in der Lage wären, inzwischen deutlich größer als die der bisher dazu berechtigten 4 Zentren in Deutschland.

Um im internationalen Wettbewerb bestehen zu können und für die Einführung der Genomsequenzierung in die genetische Routineversorgung gerüstet zu sein, möchte das Institut für Humangenetik in Mainz in naher Zukunft die Synergieeffekte des ebenfalls in Mainz ansässigen Unternehmens TRON nutzen und neben der Exomsequenzierung auch die gesamtgenomische Sequenzierung einführen. Der mit der WGS einhergehende, drastisch steigende Rechenaufwand könnte durch die Anschaffung spezialisierter Hardware wie der Illumina DRAGEN Bio-IT Plattform oder eines Grafikkartenservers kompensiert werden. Im Bereich der Auswertung und Interpretation der Daten sind, wie oben bereits erwähnt, in nicht allzu ferner Zukunft neue und verbesserte Algorithmen auf Basis von künstlicher Intelligenz sowie wachsende Populations- und Genotyp-Phänotyp-Datenbanken zu erwarten. Mit diesen kann neben der Beurteilung von protein-kodierenden auch die Pathogenitätsvorhersage nicht-kodierender und regulatorischer Varianten verbessert werden.

6.7 Ausblick

Neben der derzeit in der genetischen Diagnostik nahezu ausschließlich eingesetzten „short-read“-HDS schreitet die Entwicklung neuer Sequenzierungstechniken, die eine Sequenzierung viel längerer DNA-Fragmente erlauben, rasch voran. So ermöglicht die Oxford Nanopore Technologie [Eisenstein, 2012] [Mikheyev und Tin, 2014] oder die Single Molecule, Real Time (SMRT) Sequenzierung von Pacific Bioscience [Levene *et al.*, 2003] die Sequenzierung von DNA-Fragmenten bis zu einer Länge von 175 kbp („long-read“-HDS). Die im Vergleich zu Illuminas SBS Technik (50 – 300 bp) viel längeren Leseweiten haben deutliche Vorteile. So erlaubt die viel größere Überlappung solcher langer Sequenzen die Assemblierung sogenannter Haplotypen, also einer Zuordnung von Varianten zu einem der beiden homologen Chromosomen, was für die klinische Interpretation von Sequenzdaten von großer Bedeutung ist. Auch ge-

lingt es mithilfe dieser Technik, Genomabschnitte vollständig zu sequenzieren, die lange repetitive Sequenzen oder sehr ähnliche Tandem-Duplikationen aufweisen [Huddleston *et al.*, 2014]. Ebenso ist es damit möglich, eine Expansion kurzer repetierter Sequenzen zu erkennen, auf die Krankheiten wie die Chorea Huntington oder das Fragile X-Syndrom zurückgehen, oder die Bruchpunkte von Deletionen und Insertionen basengenau zu bestimmen. Trotz dieser Vorteile kann sich die long-read-Technologie mit ihren weiten Leselängen bisher in der klinischen Routinediagnostik nicht gegen die durch SBS generierten kurzen Sequenzen durchsetzen, da SBS klare Vorteile bezüglich der Kosten und des Durchsatzes bietet. Zudem sind die für SBS benötigten Libraries einfacher zu generieren sowie zu automatisieren und erfordern eine deutlich geringere Mengen an DNA, die bei klinischen Untersuchungen nicht selten limitierend ist. Allerdings ist es nicht unwahrscheinlich, dass die long-read-HDS bei weiter sinkenden Kosten und zunehmender Präzision dieser Methode gegenüber der short-read-HDS weiter an Boden gewinnen wird.

Nicht nur die genetische Diagnostik, sondern auch die Genomforschung dürfte von diesen neuen Methoden profitieren und die Verantwortlichen für die genomDE–Initiative sowie andere nationale Genommedizin-Programme sollten diese Entwicklungen im Auge behalten.

Literaturverzeichnis

[Abdellah *et al.*, 2004] Abdellah, Z., Ahmadi, A., Ahmed, S., Aimable, M., Ainscough, R., Almeida, J., Almond, C., Ambler, A., Ambrose, K., Ambrose, K., Andrew, R., Andrews, D., Andrews, N., Andrews, D., Apweiler, E., Arbery, H., Archer, B., Ash, G., Ashcroft, K., Ashurst, J., Ashwell, R., Atkin, D., Atkinson, A., Atkinson, B., Attwood, J., Aubin, K., Auger, K., Avis, T., Babbage, A., Babbage, S., Bacon, J., Bagguley, C., Bailey, J., Baker, A., Banerjee, R., Bardill, S., Barker, D., Barker, G., Barker, D., Barlow, K., Baron, L., Barrett, A., Bartlett, R., Basham, D., Basham, V., Bateman, A., Bates, K., Baynes, C., Beard, L., Beard, S., Beare, D., Beasley, A., Beasley, H., Beasley, O., Beck, S., Bell, E., Bellerby, D., Bellerby, T., Bemrose, R., Bennett, J., Bentley, D., Bentley, A., Berks, M., Berks, M., Bethel, G., Bird, C., Birney, E., Bissell, H., Blackburne-Maze, S., Blakey, S., Bolton, C., Bonfield, J., Bonnett, R., Border, R., Bradley, A., Brady, N., Bray, J., Bray-Allen, S., Bridgeman, A., Brook, J., Brooking, S., Brown, A., Brown, C., Brown, J., Brown, M., Brown, M., Bruskiewich, R., Bryant, J., Buck, D., Buckle, V., Budd, C., Buller, S., Burberry, J., Burford, D., Burgess, J., Burrill, W., Burrows, C., Burton, J., Burton, C., Butcher, P., Butler, A., Cairns, M., Camm, N., Campbell, C., Canning, B., Carder, C., Carder, P., Carter, N., Cavanna, T., Chalk, S., Chan, K., Chapman, J., Charles, R., Chillingworth, N., Chothia, T., Chui, C., Clack, R., Clamp, M., Clark, A., Clark, G., Clark, K., Clark, S., Clark, S., Clark, R., Clarke, B., Clarke, E., Clarke, K., Clarke, A., Clarke, L., Clee, C., Clegg, S., Clifford, K., Coates, J., Cobley, V., Coffey, A., Coggill, P., Cole, L., Collier, R., Collings, S., Collins, J., Collins, P., Colman, L., Connolly, A., Connor, R., Conquer, J., Conroy, D., Constance, D., Cook, L., Cooper, J., Cooper, R., Cooper, R., Coppola, M., Copsey, T., Corby, N., Cornell, L., Cornell, R., Cornell, C., Cottage, A., Coulson, A., Coville, G., Cox, A., Cox, T., Coxhill, R., Craig, M., Crane, T., Crawley, M., Crew, V., Cuff, J., Culley, K., Cummings, A., Cummings, K., Cummings, P., Curran, A., Curwen, V., Cutts, J., Daniels, R., Davidson, L., Davies, J., Davies, J., Davies, N., Davies, R., Davis, J., Davis, J., Davis, M., Dawson, E., Deadman, R., Dean, P., Dear, S., Dearden, F., Delgado, M., Deloukas, P., Dennis, J., Dhami, P., Dibling, C., Dobbs, R., Dobson, R., Dockree, C., Doddington, D., Dodsworth, S., Doggett, N., Down, T., Dunham, A., Dunham, I., Dunn, A., Dunn, M., Durbin, R., Durham, J., Dutta, I., Dwyer, R., Dyer, L., Earthrowl, M., Eastham, T., Eastham, E., Edwards, C., Edwards, K., Ellington, A., Elliott, D., Ellwood, M., Emberson, B., Errington, H., Evans, G., Evans, J., Evans, K., Evans, R., Eyras, E., Faulkner, L., Fellingham, C., Feltwell, T., Fennell, S., Finn, R., Flack, T., Felming,

C., Fleming, K., Flint, J., Flint, M., Floyd, Y., Footman, S., Fowler, J., Frame, D., Francis, M., Francis, S., Frankish, A., Frankland, J., Fraser, A., Fraser, D., French, L., Fricker, D., Frost, D., Frost, J., Frost, L., Frost, C., Fuller, L., Fullerton, K., Gardner, A., Garner, P., Garnett, J., Gatland, L., Gatland, L., Ghori, J., Gibbs, B., Gibson, D., Gibson, E., Gilbert, J., Gilby, L., Gillson, C., Glithero, R., Gooderham, A., Gorton, M., Grafham, D., Grant, M., Grant, S., Gray, I., Gray, E., Green, L., Greenhalgh, J., Greenhill, J., Griffiths-Jones, S., Gregg, P., Gregory, S., Gribble, S., Griffiths, C., Griffiths, E., Griffiths, M., Grocock, R., Guthrie, I., Gwilliam, R., Hall, R., Halls, K., Hall-Tamly, G., Hamlett, J., Hammond, S., Hancock, J., Harding, A., Harley, J., Harper, D., Harper, G., Harper, P., Harradence, G., Harrison, C. L., Harrison, E., Harrison, R., Hart, E., Hassan, D., Hawkins, N., Hawley, K., Hayes, K., Heath, P., Heathcott, R., Hembry, C., Henderson, C., Herd, T., Hewitt, S., Higgs, D., Hillyard, G., Hinkins, R., Ho, S. J., Hodgson, D., Hoffs, M., Holden, J., Holdgate, J., Holloway, E., Holmes, I., Holmes, S., Holroyd, S., Hooper, A., Hopewell, L., Hopkins, B., Hornett, G., Hornsby, G., Hornsby, T., Horsley, S., Horton, R., Howard, P., Howden, P., Howe, K., Howell, G., Hubbard, T., Huckle, E., Hughes, J., Hughes, J., Hull, L., Hummeric, H., Humphray, S., Humphries, M., Hunt, A., Hunt, P., Hunt, S., Hunter, G., Hyde, D., Ince, M., Isherwood, J., Iyer, V., Izatt, J., Izmajlowicz, M., Jareborg, N., Jassal, B., Jeffery, G., Jeffery, K., Jeffrey, C., Jekosch, K., Jenkins, L., Johansen, T., Johnson, C., Johnson, C., Johnson, D., Jolley, K., Jones, A., Jones, C., Jones, J., Jones, M., Jones, M., Jones, S., Joseph, S., Joy, A., Joy, L., Joy, V., Joyce, G., Jubb, M., Karunaratne, K., Kay, M., Kaye, D., Kearney, L., Keenan, S., Kelley, S., Kershaw, J., Kettleborough, R., Kidd, C., Kierstan, P., Kimberley, A., King, A., Kingsley, S., Kingswood, C., Klingle, G., Knights, A., Korf, I., Krogh, A., Lad, H., Laidlaw, P., Laing, M., Laird, G., Lambart, C., Lamble, R., Langford, C., Larke, B., Lau, T., L (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945.

[Acuna-Hidalgo *et al.*, 2016] Acuna-Hidalgo, R., Veltman, J. A., und Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. *Genome Biology*, 17(1):1–19.

[Adelaja, 2017] Adelaja, A. A. (2017). Human induced pluripotent stem cells generated neural cells behaving like brain and spinal cord cells: An insight into the involvement of retinoic acid and sonic hedgehog proteins. *International journal of health sciences*, 11(2):21–27.

[Adzhubei *et al.*, 2010] Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., und Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249.

[Afgan *et al.*, 2018] Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Ech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., und Blankenberg, D. (2018).

The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, 46(W1):W537–W544.

[Agarwal *et al.*, 2015] Agarwal, V., Bell, G. W., Nam, J. W., und Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4.

[Ali *et al.*, 2014] Ali, M. M., Li, F., Zhang, Z., Zhang, K., Kang, D. K., Ankrum, J. A., Le, X. C., und Zhao, W. (2014). Rolling circle amplification: A versatile tool for chemical biology, materials science and medicine. *Chemical Society Reviews*, 43(10):3324–3341.

[Allen *et al.*, 1992] Allen, R. C., Zoghbi, H. Y., Moseley, A. B., Rosenblatt, H. M., und Belmont, J. W. (1992). Methylation of HpaII and HhaI sites near the polymorphic CAG repeat in the human androgen-receptor gene correlates with X chromosome inactivation. *American Journal of Human Genetics*, 51(6):1229–1239.

[Alles *et al.*, 2019] Alles, J., Fehlmann, T., Fischer, U., Backes, C., Galata, V., Minet, M., Hart, M., Abu-Halima, M., Grässer, F. A., Lenhof, H. P., Keller, A., und Meese, E. (2019). An estimate of the total number of true human miRNAs. *Nucleic Acids Research*, 47(7):3353–3364.

[Altschul *et al.*, 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., und Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

[Andrews, 2010] Andrews, S. (2010). FastQC Manual. in: https://dnacore.missouri.edu/PDF/FastQC_Manual.pdf (Zugriff am 24.03.2020).

[Andrews *et al.*, 2010] Andrews, S., Krueger, F., Segonds-Pichon, Anne Biggins, L., Krueger, C., und Wingett, S. (2010). FastQC. in: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Zugriff am 24.03.2020).

[Ashburner *et al.*, 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., und Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

[Augsten, 2017] Augsten, S. (2017). Was ist Perl. in: ttps://www.dev-insider.de/was-ist-perl-a-583568/ (Zugriff am 23.11.2020).

[Auton *et al.*, 2015] Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., Marth, G. T., McVean, G. A., Nickerson, D. A., Schmidt,

J. P., Sherry, S. T., Wang, J., Wilson, R. K., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., Zhu, Y., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Gupta, N., Gharani, N., Toji, L. H., Gerry, N. P., Resch, A. M., Barker, J., Clarke, L., Gil, L., Hunt, S. E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R. E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Borodina, T. A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M. L., Fulton, L., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T. M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Davies, C. J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Campbell, C. L., Kong, Y., Marcketta, A., Yu, F., Antunes, L., Bainbridge, M., Sabo, A., Huang, Z., Coin, L. J., Fang, L., Li, Q., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Garrison, E. P., Kural, D., Lee, W. P., Leong, W. F., Stromberg, M., Ward, A. N., Wu, J., Zhang, M., Daly, M. J., DePristo, M. A., Handsaker, R. E., Banks, E., Bhatia, G., Del Angel, G., Genovese, G., Li, H., Kashin, S., McCarroll, S. A., Nemesh, J. C., Poplin, R. E., Yoon, S. C., Lihm, J., Makarov, V., Gottipati, S., Keinan, A., Rodriguez-Flores, J. L., Rausch, T., Fritz, M. H., Stütz, A. M., Beal, K., Datta, A., Herrero, J., Ritchie, G. R., Zerbino, D., Sabeti, P. C., Shlyakhter, I., Schaffner, S. F., Vitti, J., Cooper, D. N., Ball, E. V., Stenson, P. D., Barnes, B., Bauer, M., Cheetham, R. K., Cox, A., Eberle, M., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E. E., Batzer, M. A., Konkel, M. K., Walker, J. A., MacArthur, D. G., Lek, M., Herwig, R., Ding, L., Koboldt, D. C., Larson, D., Ye, K., Gravel, S., Swaroop, A., Chew, E., Lappalainen, T., Erlich, Y., Gymrek, M., Willems, T. F., Simpson, J. T., Shriver, M. D., Rosenfeld, J. A., Bustamante, C. D., Montgomery, S. B., De La Vega, F. M., Byrnes, J. K., Carroll, A. W., DeGorter, M. K., Lacroute, P., Maples, B. K., Martin, A. R., Moreno-Estrada, A., Shringarpure, S. S., Zakharia, F., Halperin, E., Baran, Y., Cerveira, E., Hwang, J., Malhotra, A., Plewczynski, D., Radew, K., Romanovitch, M., Zhang, C., Hyland, F. C., Craig, D. W., Christoforides, A., Homer, N., Izatt, T., Kurdoglu, A. A., Sinari, S. A., Squire, K., Xiao, C., Sebat, J., Antaki, D., Gujral, M., Noor, A., Ye, K., Burchard, E. G., Hernandez, R. D., Gignoux, C. R., Haussler, D., Katzman, S. J., Kent, W. J., Howie, B., Ruiz-Linares, A., Dermitzakis, E. T., Devine, S. E., Kang, H. M., Kidd, J. M., Blackwell, T., Caron, S., Chen, W., Emery, S., Fritsche, L., Fuchsberger, C., Jun, G., Li, B., Lyons, R., Scheller, C., Sidore, C., Song, S., Sliwerska, E., Taliun, D., Tan, A., Welch, R., Wing, M. K., Zhan, X., Awadalla, P., Hodgkinson, A., Li, Y., Shi,

X., Quitadamo, A., Lunter, G., Marchini, J. L., Myers, S., Churchhouse, C., Delaneau, O., Gupta-Hinch, A., Kretzschmar, W., Iqbal, Z., Mathieson, I., Menelaou, A., Rimmer, A., Xifara, D. K., Oleksyk, T. K., Fu, Y., Liu, X., Xiong, M., Jorde, L., Witherspoon, D., Xing, J., Browning, B. L., Browning, S. R., Hormozdiari, F., Sudmant, P. H., Khurana, E., Tyler-Smith, C., Albers, C. A., Ayub, Q., Chen, Y., Colonna, V., Jostins, L., Walter, K., Xue, Y., Gerstein, M. B., Abyzov, A., Balasubramanian, S., Chen, J., Clarke, D., Fu, Y., Harmanci, A. O., Jin, M., Lee, D., Liu, J., Mu, X. J., Zhang, J., Zhang, Y., Hartl, C., Shakir, K., Degenhardt, J., Meiers, S., Raeder, B., Casale, F. P., Stegle, O., Lameijer, E. W., Hall, I., Bafna, V., Michaelson, J., Gardner, E. J., Mills, R. E., Dayama, G., Chen, K., Fan, X., Chong, Z., Chen, T., Chaisson, M. J., Huddleston, J., Malig, M., Nelson, B. J., Parrish, N. F., Blackburne, B., Lindsay, S. J., Ning, Z., Zhang, Y., Lam, H., Sisu, C., Challis, D., Evani, U. S., Lu, J., Nagaswamy, U., Yu, J., Li, W., Habegger, L., Yu, H., Cunningham, F., Dunham, I., Lage, K., (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

[Avery *et al.*, 1944] Avery, O. T., Macleod, C. M., und McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *Journal of Experimental Medicine*, 79(2):137–158.

[Backenroth *et al.*, 2014] Backenroth, D., Homsy, J., Murillo, L. R., Glessner, J., Lin, E., Brueckner, M., Lifton, R., Goldmuntz, E., Chung, W. K., und Shen, Y. (2014). CANOES: Detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Research*, 42(12):e97.

[Baker *et al.*, 2019] Baker, S. W., Murrell, J. R., Nesbitt, A. I., Pechter, K. B., Balciuniene, J., Zhao, X., Yu, Z., Denenberg, E. H., DeChene, E. T., Wilkens, A. B., Bhoj, E. J., Guan, Q., Dulik, M. C., Conlin, L. K., Abou Tayoun, A. N., Luo, M., Wu, C., Cao, K., Sarmady, M., Bedoukian, E. C., Tarpinian, J., Medne, L., Skraban, C. M., Deardorff, M. A., Krantz, I. D., Krock, B. L., und Santani, A. B. (2019). Automated Clinical Exome Reanalysis Reveals Novel Diagnoses. *Journal of Molecular Diagnostics*, 21(1):38–48.

[Ballif *et al.*, 2006] Ballif, B. C., Rorem, E. A., Sundin, K., Lincicum, M., Gaskin, S., Coppinger, J., Kashork, C. D., Shaffer, L. G., und Bejjani, B. A. (2006). Detection of low-level mosaicism by array CGH in routine diagnostic specimens. *American Journal of Medical Genetics, Part A*, 140(24):2757–2767.

[Bateson und Bateson, 1928] Bateson, W. und Bateson, B. (1928). *William Bateson, F. R. S., naturalist: his essays & addresses, together with a short account of his life.* Cambridge University Press, Cambridge, UK.

[Berkovic *et al.*, 2019] Berkovic, S. F., Goldstein, D. B., Heinzen, E. L., Laughlin, B. L., Lowenstein, D. H., Lubbers, L., Stewart, R., Whittemore, V., Angione, K., Bazil, C. W., Bier, L., Bluvstein, J., Brimble, E., Campbell,

C., Cavalleri, G., Chambers, C., Choi, H., Cilio, M. R., Ciliberto, M., Cornes, S., Delanty, N., Demarest, S., Devinsky, O., Dlugos, D., Dubbs, H., Dugan, P., Ernst, M. E., Gibbons, M., Goodkin, H. P., Helbig, I., Jansen, L., Johnson, K., Joshi, C., Lippa, N. C., Marsh, E., Martinez, A., Millichap, J., Mulhern, M. S., Numis, A., Park, K., Pippucci, T., Poduri, A., Porter, B., Regan, B., Sands, T. T., Scheffer, I. E., Schreiber, J. M., Sheidley, B., Singhal, N., Smith, L., Sullivan, J., Taylor, A., Tolete, P., Afgani, T. M., Aggarwal, V., Burgess, R., Dixon-Salazar, T., Hemati, P., Milder, J., Petrovski, S., Revah-Politi, A., und Stong, N. (2019). The Epilepsy Genetics Initiative: Systematic reanalysis of diagnostic exomes increases yield. *Epilepsia*, 60(5):797–806.

[Botstein *et al.*, 1980] Botstein, D., White, R. L., Skolnick, M., und Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Gen*, 32:314–331.

[Boveri, 1902] Boveri, T. (1902). Über mehrpolige Mitosen als Mittel zur Analyse des Zellkerns. [Concerning multipolar mitoses as a means of analysing the cell nucleus.]. *C. Kabitzch, Würzburg and Verh. d. phys. med. Ges. zu Würzburg. N.F., Bd. 35.*

[Bowers *et al.*, 2009] Bowers, J., Mitchell, J., Beer, E., Buzby, P. R., Causey, M., Efcavitch, J. W., Jarosz, M., Krzymanska-Olejnik, E., Kung, L., Lipson, D., Lowman, G. M., Marappan, S., McInerney, P., Platt, A., Roy, A., Siddiqi, S. M., Steinmann, K., und Thompson, J. F. (2009). Virtual terminator nucleotides for next-generation DNA sequencing. *Nature Methods*, 6(8):593–595.

[Brenner *et al.*, 1961] Brenner, S., Jacob, F., und Meselson, M. (1961). An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 190(4776):576–581.

[Broad Institute, 2020a] Broad Institute (2020a). GATK - Base Quality Score Recalibration. in: https://gatk.broadinstitute.org/hc/en-us/articles/360035890531-Base-Quality-Score-Recalibration-BQSR- (Zugriff am 20.03.2020).

[Broad Institute, 2020b] Broad Institute (2020b). GATK - HaplotypeCaller in a nutshell. in: https://gatk.broadinstitute.org/hc/en-us/articles/360035531412-HaplotypeCaller-in-a-nutshell (Zugriff am 20.03.2020).

[Broad Institute, 2020c] Broad Institute (2020c). GATK - (How to) Call common and rare germline copy number variants. in: https://gatk.broadinstitute.org/hc/en-us/articles/360035531152 (Zugriff am 20.03.2020).

[Broad Institute, 2020d] Broad Institute (2020d). GATK Best Practice Workflow. in: https://gatk.broadinstitute.org/hc/en-us (Zugriff am 12.08.2020).

[Bruel *et al.*, 2019] Bruel, A. L., Nambot, S., Quéré, V., Vitobello, A., Thevenon, J., Assoum, M., Moutton, S., Houcinat, N., Lehalle, D., Jean-Marçais, N., Verloès, A., Karsenti, A., Goldenberg, A., Jacquette, A., Jouret, B., Laudier, B., Coubes, C., Francannet, C., Lehalle, D., Geneviève, D., Heron, D., Lacombe, D., Schaefer, E., Lacaze, E., Jacquemin, E., Prieur, F., Laffarge, F., Petit, F., Feillet, F., Morin, G., Diene, G., Lespinasse, J., Amiel, J., Melki, J., Lambert, L., Perrin, L., Pinson, L., Jacquemont, M. L., Cordier-Alex, M. P., Lebrun, M., Gérard-Blanluet, M., Willems, M., Rossi, M., Chassaing, N., Philip, N., Touraine, R., El-Chehadeh, S., Audebert-Bellanger, S., Blesson, S., Capri, Y., Chevarin, M., Jouan, T., Poë, C., Callier, P., Tisserand, E., Philippe, C., Them, F. T. M., Duffourd, Y., Faivre, L., und Thauvin-Robinet, C. (2019). Increased diagnostic and new genes identification outcome using research reanalysis of singleton exome sequencing. *European Journal of Human Genetics*, 27(10):1519–1531.

[Burrows und Wheeler, 1994] Burrows, M. und Wheeler, D. (1994). A block-sorting lossless data compression algorithm. *Algorithm, Data Compression*, (124):18.

[Carig *et al.*, 1990] Carig, A. G., Nizetic, D., Hoheisel, J. D., Zehetner, G., und Lehrach, H. (1990). Ordering of cosmid clones covering the Herpes simplex virus type I (HSV-I) genome: A test case for fingerprinting by hybridisation. *Nucleic Acids Research*, 18(9):2653–2660.

[Caron *et al.*, 2019] Caron, B., Luo, Y., und Rausell, A. (2019). NCBoost classifies pathogenic non-coding variants in Mendelian diseases through supervised learning on purifying selection signals in humans. *Genome Biology*, 20(1):1–22.

[Chargaff *et al.*, 1952] Chargaff, E., Lipshitz, R., und Green, C. (1952). Composition of the desoxypentose nucleic acids of four genera of sea-urchin. *The Journal of biological chemistry*, 195(1):155–160.

[Chow *et al.*, 1977] Chow, L. T., Roberts, J. M., Lewis, J. B., und Broker, T. R. (1977). A map of cytoplasmic RNA transcripts from lytic adenovirus type 2, determined by electron microscopy of RNA:DNA hybrids. *Cell*, 11(4):819–836.

[Cimino, 2020] Cimino, A. (2020). Here's Why Illumina Shares Barely Budged When It Gave Up On Its Takeover to The Tune Of $98 Million. in: https://www.fool.com/investing/2020/01/08/heres-why-illumina-shares-barely-budged-when-it-ga.aspx (Zugriff am 15.10.2020).

[Claverie, 2001] Claverie, J. M. (2001). Gene number: What if there are only 30,000 human genes? *Science*, 291(5507):1255–1257.

[Cock *et al.*, 2009] Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., und Rice, P. M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771.

[Collins, 1992] Collins, F. S. (1992). Positional cloning: Let's not call it reverse anymore. *Nature Genetics*, 1(1):3–6.

[Collins und Fink, 1995] Collins, F. S. und Fink, L. (1995). The Human Genome Project. *Alcohol Health and Research World*, 19(3):190.

[Cooper, 1992] Cooper, D. N. (1992). Regulatory mutations and human genetic disease. *Annals of Medicine*, 24(6):427–437.

[Cunningham *et al.*, 2019] Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., Bennett, R., Bhai, J., Billis, K., Boddu, S., Cummins, C., Davidson, C., Dodiya, K. J., Gall, A., Girón, C. G., Gil, L., Grego, T., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., Kay, M., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Marugán, J. C., Maurel, T., McMahon, A. C., Moore, B., Morales, J., Mudge, J. M., Nuhn, M., Ogeh, D., Parker, A., Parton, A., Patricio, M., Abdul Salam, A. I., Schmitt, B. M., Schuilenburg, H., Sheppard, D., Sparrow, H., Stapleton, E., Szuba, M., Taylor, K., Threadgold, G., Thormann, A., Vullo, A., Walts, B., Winterbottom, A., Zadissa, A., Chakiachvili, M., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Yates, A. D., Zerbino, D. R., und Flicek, P. (2019). Ensembl 2019. *Nucleic Acids Research*, 47(D1):D745–D751.

[Danecek *et al.*, 2011] Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., und Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158.

[D'Antonio *et al.*, 2013] D'Antonio, M., D'Onorio De Meo, P., Paoletti, D., Elmi, B., Pallocca, M., Sanna, N., Picardi, E., Pesole, G., und Castrignanò, T. (2013). WEP: A high-performance analysis pipeline for whole-exome data. *BMC Bioinformatics*, 14(SUPPL7):S11.

[Davydov *et al.*, 2010] Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., und Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology*, 6(12).

[Dayhoff, 1969] Dayhoff, M. O. (1969). Computer analysis of protein evolution. *Scientific American*, 221(1):86–95.

[Department of Health and Social Care, 2013] Department of Health and Social Care (2013). DNA mapping to better understand cancer, rare diseases and infectious diseases - GOV.UK. in: https://www.gov.uk/government/news/dna-mapping-to-better-understand-cancer-rare-diseases-and-infectious-diseases (Zugriff am 23.09.2020).

[Dolzhenko *et al.*, 2017] Dolzhenko, E., van Vugt, J. J., Shaw, R. J., Bekritsky, M. A., Van Blitterswijk, M., Narzisi, G., Ajay, S. S., Rajan, V., Lajoie, B. R., Johnson, N. H., Kingsbury, Z., Humphray, S. J., Schellevis, R. D., Brands, W. J., Baker, M., Rademakers, R., Kooyman, M., Tazelaar, G. H., van Es, M. A., McLaughlin, R., Sproviero, W., Shatunov, A., Jones, A., Khleifat, A. A., Pittman, A., Morgan, S., Hardiman, O., Al-Chalabi, A., Shaw, C., Smith, B., Neo, E. J., Morrison, K., Shaw, P. J., Reeves, C., Winterkorn, L., Wexler, N. S., Housman, D. E., Ng, C. W., Li, A. L., Taft, R. J., van den Berg, L. H., Bentley, D. R., Veldink, J. H., und Eberle, M. A. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Research*, 27(11):1895–1903.

[Donis-Keller *et al.*, 1987a] Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T. P., Bowden, D. W., Smith, D. R., Lander, E. S., Botstein, D., Akots, G., Rediker, K. S., Gravius, T., Brown, V. A., Rising, M. B., Parker, C., Powers, J. A., Watt, D. E., Kauffman, E. R., Bricker, A., Phipps, P., Muller-Kahle, H., Fulton, T. R., Ng, S., Schumm, J. W., Braman, J. C., Knowlton, R. G., Barker, D. F., Crooks, S. M., Lincoln, S. E., Daly, M. J., und Abrahamson, J. (1987a). A genetic linkage map of the human genome. *Cell*, 51(2):319–337.

[Donis-Keller *et al.*, 1987b] Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T. P., Bowden, D. W., Smith, D. R., Lander, E. S., Botstein, D., Akots, G., Rediker, K. S., Gravius, T., Brown, V. A., Rising, M. B., Parker, C., Powers, J. A., Watt, D. E., Kauffman, E. R., Bricker, A., Phipps, P., Muller-Kahle, H., Fulton, T. R., Ng, S., Schumm, J. W., Braman, J. C., Knowlton, R. G., Barker, D. F., Crooks, S. M., Lincoln, S. E., Daly, M. J., und Abrahamson, J. (1987b). A genetic linkage map of the human genome. *Cell*, 51(2):319–337.

[Dunham *et al.*, 1999] Dunham, I., Shimizu, N., Roe, B. A., Chissoe, S., Dunham, I., Hunt, A. R., Collins, J. E., Bruskiewich, R., Beare, D. M., Clamp, M., Smink, L. J., Ainscough, R., Almeida, J. P., Babbage, A., Bagguley, C., Bailey, J., Barlow, K., Bates, K. N., Beasley, O., Bird, C. P., Blakey, S., Bridgeman, A. M., Buck, D., Burgess, J., Burrill, W. D., Burton, J., Carder, C., Carter, N. P., Chen, Y., Clark, G., Clegg, S. M., Cobley, V., Cole, C. G., Collier, R. E., Connor, R. E., Conroy, D., Corby, N., Coville, G. J., Cox, A. V., Davis, J., Dawson, E., Dhami, P. D., Dockree, C., Dodsworth, S. J., Durbin, R. M., Ellington, A., Evans, K. L., Fey, J. M., Fleming, K., French, L., Garner, A. A., Gilbert, J. G., Goward, M. E., Grafham, D., Griffiths, M. N., Hall, C., Hall, R., Hall-Tamlyn, G., Heathcott, R. W., Ho, S., Holmes, S., Hunt, S. E., Jones, M. C., Kershaw, J., Kimberley, A., King, A., Laird, G. K., Langford, C. F., Leversha, M. A., Lloyd, C., Lloyd, D. M., Martyn, I. D., Mashreghi-Mohammadi, M., Matthews, L., Mccann, O. T., Mcclay, J., Mclaren, S., Mcmurray, A. A., Milne, S. A., Mortimore, B. J., Odell, C. N., Pavitt, R., Pearce, A. V., Pearson, D., Phillimore, B. J., Phillips, S. H., Plumb, R. W., Ramsay, H., Ramsey, Y., Rogers, L., Ross, M. T., Scott, C. E., Sehra, H. K., Skuce, C. D., Smalley, S., Smith, M. L.,

Soderlund, C., Spragon, L., Steward, C. A., Sulston, J. E., Swann, R. M., Vaudin, M., Wall, M., Wallis, J. M., Whiteley, M. N., Willey, D., Williams, L., Williams, S., Williamson, H., Wilmer, T. E., Wilming, L., Wright, C. L., Hubbard, T., Bentley, D. R., Beck, S., Rogers, J., Shimizu, N., Minoshima, S., Kawasaki, K., Sasaki, T., Asakawa, S., Kudoh, J., Shintani, A., Shibuya, K., Yoshizaki, Y., Aoki, N., Mitsuyama, S., Roe, B. A., Chen, F., Chu, L., Crabtree, J., Deschamps, S., Do, A., Do, T., Dorman, A., Fang, F., Fu, Y., Hu, P., Hua, A., Kenton, S., Lai, H., Lao, H. I., Lewis, J., Lewis, S., Lin, S. P., Loh, P., Malaj, E., Nguyen, T., Pan, H., Phan, S., Qi, S., Qian, Y., Ray, L., Ren, Q., Shaull, S., Sloan, D., Song, L., Wang, Q., Wang, Y., Wang, Z., White, J., Willingham, D., Wu, H., Yao, Z., Zhan, M., Zhang, G., Chissoe, S., Murray, J., Miller, N., Minx, P., Fulton, R., Johnson, D., Bemis, G., Bentley, D., Bradshaw, H., Bourne, S., Cordes, M., Du, Z., Fulton, L., Goela, D., Graves, T., Hawkins, J., Hinds, K., Kemp, K., Latreille, P., Layman, D., Ozersky, P., Rohlfing, T., Scheet, P., Walker, C., Wamsley, A., Wohldmann, P., Pepin, K., Nelson, J., Korf, I., Bedell, J. A., Hillier, L., Mardis, E., Waterston, R., Wilson, R., Emanuel, B. S., Shaikh, T., Kurahashi, H., Saitta, S., Budarf, M. L., Mcdermid, H. E., Johnson, A., Wong, A. C., Morrow, B. E., Edelmann, L., Kim, U. J., Shizuya, H., Simon, M. I., Dumanski, J. P., Peyrard, M., Kedra, D., Seroussi, E., Fransson, I., Tapia, I., Bruder, C. E., und O'Brien, K. P. (1999). The DNA sequence of human chromosome 22. *Nature*, 402(6761):489–495.

[Durbin *et al.*, 1998] Durbin, R., Eddy, S. R., Krogh, A., und Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids (1999)*, volume 1. New York : Cambridge University Press, Cambridge, UK.

[ecSeq Bioinformatics, 2017] ecSeq Bioinformatics (2017). What is mate pair sequencing for? in: https://www.ecseq.com/support/ngs/what-is-mate-pair-sequencing-useful-for (Zugriff am 02.03.2020).

[Ehrhart *et al.*, 2019] Ehrhart, F., Willighagen, E. L., Kutmon, M., van Hofften, M., Sangani, N. B., Curfs, L. G., und Evelo, C. T. (2019). History of rare diseases and their genetic causes - A data driven approach. *bioRxiv*, page 595819.

[Eisenstein, 2012] Eisenstein, M. (2012). Oxford Nanopore announcement sets sequencing sector abuzz. *Nature Biotechnology*, 30(4):295–296.

[EMBL-EBI, 2014] EMBL-EBI (2014). About the Ensembl Project. in: https://www.ensembl.org/info/about/index.html (Zugriff am 19.04.2020).

[ENSEMBL, 2020a] ENSEMBL (2020a). ENSEMBL - BED File Format. in: https://www.ensembl.org/info/website/upload/bed.html (Zugriff am 17.03.2020).

[ENSEMBL, 2020b] ENSEMBL (2020b). Transcript Support Level (TSL) Homo sapiens - Ensembl genome browser 99. in: http://www.ensembl.org/Help/Glossary?id=492 (Zugriff am 27.03.2020).

[EURORDIS, 2020a] EURORDIS (2020a). Nicht diagnostizierte seltene Erkrankungen. in: https://www.eurordis.org/de/content/nicht-diagnostizierte-seltene-erkrankungen#2 (Zugriff am 27.10.2020).

[EURORDIS, 2020b] EURORDIS (2020b). Über seltene Krankheiten. in: https://www.eurordis.org/de/seltene-krankheiten (Zugriff 27.10.2020).

[Ezkurdia *et al.*, 2014] Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., und Tress, M. L. (2014). Multiple evidence strands suggest that theremay be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*, 23(22):5866–5878.

[Farnaes *et al.*, 2018] Farnaes, L., Hildreth, A., Sweeney, N. M., Clark, M. M., Chowdhury, S., Nahas, S., Cakici, J. A., Benson, W., Kaplan, R. H., Kronick, R., Bainbridge, M. N., Friedman, J., Gold, J. J., Ding, Y., Veeraraghavan, N., Dimmock, D., und Kingsmore, S. F. (2018). Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. *npj Genomic Medicine*, 3(1):10.

[Ferragina und Manzini, 2000] Ferragina, P. und Manzini, G. (2000). Opportunistic data structures with applications. In *Annual Symposium on Foundations of Computer Science - Proceedings*, pages 390–398. IEEE.

[Francioli *et al.*, 2015] Francioli, L. C., Polak, P. P., Koren, A., Menelaou, A., Chun, S., Renkens, I., Van Duijn, C. M., Swertz, M., Wijmenga, C., Van Ommen, G., Slagboom, P. E., Boomsma, D. I., Ye, K., Guryev, V., Arndt, P. F., Kloosterman, W. P., De Bakker, P. I., und Sunyaev, S. R. (2015). Genome-wide patterns and properties of de novo mutations in humans. *Nature Genetics*, 47(7):822–826.

[Franek *et al.*, 2011] Franek, K. J., Butler, J., Johnson, J., Simensen, R., Friez, M. J., Bartel, F., Moss, T., Dupont, B., Berry, K., Bauman, M., Skinner, C., Stevenson, R. E., und Schwartz, C. E. (2011). Deletion of the immunoglobulin domain of IL1RAPL1 results in nonsyndromic X-linked intellectual disability associated with behavioral problems and mild dysmorphism. *American Journal of Medical Genetics, Part A*, 155(5):1109–1114.

[Franklin und Gosling, 1953] Franklin, R. E. und Gosling, R. G. (1953). Molecular configuration in sodium thymonucleate. *Nature*, 171(4356):740–741.

[Fritz *et al.*, 2011] Fritz, M. H. Y., Leinonen, R., Cochrane, G., und Birney, E. (2011). Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Research*, 21(5):734–740.

[Fromer *et al.*, 2012] Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., Handsaker, R. E., McCarroll, S. A., O'Donovan, M. C., Owen, M. J., Kirov, G., Sullivan, P. F., Hultman, C. M., Sklar, P., und Purcell, S. M. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *American Journal of Human Genetics*, 91(4):597–607.

[Gajecka, 2016] Gajecka, M. (2016). Unrevealed mosaicism in the next-generation sequencing era. *Molecular Genetics and Genomics*, 291(2):513–530.

[Genomics England, 2018] Genomics England (2018). The 100,000 Genomes Project by numbers — Genomics England. in: https://www.genomicsengland.co.uk/the-100000-genomes-project-by-numbers/ (Zugriff am 23.09.2020).

[Gilissen *et al.*, 2014] Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., Van De Vorst, M., Van Bon, B. W., Willemsen, M. H., Kwint, M., Janssen, I. M., Hoischen, A., Schenck, A., Leach, R., Klein, R., Tearle, R., Bo, T., Pfundt, R., Yntema, H. G., De Vries, B. B., Kleefstra, T., Brunner, H. G., Vissers, L. E., und Veltman, J. A. (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature*, 511(7509):344–347.

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

[Goldmann *et al.*, 2016] Goldmann, J. M., Wong, W. S., Pinelli, M., Farrah, T., Bodian, D., Stittrich, A. B., Glusman, G., Vissers, L. E., Hoischen, A., Roach, J. C., Vockley, J. G., Veltman, J. A., Solomon, B. D., Gilissen, C., und Niederhuber, J. E. (2016). Parent-of-origin-specific signatures of de novo mutations. *Nature Genetics*, 48(8):935–939.

[Gonzaga-Jauregui *et al.*, 2012] Gonzaga-Jauregui, C., Lupski, J. R., und Gibbs, R. A. (2012). Human Genome Sequencing in Health and Disease. *Annual Review of Medicine*, 63(1):35–61.

[Griffith, 1928] Griffith, F. (1928). The Significance of Pneumococcal Types. *Journal of Hygiene*, 27(2):113–159.

[Griffiths-Jones, 2004] Griffiths-Jones, S. (2004). The microRNA Registry. *Nucleic Acids Research*, 32(90001):109D–111.

[Grozeva *et al.*, 2015] Grozeva, D., Carss, K., Spasic-Boskovic, O., Tejada, M. I., Gecz, J., Shaw, M., Corbett, M., Haan, E., Thompson, E., Friend, K., Hussain, Z., Hackett, A., Field, M., Renieri, A., Stevenson, R., Schwartz, C., Floyd, J. A., Bentham, J., Cosgrove, C., Keavney, B., Bhattacharya, S., Hurles, M., und Raymond, F. L. (2015). Targeted Next-Generation Sequencing Analysis of 1,000 Individuals with Intellectual Disability. *Human Mutation*, 36(12):1197–1204.

[Gusella *et al.*, 1983] Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. R., Sakaguchi, A. Y., Young, A. B., Shoulson, I., Bonilla, E., und Martin, J. B. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, 306(5940):234–238.

[Hamosh *et al.*, 2005] Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., und McKusick, V. A. (2005). Online Mendelian Inheritance in Man

(OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(DATABASE ISS.):514–517.

[Hancock *et al.*, 2004] Hancock, J. M., Zvelebil, M. J., Griffith, O. L., und Griffith, M. (2004). OMIM (Online Mendelian Inheritance in Man). In *Dictionary of Bioinformatics and Computational Biology*.

[Herold, 2003] Herold, H. (2003). *Linux-Unix-Shells*. Addison-Wesley, München.

[Höchsmann *et al.*, 2003] Höchsmann, M., Töller, T., Giegerich, R., und Kurtz, S. (2003). Local similarity in RNA secondary structures. *Proceedings of the 2003 IEEE Bioinformatics Conference, CSB 2003*, pages 159–168.

[Horner *et al.*, 2009] Horner, D. S., Pavesi, G., Castrignano', T., de Meo, P. D., Liuni, S., Sammeth, M., Picardi, E., und Pesole, G. (2009). Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefings in Bioinformatics*, 11(2):181–197.

[Hu *et al.*, 2014] Hu, H., Wienker, T. F., Musante, L., Kalscheuer, V. M., Kahrizi, K., Najmabadi, H., und Ropers, H. H. (2014). Integrated sequence analysis pipeline provides one-stop solution for identifying disease-causing mutations. *Human Mutation*, 35(12):1427–1435.

[Huang *et al.*, 2009] Huang, D. W., Sherman, B. T., und Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57.

[Huddleston *et al.*, 2014] Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P. H., Graves, T. A., Alkan, C., Dennis, M. Y., Wilson, R. K., Turner, S. W., Korlach, J., und Eichler, E. E. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Research*, 24(4):688–696.

[Huse *et al.*, 2007] Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., und Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, 8(7).

[Illumina, 2012] Illumina (2012). Reducing Whole-Genome Data Storage Footprint. in: https://www.illumina.com/ documents/products/whitepapers/whitepaper_datacompression.pdf (Zugriff am 05.06.2020).

[Illumina, 2013a] Illumina (2013a). Illumina Sequencing Technology. in: https://www.illumina.com/documents/products/techspotlights/ techspotlight_sequencing.pdf (Zugriff am 16.04.2020).

[Illumina, 2013b] Illumina (2013b). MiSeq Personal Sequncer - Illumina. in: http://www.illumina.com/%0Asystems/miseq.ilmn (Zugriff am 25.01.2013).

[Illumina, 2017a] Illumina (2017a). MiSeq Specifications: Key performance parameters. in: https://emea.illumina.com/systems/sequencing-platforms/miseq/specifications.html (Zugriff am 18.11.2020).

[Illumina, 2017b] Illumina (2017b). Sequencing Sample Sheet. in: https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/sequencing-sheet-format-specifications-technical-note-970-2017-004.pdf (Zugriff am 05.06.2020).

[Illumina, 2017c] Illumina (2017c). TruSeq DNA PCR-Free. in: https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_truseq_dna_pcr_free_sample_prep.pdf (Zugriff am 06.06.2020).

[Illumina, 2018] Illumina (2018). Illumina CMOS Chip and One-Channel SBS Chemistry. in: https://www.illumina.com/content/dam/illumina-marketing/documents/products/techspotlights/cmos-tech-note-770-2013-054.pdf (Zugriff am 20.05.2020).

[Illumina, 2019a] Illumina (2019a). bcl2fastq2 Conversion Software v2.20. in: https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/bcl2fastq/bcl2fastq2-v2-20-software-guide-15051736-03.pdf (Zugriff am 15.06.2020).

[Illumina, 2019b] Illumina (2019b). Illumina DRAGEN Bio-IT Platorm. in: https://science-docs.illumina.com/documents/Informatics/dragen-overview-data-sheet-970-2018-002/dragen-overview-data-sheet-970-2018-002-pdf.pdf (Zugriff am 25.05.2020).

[Illumina, 2020a] Illumina (2020a). MiSeq Reporter Software (MSR). in: https://emea.illumina.com/systems/sequencing-platforms/miseq/products-services/miseq-reporter.html (Zugriff am 20.05.2020).

[Illumina, 2020b] Illumina (2020b). NextSeq Series Specifications — Key performance parameters. in: https://emea.illumina.com/systems/sequencing-platforms/nextseq/specifications.html (Zugriff am 18.11.2020).

[Ingram, 1956] Ingram, V. M. (1956). A specific chemical difference between the globins of normal human and sickle-cell anæmia hæmoglobin. *Nature*, 178(4537):792–794.

[Irimia *et al.*, 2014] Irimia, M., Weatheritt, R. J., Ellis, J. D., Parikshak, N. N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallières, M., Tapial, J., Raj, B., O'Hanlon, D., Barrios-Rodiles, M., Sternberg, M. J., Cordes, S. P., Roth, F. P., Wrana, J. L., Geschwind, D. H., und Blencowe, B. J. (2014). A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell*, 159(7):1511–1523.

[Itakura *et al.*, 1977] Itakura, K., Hirose, T., Crea, R., Riggs, A. D., Heyneker, H. L., Bolivar, F., und Boyer, H. W. (1977). Expression in Escherichia coli of a chemically synthesized gene for the hormone somatostatin. *Science*, 198(4321):1056–1063.

[Jackson *et al.*, 1972] Jackson, D. A., Symons, R. H., und Berg, P. (1972). Biochemical method for inserting new genetic information into DNA of Simian

Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 69(10):2904–2909.

[Jaganathan *et al.*, 2019] Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., Chow, E. D., Kanterakis, E., Gao, H., Kia, A., Batzoglou, S., Sanders, S. J., und Farh, K. K. H. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, 176(3):535–548.e24.

[Jiang *et al.*, 1995] Jiang, T., Wang, L., und Zhang, K. (1995). Alignment of trees - an alternative to tree edit. *Theoretical Computer Science*, 143(1):137–148.

[Jiang *et al.*, 2015] Jiang, Y., Oldridge, D. A., Diskin, S. J., und Zhang, N. R. (2015). CODEX: A normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Research*, 43(6):e39.

[Johannsen, 1909] Johannsen, W. (1909). Elemente der Exakten Erblichkeitslehre. *Science*, 30(780):851–853.

[John *et al.*, 2004] John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., und Marks, D. S. (2004). Human microRNA targets. *PLoS Biology*, 2(11).

[Jorde und Wooding, 2004] Jorde, L. B. und Wooding, S. P. (2004). Genetic variation, classification and 'race'. *Nature Genetics*, 36(11 SUPPL. 1):S28–S33.

[Karczewski *et al.*, 2020] Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J. X., Samocha, K. E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A. H., Minikel, E. V., Weisburd, B., Lek, M., Ware, J. S., Vittal, C., Armean, I. M., Bergelson, L., Cibulskis, K., Connolly, K. M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M. E., Consortium, T. G. A. D., Neale, B. M., Daly, M. J., und MacArthur, D. G. (2020). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *Nature*, 581:434–443.

[Karimzadeh *et al.*, 2018] Karimzadeh, M., Ernst, C., Kundaje, A., und Hoffman, M. M. (2018). Umap and Bismap: Quantifying genome and methylome mappability. *Nucleic Acids Research*, 46(20):1–13.

[Keleher, 1993] Keleher, C. (1993). Translating the genetic library: The goals, methods, and applications of the human genome project. In *Bulletin of the Medical Library Association*, volume 81, pages 274–277. Medical Library Association.

[Kelly *et al.*, 1989] Kelly, R., Bulfield, G., Collick, A., Gibbs, M., und Jeffreys, A. J. (1989). Characterization of a highly unstable mouse minisatellite locus: Evidence for somatic mutation during early development. *Genomics*, 5(4):844–856.

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

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

[Kingsmore, 2012] Kingsmore, S. (2012). Comprehensive carrier screening and molecular diagnostic testing for recessive childhood diseases. *PLoS Currents*, 4(MAY 2012).

[Kircher *et al.*, 2014] Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., und Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315.

[Kong *et al.*, 2012] Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Wong, W. S., Sigurdsson, G., Walters, G. B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D. F., Helgason, A., Magnusson, O. T., Thorsteinsdottir, U., und Stefansson, K. (2012). Rate of de novo mutations and the importance of father-s age to disease risk. *Nature*, 488(7412):471–475.

[Kosugi *et al.*, 2019] Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., und Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*, 20(1):117.

[Kozak *et al.*, 1993] Kozak, L., Chiurazzi, P., Genuardi, M., Pomponi, M. G., Zollino, M., und Neri, G. (1993). Mapping of a gene for non-specific X linked mental retardation: Evidence for linkage to chromosomal region Xp21. 1-Xp22.3. *Journal of Medical Genetics*, 30(10):866–869.

[Krüger und Rehmsmeier, 2006] Krüger, J. und Rehmsmeier, M. (2006). RNAhybrid: MicroRNA target prediction easy, fast and flexible. *Nucleic Acids Research*, 34(WEB. SERV. ISS.):451–454.

[Krumm *et al.*, 2012] Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., Quinlan, A. R., Nickerson, D. A., und Eichler, E. E.

(2012). Copy number variation detection and genotyping from exome sequence data. *Genome Research*, 22(8):1525–1532.

[Lander *et al.*, 2001] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Hong, M. L., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., De La Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A.,

Athanasiou, M., Schultz, R., Patrinos, A., und Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[Landrum und Kattman, 2018] Landrum, M. J. und Kattman, B. L. (2018). ClinVar at five years: Delivering on the promise. *Human Mutation*, 39(11):1623–1630.

[Landrum *et al.*, 2014] Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., und Maglott, D. R. (2014). ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(D1):980–985.

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

[Lee *et al.*, 2020] Lee, C. E., Singleton, K. S., Wallin, M., und Faundez, V. (2020). Rare Genetic Diseases: Nature's Experiments on Human Development. *iScience*, 23(5):101123.

[Lehninger, 2011] Lehninger, A. L. (2011). *Grundkurs Biochemie*. De Gruyter.

[Lejeune *et al.*, 1959] Lejeune, J., Gauthier, M., und Raymond, T. (1959). Les chromosomes humains en culture de tissus. *Comptes Rendus de l'Académie des sciences*, 248(4):602–603.

[Lennon und Lehrach, 1991] Lennon, G. G. und Lehrach, H. (1991). Hybridization analyses of arrayed cDNA libraries. *Trends in Genetics*, 7(10):314–317.

[Levene *et al.*, 2003] Levene, H. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., und Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, 299(5607):682–686.

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

[Li, 2012] Li, H. (2012). SAM tools Mailing List / Re: [Samtools-devel] CRAM 0.7. in: https://sourceforge.net/p/samtools/mailman/message/28989211/ (Zugriff am 12.03.2020).

[Li und Durbin, 2009] Li, H. und Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

[Li *et al.*, 2009a] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., und Durbin, R. (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

[Li und Wang, 2017] Li, Q. und Wang, K. (2017). InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *American Journal of Human Genetics*, 100(2):267–280.

[Li *et al.*, 2009b] Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., und Wang, J. (2009b). SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967.

[Liu *et al.*, 2019a] Liu, L., Sanderford, M. D., Patel, R., Chandrashekar, P., Gibson, G., und Kumar, S. (2019a). Biological relevance of computationally predicted pathogenicity of noncoding variants. *Nature Communications*, 10(1).

[Liu *et al.*, 2019b] Liu, P., Meng, L., Normand, E. A., Xia, F., Song, X., Ghazi, A., Rosenfeld, J., Magoulas, P. L., Braxton, A., Ward, P., Dai, H., Yuan, B., Bi, W., Xiao, R., Wang, X., Chiang, T., Vetrini, F., He, W., Cheng, H., Dong, J., Gijavanekar, C., Benke, P. J., Bernstein, J. A., Eble, T., Eroglu, Y., Erwin, D., Escobar, L., Gibson, J. B., Gripp, K., Kleppe, S., Koenig, M. K., Lewis, A. M., Natowicz, M., Mancias, P., Minor, L. K., Scaglia, F., Schaaf, C. P., Streff, H., Vernon, H., Uhles, C. L., Zackai, E. H., Wu, N., Reid Sutton, V., Beaudet, A. L., Muzny, D., Gibbs, R. A., Posey, J. E., Lalani, S., Shaw, C., Eng, C. M., Lupski, J. R., und Yang, Y. (2019b). Reanalysis of clinical exome sequencing data. *New England Journal of Medicine*, 380(25):2478–2480.

[Lockhart *et al.*, 1996] Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., und Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13):1675–1680.

[Lorenz *et al.*, 2011] Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., und Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26.

[Louadi *et al.*, 2019] Louadi, Z., Oubounyt, M., Tayara, H., und To Chong, K. (2019). Deep splicing code: Classifying alternative splicing events using deep learning. *Genes*, 10(8).

[Magi *et al.*, 2014] Magi, A., Tattini, L., Palombo, F., Benelli, M., Gialluisi, A., Giusti, B., Abbate, R., Seri, M., Gensini, G. F., Romeo, G., und Pippucci, T. (2014). H3M2: detection of runs of homozygosity from whole-exome sequencing data. *Bioinformatics (Oxford, England)*, 30(20):2852–2859.

[Margulies *et al.*, 2005] Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B.,

McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., und Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380.

[Mathews *et al.*, 1998] Mathews, D. H., Andre, T. C., Kim, J., Turner, D. H., und Zuker, M. (1998). An Updated Recursive Algorithm for RNA Secondary Structure Prediction with Improved Thermodynamic Parameters. *ACS Symposium Series*, 682:246–257.

[Mathews *et al.*, 2004] Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., und Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19):7287–7292.

[Maxam und Gilbert, 1977] Maxam, A. M. und Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–564.

[McCombie *et al.*, 2019] McCombie, W. R., McPherson, J. D., und Mardis, E. R. (2019). Next-generation sequencing technologies. *Cold Spring Harbor Perspectives in Medicine*, 9(11):a036798.

[McKenna *et al.*, 2010] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., und DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303.

[McKusick, 1966] McKusick, V. A. (1966). *Mendelian Inheritance in Man*. Elsevier.

[McLay *et al.*, 2011] McLay, R., Schulz, K. W., Barth, W. L., und Minyard, T. (2011). Best practices for the deployment and management of production HPC clusters. In *State of the Practice Reports, SC'11*, page 1, New York, New York, USA. ACM Press.

[Mendel, 1866] Mendel, G. (1866). *Versuche über Pflanzen-Hybriden*. Verhandlungen des Naturforschenden Vereines in Brünn. Naturforschender Verein Brünn, Brünn, 1. auflage edition.

[Michaelson *et al.*, 2012] Michaelson, J. J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A., Wu, W., Corominas, R., Peoples, Á., Koren, A., Gore, A., Kang, S., Lin, G. N., Estabillo, J., Gadomski, T., Singh, B., Zhang, K., Akshoomoff, N., Corsello, C., McCarroll, S., Iakoucheva, L. M., Li, Y., Wang, J., und Sebat, J. (2012). Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*, 151(7):1431–1442.

[Mighell *et al.*, 2000] Mighell, A. J., Smith, N. R., Robinson, P. A., und Markham, A. F. (2000). Vertebrate pseudogenes. *FEBS Letters*, 468(2-3):109–114.

[Mikheyev und Tin, 2014] Mikheyev, A. S. und Tin, M. M. (2014). A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources*, 14(6):1097–1102.

[Miller *et al.*, 2010] Miller, D. T., Adam, M. P., Aradhya, S., Biesecker, L. G., Brothman, A. R., Carter, N. P., Church, D. M., Crolla, J. A., Eichler, E. E., Epstein, C. J., Faucett, W. A., Feuk, L., Friedman, J. M., Hamosh, A., Jackson, L., Kaminsky, E. B., Kok, K., Krantz, I. D., Kuhn, R. M., Lee, C., Ostell, J. M., Rosenberg, C., Scherer, S. W., Spinner, N. B., Stavropoulos, D. J., Tepperberg, J. H., Thorland, E. C., Vermeesch, J. R., Waggoner, D. J., Watson, M. S., Martin, C. L., und Ledbetter, D. H. (2010). Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies. *American Journal of Human Genetics*, 86(5):749–764.

[Mooney, 2015] Mooney, S. D. (2015). Progress towards the integration of pharmacogenomics in practice. *Human Genetics*, 134(5):459–465.

[Mullany *et al.*, 2015] Mullany, L. E., Wolff, R. K., und Slattery, M. L. (2015). Effectiveness and usability of bioinformatics tools to analyze pathways associated with miRNA expression. *Cancer Informatics*, 14:121–130.

[Müller und Röder, 2004] Müller, H.-J. und Röder, T. (2004). *Der Experimentator: Microarrays*. Springer Spektrum, 1 edition.

[Mullis *et al.*, 1986] Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., und Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, 51(1):263–273.

[Murdock *et al.*, 2020] Murdock, D. R., Dai, H., Burrage, L. C., Rosenfeld, J. A., Ketkar, S., Müller, M. F., Yépez, V. A., Gagneur, J., Liu, P., Chen, S., Jain, M., Zapata, G., Bacino, C. A., Chao, H.-T., Moretti, P., Craigen, W. J., Hanchard, N. A., und Lee, B. (2020). Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. *Journal of Clinical Investigation*.

[Murray *et al.*, 1994] Murray, J. C., Buetow, K. H., Weber, J. L., Ludwigsen, S., Scherpbier-Heddema, T., Manion, F., Quillen, J., Sheffield, V. C., Sunden, S., Duyk, G. M., Weissenbach, J., Gyapay, G., Dib, C., Morrissette, J., Lathrop, G. M., Vignal, A., White, R., Matsunami, N., Gerken, S., Melis, R., Albertsen, H., Plaetke, R., Odelberg, S., Ward, D., Dausset, J., Cohen, D., und Cann, H. (1994). A comprehensive human linkage map with centimorgan density. *Science*, 265(5181):2049–2054.

[Najmabadi *et al.*, 2011] Najmabadi, H., Hu, H., Garshasbi, M., Zemojtel, T., Abedini, S. S., Chen, W., Hosseini, M., Behjati, F., Haas, S., Jamali, P.,

Zecha, A., Mohseni, M., Püttmann, L., Vahid, L. N., Jensen, C., Moheb, L. A., Bienek, M., Larti, F., Mueller, I., Weissmann, R., Darvish, H., Wrogemann, K., Hadavi, V., Lipkowitz, B., Esmaeeli-Nieh, S., Wieczorek, D., Kariminejad, R., Firouzabadi, S. G., Cohen, M., Fattahi, Z., Rost, I., Mojahedi, F., Hertzberg, C., Dehghan, A., Rajab, A., Banavandi, M. J. S., Hoffer, J., Falah, M., Musante, L., Kalscheuer, V., Ullmann, R., Kuss, A. W., Tzschach, A., Kahrizi, K., und Ropers, H. H. (2011). Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature*, 478(7367):57–63.

[Narasimhan *et al.*, 2016] Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., und Durbin, R. (2016). BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, 32(11):1749–1751.

[National Cancer Institute, 2014] National Cancer Institute (2014). SureSelectXT Automated Target Enrichment for Illumina Paired-End Sequencing Library SureSelectXT Automated Target Enrichment for Illumina Paired-End Sequencing Library. Technical report.

[National Human Genome Research Institute, 1990] National Human Genome Research Institute (1990). Understanding Our Genetic Inheritance: The First Five Years: Fiscal Years 1991-1995. Technical Report 20 April 2011, Technical Information Center, U.S. Department of Energy.

[Needleman und Wunsch, 1970] Needleman, S. B. und Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

[Neumeister *et al.*, 2018] Neumeister, B., Boehm, B., und Beneke, H. (2018). *Klinikleitfaden Labordiagnostik*. Urban & Fischer Verlag/Elsevier GmbH, 5. auflage edition.

[Ng *et al.*, 2009] Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E., Bamshad, M., Nickerson, D. A., und Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276.

[NHLBI, 2014] NHLBI (2014). Exome Variant Server. in: https://evs.gs.washington.edu/EVS/ (Zugriff am 27.03.2020).

[Nielsen *et al.*, 2011] Nielsen, R., Paul, J. S., Albrechtsen, A., und Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451.

[Nirenberg *et al.*, 1966] Nirenberg, M., Caskey, T., Marshall, R., Brimacombe, R., Kellogg, D., Doctor, B., Hatfield, D., Levin, J., Rottman, F., Pestka, S., Wilcox, M., und Anderson, F. (1966). The RNA code and protein synthesis. *Cold Spring Harbor symposia on quantitative biology*, 31:11–24.

[Nuland, 2003] Nuland, S. B. (2003). Doctors and Discoveries: Lives That Created Today's Medicine (review). *Bulletin of the History of Medicine*, 77(3):724–726.

[Nyrén *et al.*, 1993] Nyrén, P., Pettersson, B., und Uhlén, M. (1993). Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Analytical Biochemistry*, 208(1):171–175.

[OMIM, 2020] OMIM (2020). OMIM Gene Map Statistics. in: https://www.omim.org/statistics/geneMap (Zugriff am 14.10.2020).

[Orphanet, 2020] Orphanet (2020). Orphanet: Über Orphanet. in: https://www.orpha.net/consor/cgi-bin/Education_AboutOrphanet.php?lng=DE (Zugriff am 26.10.2020).

[Ostrow, 2016] Ostrow, S. R. (2016). CS 262 Lecture 4 Notes The Burrows-Wheeler Transform. Technical report.

[Packer *et al.*, 2016] Packer, J. S., Maxwell, E. K., O'Dushlaine, C., Lopez, A. E., Dewey, F. E., Chernomorsky, R., Baras, A., Overton, J. D., Habegger, L., und Reid, J. G. (2016). CLAMMS: A scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics*, 32(1):133–135.

[Pearson *et al.*, 2007] Pearson, B. M., Gaskin, D. J., Segers, R. P., Wells, J. M., Nuijten, P. J., und Van Vliet, A. H. (2007). The complete genome sequence of Campylobacter jejuni strain 81116 (NCTC11828). *Journal of Bacteriology*, 189(22):8402–8403.

[Pearson und Lipman, 1988] Pearson, W. R. und Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8):2444–2448.

[Petryszak *et al.*, 2016] Petryszak, R., Keays, M., Tang, Y. A., Fonseca, N. A., Barrera, E., Burdett, T., Füllgrabe, A., Fuentes, A. M. P., Jupp, S., Koskinen, S., Mannion, O., Huerta, L., Megy, K., Snow, C., Williams, E., Barzine, M., Hastings, E., Weisser, H., Wright, J., Jaiswal, P., Huber, W., Choudhary, J., Parkinson, H. E., und Brazma, A. (2016). Expression Atlas update - An integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Research*, 44(D1):D746–D752.

[Piton *et al.*, 2008] Piton, A., Michaud, J. L., Peng, H., Aradhya, S., Gauthier, J., Mottron, L., Champagne, N., Lafrenière, R. G., Hamdan, F. F., Joober, R., Fombonne, E., Marineau, C., Cossette, P., Dubé, M. P., Haghighi, P., Drapeau, P., Barker, P. A., Carbonetto, S., und Rouleau, G. A. (2008). Mutations in the calcium-related gene IL1RAPL1 are associated with autism. *Human Molecular Genetics*, 17(24):3965–3974.

[Poplin *et al.*, 2017] Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., der Auwera, G. A. V., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S.,

Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., und Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, page 22.

[Pounraja *et al.*, 2019] Pounraja, V. K., Jayakar, G., Jensen, M., Kelkar, N., und Girirajan, S. (2019). A machine-learning approach for accurate detection of copy number variants from exome sequencing. *Genome Research*, 29(7):1134–1143.

[QIAGEN, 2020] QIAGEN (2020). HGMD 2020.1 Release. in: https://digitalinsights.qiagen.com/news/blog/clinical/hgmd-spring-2020/ (Zugriff am 28.05.2020).

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

[Rédei, 2008] Rédei, G. P. (2008). Chromosome Walking. In *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*, pages 361–362. Springer Netherlands.

[Reuter und Mathews, 2010] Reuter, J. S. und Mathews, D. H. (2010). RNAstructure: Software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11(1):129.

[Richards *et al.*, 2015] Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., und Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5):405–424.

[Riffo-Campos *et al.*, 2016] Riffo-Campos, Á. L., Riquelme, I., und Brebi-Mieville, P. (2016). Tools for sequence-based miRNA target prediction: What to choose? *International Journal of Molecular Sciences*, 17(12):18.

[Robison, 2013] Robison, K. (2013). Ripples from 454s Shutdown Announcement. in: http://omicsomics.blogspot.com/2013/10/ripples-from-454s-shutdown-announcment.html (Zugriff am 14.10.2020).

[Rodenburg, 2018] Rodenburg, R. J. (2018). The functional genomics laboratory: functional validation of genetic variants. *Journal of Inherited Metabolic Disease*, 41(3):297–307.

[Rodriguez *et al.*, 2013] Rodriguez, J. M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J. J., Lopez, G., Valencia, A., und Tress, M. L. (2013). APPRIS: Annotation of principal and alternative splice isoforms. *Nucleic Acids Research*, 41(D1):110–117.

[Ronaghi *et al.*, 1996] Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., und Nyrén, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, 242(1):84–89.

[Ronaghi *et al.*, 1998] Ronaghi, M., Uhlén, M., und Nyrén, P. (1998). A sequencing method based on real-time pyrophosphate. *Science*, 281(5375):363–365.

[Rozen und Skaletsky, 2000] Rozen, S. und Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods in molecular biology (Clifton, N.J.)*, 132:365–386.

[Salfati *et al.*, 2019] Salfati, E. L., Spencer, E. G., Topol, S. E., Muse, E. D., Rueda, M., Lucas, J. R., Wagner, G. N., Campman, S., Topol, E. J., und Torkamani, A. (2019). Re-analysis of whole-exome sequencing data uncovers novel diagnostic variants and improves molecular diagnostic yields for sudden death and idiopathic diseases. *Genome Medicine*, 11(1):83.

[Samtools Organisation, 2020] Samtools Organisation (2020). The Variant Call Format Specification v4.3. in: https://samtools.github.io/hts-specs/VCFv4.3.pdf (Zugriff am 05.07.2020).

[Sanger *et al.*, 1977] Sanger, F., Nicklen, S., und Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467.

[Scheid, 2018] Scheid, C. (2018). Aktie der Woche. in: https://www.capital.de/geld-versicherungen/aktie-der-woche-illumina (Zugriff am 18.11.2020).

[Schena *et al.*, 1995] Schena, M., Shalon, D., Davis, R. W., und Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470.

[Schwarz *et al.*, 2014] Schwarz, J. M., Cooper, D. N., Schuelke, M., und Seelow, D. (2014). Mutationtaster2: Mutation prediction for the deep-sequencing age. *Nature Methods*, 11(4):361–362.

[Sherry *et al.*, 1999] Sherry, S. T., Ward, M., und Sirotkin, K. (1999). dbSNP - database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Research*, 9(8):677–679.

[Singer, 1982] Singer, M. F. (1982). SINEs and LINEs: Highly repeated short and long interspersed sequences in mammalian genomes. *Cell*, 28(3):433–434.

[Smith *et al.*, 1986] Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B., und Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071):674–679.

[Smith und Waterman, 1981] Smith, T. F. und Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

[Stenson *et al.*, 2017] Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A. D., und Cooper, D. N. (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics*, 136(6):665–677.

[Stosser *et al.*, 2018] Stosser, M. B., Lindy, A. S., Butler, E., Retterer, K., Piccirillo-Stosser, C. M., Richard, G., und McKnight, D. A. (2018). High frequency of mosaic pathogenic variants in genes causing epilepsy-related neurodevelopmental disorders. *Genetics in Medicine*, 20(4):403–410.

[Sutton, 1903] Sutton, W. S. (1903). THE CHROMOSOMES IN HEREDITY. *The Biological Bulletin*, 4(5):231–250.

[Taliun *et al.*, 2019] Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S.-b., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., Shetty, A. C., Blackwell, T. W., Wong, Q., Aguet, F., Albert, C., Alonso, A., Ardlie, K. G., Aslibekyan, S., Auer, P. L., Barnard, J., Barr, R. G., Becker, L. C., Beer, R. L., Benjamin, E. J., Bielak, L. F., Blangero, J., Boehnke, M., Bowden, D. W., Brody, J. A., Burchard, E. G., Cade, B. E., Casella, J. F., Chalazan, B. L., Chen, Y.-D. I., Cho, M. H., Choi, S. H., Chung, M. K., Clish, C. B., Correa, A., Curran, J., Custer, B., Darbar, D., Daya, M., de Andrade, M., DeMeo, D. L., Dutcher, S. K., Ellinor, P. T., Emery, L. S., Fatkin, D., Forer, L., Fornage, M., Franceschini, N., Fuchsberger, C., Fullerton, S. M., Germer, S., Gladwin, M. T., Gottlieb, D. J., Guo, X., Hall, M. E., He, J., Heard-Costa, N. L., Heckbert, S. R., Irvin, M. R., Johnsen, J. M., Johnson, A. D., Kardia, S. L. R., Kelly, T., Kelly, S., Kenny, E. E., Kiel, D. P., Klemmer, R., Konkle, B. A., Kooperberg, C., Köttgen, A., Lange, L. A., Lasky-Su, J. A., Levy, D., Lin, X., Lin, K.-H., Liu, C., Loos, R. J. F., Garman, L. D., Gerszten, R., Lubitz, S. A., Lunetta, K. L., Mak, A. C. Y., Manichaikul, A., Manning, A. K., Mathias, R. A., McManus, D. D., McGarvey, S. T., Meigs, J. B., Meyers, D. A., Mikulla, J. L., Minear, M. A., Mitchell, B., Mohanty, S., Montasser, M. E., Montgomery, C., Morrison, A. C., Murabito, J. M., Natale, A., Natarajan, P., Nelson, S. C., North, K. E., O'Connell, J., Palmer, N. D., Pankratz, N., Peloso, G. M., Peyser, P. A., Post, W. S., Psaty, B. M., Rao, D., Redline, S., Reiner, A. P., Roden, D., Rotter, J. I., Ruczinski, I., Sarnowski, C., Schoenherr, S., Seo, J.-S., Seshadri, S., Sheehan, V. A., Shoemaker, M. B., Smith, A. V., Smith, N. L., Smith, J. A., Sotoodehnia,

N., Stilp, A. M., Tang, W., Taylor, K. D., Telen, M., Thornton, T. A., Tracy, R. P., Berg, D. V. D., Vasan, R. S., Viaud-Martinez, K. A., Vrieze, S., Weeks, D. E., Weir, B. S., Weiss, S. T., Weng, L.-C., Willer, C. J., Zhang, Y., Zhao, X., Arnett, D. K., Ashley-Koch, A. E., Barnes, K. C., Boerwinkle, E., Gabriel, S., Gibbs, R., Rice, K. M., Rich, S. S., Silverman, E. K., Qasba, P., Gan, W., Papanicolaou, G. J., Nickerson, D. A., Browning, S. R., Zody, M. C., Zöllner, S., Wilson, J. G., Cupples, L. A., Laurie, C. C., Jaquish, C. E., Hernandez, R. D., O'Connor, T. D., und Abecasis, G. R. (2019). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv*, pages 1–46.

[Tate *et al.*, 2019] Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., Speedy, H. E., Stefancsik, R., Thompson, S. L., Wang, S., Ward, S., Campbell, P. J., und Forbes, S. A. (2019). COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1):D941–D947.

[Tawfik und Griffiths, 1998] Tawfik, D. S. und Griffiths, A. D. (1998). Man-made cell-like compartments for molecular evolution. *Nature Biotechnology*, 16(7):652–656.

[The Monarch Initiative, 2020] The Monarch Initiative (2020). Variants and Diseases. in: https://archive.monarchinitiative.org/202003/tsv/ variant_associations/variant_disease.9606.tsv (Zugriff am 26.05.2020).

[Tishkoff und Kidd, 2004] Tishkoff, S. A. und Kidd, K. K. (2004). Implications of biogeography of human populations for 'race' and medicine. *Nature Genetics*, 36(11S):s21–s27.

[Tjio und Levan, 1956] Tjio, J. H. und Levan, A. (1956). THE CHROMOSOME NUMBER OF MAN. *Hereditas*, 42(1-2):1–6.

[Tzur *et al.*, 2020] Tzur, S., Ph, D., und Officer, C. S. (2020). Pathorolo : Automating the reanalysis of accumulating unsolved exome cases using an AI model.

[University of California Santa Cruz, 2019] University of California Santa Cruz (2019). Genome Browser - Bed File Format. in: http://genome.ucsc.edu/FAQ/FAQformat#format1 (Zugriff am 17.03.2020).

[Untergasser *et al.*, 2012] Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., und Rozen, S. G. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Research*, 40(15).

[Uppsala University, 2020] Uppsala University (2020). Using CRAM to compress BAM files - Uppsala Multidisciplinary Center for Advanced Computational Science. in: https://www.uppmax.uu.se/support/user-guides/using-cram-to-compress-bam-files/ (Zugriff am 12.03.2020).

[Varma *et al.*, 2019] Varma, M., Paskov, K. M., Jung, J. Y., Chrisman, B. S., Stockham, N. T., Washington, P. Y., und Wall, D. P. (2019). Outgroup machine learning approach identifies single nucleotide variants in noncoding DNA associated with autism spectrum disorder. *Pacific Symposium on Biocomputing*, 24(2019):260–271.

[Vaser *et al.*, 2016] Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., und Ng, P. C. (2016). SIFT missense predictions for genomes. *Nature Protocols*, 11(1):1–9.

[Venter *et al.*, 2001] Venter, C. J., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Yuan Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M. H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Lai Cheng, M., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Ni Tint, N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Deslattes Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C.,

Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., und Zhu, X. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

[Voelkerding *et al.*, 2009] Voelkerding, K. V., Dames, S. A., und Durtschi, J. D. (2009). Next-generation sequencing:from basic research to diagnostics. *Clinical Chemistry*, 55(4):641–658.

[Wai Kan und Dozy, 1978] Wai Kan, Y. und Dozy, A. M. (1978). Antenatal Diagnosis of Sickle-Cell Anæmia By D.N.a. Analysis of Amniotic-Fluid Cells. *The Lancet*, 312(8096):910–912.

[Wang *et al.*, 2010] Wang, K., Li, M., und Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):1–7.

[Watson und Crick, 1953] Watson, J. D. und Crick, F. H. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

[Weber und Myers, 1997] Weber, J. L. und Myers, E. W. (1997). Human whole-genome shotgun sequencing. *Genome Research*, 7(5):401–409.

[Weise *et al.*, 2014] Weise, A., Mrasek, K., und Liehr, T. (2014). Zytogenetische und molekularzytogenetische Methoden in der Pränataldiagnostik. *Medizinische Genetik*, 26(4):391–397.

[Wikipedia Die freie Enzyklopädie., 2020] Wikipedia Die freie Enzyklopädie. (2020). Java (Programmiersprache) – Wikipedia. in: https://de.wikipedia.org/w/index.php?title=Java_(Programmiersprache) (Zugriff am 28.02.2020).

[Wikipedia Die freie Enzyklopädie, 2020] Wikipedia Die freie Enzyklopädie (2020). Java-Technologie – Wikipedia. in: https://de.wikipedia.org/w/index.php?title=Java-Technologie (Zugriff am 28.02.2020).

[Wikipedia The Free Encyclopedia, 2019] Wikipedia The Free Encyclopedia (2019). FASTQ format - Wikipedia. in: https://en.wikipedia.org/wiki/FASTQ_format (Zugriff am 12.03.2020).

[Yepez *et al.*, 2020] Yepez, V., Gusic, M., Mertes, C., Kopajtich, R., Smith, N., Prokisch, H., und Gagneur, J. (2020). No Title. In *The added value of RNA sequencing over WES for variant interpretation and diagnosis of patients with rare genetic disorders*. ASHG.

[Zhang *et al.*, 2020] Zhang, X., Wakeling, M., Ware, J., und Whiffin, N. (2020). Annotating high-impact 5'untranslated region variants with the UTRannotator. *Bioinformatics*, btaa783.

[Zou *et al.*, 2019] Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., und Telenti, A. (2019). A primer on deep learning in genomics. *Nature Genetics*, 51(1):12–18.

Abbildungsverzeichnis

Tabellenverzeichnis

Anhangsverzeichnis

Anhang I

Hardware

Als Hardware zur Analyse und Kurzzeitspeicherung der NGS-Daten steht am Institut für Humangenetik Mainz ein Server (Primärserver) zur Verfügung. Dieser besitzt 2 Intel Xeon E5-2630 v4 (Broadwell-EP) CPUs mit jeweils 10 Kernen, 256 GB DDR4-2400 DIMM Arbeitsspeicher (engl. Read Access Memory, RAM) und 12 2 TB (Terrabyte) Festplatten (engl. Hard Drive Disk, HDD) sowie 2 jeweils 960 GB (engl. Gigabyte) umfassende Halbleiterlaufwerke (engl. Solid-State-Drive, SSD). Als Backup und zur Spiegelung der Daten wird ein zweiter Server (Sekundärserver) verwendet. Er beinhaltet 2 Intel Xeon E5-2620 v4 (Broadwell-EP) CPUs mit jeweils 8 Kernen, ebenfalls 256 GB DDR4-2400 DIMM RAM und 12 2 TB HDDs. Über eine SAS (engl. Serial Attached SCSI) Schnittstelle ist an jeden Server ein DELL PowerVault MD3460 SAS zur Speichererweiterung und Verwendung als Langzeitspeicher angeschlossen. Jedes dieser Geräte schließt Platz für 60 HDD oder SSD Speichermedien ein. In der im Institut für Humangenetik der Universitätsmedizin Mainz derzeit genutzten Konfiguration sind pro PowerVault 20 Steckplätze mit je 8 TB großen HDDs bestückt.

Auf dem Primärserver werden jeweils 6 HDDs durch eine redundante Anordnung unabhängiger Festplatten (engl. Redundant Array of Independent Disks, RAID) im Level 6 (RAID 6) zur Erhöhung der Ausfallsicherheit (Redundanz) zu jeweils etwa 7 TB großen logischen Laufwerken zusammengeschlossen. Diese im ext4 Dateisystem formatierten Laufwerke erhalten den Namen „Rohdaten“ bzw. „Ergebnisse“. Die beiden SSDs sind durch ein RAID 0, welches die Datenübertragungsrate erhöht, aber keine Redundanz bildet, zu einem etwa 1,8 TB großen logischen Laufwerk kombiniert. Dieses ist ebenfalls im ext4 Dateisystem formatiert. Es erhält den Namen „Berechnungen“. Da auf diesem Laufwerk nur temporäre Daten von aktuell durchgeführten Analysen abgelegt werden, lässt sich hier mit dem RAID 0 eine hohe Übertragungsrate erreichen und auf eine Redundanz verzichten. Die 12 HDDs des Sekundärservers sind wie auf dem Primärserver mittels RAID 6 zu 2 jeweils 7 TB großen Laufwerken mit ext4 Dateisystemen und den Namen „Rohdaten“ und „Ergebnisse“ zusammengefasst.

Um auf den Massenspeichergeräten (MD PowerVault) eine höhere Flexibilität

bezüglich der Speicherplatzanforderungen zu erhalten und dynamisch auf das Wachsen der Datenbestände reagieren zu können, wird ein sogenannter „Logical Volume Manager“ (LVM) eingesetzt. Dieser bietet die Möglichkeit mehrere physische Partitionen zu einer logischen Gruppe (engl. Volume Group) zusammenzufassen. Einer Volume Group lassen sich später weitere Speichermedien dynamisch hinzufügen. Innerhalb einer solchen Gruppe können logische Partitionen (engl. Logical Volumes) angelegt werden, welche sich im laufenden Betrieb dynamisch vergrößern und verkleinern lassen. Zum Zeitpunkt der Anfertigung der vorliegenden Dissertationsschrift standen auf den jeweiligen Geräten jeweils 19 TB zur Archivierung der Rohdaten (Archiv_Rohdaten) und 25 TB zur Archivierung der Ergebnisse (Archiv_Ergebnisse) zur Verfügung. Rohdaten und Ergebnisse, die älter als 6 Monate sind, werden durch ein Skript automatisch ins jeweilige Archiv verschoben. Eine Spiegelung aller Daten der einzelnen Laufwerke des Primärservers auf den Sekundärserver geschieht einmal täglich um 20:00 Uhr (Rohdaten und Ergebnisse) bzw. um 02:00 Uhr (Archiv_Rohdaten und Archiv_Ergebnisse) (siehe Abbildung I.1).

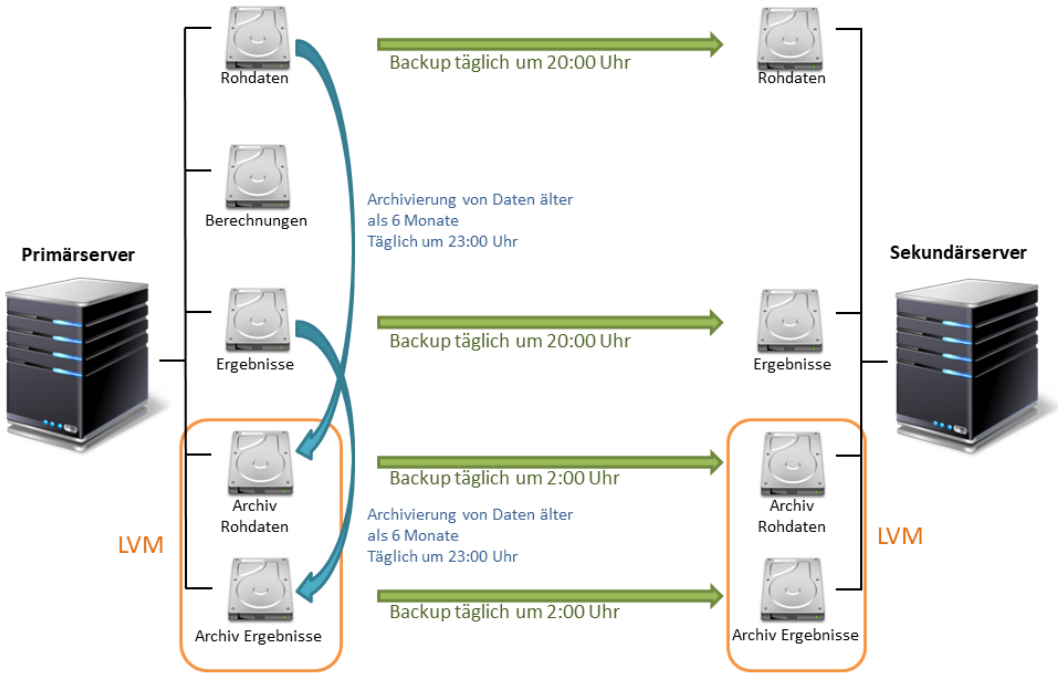


Abbildung I.1: *Konfiguration der Server.*
Datensätze auf der `Rohdaten` *sowie* `Ergebnisse` *Partition, die älter als 6 Monate sind werden automatisch durch ein Skript auf die Partition* `Archiv Rohdaten` *bzw.* `Archiv Ergebnisse` *verschoben. Um 20:00 bzw. um 02:00 Uhr wird ein tägliches Backup der Daten des Primärservers auf den Sekundärserver geschrieben*

Auf beiden Servern ist der Ubuntu Server 18.04.4 LTS (GNU/Linux 4.15.0-88-generic x86_64) als Betriebssystem installiert. Per Putty-Software[1] können

[1] `https://www.putty.org`

durch die bereitgestellte textorientierte Terminalsitzung Befehle auf dem jeweiligen Server direkt ausgeführt werden. Die Verwaltung der für die Auswertung verwendete Software auf dem Primärserver ist durch Lmod [McLay *et al.*, 2011] realisiert. Lmod erlaubt es durch Veränderung des Systempfads, Softwarepakete zu aktivieren und wieder zu deaktivieren. Dadurch können verschiedene Versionen von Programmen nebeneinander auf dem gleichen System installiert sein und je nach Bedarf ausgeführt werden.

Anhang II

Dateiformate

Ein Dateiformat definiert die Strukturierung von Informationen innerhalb einer Datei. Für viele Programme ist die Kenntnis des Dateiformates für die Interpretation der abgelegten Daten essentiell.

II.I Illumina SampleSheet

Das SampleSheet ist ein von Illumina entwickeltes komma-separiertes Dateiformat zur Speicherung von Probeninformationen und weiteren Metadaten eines Sequenzierungsexperiments. Es dient zum Beispiel als Eingabedatei für die `bcl2fastq`-Software (siehe Anhang IV.V) und enthält mehrere Abschnitte, die unter anderem das experimentelle Setup und die Angaben zum Demultiplexen einschließen. Die einzelnen Bereiche innerhalb der Textdatei werden durch ein Kennwort eingeleitet, welches in eckige Klammern eingefasst alleine in einer Zeile steht. Die nachfolgenden Zeilen sind alle dem so eingeleiteten Abschnitt zugeordnet, bis ein neuer Abschnitt folgt. Neben den essentiellen *Header*- und *Data*-Abschnitten gibt es weitere optionale Abschnitte wie *Settings*, *Reads* und *Manifests*, die von Illumina verwendet werden. Der *Header* muss die SampleSheet-Datei einleiten, der *Data*-Abschnitt ist stets der letzte. Zwischen diesen beiden Bereichen können die optionalen sowie benutzerdefinierte Abschnitte in beliebiger Reihenfolge liegen [Illumina, 2017b].

Im *Header* werden Informationen über den Kontext eines Sequenzierungslaufs, wie zum Beispiel das Datum, das verwendete „Lirary Preparation"-Kit, die Sequenzierchemie, ein Name für den Sequenzierungslauf, die zuständige Person und eine Beschreibung des Experiments als komma-separierte Schlüssel-Wert-Paare angegeben. Der optionale *Settings*-Abschnitt kann Einstellungen und Parameter für die Konvertierung der Base Calls zu FASTQ-Dateien beinhalten. Meistens sind hier die durch `bcl2fastq` zu entfernenden Adaptersequenzen als Schlüssel-Wert-Paare notiert. Die Anzahl der sequenzierten Basen pro Leserichtung wird als einzelne Zahl im *Reads*-Abschnitt festgehalten. Bei einem PE-Sequenzierexperiment enthält dieser zwei Zeilen. Die erste steht für die Anzahl an Basen des Forward- und die zweite für die Basenanzahl des Reverse-Reads. Die Verwendung des *Manifests*-Abschnitts beschränkt sich auf einige Illumina Analyse-Softwares. In ihm ist der Pfad zu einer Datei mit ROIs für ein TES-

Experiment abgelegt. Der *Data*-Abschnitt stellt eine komma-separierte Tabelle mit probenspezifischen Metadaten dar. Die erste Zeile dieser Tabelle trägt die Spaltennamen, wobei nur die Spalte *Sample_ID* essentiell ist. Weitere Spalten können die verwendeten Indizes zum Demultiplexen und eine Beschreibung der Probe beinhalten. In den folgenden Zeilen sind jeweils die entsprechenden Daten für jede Probe abgelegt [Illumina, 2017b].

Es ist möglich, ein SampleSheet mit Hilfe der „Illumina Experiment Manager"-Software[1] zu erstellen. Die in dieser Arbeit verwendeten SampleSheets wurden jedoch über ein in der internen Patientendatenbank integriertes Skript angelegt. So lassen sich Patienten-IDs direkt über die Datenbank abfragen und ohne Fehler mit Zuordnung der korrekten Untersuchung und Anreicherungsmethode (*Description*-Spalte der *Data*-Sektion) ins SampleSheet übertragen. Eine Notiz über die verwendete Anreicherungsmethode und Untersuchung ist für die automatisierte Analyse wichtig. Abbildung II.1 zeigt ein beispielhaftes SampleSheet.

```
[Header]
IEMFileVersion,4
Investigator Name,Max Mustermann
Experiment Name,Exom Run 1
Date,23.03.2020
Workflow,GenerateFASTQ
Application,Humangenetik Mainz
Assay,Our Panels
Description,NextSeq
Chemistry,Amplicon

[Reads]
126
126

[Settings]
Adapter,CTGTCTCTTATACACATCT+AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
AdapterRead2,AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

[Data]
Sample_ID,I7_Index_ID,index,I5_Index_ID,index2,Description
0001-01,Agi_1,TAAGGCGA,Agi_14,ATAGAGAG,Exom:Agilent Human All Exon V7
0002-01,Agi_3,AGGCAGAA,Agi_14,ATAGAGAG,Exom:Agilent Human All Exon V7
```

Abbildung II.1: *Beispiel eines Illumina SampleSheets.*
Es enthält das experimentelle Setup und die Angaben zum Demultiplexen. Einzelne Abschnitte werden durch ein Kennwort in eckigen Klammern eingeleitet. Der Header- und der Datenabschnitt sind essentiell. Illumina verwendet zudem teilweise die optionalen Abschnitte Settings, Reads und Manifests. Im Header werden Informationen über den Kontext eines Sequenzierungslaufs abgelegt. Der Data-Abschnitt stellt eine kommaseparierte Tabelle mit probenspezischen Metadaten dar

[1] `https://emea.support.illumina.com/sequencing/sequencing_software/experiment_manager.html?langsel=/de/`

II.II FASTQ

Das FASTQ-Format ist ein textbasiertes Format, in dem Nukleotidsequenzdaten zusammen mit deren Qualitätswerten (Quality Scores) gespeichert werden. Die Sequenz- sowie Qualitätswerte sind als einzelne ASCII (American Standard Code for Information Interchange)-Buchstaben dargestellt [Wikipedia The Free Encyclopedia, 2019]. Ursprünglich wurde FASTQ am Wellcome Trust Sanger-Institut entwickelt. Allerdings hat es sich unlängst als Ausgabeformat von Next Generation Sequencing-Systemen etabliert. Es gibt keine Standarddateiendung für FASTQ-Dateien. Am häufigsten wird jedoch `*.fq` oder `*.fastq` verwendet.

Das FASTQ-Format nutzt vier Zeilen zur Darstellung einer Sequenz mit den jeweils zugehörigen Qualitätswerten. Die erste Zeile beginnt mit einem „@“-Zeichen. Ihm folgt eine Sequenzbezeichnung (Identifier). Die zweite Zeile enthält die Basenabfolge in Großbuchstaben. Zeile drei wird mit einem „+“ eingeleitet und verfügt optional über zusätzliche Informationen zur Sequenz. Der Quality Score für jede Base befindet sich in Zeile vier. Da er wie die Basen als Einzelbuchstabe des ASCII-Codes vorliegt, müssen Zeile zwei und vier gleich viele Zeichen enthalten (siehe Abbildung II.2) [Cock *et al.*, 2009]. Andernfalls wird die Datei von vielen Programmen als beschädigt eingestuft.

```
@SRR002055.11186336 207BOAAXX:6:324:855:490
AAAATTCTTAGGCCTTTTCTCAAACAGGGGATT
+
FI:DHIIII.1?==IIIA:I3..*-.)%'#'&&
@SRR002055.11186337 207BOAAXX:6:324:933:7
TGGTAATGTGTTTTAAGCTCTTTGCGTTTNNNG
+
I=<97;IF;>3IAF+)3-+*0-+.%/&&-"""&
@SRR002055.11186338 207BOAAXX:6:324:375:880
CAGAAACCACCACAGACAGACAGACAGACGGAC
+
III;IIII8II2I3I+I)I+8(B'=(0%-+0#&
@SRR002055.11186339 207BOAAXX:6:324:682:940
CCGATTTCTCCAGCTCAGCCAGCACATTGGCCT
+
III+IIIIIII46IIIBIIG&IB-B-I:0++9H
```

Abbildung II.2: *Vier Sequenzen mit zugehörigem Quality Score in einer FASTQ-Datei.*
Die nachfolgenden Zeilenerläuterungen gelten für jede dargestellte Sequenz. Zeile 1 beginnt mit einem „@“-Zeichen gefolgt von dem Sequenzidentifier. In Zeile 2 steht die Basenabfolge. Zeile 3 wird mit einem „+“ eingeleitet und trägt optional zusätzliche Informationen über die Sequenz. Die als ASCII-Zeichen kodierten Qualitätswerte sind in Zeile 4 notiert.

Der Quality Score einer Base gibt an, mit welcher Wahrscheinlichkeit diese falsch sequenziert oder beim Basecalling falsch interpretiert wurde. Die Errech-

nung des Qualitätswerts Q erfolgt in allen Systemen nach der Sanger-Gleichung (Gleichung II.1) aus p. Dabei ist p die Wahrscheinlichkeit dafür, dass die genannte Base beim Basecalling falsch identifiziert wurde. Die Qualitätswerte von 0 bis 93 werden mit den ASCII-Zeichen 33 bis 126 kodiert.

$$Q = -10\, log_{10}(p) \tag{II.1}$$

In die Abschätzung für die Wahrscheinlichkeit p fließen unter anderem folgende Prädiktoren ein:

1. Ist das Signal für die betrachtete Base viel heller als das von anderen Basen?
2. Wird der Spot dunkler, verglichen mit dem Anfang der Sequenzierung?
3. Ist das Signal in den vorherigen und nachfolgenden Zyklen eindeutig differenzierbar?

Eine Auflösung der Gleichung II.1 nach p (siehe Gleichung II.2) ergibt für einen Qualitätswert von 40, im Sanger Format kodiert durch das ASCII Zeichen 73 („I“), eine Fehlerwahrscheinlichkeit von 0,01% für die Sequenzierung der entsprechende Base.

$$p = 10^{\frac{-Q}{10}} \tag{II.2}$$

Tabelle II.1 zeigt verschiedene Qualitätswerte nach Sanger mit deren Bedeutung.

Tabelle II.1: *Qualitätswerte und die zugehörige Base-Calling Genauigkeit*

Qualitätswert	**Wahrscheinlichkeit für einen fehlerhaften Base-Call**	**Genauigkeit des Basecalls**
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1.000	99,9%
40	1 in 10.000	99,99%
50	1 in 100.000	99,999%

II.III SAM, BAM und CRAM

Das Sequence Alignment/Map (SAM)-Dateiformat speichert Alignments von Sequenzen an einem Referenzgenom. Es unterstützt lange und kurze Sequenzbruchstücke von diversen Next Generation Sequencing-Plattformen und schafft somit eine gut definierte Schnittstelle zwischen dem Alignment und den nachfolgenden Analysen [Li *et al.*, 2009a].

Das SAM-Format beinhaltet einen „Header“ und einen „Alignmentabschnitt“. Im Gegensatz zum Alignmentabschnitt beginnen die Zeilen des Headerabschnittes mit einem „@“. Die Zeilen des Alignments bestehen aus 11 Pflichtspalten, welche jeweils durch einen Tabulator getrennt sind. Tabelle II.2 zeigt eine Liste dieser Felder mit einer kurzen Beschreibung. Wenn keine Informationen für ein Pflichtfeld vorliegen, kann „*“ beziehungsweise „0“ eingetragen werden. Zudem besteht die Möglichkeit weitere Informationen als Schlüssel-Wert-Paare im Format `TAG:TYPE:VALUE` hinzuzufügen.

Tabelle II.2: *Pflichtfelder des Alignment-Abschnittes des SAM-Formats mit Spaltennummer, Name und kurzer Beschreibung (Nach [Li et al., 2009a]).*

#	Name	Beschreibung
1	`QNAME`	Name des Reads
2	`FLAG`	Bitschalter
3	`RNAME`	Name der Referenzsequenz
4	`POS`	Anfangsposition an der Referenzsequenz
5	`MAPQ`	Qualität, mit der der Read zur Referenzsequenz passt
6	`CIGAR`	Erweiterte CIGAR-Zeichenkette
7	`MRNM`	Name des zugehörigen Reads („=“ wenn gleich wie `RNAME`)
8	`MPOS`	`POS` Wert des zugehörigen Reads
9	`ISIZE`	Länge der eingefügten Sequenz
10	`SEQ`	Basenabfolge der eingefügten Sequenz
11	`QUAL`	Qualitätswerte der eingefügten Sequenz

Um die Leistung zu verbessern, wurde das Binary Alignment/Map-Dateiformat (BAM) entwickelt. Es enthält exakt die gleichen Informationen wie das SAM-Format. Durch seine binäre Kodierung ist die BAM-Datei allerdings um ein Vielfaches kleiner als die SAM-Datei und lässt sich von den Programmen schneller verarbeiten.

Im Jahr 2012 wurde basierend auf einer Veröffentlichung von Hsi-Yang Fritz *et al.* [Fritz *et al.*, 2011] mit dem CRAM Format vom Europäischen Institut für Bioinformatik (engl. European Bioinformatics Institute, EBI) ein Dateiformat vorgestellt, mit dem sich Mapping Informationen durch einen effizienten referenzbasierten Ansatz verlustfrei (engl. lossless) stärker komprimieren lassen, als mit dem BAM-Format. Mit der Möglichkeit einer verlustbehafteten (engl. lossy) Komprimierung lässt sich eine SAM-Datei sogar 40-50% stärker komprimieren als mit dem BAM-Format. In diesem Modus kann eingestellt werden, wie mit den Qualitätswerten umzugehen ist. So ist es zum Beispiel möglich das 8 Klassen Einteilungssystem (engl. 8-bin schema) von Illumina [Illumina, 2012] auf die Qualitätswerte anzuwenden, wobei die stärkste Komprimierung durch die vollständige Löschung des Qualitätswerts zu erreichen ist (siehe Tabelle II.3).

Tabelle II.3: *CRAM Komprimierung [Uppsala University, 2020]*

Dateiformat	**Dateigröße [GB]**
SAM	7,4
BAM	1,9
CRAM lossless	1,4
CRAM 8 bin	0,8
CRAM keine Qualitätswerte	0,26

Da die meisten Software-Tools dieses CRAM-Format noch nicht unterstützen und das Lesen der Datei mehr Zeit in Anspruch nimmt als das Lesen einer BAM-Datei, wird das CRAM-Dateiformat bisher meist nur zur Langzeitarchivierung von Alignmentdaten verwendet [Li, 2012].

II.IV VCF

Ein Dateiformat zur Speicherung der gefunden Varianten ist das sogenannte „Variant Call Format“ (VCF). Es wurde ursprünglich für das 1000 Genom Projekt entwickelt und von Danecek *et al.* 2011 in deren Veröffentlichung als das Standardformat zur Speicherung der am häufigsten vorkommenden Arten von Sequenzvarianten und deren Annotationen empfohlen [Danecek *et al.*, 2011].

Abbildung II.3 zeigt beispielhaft den Aufbau einer VCF-Datei. Sie besteht aus einem Header-Bereich mit Meta-Informationen und einem Daten-Bereich, welcher eine Liste mit Varianten beinhaltet. Zeilen mit Meta-Informationen werden durch `##` eingeleitet und schließen standardisierte Beschreibungen der im Datenabschnitt verwendeten Tags und Annotationen ein. Ebenso kann dieser Bereich Informationen über die Version des Dateiformats, das Erstellungsdatum, die verwendete Referenzsequenz und der zum Auffinden der Varianten und deren Annotation genutzten Software enthalten.

Der Header-Block endet mit einer durch `#` angeführten tabseparierten Zeile, welche die Felder des Datenbereichs beschreibt. Diese Zeile und der darauffolgende Datenblock beinhalten 8 Pflichtfelder, ein optionales `Format` Feld und eine beliebige Anzahl an Feldern mit Daten zu verschiedenen Proben. Die Pflichtfelder umfassen die Definition des Chromosoms (`CHROM`), die Startposition der Variante (`POS`), sowie einen eindeutigen Identifier (`ID`). Diesen Feldern folgt die Angabe des Referenzallels (`REF`), eine kommaseparierte Liste mit alternativen Allelen (`ALT`), ein Qualitätswert (`QUAL`), Informationen über gegebenenfalls verwendete Filter (`FILTER`) und eine durch Semikolon getrennte Liste mit Zusatzinformationen, wie zum Beispiel Annotationen (`INFO`). Werden in der VCF-Datei Probendaten, also die Zuordnung der Varianten zu Patientenproben abgelegt, so folgen diesen Pflichtfeldern ein `FORMAT`-Feld und pro Probendatensatz ein `SAMPLE`-Feld. Dabei dient Ersteres zur Definition der in den darauffolgenden Spalten mit Probendaten enthaltenen Informationen. Die einzelnen Informationen in diesen `FORMAT`- und `SAMPLE`-Feldern sind jeweils durch

einen Doppelpunkt getrennt. Ein Eintrag von `GT:GQ:DP` im `FORMAT`-Feld weist zum Beispiel darauf hin, dass die folgenden Probendatensätze jeweils Informationen über den Genotypen, die Qualität dieses Genotyps und die Lesetiefe dieser Position einschließen (siehe Abbildung II.3 letzte Zeile). Beinhaltet ein Feld der VCF-Datei keine Daten, so wird dies mit einem „.“ notiert. Im September 2019 erschien die Version 4.3 des VCF-Formats [Samtools Organisation, 2020]. Die in dieser Thesis verwendeten VCF-Dateien basieren alle auf den VCF-Format Definitionen der Version 4.0 bis 4.3.

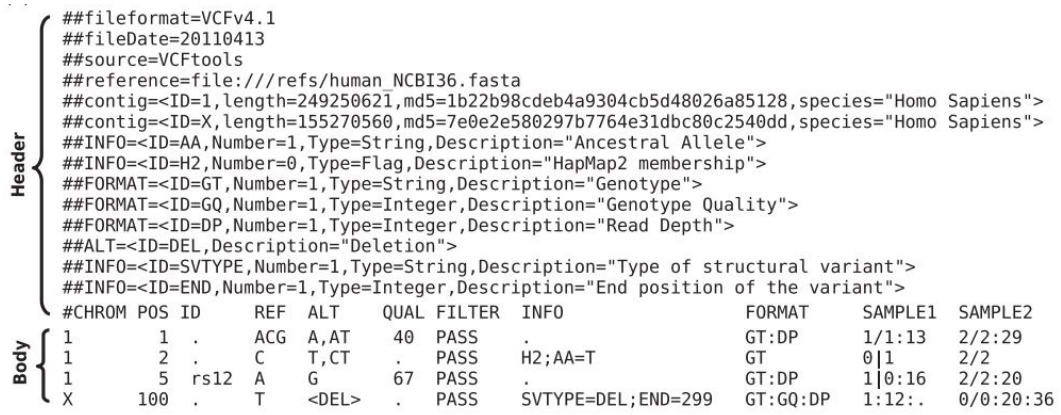


Abbildung II.3: *VCF-Format.*
Eine VCF-Datei besteht aus einem Header-Bereich mit Meta-Informationen und einem Daten-Bereich. Metainformationen werden durch `##` *eingeleitet. Der Datenbereich beginnt durch eine mit* `#` *angeführte Zeile mit der Beschreibung der folgenden Datenfeder. Die 8 Pflichtfelder umfassen die Definition des Chromosoms (*`CHROM`*), die Startposition der Variante (*`POS`*), einen eindeutigen Identifier (*`ID`*), die Angabe des Referenzallels (*`REF`*), eine komma-separierte Liste mit alternativen Allelen (*`ALT`*), einen Qualitätswert (*`QUAL`*), Informationen über verwendete Filter (*`FILTER`*) und eine semikolon-getrennte Liste mit Zusatzinformationen (*`INFO`*). (nach [Danecek et al., 2011])*

II.V BED

Das BED (engl. Browser-Extensible-Data)-Format wurde von der University of California, Santa Cruz (UCSC) definiert und bietet eine flexible Möglichkeit genomische Regionen und deren Merkmale abzuspeichern. Diese Dateien können zur Visualisierung als sogenannte „Tracks“ zum Beispiel in den UCSC Genome Browser[2] [Kent *et al.*, 2002] geladen werden. Das Format besteht aus einer Zeile pro Feature mit jeweils 3-12 leerzeichen- oder tabulatorgetrennten Datenspalten, sowie optionalen Definitionszeilen (Trackzeilen) für die Darstellung der in der Datei abgelegten Regionen.

[2] `http://genome.ucsc.edu/index.html`

Jede dieser Datenzeilen beinhaltet mit `chrom`, `chromStart` und `chromEnd` 3 obligatorische und 9 weitere optionale Felder (siehe Tabelle II.4), wobei die Reihenfolge dieser Felder eingehalten werden muss. Wird ein optionales Feld verwendet, so müssen auch seine vorangehenden Felder definiert sein. Besitzt ein Feld keinen Inhalt, ist dies mit einem „.“ anzugeben. Die Reihenfolge und Definition der einzelnen Felder ist in Tabelle II.4 dargestellt [University of California Santa Cruz, 2019]. Mit Hilfe der optionalen Track-

Tabelle II.4: *BED Dateiformat nach [University of California Santa Cruz, 2019]*

Art	Feldname	Definition
verpflichtend	`chrom`	Bezeichnung des Chromosoms der Referenzsequenz
	`chromStart`	Startposition der Region auf dem Chromosom (0-basiert)
	`chromEnd`	Endposition der Region auf dem Chromosom
optional	`name`	Bezeichnung bzw. Name der definierten Region
	`score`	Wert zwischen 0 und 1000. Wenn die Variable *useScore = 1* definiert ist, bestimmt dieser Wert die Graustufe in der die Region angezeigt wird.
	`strand`	Definiert den DNA Strang. „+“ oder „-“
	`thickStart`	Position ab der die Region breiter dargestellt werden soll
	`thickEnd`	Endposition der breiter dargestellten Region
	`itemRgb`	RGB Werte. Wenn die Variable *itemRgb = On* definiert ist,wird die Region (z.B. Exons) in der festgelegten Farbe dargestellt
	`blockCount`	Anzahl einzelner Elemente in der Region
	`blockSizes`	kommaseparierte Liste mit Größen der Elemente
	`blockStarts`	kommaseparierte Liste mit Startpositionen der einzelnen Elemente relativ zu chromStart

zeile besteht die Möglichkeit die Visualisierung der ihr folgenden Regionen zu konfigurieren. Eine solche Zeile beginnt zwingend mit dem Wort „track“, gefolgt von durch Leerzeichen getrennten Schlüssel-Wert-Paaren im Format `Schlüssel=Wert`. So ist eine Option zum Beispiel mit dem Schlüssel `name` einen eindeutigen Namen und mit `description` eine Beschriftung des Tracks zu definieren. Zudem kann in dieser Zeile, wie in Tabelle II.4 erwähnt mit den Schlüsseln `useScore` bzw. `itemRgb` festgelegt werden, ob die Region in der entsprechenden Graustufe oder angegebenen Farbe anzuzeigen ist [ENSEMBL, 2020a].

Anhang III

Datenbanken

Eine Datenbank (DB) beschreibt ein System zur Verwaltung elektronischer Daten. Sie dient dazu große Datenmengen effizient und dauerhaft zu speichern. Zudem können benötigte Teilmengen der Daten in geeigneter Darstellungsform für den Benutzer oder für eine Anwendung bereitgestellt werden.

III.I HGMD®

Die Human Gene Mutation Database (HGMD®) wurde 1996 für die Öffentlichkeit zur Verfügung gestellt. Sie sollte ursprünglich für die wissenschaftliche Untersuchung der Mutagenese in menschlichen Genen dienen. Heute beinhaltet sie eine umfassende Kollektion von Varianten, die ursächlich für genetische Krankheiten sind oder mit diesen in Verbindung gebracht werden. Somit stellt die HGMD® ein mächtiges Werkzeug für Ärzte, genetische Berater und die wissenschaftliche Forschung dar.

Im Jahr 2017 beinhaltete die HGMD® mehr als 203.000 verschiedene Einträge in mehr als 8000 unterschiedlichen Genen. Pro Jahr wächst die Datenbank um etwa 17.000 Einträge [Stenson *et al.*, 2017]. So befinden sich im Release 2020.1 282.895 verschiedene Varianten [QIAGEN, 2020]. Alle in der Literatur publizierten krankheitsverursachenden, sowie krankheitsassoziierten und funktionseinschränkenden Polymorphismen werden durch die HGMD® bereitgestellt. Diese Daten umfassen den Austausch einzelner Basen (Missense- und Nonsense-Mutationen) in kodierenden und regulatorischen Sequenzen, sowie splicing-relevante Mutationen, Mikrodeletionen, Mikroduplikationen, repetitive Elemente, Insertionen und Deletionen.

Für die akademische und nicht kommerzielle Verwendung werden alle Mutations-Daten drei Jahre nach ihrem Eintrag in die kommerzielle Datenbank für registrierte Benutzer kostenlos unter `http://www.hgmd.org` zugänglich gemacht. Die Lizenz für den Zugang zur aktuellen Version der HGMD® muss für kommerzielle, sowie akademische Zwecke bei der BIOBASE GmbH erworben werden.

III.II dbSNP

Die Single-Nucleotide-Polymorphism-Database (dbSNP) ist eine kostenlose öffentliche Datenbank für genetische Varianten in diversen Spezies. Sie wurde 1998 am National Center for Biotechnology Information (NCBI) zur Ergänzung der eigenen Nucleotid-Sequenz-Datenbank (GenBank) entworfen. Die dbSNP beinhaltet Einzel-Nucleotid-Polymorphismen (SNP's), kurze Insertionen und Deletionen (Indels), Mikrosatelliten-Marker beziehungsweise Short Tandem Repeats (STR's), Multi-Nucleotid-Polymorphismen sowie homozygote Sequenzen [Sherry *et al.*, 1999].

Die Datenbank soll die biologische Forschung als Online-Ressource unterstützen. Sie verfolgt das Ziel alle bekannten genetischen Varianten in einer Datenbank zu vereinen. Dies kann genutzt werden, um genetische Phänomene wie zum Beispiel die Populationsgenetik oder evolutionäre Zusammenhänge zu untersuchen. Zudem ist es möglich Assoziationen der genetischen Varianten mit phänotypischen Merkmalen schnell zu erkennen.

Die Veröffentlichung neuer Daten findet in unbestimmten Zeitabständen in sogenannten „builds“ statt.

III.III ClinVar

ClinVar wurde vom National Center for Biotechnology Information (NCBI) am National Institute of Health (NIH) entwickelt und im April 2013 offiziell publiziert. Es ist eine öffentliche, frei zugängliche Datenbank mit Informationen über die Beziehungen zwischen Varianten im menschlichen Genom und phänotypischen Erscheinungsbildern. Ein Ziel der Datenbank ist es, den Zugang und die wissenschaftliche Diskussion über diese Genotyp-Phänotyp-Beziehungen zu erleichtern. ClinVar ist eng mit der dbSNP (siehe Anhang III.II) verbunden. Jeder Eintrag in ClinVar bezieht sich auf eine Variante aus der dbSNP Datenbank. Ist die Variante in dbSNP bisher nicht bekannt, wird sie neu angelegt. Akzeptierte Übermittlungen von Daten kommen aus genetischen Laboratorien, aus der Forschung und aus Literaturrecherchen. Dabei werden die übermittelten Daten von verschiedenen einsendenden Institutionen pro Variante zusammengefasst. Bis Mitte des Jahres 2018 übermittelten 1000 verschiedene Institutionen aus 65 Ländern insgesamt über 600.000 Interpretationen von 430.000 Varianten [Landrum und Kattman, 2018]. Da verschiedene Einsender unterschiedliche Interpretationen von Varianten bezüglich ihrer Korrelation zu einem Phänotyp und der damit verbundenen klinischen Signifikanz haben können, gibt es einen sogenannten Rezensionsstatus, welcher durch eine unterschiedliche Anzahl an Sternen dargestellt ist. Wiedersprechen sich zwei Einsender in ihrer Interpretation, so wird die Variante als „Interpretation mit Konflikten“ markiert. Ein Einsender kann seine übermittelten Daten jederzeit aktualisieren. Der alte Datensatz wird archiviert und der aktuelle erhält eine neue Versionsnummer [Landrum *et al.*, 2014].

ClinVar ist als Webseite im Internet[1] verfügbar oder kann als XML, VCF oder tabulatorgetrennte Textdatei heruntergeladen werden.

III.IV OMIM

Online Mendelian Inheritance in Man (OMIM) ist eine Datenbank für Krankheiten, die einen genetischen Hintergrund besitzen. Falls verfügbar, werden den Krankheiten ursächliche Gene und zusätzliche Literaturquellen für weitergehende Forschungen zugeordnet [Hamosh *et al.*, 2005].

In den 1960er Jahren wurde das Projekt Mendelian in Man von Dr. Viktor A. McKusick an der Johns Hopkins Universität (California, USA) gegründet. Es liegt heute in der zwölften Edition als Buch vor. Die Online-Version dieses Schriftwerkes, die OMIM-Datenbank, wurde 1987 entwickelt und im Jahre 1998 durch das NCBI[2] im World Wide Web für die Öffentlichkeit zur Verfügung gestellt.

Die Datenbank beinhaltet alle bekannten genetischen Erkrankungen in insgesamt 25.357 Einträgen mit über 10.000 assoziierten Genen (Stand 12. März 2020) [Hancock *et al.*, 2004]. Sie wird täglich aktualisiert und dient Biologen, Ärzten und Forschern vor allem als Nachschlagewerk für Informationen über hereditäre Krankheiten.

III.V Orphanet

Orphanet wurde 1997 in Frankreich gegründet, um das Wissen über seltene Krankheiten zu sammeln und so die Diagnose, Versorgung und Behandlung von Patienten mit seltenen Krankheiten zu verbessern. Ab dem Jahr 2000 wandelte sich diese Initiative zu einem europaweiten Anliegen, welches ab dann durch Zuschüsse der Europäischen Kommission unterstützt wurde. Orphanet entwickelte sich so zu einem Netzwerk von 41 Ländern. Neben vielen europäischen Ländern zählt Orphanet mittlerweile auch Länder rund um den Globus zu seinen Mitgliedern [Orphanet, 2020].

Orphanet hat sich als das europäische Informationssystem für seltene Krankheiten entwickelt. Durch seine Webseite bietet die Orphanet-Initiative allen Zielgruppen (1/3 Patienten, 2/3 Ärzte, Wissenschaftler und Studierende) gleichermaßen den Zugang zu hochwertigen Informationen über Krankheiten, Fachkliniken, klinische Speziallabore, Forschungsprojekte und Patientenorganisationen. So beinhaltet die Orphanet Homepage Informationen über mehr als 6100 seltene Erkrankungen und 5400 Gene [Orphanet, 2020] (Stand Oktober 2020). Die reichhaltige Datenbank ist unter www.orpha.net im Netz erreichbar.

[1] `https://www.ncbi.nlm.nih.gov/clinvar/`
[2] `http://www.ncbi.nlm.nih.gov/omim`

III.VI GnomAD

Die Genom Aggregationsdatenbank (engl. Genome Aggregation Database, gnomAD) ist durch eine Koalition von Forschern entstanden, die versuchen Exom- und Genomsequenzierungsdaten aus einer Vielzahl von großen Sequenzierungsprojekten zu aggregieren und zu harmonisieren, um die zusammengefassten Daten einer breiten wissenschaftlichen Gemeinschaft zur Verfügung zu stellen. Eine frühere Veröffentlichung beinhaltete ausschließlich Exomdaten, welche vom Exom Aggregations Konsortium (engl. Exome Aggregation Consortium, ExAC) zusammengetragen und im ExAC-Browser zur Verfügung gestellt wurde.

Zur Erstellung der gnomAD-Datenbank (Version 2.1) trug das Konsortium WES-Daten von 199.558 und WGS-Daten von weiteren 20.314 Individuen zusammen. Diese Daten stammen fast ausschließlich aus Fall-Kontroll-Studien von im Erwachsenenalter auftretenden Erkrankungen, wie zum Beispiel kardiovaskulären Erkrankungen, Diabetes Typ 2 oder psychiatrischen Störungen. Jeder der Datensätze wurde durch eine standardisierte BWA-Picard-GATK-Pipeline einheitlich prozessiert. Datensätze mit niedriger Sequenzierungsqualität sind von der Analyse ausgeschlossen. Exkludiert wurden zudem Personen und deren Verwandte ersten Grades, von denen bekannt war, dass sie an einer schweren Erkrankung im Kindesalter litten. So beinhaltet die Veröffentlichung von Version 2.1 genetische Varianten von 125.748 Exomen und 15.708 Genomen aus hoch qualitativen Datensätzen nicht verwandter Individuen aus sechs globalen sowie 8 subkontinentalen ethnischen Gruppen. Die 17,2 Millionen (Exom) bzw. 261,9 Millionen (Genom) gefundenen Varianten wurden anschließend auf ihre Qualität gefiltert, was in einem Datensatz mit 14,9 Millionen für das Exom bzw. 229,9 Millionen Varianten Genom resultierte [Karczewski *et al.*, 2020]. Das Mapping erfolgte gegen das Referenzgenom GRCh37 bzw. HG19.

Über die Häufigkeit beziehungsweise das Fehlen von Varianten in dieser Kohorte von „gesunden“ Individuen kann auf die funktionelle Neutralität oder Pathogenität einer Variante geschlossen werden. Allerdings ist zu beachten, dass dies primär für dominante Varianten gilt. Heterozygote Anlageträger für rezessive Erkrankungen sind meist symptomfrei, wodurch diese Varianten ebenfalls in der Datenbank zu finden sein können.

Die Daten der Version 2.1 stehen im Downloadbereich der gnomAD Homepage[3] zur Verfügung. Zudem wird dort mit der Version 3 ein weiterer Datensatz bereitgestellt, welcher 71.702 Genome von nicht verwandten Personen, die im Rahmen verschiedener krankheitsspezifischer und populationsgenetischer Studien sequenziert wurden umfasst. Dieser Datensatz ist an der GRCh38-Referenz ausgerichtet.

[3] `https://gnomad.broadinstitute.org/downloads`

III.VII ENSEMBL

Das Ensembl-Projekt wurde 1999 am Europäischen Institut für Bioinformatik (engl. European Bioinformatics Institute, EBI) einige Jahre vor der Fertigstellung des Entwurfs des menschlichen Genoms gegründet. Schon zu diesem frühen Zeitpunkt war ersichtlich, dass eine manuelle Annotation des gesamten menschlichen Genoms mit seinen 3 Milliarden Basenpaaren nicht zeitnah bereitgestellt und in angemessenen Zeiträumen aktualisiert werden kann. Das Ziel von Ensembl war es daher das Genom automatisiert zu annotieren, diese Annotationen in andere verfügbare biologische Daten zu integrieren und all dies über das Internet öffentlich zugänglich zu machen. Seit dem Start der Website[4] im Juli 2000 wurden der Ensembl-Datenbank neben dem menschlichen Genom, Genome weiterer Spezies hinzugefügt und die vergleichende Genomik sowie Variations- und Genregulationsdaten in das Spektrum der verfügbaren Daten inkludiert. Neben der Annotation und Bereitstellung von Genomdaten werden Softwarepakete wie zum Beispiel der Variant Effect Predictor zur Annotation und funktionellen Analyse von Variantendaten oder der BioMart zum Export von benutzerdefinierten Datensätzen aus der Ensembl Datenbank entwickelt [EMBL-EBI, 2014].

Viermal im Jahr erfolgt eine Aktualisierung der Daten, Websites, APIs und Tools mit den neuesten Genomdaten. Dabei werden Primärdaten, wie Assemblies und neu entdeckte Varianten hinzugefügt, woraufhin Gene, Transkripte, Varianten und regulatorische Regionen neu annotiert werden. Im Jahr 2019 ließ sich so die Anzahl an Varianten innerhalb der Datenbank durch Integration der Daten des gnomAD (siehe Anhang III.VI) und der TOPMed Projekte [Taliun *et al.*, 2019] auf über 600 Millionen verdoppeln [Cunningham *et al.*, 2019].

III.VIII miRBase

MicroRNAs (miRNAs) sind etwa 22 Nukleotide lange, nicht-kodierende, regulatorische RNA-Sequenzen. Sie binden an 3'UTRs von mRNA-Sequenzen und spielen somit eine wichtige Rolle in der posttranskriptionalen Regulierung der Genexpression. Die miRBase Datenbank wurde 2003 von Sam Griffith-Jones etabliert und dient als Archiv von microRNA-Sequenzen und Annotationen. In der im Jahr 2003 publizierten Version 2.0 enthielt die Datenbank 506 miRNAs von 6 verschiedenen Spezies (*C.elegans*, *Caenorhabditis briggsae*, *D.melanogaster*, des Menschen, der Maus und *Arabidopsis thaliana*). Die miRBase-Datenbank bietet zudem ein zentrales System für die Zuweisung neuer Namen zu mikroRNA-Genen. Nachdem ein wissenschaftlicher Artikel mit der Beschreibung einer neuen miRNA angenommen wurde, kann ein Name für die miRNA beantragt werden. Dieser ist dann in die finale Version der Publikation einzubinden. Die Namen der miRNAs werden dabei mit fortlaufenden Nummern vergeben und beruhen auf Sequenzähnlichkeit. miRNAs mit gleicher

[4] `https://www.ensembl.org`

Sequenz aber unterschiedlichem *Loci* im Genom bekommen einen numerischen Suffix (z.B. *mir-6-1* und *mir-6-2*). Weichen die Sequenzen in ein oder zwei Basen voneinander ab, so erhalten sie einen lexikografischen Suffix (z.B. *mir-181a* und *mir-181b*). Solange nicht bekannt ist, ob die mature-miRNA des 3'- oder 5'-Strangs überwiegend transkribiert wird, sind die jeweiligen mature-miRNAs mit dem Suffix „5p“ beziehungsweise „3p“ zu versehen [Griffiths-Jones, 2004].

Die in dieser Arbeit verwendete Version 21 der miRBase erschien im Juni 2014 und enthält 28.645 Haarnadel pre-miRNAs aus 223 Spezies, die 35.828 mature-miRNAs bilden. Die miRBase ist im Internet[5] frei verfügbar. Alle Sequenzen können zudem als FASTA-Datei heruntergeladen werden. Die Annotationen der miRNAs stehen im EMBL-Format zur Verfügung.

[5] `http://www.mirbase.org/`

Anhang IV

Software

In diesem Teil der Arbeit werden die zur Untersuchung der Labordaten herangezogenen Softwarepakete und Programmiersprachen beschrieben.

IV.I Perl

Perl ist eine 1987 von Larry Wall entworfene, plattformunabhängige und interpretierte Programmiersprache. Sie ist unter anderem maßgeblich von den Sprachen C und Pascal beeinflusst worden und diente ursprünglich als Werkzeug für die Verarbeitung und Manipulation von Textdateien. Neben der Anwendung im Bereich der System- und Netzwerkadministration hat sich Perl im Laufe der Jahre weitgehend in der Bioinformatik etabliert. Perl ist für Linux-, Mac- und Windows-Systeme frei verfügbar[1]. Die Sprache kann mit vielen weiteren Paketen aus dem Comprehensive Perl Archive Network (CPAN) in ihrem Funktionsumfang erweitert werden. Die Perl-Skripte aus dieser Arbeit wurden mit Perl in der Version 5.26.1 geschrieben und aus dem Terminal gestartet [Augsten, 2017].

IV.II Java

Die erste Version von Java, damals Oak (Object Application Kernel) genannt, wurde von Frühjahr 1991 bis Sommer 1992 unter der Leitung von James Gosling im Auftrag des US-amerikanischen Computerherstellers Sun Microsystems unter dem Namen *The Green Project* entwickelt. Rechtliche Probleme führten dazu, dass der Name Oak durch Java ersetzt werden musste. Am 23. Mai 1995 fand die offizielle Vorstellung von Java statt [Wikipedia Die freie Enzyklopädie, 2020]. Java kann kostenlos für viele verschiedene Betriebssysteme (u.a. Windows, Linux und Mac) heruntergeladen werden[2].

Java ist eine objektorientierte Programmiersprache und Bestandteil der Java-Technologie. Diese Technologie setzt sich aus dem Java-Entwicklungswerkzeug

[1] `http://www.perl.org/get.html`
[2] `http://www.java.com/de/download/`

(engl. Java Developement Kit, JDK) zur Anfertigung eigener Programme und der Java-Laufzeitumgebung (engl.: Java Runtime Environment; Abk.: JRE) zur Ausführung von Java-Programmen zusammen. Der aus menschenverständlichem Text bestehende Java-Quellcode ist jedoch zur Ausführung vom Java-Compiler in maschinenverständlichen Code, den sogenannten Bytecode, zu übersetzen. Dieser Bytecode wird nicht direkt durch Hardware, sondern durch eine Java-Virtuelle-Maschine (eng.: Java-Virtual-Machine; Abk.: JVM) ausgeführt. Grund dieser Virtualisierung ist die Konstruktion einer Plattformunabhängigkeit. So kann ein Java Programm auf jeder Rechnerarchitektur ausgeführt werden, auf der eine passende Laufzeitumgebung installiert ist [Wikipedia Die freie Enzyklopädie., 2020].

Die in dieser Arbeit in Java entwickelte Software wurde in Eclipse Oxygen (Version 4.7.1) mit der Java-Version 1.8.0_201 erstellt.

IV.III Unix-Shell

Die Unix-Shell (im Weiteren nur noch als Shell bezeichnet) stellt die traditionelle Benutzerschnittstelle in Unix Betriebssystemen (z.B.: Linux und Mac) dar. Sie ist ein Kommandozeileninterpreter, da sie die vom Benutzer in der Eingabezeile eingegebenen Kommandos direkt ausführt. Der Unterschied zwischen der Shell und einer reinen Programmiersprache, wie zum Beispiel Perl besteht darin, dass sie über besondere Mittel zum Dialog mit dem Anwender verfügt [Herold, 2003].

Es existieren verschiedene Varianten von Shells. Die Bekanntesten sind die Bash-Shell, die Korn-Shell oder die C-Shell. Es besteht die Möglichkeit mehrere Shell-Befehle hintereinander in eine Textdatei zu schreiben und diese als Skript im Terminal auszuführen. Diese kleinen Programme sind dann sinnvoll, wenn eine sukzessive Abarbeitung von immer gleich bleibenden Shell-Befehlen mehrfach durchgeführt werden soll.

In dieser Arbeit wird die Bash-Shell in der Version 4.4.20(1) verwendet.

IV.IV MERAP

MERAP (medical resequencing analysis pipeline) ist ein von Cougar Hao Hu *et al.* am Max-Planck-Institut für molekulare Genetik in Berlin entwickeltes und 2014 veröffentlichtes Software-Paket zur Datenanalyse von NGS-Projekten [Hu *et al.*, 2014]. Die einzelnen Softwareprogramme sind als Perlskripte implementiert, rufen teilweise externe Programme zur Datenanalyse auf und müssen aus dem Terminal gestartet werden. Vor der Datenanalyse mit dem MERAP-Softwarepaket ist es notwendig ein Mapping mit dem SOAP2-Aligner [Li und Durbin, 2009] durchzuführen. Anschließend werden die gemappten Bereiche mit dem Perlskript `SNVFinder1.02.pl` auf Einzel-Nukleotid-Varianten (SNVs) untersucht. Das Skript `Coverage1.01.pl` berechnet die Abdeckung

der angereicherten Regionen und gibt dadurch einen Hinweis auf die Effizienz der Studie. In Schritt vier werden durch das Skript `IndelFinder1.02.pl` mit Hilfe des Programms BLAT [Kent, 2002] Insertionen und Deletionen (Indels) detektiert. Liegt in verschiedenen angereicherten Genomregionen eine signifikant unterschiedliche Coverage vor, so kann dies durch das Programm `CNVFinder1.01.pl` im fünften Schritt erkannt und als mögliche Kopienzahlveränderung (eng.: Copy Number Variation; Abk.: CNV) eingestuft werden. Gefundene SNV's, Indels und CNV's fasst das Skript `VariantPacker.pl` mit entsprechender Patientenkennung in einer Datei zusammen. Das Skript `SSFinder1.01.pl` sucht im folgenden Schritt innerhalb dieser Varianten diejenigen heraus, die vermutlich eine kryptische Splice-Site auslösen. Abschließend werden alle gefundenen Varianten mit dem Skript `Annotation1.09.pl` durch Datenbank- und Literaturabgleiche annotiert und mit Hilfe des Skriptes `Prioritize1.04.pl` priorisiert.

IV.V BCL2FASTQ

Illuminas `bcl2fastq` Software konvertiert die von der RealTimeAnalysis Software des Illumina Geräts produzierten Base Call (BCL) Dateien eines Sequenzierungslaufs in FASTQ-Dateien (siehe Anhang II.II). Um mehrere Proben gleichzeitig sequenzieren zu können, wurden jeder Probe während der Library Preparation im Multiplexing-Schritt eindeutige Sequenzindizes hinzugefügt. Die Sequenzabfolgen dieser Indizes sind im SampleSheet (siehe Anhang II.I) angegeben und helfen dem Programm beim Zuordnen der Sequenzfragmente zur entsprechenden Probe (demultiplexen). Für jedes erfolgreich sequenzierte Cluster der Flow Cell wird ein Eintrag in die entsprechende FASTQ Datei geschrieben. Für einen PE Lauf entstehen so pro Patient zwei FASTQ-Dateien (eine für die Forward- und eine für die Reverse-Reads). Es kommt vor, dass ein Sequenzfragment über das Ende des DNA-Inserts hinaus sequenziert wird. Dies führt dazu, dass der Read neben der DNA-Sequenz ein Teil der Adaptersequenz enthält. `bcl2fastq` erkennt die Adaptersequenzen und entfernt sie aus den resultierenden Reads. Über den Parameter `--no-lane-splitting` lassen sich die Ergebnisse der einzelnen Lanes pro Patient zusammenfassen [Illumina, 2019a].

Die benötigten Angaben über die zu trimmenden Adaptersequenzen sowie die einzelnen Proben-IDs und deren zugehörige Indizes sind in der SampleSheet-Datei (siehe Anhang II.I) abgespeichert. Liegt die Datei `SampleSheet.csv` im Stammverzeichnis des Rohdaten-Ordners (`-R`), so liest die Software diese automatisch ein. Trägt die SampleSheet-Datei jedoch einen anderen Namen oder ist sie in einem abweichenden Verzeichnis abgespeichert, so ist es möglich den Pfad über `--sample-sheet` zu definieren.

Einige wichtige und in dieser Arbeit verwendete Parameter für die `bcl2fastq`-Software (Version v2.17.1.14) sind in Tabelle IV.1 gelistet.

Tabelle IV.1: `bcl2fastq` *Parameter*

Parameter	**Erläuterung**
`-R`	Pfad zum Rohdaten-Ordner des Laufs
`-o`	Pfad zum Ordner für die FASTQ-Dateiausgabe
`--no-lane-splitting`	Lanes zusammenfassen
`--sample-sheet`	Pfad zum SampleSheet

IV.VI Burrows-Wheeler Alignment

Das Burrows-Wheeler Alignment Tool (BWA) ist ein Softwarepaket zur Kartierung von Sequenzen mit geringer Divergenz gegen eine große Referenzsequenz wie das menschliche Genom. Das Paket beinhaltet drei Algorithmen: BWA-Backtrack, BWA-SW und BWA-MEM. Der erste Algorithmus ist für Illumina Reads bis zu 100 bp optimiert, während die beiden anderen für längere Sequenzen zwischen 70 bp und 1 Mbp ausgelegt sind. BWA-MEM und BWA-SW haben ähnliche Funktionen wie Long-Read-Unterstützung und Split-Alignment. BWA-MEM, als neuester Programmteil, wird jedoch im Allgemeinen für qualitativ hochwertige Mappings empfohlen, da es schneller und genauer ist. Daher findet in dieser Arbeit ausschließlich der BWA-MEM Algorithmus Verwendung.

BWA basiert dabei auf der Burrows-Wheeler Transformation in Verbindung mit Suffix-Arrays und kann so auch lückenhafte (engl. gapped) Alignments durchführen.

Suffix-Arrays

Ein Suffix-Array beschreibt ein Feld (Array), in dem alle möglichen Endsilben (Suffixe) einer Zeichenkette (String) in lexikographischer Reihenfolge abgelegt sind.

Das Suffix-Array einer Zeichenfolge $T = t_1 t_2 t_3 \dots t_n$ ist als Permutation $S_1, S_2, \dots, S_n$ definiert, sodass jedes Suffix von T, das an Position S_i startet lexikographisch kleiner ist als, ein Suffix, welches an Position S_j beginnt (gilt für alle $i < j$). Nach der Sortierung stellen die ursprünglichen Indizes das Suffix-Array dar. So entsteht zum Beispiel das Suffix-Array $\{2, 4, 1, 6, 3, 5\}$ nach der lexikographischen Sortierung aller Suffixe des Strings `BANANE` (siehe Abbildung IV.1) [Ostrow, 2016].

Um einen Suchstring (Pattern) P im Text T zu finden, erfolgt die Durchsuchung der Suffix-Matrix nach Endsilben, die mit P starten mit Hilfe der Binären-Suche. Dazu wird das Interval $[s, e]$ als $[1, n]$ definiert und anschließend der Mittelpunkt $k = \lfloor (s+e)/2 \rfloor$ bestimmt. Startet der Suffix S_k mit P so wurde der Suchtext P gefunden. Ist der S_k-te Suffix lexikographisch kleiner als P, so ist es notwendig die Suche im Interval $[s, e]$ mit $s = k+1$, andernfalls mit $e = k-1$ fortzusetzen. Der zum Suffix S_k gehörige k-te Eintrag im Suffix-Array gibt die Startposition des Pattern P im Text T an.

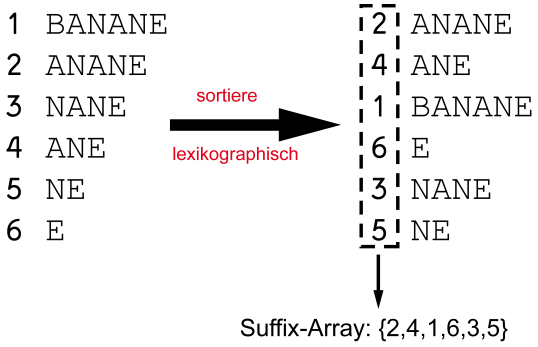


Abbildung IV.1: *Suffix-Array des Strings* `BANANE`*.*
Nach lexikographischer Sortierung aller Suffixe des Wortes `BANANE` *stellen die ursprünglichen Indizes das Suffix-Array* $\{2, 4, 1, 6, 3, 5\}$ *dar.*

Burrows-Wheeler-Transformation

Die Burrows-Wheeler-Transformation (BWT) ist eine invertierbare Permutation einer Zeichenkette, die ursprünglich in der Datenkomprimierung Anwendung fand. Dadurch, dass die BWT weniger Arbeitsspeicher benötigt als Suffix-Arrays, wurde sie in einige Mapping-Algorithmen wie BWA, Bowtie und SOAP2 integriert. Die BWT-Matrix eines Textes $T = t_1 t_2 t_3 \dots t_n$ mit $T' = T t_{n+1}$, wobei $t_{n+1} = \$$ und $\$$ lexikographisch kleiner ist, als jeder andere Buchstabe aus T, definiert sich als eine sortierte Liste aller Rotationen von T'. Jede Zeile der Matrix beschreibt eine dieser Rotationen. Die letzte Spalte der Matrix bildet von oben nach unten gelesen die BWT des Strings T. Die BWT $B = b_1 \dots b_{n+1}$ des oben genannten Beispiels `BANANE` lautet somit `EBN$NAA` (siehe Abbildung IV.2).

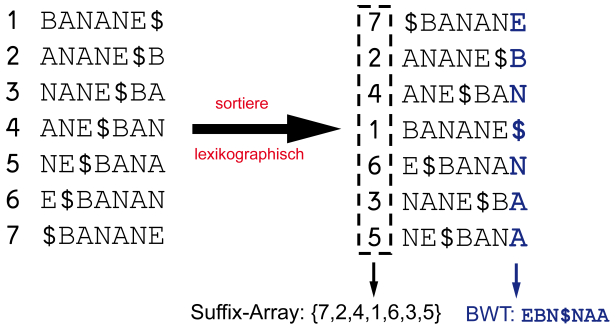


Abbildung IV.2: *Burrows-Wheeler-Transformation des Strings* `BANANE`*.*
Nach Anfügen eines Textbegrenzungszeichens (\$) und anschließender lexikographischer Sortierung aller Rotationen des Strings `BANANE$`*, stellt die letzte Spalte der entstandenen Matrix den BWT dar.*

Um mit Hilfe der BWT und des Suffix-Arrays einen Pattern P im Text T zu finden, müssen folgende Funktionen definiert werden. $F(x, i)$ beschreibt die absolute Häufigkeit des Buchstaben x in der BWT von T von Index 1 bis i. Die Funktion $C(x)$ gibt die Anzahl von Symbolen in B wieder, die lexikographisch kleiner als x sind. Der kleinste Index k, bei dem die k-te Zeile der BWT-Matrix

mit der Vorsilbe (Präfix) W beginnt, wird als $L(W)$ bezeichnet. Auf ähnliche Weise ist $U(W)$ als der größte Zeilenindex definiert, in der der Präfix W in der BWT-Matrix vorkommt.

Gelingt es $L(P)$ und $U(P)$ des Pattern $P = p_1 p_2 \dots p_m$ zu finden, so lässt sich die Position dieses Pattern im Text T bestimmen. Ein leerer String ϵ ist per Definition ein Präfix jedes beliebigen Strings. So kann W als leerer String initialisiert werden, wodurch $L(\epsilon) = 1$ und $U(\epsilon) = n+1$ gilt. Anschließend wird das Pattern rückwärts durchlaufen (i von m zu 1) und $L(p_i W)$ sowie $U(p_i W)$ (Gleichung IV.1 und IV.2) für jedes i aus der Funktion $L(W)$ und $U(W)$ mit Hilfe von $F(x, i)$ berechnet. Als additiver Faktor wird $C(x)$ hinzugefügt. Im nächsten Schritt ist W durch $p_i W$ zu ersetzen [Ostrow, 2016].

$$L(p_i W) = C(p_i) + F(p_i, L(W) - 1) + 1 \quad \text{(IV.1)}$$
$$U(p_i W) = C(p_i) + F(p_i, U(W)) \quad \text{(IV.2)}$$

Wird der oben beschriebene Algorithmus für das Beispielpattern $P = ANE$ und den Beispieltext $T = BANANE$ angewendet, so endet dieser, wie im Folgenden gezeigt, mit $L(ANE\epsilon) = 3$ und $U(ANE\epsilon) = 3$:

$$\begin{aligned} L(\epsilon) &= 1 & U(\epsilon) &= 7 \\ L(E\epsilon) &= 5 & U(E\epsilon) &= 5 \\ L(NE\epsilon) &= 7 & U(NE\epsilon) &= 7 \\ L(ANE\epsilon) &= 3 & U(ANE\epsilon) &= 3 \end{aligned}$$

Dies bedeutet, dass nur die dritte Reihe der BWT-Matrix von T mit dem Pattern ANE beginnt. Über den dritten Eintrag im Suffix-Array $S = \{7, 2, 4, 1, 6, 3, 5\}$ der BWT lässt sich die exakte Startposition von P mit $S_3 = 4$ im originalen Text T bestimmen.

Vor dem ersten Mapping von Sequenzdaten ist es notwendig mit dem `bwa index`-Befehl einen sogenannten FM-Index des Referenzgenoms zu erstellen. Dieser FM-Index wurde von Paolo Ferragina und Giovanni Manzini entworfen. Er basiert auf der Burrows-Wheeler Transformation und hat Ähnlichkeiten zu dem bereits erwähnten Suffix Array [Ferragina und Manzini, 2000]. Da die Weiterverarbeitung des Aligments durch die GATK die Zuordnung jedes Reads in eine Gruppe (engl. Read Group) fordert, ist die Definition einer solchen durch den Parameter `-R` zwingend erforderlich. Dabei müssen die Felder für den Identifier (`ID`), die Probenbezeichnung (`SM`), den Identifer für die Library Preparation (`LB`), den Hersteller des Sequenziergeräts (`PL`) und die Sequenzierplattform (`PU`) definiert sein. Das Read Group Tag wird durch ein `@RG` eingeleitet und führt die einzelnen Felder getrennt durch einen Tabulator auf (siehe Abbildung IV.3). Die Durchführung des Mappings erfolgt mit dem `bwa mem`-Befehl und den in Tabelle IV.2 aufgelisteten Parametern. Da das Tool das Alignment im SAM-Format direkt ins Terminal ausgibt, lässt sich die Standardausgabe durch Anhängen von `> <<NAME>>.sam` an den Terminal-Befehl in eine Datei umleiten.

Abbildung IV.3: *BWA Read Group*

Tabelle IV.2: `bwa mem` *Parameter*

Parameter	Erläuterung
`<<INDEX>>`	Pfad zum erstellten Index des Referenzgenoms
`-t`	Anzahl der zu verwendenden CPU Kerne
`-R`	Angabe der Read Group
`<<FASTQ1>>`	Pfad zur FASTQ-Datei mit FWD Sequenzen
`<<FASTQ2>>`	Pfad zur FASTQ-Datei mit REV Sequenzen

IV.VII FastQC

Die `fastqc`-Software liest FASTQ-Dateien ein und berechnet Metriken zur Qualitätskontrolle des Sequenzierlaufs und der Library Preparation. Diese können einen schnellen Überblick geben, ob die Rohdaten für eine weitere Analyse nutzbar sind. Über eine graphische Benutzeroberfläche ist es möglich einen FASTQ-Datensatz zu laden und die Ergebnisse nach der Analyse direkt in dieser auszuwerten. Zur Integration in eine Analysepipeline kann das Programm mit den entsprechenden Parametern (siehe Tabelle IV.3) auf dem Terminal gestartet werden. Die Software schreibt die Ergebnisse dann in Form einer HTML-Datei mit entsprechenden Abbildungen in das spezifizierte Ausgabeverzeichnis [Andrews *et al.*, 2010].

FastQC erstellt eine kleine Statistik über die Gesamtanzahl an Reads, die Anzahl der gefilterten Reads, die Sequenzlänge der Reads und den prozentualen GC-Gehalt des Eingabedatensatzes. Des Weiteren wird eine Abbildung mit Box-Plots des Qualitätswerts für jede Position der Reads sowie ein Report über den Qualitätswert pro Sequenz angezeigt. Der „Per Base Sequence Content“-Plot stellt die prozentuale Häufigkeit der vier verschiedenen Basen an jeder Readposition dar. Zur Analyse des GC-Gehalts stehen zwei Metriken zur Verfügung. Im „Per Base GC Content“-Plot ist der GC-Gehalt über alle Basen abgebildet. Die Abbildung „Per Sequence GC Content“ zeigt hingegen ein Histogramm des GC-Gehalts über alle Sequenzen. Über die Verteilung der Read-Längen lässt sich herausfinden, ob es eine Submenge an Reads gibt, die eine Abweichung von der erwarteten Sequenzierlänge aufweisen. Durch die graphische Darstellung der duplizierten Reads besteht die Möglichkeit einen schnellen Überblick über eine eventuell vorliegende PCR-Überamplifikation zu erhalten. Mithilfe der Angabe von überrepräsentativen Sequenzen kann zum Beispiel auf eine potentiell vorliegende Kontamination der Library geschlossen werden [Andrews, 2010].

Tabelle IV.3: `fastqc` *Parameter*

Parameter	Erläuterung
`-t`	Anzahl der zu verwendenden CPU Kerne
`-outdir`	Pfad zum Ausgabeverzeichnis
`<<FASTQ1>>`	Pfad zur FASTQ-Datei mit FWD Sequenzen
`<<FASTQ2>>`	Pfad zur FASTQ-Datei mit REV Sequenzen

IV.VIII GATK

Das Genome Analysis Toolkit (GATK) wurde am Broad Institute of MIT and Harvard entwickelt und im Jahr 2010 von McKenna *et al.* publiziert [McKenna *et al.*, 2010]. In den folgenden Jahren hat es sich als Industriestandard zur Identifizierung von SNVs sowie kleinen Insertionen und Deletionen (Indels) in NGS-Daten aus Keimbahn-DNA-Proben etabliert. Derzeit wird der Anwendungsbereich um die Detektion von somatischen Varianten sowie die Behandlung von Kopienzahlveränderungen und strukturellen Variationen (engl. Structual Variations, SVs) erweitert. Zusätzlich zu den Algorithmen zur Variantendetektion enthält das GATK viele Programme zur Ausführung verwandter Aufgaben, wie die Verarbeitung und Qualitätskontrolle von Hochdurchsatz-Sequenzierungsdaten. Ab Version 4 (GATK4) ist ebenfalls das Picard-Toolkit zur Bearbeitung und Qualitätskontrolle von NGS-Daten enthalten. Alle Picard-Tools sind direkt über die GATK-Befehlszeile aufrufbar und verfügen über eine harmonisierte Befehlssyntax und ein zusammengelegtes Benutzerhandbuch.

Die im GATK enthaltenen Werkzeuge (engl. Tools) wurden hauptsächlich entwickelt, um mit Illumina Geräten produzierte NGS-Daten aus Exom und Genom Sequenzierungen zu verarbeiten. Sie lassen sich jedoch anpassen um eine Vielzahl anderer Technologien und experimenteller Designs zu behandeln. Durch Erweiterungen kann das ursprünglich für die Humangenetik entwickelte Softwarepaket auch auf Genomdaten jedes beliebigen Organismus und diverse Ploidiegrade angewandt werden. Das Genom Analysis Toolkit ist unter Linux und anderen POSIX-kompatiblen Plattformen, einschließlich MacOS X, ausführbar. Die wichtigste Systemanforderung ist Java 1.8, allerdings haben einige Tools zusätzliche R- oder Python-Abhängigkeiten. Mit der Version 4 des GATK werden neben herkömmlichen Computerumgebungen, wie lokalen Clustern auch Cloud-Umgebungen und Spark-Architekturen unterstützt.

Das Herzstück des GATK ist ein Framework, welches die Datenzugriffe, Datenkonvertierungen und Datentraversierungen, sowie leistungsstarke Berechnungen übernimmt. Dies umfasst die Parallelisierung mit Apache Spark und die optimierte Nutzung der Cloud-Infrastruktur. Die Kernfunktionen werden von einer Vielzahl spezialisierter Tools genutzt, die einzeln oder verkettet in Workflows Anwendung finden können. Eine vollständige Liste aller zur Verfügung

stehenden Tools und deren Funktionen ist in der Tool-Dokumentation[3] auf der GATK Homepage zu finden.

Die Entwickler des Toolkits erarbeiteten eine Empfehlung für einen optimalen Workflow zur Anwendung der verschiedenen Software-Tools, welcher die Verarbeitung von den Rohdaten bis hin zu fertigen Ergebnissen in Form von Variantenlisten beschreibt. Dieser sogenannte „Best Practice Workflow“ wurde in der Produktion am Broad Institute erprobt und optimiert, um die genauesten Ergebnisse mit der höchsten Recheneffizienz zu erzielen. Abbildung IV.4 zeigt den Workflow, welcher für alle Hauptvariantenkategorien von Genom-, Exom- und Genpanel-Analysen genutzt werden kann.

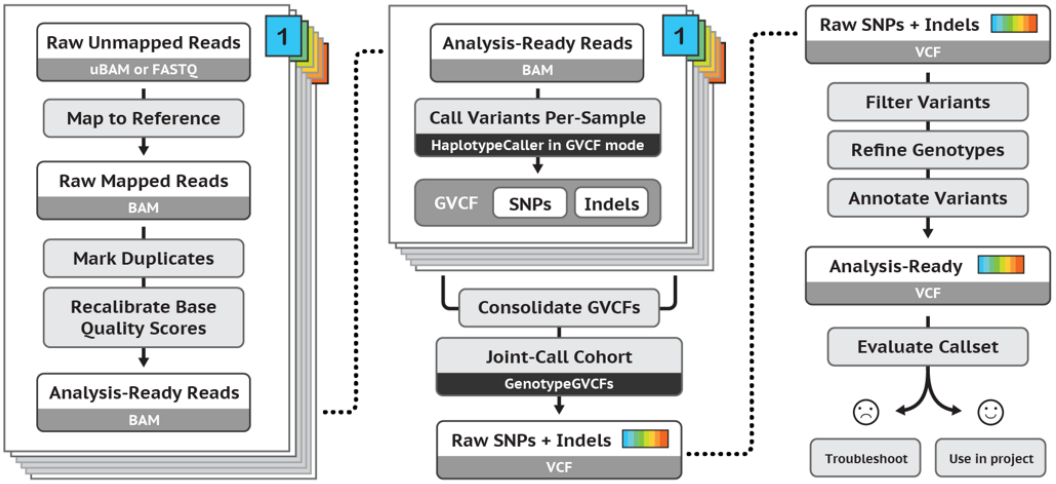


Abbildung IV.4: *GATK Best Practice Workflow zur Detektion von SNVs und Indels in Keimbahn-Proben.*
Dargestellt ist die Empfehlung der Entwickler des GATK für eine optimale Anwendung der verschiedenen Software-Tools, welche die Verarbeitung von den Rohdaten bis hin zu fertigen Ergebnissen in Form von Variantenlisten beschreibt. Dieser Workflow wurde in der Produktion am Broad Institute erprobt und optimiert. [Broad Institute, 2020d]

Im Folgenden werden die in dieser Arbeit genutzten Software-Tools mit den angewandten Parametern vorgestellt. Eine komplette Liste aller Parameter kann in der jeweiligen Online-Dokumentation oder über den Aufruf des Tools mit dem Parameter `-h` eingesehen werden.

IV.VIII.I Mark Duplicates

Als doppelter Read werden Sequenzen bezeichnet, die aus einem gleichen DNA-Fragment entstanden sind. Duplikate können durch Einsatz einer PCR zur Vervielfältigung der Fragmentbibliothek während der Library Preparation (PCR Duplikate) oder auch aus einem einzelnen Amplifikationscluster resultieren, welches der optische Sensor des Sequenzierers fälschlicherweise als mehrere

[3] `https://gatk.broadinstitute.org/hc/en-us/articles/360037224712--Tool-Documentation-Index`

Cluster erkennt. Diese Duplikationsartefakte werden als optische Duplikate bezeichnet. `MarkDuplicates` sucht und markiert diese doppelten Reads in einer BAM- oder SAM-Datei.

Das `MarkDuplicates`-Tool vergleicht Sequenzen an den 5‘ Positionen von Reads in einer SAM/BAM-Datei. Wurden bei der Bibliotheksherstellung eindeutige molekulare Barcodes verwendet, sind diese optional mit der `BARCODE_TAG`-Option anzugeben, um Duplikate zu finden und zu markieren. Anschließend hilft die Summe der Basenqualitätswerte eines Reads zur Unterscheidung des primären vom duplizierten Read. Das Tool gibt eine neue SAM- oder BAM-Datei aus, in der jeder als Duplikat identifizierte Read mit dem Hexadezimalwert 0x0400 (Dezimalwert 1024) im Flag-Feld markiert ist. Dieser Flag enthält allerdings keine Information über den Typ des Dukplikats. Zu diesem Zweck wird ein neues Tag mit dem Namen „Duplicate Type“ (DT) als optionale Ausgabe hinzugefügt. Durch Aufrufen der Option `TAGGING_POLICY` ist es möglich entweder alle Duplikate (All), nur die optischen Duplikate (OpticalOnly) oder keine Duplikate (DontTag) zu markieren. PCR-generierte Duplikate erhalten den Wert `LB` und sequenzierplattformabhängige Duplikate werden mit einem `SQ` im `DT`-Feld notiert. `MarkDuplicates` erstellt auch eine Datei mit einer Metrik, in der die Anzahl der Duplikate für Single- und Paired-End Reads angegeben ist.

Falls gewünscht lassen sich Duplikate mit den Optionen `REMOVE_DUPLICATE` und `REMOVE_SEQUENCING_DUPLICATES` entfernen. Das `MarkDuplikatesSpark`-Tool ist eine Spark-Implementierung von `MarkDuplikates`, mit welcher der Algorithmus auf mehreren Kernen eines lokalen Rechners oder auf einem Spark-Rechencluster ausgeführt werden kann. Die Ausgabe von `MarkDuplikatesSpark` und `Mark- Duplikates` unterscheidet sich nicht. Die parallele Ausführung mit zwei Kernen ist allerdings um 15% schneller. Dies lässt sich bis zu 16 Kernen linear skalieren.

Nachfolgend sind die für `MarkDuplicatesSpark` verwendeten Parameter gelistet und kurz erläutert (siehe Tabelle IV.4).

Tabelle IV.4: `MarkDuplicates` *Parameter*

Parameter	**Erläuterung**
`-I`	Eingabedatei im BAM/SAM Format
`-O`	Ausgabedatei im BAM/SAM Format
`--spark-master local[X]`	X = Anzahl der zu verwendenden CPU Kerne
`--tmp-dir`	Pfad zu einem temporären Ordner
`-OBI`	erstellt einen Index für die Ausgabedatei

IV.VIII.II Base Qualitiy Score Recalibration

Leider unterliegen die von den Sequenziermaschinen erzeugten Basenqualitätswerte verschiedenen Quellen systematischer technischer Fehler, was zu über- oder unterschätzten Qualitätswerten in den Daten führt. Einige dieser Fehler sind auf die Funktionsweise der Sequenzierungsreaktion zurückzuführen, andere wiederum auf Herstellungsfehler in den zugehörigen Verbrauchsmaterialien. Bei der Rekalibrierung der Basenqualitätswerte (engl. Base Quality Score Recalibration, BQSR) handelt es sich um einen Datenvorverarbeitungsschritt, der diese Fehler bei der Abschätzung der jeweiligen Basenqualität findet und anpasst. Die korrekte Angabe dieser Qualitätswerte ist wichtig, da Algorithmen von nachfolgenden Analyseschritten wie dem Auffinden von Varianten (engl. Variant Calling) stark von diesen Qualitätsfaktoren abhängen.

Bei der BSQR wird der Prozess des maschinellen Lernens angewendet, um diese Fehler empirisch zu modellieren und die Qualitätswerte entsprechend anzupassen. Zum Beispiel kann für einen bestimmten Lauf festgestellt werden, dass wenn zwei A-Nukleotide hintereinander folgen, die nächste Base eine um 1% höhere Fehlerrate aufweist. Daher sollte bei jeder Base, die nach AA in einem Read folgt, der Qualitätswert um 1% reduziert sein. Die Berechnung einer solchen Anpassung geschieht über mehrere verschiedene Kovarianten, wobei hauptsächlich der Sequenzkontext und die Position im jeweiligen Read auf additive Weise einfließt. Dies ermöglicht es insgesamt exaktere Basisqualitäten zu erhalten, was wiederum die Genauigkeit des Variant Callings verbessert [Broad Institute, 2020a].

Im ersten Schritt erzeugt das Tool `BaseRecalibrator` Tabellen, die auf einem aus den Alignment Daten (`-I`) und den angegebenen bekannten Polymorphen Stellen (`-known-sites`) unter Berücksichtigung der angegebenen Kovarianten erstellten Modell basieren. Als Datenbanken für bekannte polymorphe Stellen im Genom können zum Beispiel Daten aus der dbSNP Datenbank oder dem gnomAD Projekt verwendet werden. Im Folgenden sind einige wichtige Parameter des `BaseRecalibrator`-Tools in Tabelle IV.5 vorgestellt:

Tabelle IV.5: `BaseRecalibrator` *Parameter*

Parameter	**Erläuterung**
`-I`	Eingabedatei im BAM/SAM-Format
`-O`	Ausgabe Datei mit Rekalibrierungstabelle
`-R`	Pfad zum Referenzgenom im FASTA-Format
`-known-sites`	Datenbank bekannter polymorpher Stellen. (Mehrfache Aufführung möglich)
`-L`	Datei mit Regionen, die analysiert werden sollen
`--interval-padding`	Anzahl der Basen, um die ein Interval an jeder Seite erweitert werden soll

Das `ApplyBQSR`-Tool adjustiert im zweiten Schritt die Basenqualitätswerte durch die im vorherigen Schritt entstandene Rekalibrierungsdatei und produ-

ziert eine neue Ausgabedatei (`-O`) im SAM- oder BAM-Format. Die essentiellen Optionen zum Aufruf des `ApplyBQSR`-Tools sind in Tabelle IV.6 nachstehend beschrieben.

Tabelle IV.6: `ApplyBQSR` *Parameter*

Parameter	Erläuterung
`-I`	Eingabedatei im BAM/SAM Format
`-O`	Ausgabedatei im BAM/SAM Format
`-R`	Pfad zum Referenzgenom im FASTA Format
`-bqsr`	Pfad zur Rekalibrierungstabelle aus dem BaseRecalibrator Schritt
`-OBI`	erstellt einen Index für die Ausgabedatei
`--tmp-dir`	Pfad zu einem temporären Ordner

IV.VIII.III HaplotypeCaller

Der `HaplotypeCaller` ist im Stande SNPs und Indels gleichzeitig über die lokale *de-novo*-Assemblierung von Haplotypen in sogenannten aktiven Region zu detektieren. Aktive Regionen sind Bereiche im Alignment, die eine Abweichung zum Referenzgenom zeigen. Dabei verwirft die Software das vorhandene Mapping dieses Bereichs und reassembliert die betroffenen Reads. Dies ermöglicht dem Algorithmus schwer zu sequenzierende Areale des Genoms genauer auf Varianten zu untersuchen. Außerdem kann der `HaplotypeCaller` mit diesem Ansatz kleine Insertionen und Deletionen besser detektieren als positionsbasierte Algorithmen, wie zum Beispiel der `UnifiedGenotyper` aus früheren GATK Versionen [Poplin *et al.*, 2017].

Die von `HaplotypeCaller` ausgeführten Operationen lassen sich in die vier Hauptschritte Identifizierung von aktiven Regionen, Assemblierung plausibler Haplotypen, Bestimmung der Wahrscheinlichkeiten für eine Variante pro Read und die finale Genotypisierung unterteilten (siehe Abbildung IV.5). Die einzelnen Schritte werden im Folgenden kurz erläutert [Broad Institute, 2020b].

Identifizierung von aktiven Regionen In diesem Teil des Programms durchläuft ein Algorithmus das Alignment, um Regionen des Genoms zu identifizieren, in denen die zu analysierende Probe wesentliche Anzeichen für eine Abweichung zum Referenzgenom und damit eine Variation zeigen. Diese Bereiche werden aktive Regionen genannt (siehe Abbildung IV.5 oben links). Nur die aktiven Regionen werden an den nächsten Schritt weitergegeben.

Assemblierung plausibler Haplotypen Ziel dieses Schrittes ist die Sequenz der physischen DNA-Segmente des Probenorganismus zu rekonstruieren. Dazu durchläuft das Programm jede aktive Region und verwendet die in diesem Bereich ursprünglich gemappten Reads, um vollständige Sequenzen über die Gesamtlänge der aktiven Region herzustellen. Diese werden als Haplotypen

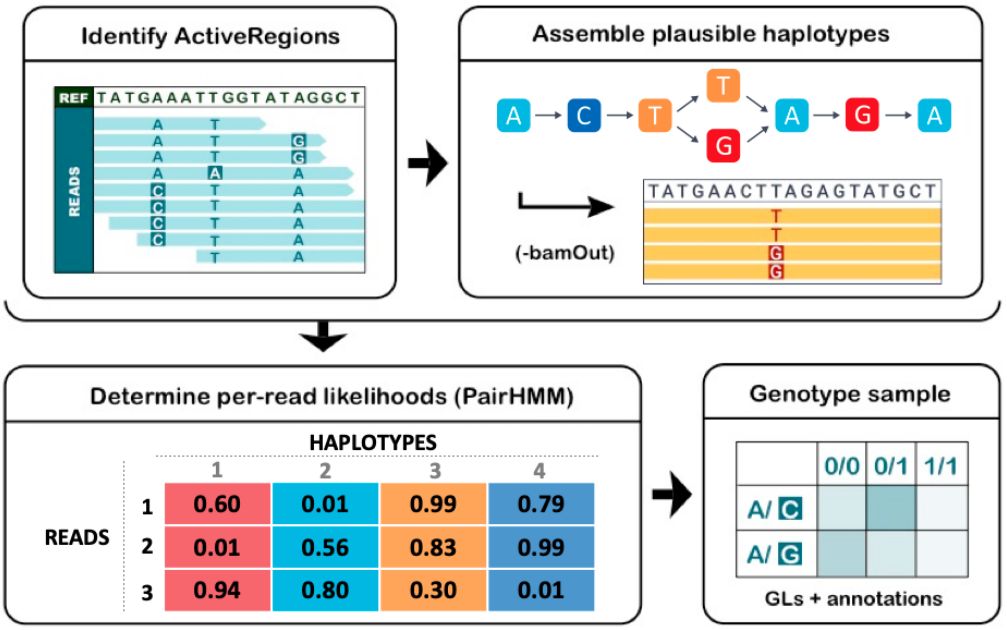


Abbildung IV.5: *GATK Haplotype Caller*
Basierend auf Informationen über das Vorliegen einer Variante bestimmt das Programm die zu bearbeitenden Regionen (oben links). Für jede dieser Regionen erstellt es einen De Bruijn-ähnlichen Graphen und identifiziert die in den Daten vorhandenen möglichen Haplotypen. Die Identifizierung von potentiellen Varianten erfolgt mit Hilfe des Smith-Waterman-Alignments (oben rechts). Im nächsten Schritt wird ein paarweises Alignment jedes Reads der betroffenen Region gegen jeden bestimmten Haplotypen durchgeführt. Die resultierende Tabelle findet Verwendung, um die Wahrscheinlichkeit eines Allels pro Read zu bestimmen (unten links). Abschließend wird mit Hilfe der Bayes'schen Regeln der Probe der wahrscheinlichste Genotyp zugeordnet (unten rechts). [Broad Institute, 2020b]

bezeichnet. Durch die Vielfalt bei polyploiden Daten, verschiedene Kombinationsmöglichkeiten von nicht koppelbaren Allelen und Sequenzier- sowie Mappingfehler entstehen verschiedene Haplotypen für jede aktive Region. In einem De Bruijn-ähnlichen Graphen für jede aktive Region wird unter Verwendung der Referenzsequenz als Vorlage versucht, jeden Read nach und nach einem Segment des Graphen zuzuordnen. Wenn ein Teil eines Reads nicht mit dem Graphen übereinstimmt, erfolgt das Hinzufügen eines neuen Knotens, um die Nichtübereinstimmung (engl. missmatch) zu berücksichtigen. Dies resultiert in einem komplexen Graphen mit vielen möglichen Pfaden. Da der Algorithmus notiert wie viele Reads einen Pfad im Graphen unterstützen, können so die wahrscheinlichsten (best unterstützten) Pfade ausgewählt werden. Aus diesen lassen sich dann die Haplotypsequenzen ableiten.

Das Programm aligniert anschließend jede Haplotypsequenz unter Verwendung des Smith-Waterman-Algorithmus [Smith und Waterman, 1981] gegen das Re-

ferenzgenom, um potenzielle Varianten zu registrieren (siehe Abbildung IV.5 oben rechts).

Bestimmung der Wahrscheinlichkeiten für eine Variante pro Read Im nun folgenden Schritt muss bewertet werden, wie viele Daten die Haplotypen aus dem vorherigen Schritt unterstützen (siehe Abbildung IV.5 unten links). Dazu betrachtet das Programm jeden einzelnen Read der aktiven Region und aligniert diesen unter Verwendung eines paarweise verdeckten Markowmodells (engl. Pair Hidden Markow Model, PairHMM) [Durbin *et al.*, 1998] nacheinander an jeden Haplotypen (einschließlich des Referenzhaplotypen). Dieses gibt eine Bewertung für jede Read-Haplotyp-Paarung aus, welche die Wahrscheinlichkeit beschreibt diesen Haplotyp bei einem Read zu beobachten. Die Werte werden dann verwendet, um die Evidenz des Vorliegens eines Allels an der im vorhergehenden Schritt identifizierten Kandidatenstelle zu berechnen. Dieser Prozess heißt „Marginalisierung der Allele“ und erzeugt die Werte, die schließlich im nächsten Schritt verwendet werden, um der Probe einen Genotyp zuzuweisen.

Genotypisierung Der vorherige Schritt erzeugte eine Tabelle, welche die Allelwahrscheinlichkeiten pro Read für jeden Varianten-Kandidaten angibt. Im Folgenden müs-sen diese Wahrscheinlichkeiten im Ganzen bewertet werden, um zu bestimmen, welcher Genotyp an der entsprechenden Position in der Probe am wahrscheinlichsten ist. Dies erfolgt durch Anwendung des Bayes-Theorems zur Berechnung der Wahrscheinlichkeiten jedes möglichen Genotyps und Auswahl des wahrscheinlichsten. Daraus ergibt sich der Genotyp sowie eine Vielzahl verschiedener Metriken, die in der VCF-Datei als Annotation zur entsprechenden Variante ausgegeben werden (siehe Abbildung IV.5 unten rechts).

Die in dieser Arbeit genutzten Parameter des `HaplotypeCaller`-Tools sind Tabelle IV.7 zu entnehmen.

Tabelle IV.7: `HaplotypeCaller` *Parameter*

Parameter	**Erläuterung**
`-I`	Eingabedatei im BAM/SAM-Format
`-O`	Ausgabedatei im VCF-Format
`-R`	Pfad zum Referenzgenom im FASTA-Format
`-L`	Datei mit Regionen, die analysiert werden sollen
`--interval-padding`	Anzahl der Basen, um die ein Intervall an jeder Seite erweitert werden soll

IV.VIII.IV Tools zur CNV-Analyse

Zur Detektion von Keimbahn CNVs in NGS-Daten mit Hilfe der Sequenziertiefe werden verschiedene Tools des GATK sequenziell verwendet. Zudem ist

es bei dieser Methode notwendig vor der Analyse einzelner Datensätze ein Modell aus mindestens 100 Proben zu erstellen. Daher werden im Folgenden für einige Software-Tools zwei Modi beschrieben. Der Kohortenmodus (engl. cohort mode) und der Fallmodus (engl. case mode). Der Kohortenmodus generiert ein Modell und sucht gleichzeitig nach CNVs in jeder Probe der Kohorte. Der Fallmodus analysiert eine einzelne Probe anhand eines bereits durch eine Kohorte erstellten Modells. Aufgrund der hohen Rechenintensität lohnt es sich einmalig ein Modell zu erstellen und neu sequenzierte Proben anhand dieses Modells zu analysieren. Für jede Anreicherungsmethode muss hierbei ein eigenes Modell erstellt werden.

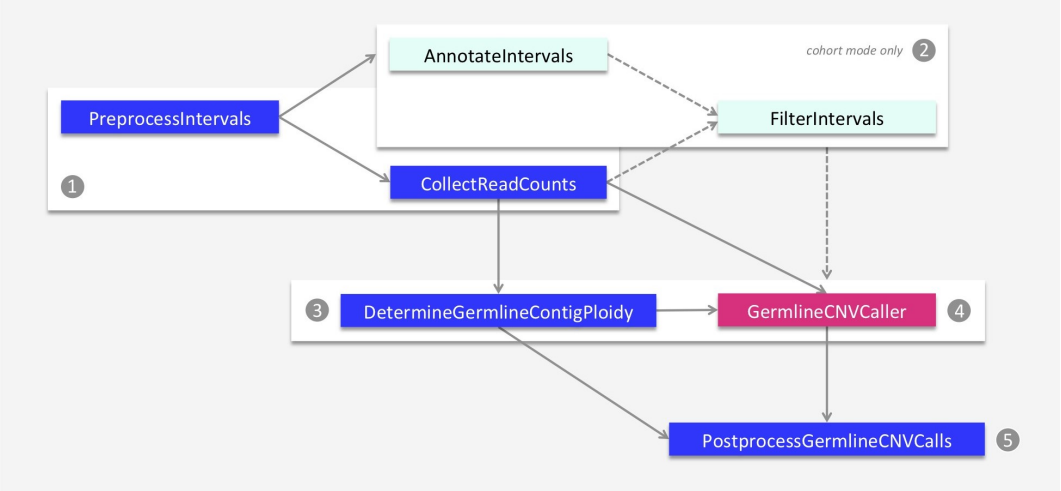


Abbildung IV.6: *GATK gCNV Workflow.*
Abschnitt 1 erstellt eine Intervallliste und zählt die Anzahl an Reads, die die Intervalle überlappen. Die optionalen Schritte in Abschnitt 2 dienen der Annotation der Intervalle mit Kovarianten. Diese werden zur Filterung der Intervalle sowie zur Erstellung des Modells verwendet. Zudem entfernt dieser Schritt Ausreißer innerhalb der Read Count Daten. Im Anschluss wird die Ploidie jedes Chromosoms bestimmt und im Kohortenmodus ein Modell erstellt (Abschnitt 3). Schritt 4 dient der Bestimmung einer Kopienzahl für jedes einzelne Intervall. Im Kohortenmodus erfolgt schließlich die Anfertigung eines Modells zur späteren Anwendung auf Einzelproben. Im letzten Schritt werden benachbarte Intervalle mit gleicher Kopienzahl zu Segmenten zusammengefasst und als VCF Datei ausgegeben. [Broad Institute, 2020c]

Abbildung IV.6 zeigt die für die CNV-Analyse notwendigen Software-Tools und Schritte. Abschnitt 1 erstellt eine Intervall-Liste (`PreprocessIntervals`) und zählt die Reads, die die Intervalle überlappen (`CollectReadCounts`). Abschnitt 2 beschreibt optionale, aber empfohlene Schritte zur Annotation von Intervallen mit Kovarianten (`AnnotateIntervals`) zur späteren Filterung. Zudem werden Intervalle mit Ausreißern aus den Ergebnissen von `CollectReadCounts` entfernt (`FilterInterals`). Abschnitt 3 generiert globale Ploidie-Basislevel für jedes Chromosom pro Probe (`DetermineGermlineContigPloidy`). In Abschnitt 4 wird die Kopienzahl pro Intervall bestimmt (`GermlineCNVCaller`).

Dieser Teil lässt sich ebenfalls wie Abschnitt 3 im Kohorten- sowie Fallmodus ausführen. Schließlich wird in Abschnitt 5 durch das `PostprocessGermlineCNVCalls`-Tool die Kopienanzahl jeder Probe pro Intervall bestimmt. Zusammen mit diesen Ergebnissen erfolgt die Generierung einer Datei, in der aufeinanderfolgende Regionen mit gleicher Kopienzahl als zusammenhängende Segmente ausgegeben werden. Die Ergebnisse liegen im VCF-Format vor. Nachfolgend werden die einzelnen Tools und deren Parameter kurz vorgestellt.

Präprozessierung von Regionen Das `PreprocessIntervals`-Tool traversiert die mit `-L` übergebenen Regionen, erweitert sie auf beiden Seiten mit der durch `--padding` angegebenen Anzahl an Basen und verbindet überlappen-de Intervalle. Wenn erwünscht, können diese Intervalle anschließend in kleinere Segmente, sogenannte „bins“, unterteilt werden. Für TES und WES Experimente ist es allerdings ratsam die Regionen nicht weiter zu teilen. Dies kann durch Setzen des Parameters `--bin-length 0` erreicht werden. Eine kurze Erläuterung der Parameter ist in Tabelle IV.8 dargestellt.

Tabelle IV.8: `PreprocessIntervals` *Parameter*

Parameter	**Erläuterung**
`-R`	Pfad zum Referenzgenom im FASTA-Format
`-L`	Datei mit Regionen, die analysiert werden sollen
`--padding`	Anzahl der Basen, um die ein Interval an jeder Seite erweitert werden soll
`--bin-length`	Größe der Segmente; 0 = keine Unterteilung in Segmente
`-O`	Ausgabedatei der präprozessierten Regionen

Zählen der Reads pro Intervall `CollectReadCounts` zählt wie viele Reads eines Alignments in durch den Parameter `-L` festgelegten Intervallen starten und gibt die Anzahl (engl. Read Counts) pro Intervall in einer HDF5 Datei aus. Tabelle IV.9 zeigt die entsprechenden Parameter.

Tabelle IV.9: `CollectReadCounts` *Parameter*

Parameter	**Erläuterung**
`-I`	Eingabedatei im BAM/SAM-Format
`-O`	Ausgabedatei mit Read Counts im HDF5-Format
`-L`	Datei mit Regionen, die analysiert werden sollen
`-imr`	Regel zum Zusammenführen von nebeneinander liegenden Intervallen

Annotation von Regionen Speziell bei TES und WES, die beide auf einer Library Preparation durch Anreicherung basieren, kann es sehr hilfreich

sein Regionen auszuschließen, die schlecht sequenzierbar sind. Um im folgenden Schritt die Daten korrekt filtern zu können, fügt `AnnotateIntervals` den übergebenen Regionen (`-L`) den entsprechenden GC-Gehalt sowie Daten über die sogenannte „Mappability“, welche die Wahrscheinlichkeit beschreibt eine eindeutige Position für einen Read im Genom zu finden, hinzu. Eine Datei mit diesen Informationen kann von der Homepage des Hoffman Labs des Princess Margaret Cancer Centre in Toronto[4] heruntergeladen werden [Karimzadeh *et al.*, 2018]. Tabelle IV.10 zeigt die in dieser Thesis verwendeten Parameter.

Tabelle IV.10: `AnnotateIntervals` *Parameter*

Parameter	**Erläuterung**
`-R`	Pfad zum Referenzgenom im FASTA-Format
`-O`	Ausgabedatei mit annotierten Regionen
`-L`	Datei mit Regionen, die annotiert werden sollen
`-imr`	Regel zum Zusammenführen von nebeneinander liegenden Intervallen
`--mappability-track`	Pfad zur Mappability Datei des Hoffman Labs

Filtern von Regionen Das `FilterIntervals`-Tool bildet ein Subset der durch `-L` übergebenen Regionen. Dabei nutzt das Programm die mit Annotationen versehene Ausgabedatei des `AnnotateIntervals`-Programms sowie die von `CollectReadCounts` ausgegebenen HDF5-Dateien. Die jeweiligen Schwellenwerte zur Filterung der Regionen können durch Setzen der entsprechenden Parameter bestimmt werden. Das Resultat ist eine gefilterte Liste mit Regionen im Picard-Stil. Tabelle IV.11 erklärt die verwendeten Parameter.

Tabelle IV.11: `FilterIntervals` *Parameter*

Parameter	**Erläuterung**
`-L`	Datei mit Regionen, die gefiltert werden sollen
`-I`	HDF5-Datei mit ReadCounts (Mehrfache Anwendung möglich)
`-O`	Ausgabedatei mit gefilterten Regionen
`--annotated-intervals`	Datei mit annotierten Regionen (Ausgabedatei von `AnnotateIntervals`)
`-imr`	Regel zum Zusammenführen von nebeneinander liegenden Intervallen

Ploidie bestimmen Das `DetermineGermlineContigPloidy`-Tool bestimmt den Ploidiezustand aller angegebenen Chromosomen in Keimbahnproben. Hierzu nutzt der Algorithmus die aus dem `CollectReadCounts`-Programm resultierenden Read Count Daten. Diese Keimbahn-Karyotypisierung wird häufig

[4] `https://bismap.hoffmanlab.org/`

in bioinformatischen Auswertungspiplines verwendet, um eine Geschlechterbestimmung oder ein Aneuploidiescreening durchzuführen. Zudem ist es der erste Schritt der GATK-Keimbahn-CNV-Pipeline und stellt ein CNV Basislevel inklusive Wahrscheinlichkeiten für alternative Kopienzahlzustände pro Chromosom bereit.

Eine Karyotypisierung auf Basis von Read-Counts erfordert die Modellierung des systematischen Fehlers und dessen Streuung in Bezug auf die technische Abdeckung der angereicherten Genomregionen jedes Chromosoms. Durch das im vorliegenden Algorithmus verwendete Bayes'sche Modell mit dem dazugehörigen Inferenzschema kann der Großteil der technischen Varianz erklärt und abgeleitet werden. Die von diesem Tool durchgeführte Berechnung lässt sich, abgesehen vom Parsen und Validieren von Eingabedaten, außerhalb der Javaumgebung und unter Verwendung des Python-Moduls `gcnvkernel` durchführen. Dazu ist es notwendig eine Python-Conda-Umgebung mit `gcnvkernel` und seinen Abhängigkeiten ordnungsgemäß einzurichten.

Dieses Tool lässt sich sowohl im Kohorten- als auch im Fallmodus ausführen. Soll die Karyotypisierung sukzessive auf neue Proben angewandt werden, so lohnt es sich einmalig im Kohortenmodus ein Modell zu berechnen, mit Hilfe dessen die einzelnen Proben dann im Fallmodus analysiert werden.

Ist dem Programm kein Pfad zu einem bereits berechneten Modell übergeben worden, so wird es im Kohortenmodus ausgeführt. In diesem Modus erfolgt die Bestimmung der Parameter des Modells und der Karyotyp jeder verwendeten Probe. Je mehr Proben zur Berechnung des Modells zur Verfügung stehen, desto besser lassen sich technische Fehler modellieren. Zudem besteht die Notwendigkeit eine Tabelle mit A-priori-Wahrscheinlichkeiten für jeden Ploidiestatus jedes Chromosoms zu spezifizieren. Dabei müssen die Bezeichnungen der Chromosomen in dieser Tabelle mit denen der Read-Count Daten übereinstimmen. Darüber hinaus ist es erforderlich, dass jedes Chromosom aus der HDF5-Datei in der Tabelle vertreten ist. Um einen Ploidiestatus explizit auszuschließen, kann der entsprechende A-priori-Wahrscheinlichkeitswert auf 0 gesetzt werden. Die in dieser Arbeit verwendeten A-priori-Wahrscheinlichkeiten der Ploidien für jedes Chromosom sind in Tabelle IV.12 dargestellt.

Als Ausgabe entstehen zwei Unterverzeichnisse, von denen eines mit „-model“ und das andere mit „-calls“ endet. Das Modell-Unterverzeichnis enthält die abgeleiteten Parameter des Ploidiemodells, die später zur Karyotypisierung weiterer Proben im Fallmodus verwendet werden können (siehe unten). Das Unterverzeichnis „-calls“ schließt wiederum ein Unterverzeichnis pro Probe ein, in dem verschiedene Werte, wie die globale Lesetiefe, die durchschnittliche Ploidie, die Basisploidien pro Chromosom und die Varianzschätzungen für die Abdeckung pro Chromosom aufgeführt sind. Die in dieser Arbeit verwendeten Parameter zur Erstellung eines Modells zur Bestimmung der Ploidien sind in Tabelle IV.13 aufgelistet.

Tabelle IV.12: *A-priori Wahrscheinlichkeiten der Ploidien*

CONTIG_ NAME	PLOIDY_ PRIOR_0	PLOIDY_ PRIOR_1	PLOIDY_ PRIOR_2	PLOIDY_ PRIOR_3
chr1	0.1	0.1	0.97	0.1
chr2	0.1	0.1	0.97	0.1
chr3	0.1	0.1	0.97	0.1
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
chrX	0.1	0.49	0.49	0.1
chrY	0.5	0.5	0	0

Tabelle IV.13: `DetermineGermlineContigPloidy (Cohort Mode)` *Parameter*

Parameter	Erläuterung
`-L`	Datei mit (gefilterten) Regionen
`-I`	Datei mit ReadCounts (Mehrfache Anwendung möglich)
`-imr`	Regel zum Zusammenführen von nebeneinander liegenden Intervallen
`--contig-ploidy-priors`	Pfad zur Tabelle mit A-priori-Wahrscheinlichkeiten für jedes Chromosom
`--output`	Pfad zum Ausgabeverzeichnis
`--output-prefix`	Präfix für die Unterverzeichnisse der Ausgabe

Wenn ein Pfad zu einem zuvor erstellten Ploidiemodell (`--model`) angegeben ist, beginnt die Ausführung des Tools im Fallmodus. In diesem werden die Parameter des Ploidiemodells aus dem bereitgestellten Verzeichnis geladen und mit Hilfe dessen aus den probenspezifischen Werten die Ploidie jedes Chromosoms der angegebenen Probe bestimmt. Die Ausgabe des Fallmodus beschränkt sich auf das Unterverzeichnis „-calls“, welches analog zum entsprechenden Ordner im Kohortenmodus organisiert ist. Die verwendeten Parameter sind in Tabelle IV.14 veranschaulicht.

Tabelle IV.14: `DetermineGermlineContigPloidy (Case Mode)` *Parameter*

Parameter	Erläuterung
`--model`	Pfad zum vorher erstellten Modell
`-I`	Datei mit Read-Counts
`--output`	Pfad zum Ausgabeverzeichnis
`--output-prefix`	Präfix für die Unterverzeichnisse der Ausgabe

Auffinden von Keimbahn-CNVs Beim Auffinden von Keimbahn-CNVs mit dem `GermlineCNVCaller`-Tool werden auf Basis der Read Counts und der entsprechenden Ploidie pro Chromosom (Ergebnis des vorherigen Schritts) Kopienzahlveränderungen für jede gegebene Region berechnet. Diese Berechnungen werden analog dem Tool `DetermineGermlineContigPloidy` unter Verwendung des `gcnvkernel` Python Moduls außerhalb der Javaumgebung ausgeführt.

Um CNVs verlässlich anhand von Coverage Daten aus TES, WES und WGS Experimenten detektieren zu können, muss ebenso wie im vorherigen Schritt zur Poidiebestimmung durch den `DetermineGermlineContigPloidy`-Algorithmus im Kohortenmodus (`--run-mode COHORT`) ein umfassendes Modell zur Berücksichtigung unterschiedlicher technischer Fehler bei der Library Preparation und Sequenzierung erstellt werden. Dieses ist auch in diesem Algorithmus durch ein Bayes'sches Modell zur Darstellung eines Großteils der technischen Varianz realisiert. Das Modell kann vom `GermlineCNVCaller`-Tool automatisch erstellt werden. Hierzu besteht die Notwendigkeit Read-Count-Daten einer Kohorte von Keimbahnproben bereitzustellen, die unter Verwendung derselben Sequenzierungsplattform und des gleichen „Library Preparation"-Protokolls sequenziert wurden. Je nach Größe des verfügbaren Arbeitsspeichers kann es erforderlich sein, die Berechnung des Modells in mehrere Prozesse zu unterteilen. Zu diesem Zweck lässt sich mit dem `-L`-Parameter eine Liste mit Regionen übergeben. Die Ausgabe des Kohortenmodus generiert ähnlich wie im vorhergehenden Schritt zur Bestimmung der Ploidie einen Ordner mit Daten des Modells (-model) und ein Verzeichnis mit CNV Ergebnissen pro Probe (-calls). Die Parameter zur Erstellung des Modells zur CNV Detektion sind in Tabelle IV.15 aufgeführt.

Tabelle IV.15: `GermlineCNVCaller (Cohort Mode)` *Parameter*

Parameter	**Erläuterung**
`--run-mode`	Kohorten- (COHORT) oder Fallmodus (CASE)
`-L`	Datei mit (gefilterten) Regionen
`-I`	Datei mit ReadCounts
`--contig-ploidy-calls`	Pfad zum „calls" Ausgabeordner von `DetermineGermlineContigPloidy`
`--annotated-intervals`	Datei mit (annotierten) Regionen
`-imr`	Regel zum Zusammenführen von nebeneinander liegenden Intervallen
`--output`	Pfad zum Ausgabeverzeichnis
`--output-prefix`	Präfix für die Unterverzeichnisse der Ausgabe
`--tmp-dir`	Pfad zu einem temporären Ordner

Das parametrisierte Modell kann im Fallmodus (`--run-mode CASE`) für die Bestimmung von CNVs zukünftiger Proben verwendet werden, sofern diese in Bezug auf den Gewebetyp, die „Library Preparation" und das Sequenzierungsprotokoll streng mit der Kohorte kompatibel sind, die zur Generierung

der Modellparameter verwendet wurde. Dazu ist der Pfad zum entsprechenden Verzeichnis, welches die Daten zum Modell beinhaltet, über `-model` anzugeben. Die Ausgabe im Fallmodus beinhaltet nur den Ordner `-calls`, welcher analog zum `-calls` Ordner im Kohortenmodus, die Daten über Kopienzahlveränderungen der untersuchten Probe enthält. Im Folgenden sind die verwendeten Parameter aufgeführt (siehe Tabelle IV.16).

Tabelle IV.16: `GermlineCNVCaller (Case Mode)` *Parameter*

Parameter	**Erläuterung**
`--run-mode`	Kohorten- (COHORT) oder Fallmodus (CASE)
`-model`	Pfad zum vorher erstellten Modell
`--contig-ploidy-calls`	Pfad zum Ausgabeordner von `DetermineGermlineContigPloidy`
`-I`	Datei mit ReadCounts
`--output`	Pfad zum Ausgabeverzeichnis
`--output-prefix`	Präfix für die Unterverzeichnisse der Ausgabe

Nachbearbeitung der gefundenen CNVs Das `PostprocessGermlineCNVCalls`-Tool verarbeitet die Ausgabe von `GermlineCNVCaller` und generiert zwei VCF-Dateien. Eine VCF-Datei enthält die Angabe der Kopienzahlveränderungen jeder einzelnen Region mit Informationen über die Qualität und den Genotyp der Veränderung sowie Wahrscheinlichkeiten für jede mögliche Kopienzahl. Angesichts der Tatsache, dass CNV-Ereignisse häufig mehrere aufeinanderfolgende Intervalle umfassen, ist es sinnvoll zusammenhängende Intervalle mit der gleichen Kopienzahl zu konstanten Segmenten zusammenzuführen. Die Ablage dieser Segmente erfolgt in einer zweiten VCF-Datei. Für die Segmentierung und Genotypisierung wird auch hier das Pythonmodul `gcnvkernel` verwendet.

In beiden VCF-Ausgaben wird das alternative Allel für eine Kopienzahlveränderung mit `<DEL>` oder `<DUP>` aufgeführt. Dies hängt davon ab, ob die wahrscheinlichste Kopienzahl unter oder über der Referenzkopienzahl des Chromosoms liegt. Eine Liste von allosomalen Chromosomen kann über das Argument`--allosomal-contig` angegeben werden. Alle nicht aufgelisteten Chromosomen behandelt der Algorithmus als autosomal. Die Referenzkopienzahl auf einem allosomalen Chromosom wird durch den Geschlechtskaryotyp der Probe bestimmt und auf den im Ergebnis von `DetermineGermlineContigPloidy` bestimmten Zustand gesetzt.

Die zur Prozessierung verwendeten Parameter sind Tabelle IV.17 zu entnehmen.

Tabelle IV.17: `PostprocessGermlineCNVCalls` *Parameter*

Parameter	**Erläuterung**
`--calls-shard-path`	Pfad zu `-calls` Verzeichnis des `GermlineCNVCaller`-Tools (mehrfache Anwendung möglich)
`--model-shard-path`	Pfad zu Verzeichnis mit dem Modell zur CNV-Analyse (mehrfache Anwendung möglich)
`--contig-ploidy-calls`	Pfad zum Ausgabeordner von `DetermineGermlineContigPloidy`
`--autosomal-ref-copy-number`	Referenzkopienzahl für autosomale Chromosomen
`--allosomal-contig`	Angabe der allosomalen Chromosomen (mehrfache Anwendung möglich)
`--output-genotyped-intervals`	Dateiname für intervallbezogene Ausgabe
`--output-genotyped-segments`	Dateiname für segmentbezogene Ausgabe
`--output-denoised-copy-ratios`	Dateiname für die entrauschten Daten

IV.IX Samtools

Samtools ist ein Paket von Programmen für die Bearbeitung von Hochdurchsatzsequenzierungsdaten im SAM, BAM oder CRAM Dateiformat (siehe Anhang II.III), welches gemeinsam mit diesen Formaten von Heng Li *et al.* publiziert wurde [Li *et al.*, 2009a]. Das Softwarepaket beinhaltet Programme zur Sortierung, Indizierung, Konvertierung sowie der Extraktion eines Subsets an Daten aus einer SAM, BAM oder CRAM Datei. Zudem wurden im Laufe der Zeit weitere Programme zum Auffinden von Varianten und der Analyse von homozygoten Regionen (engl. Regions of Homozygousity, ROH) hinzugefügt [Li, 2011] [Narasimhan *et al.*, 2016].

IV.X InterVar

Das Amerikanische College für Medizinische Genetik und Genomik (engl. American College of Medical Genetics and Genomics, ACMG) und der Verband für molekulare Pathologie (engl. Association for Molecular Pathology, AMP) haben 2015 eine Aktualisierung der Richtlinien für die klinische Interpretation von Sequenzvarianten in Bezug auf Erkrankungen des Menschen anhand von 28 Kriterien veröffentlicht [Richards *et al.*, 2015]. Die Einteilung der Varianten beruht dabei auf der Klassifizierung in fünf verschiedene Kategorien (siehe Tabelle IV.18).

Die 28 Kriterien, anhand derer eine Variante eingeteilt wird, beruhen auf Informationen aus Populationsdatenbanken, funktionellen Daten, *in silico* Patho-

Tabelle IV.18: *Klassifizierung von Varianten*

Stufe	Bezeichnung	Erläuterung
1	gutartig	Normvariante ohne klinische Relevanz
2	wahrscheinlich gutartig	Wahrscheinlich eine Normvariante
3	Variante unklarer Signifikanz (VUS)	Keine Zuordnung der klinischen Signifikanz möglich
4	wahrscheinlich pathologisch	Variante könnte eine Erkrankung auslösen
5	pathologisch	Variante löst sicher eine Erkrankung aus

genitätsvorhersagen und Segregationsanalysen. Die ACMG-AMP Richtlinien schlagen in ihrer Veröffentlichung Bewertungsregeln vor, wie diese Daten zu kombinieren sind, um eine Variante einer Kategorie im fünfstufigen Klassifizierungssystem zuzuordnen. Da die Richtlinien keine speziellen Ressourcen zur Datengewinnung empfehlen, differieren die Klassifizierungsergebnisse je nach verwendeten Datenbanken und Algorithmen. Das von Quan Li und Kai Wang entwickelte Programm `InterVar` adressiert dieses Problem und hilft dem Anwender die klinische Bedeutung von Varianten einheitlich und reproduzierbar zu interpretieren. Dazu liest InterVar eine VCF-Datei ein, annotiert diese mit Hilfe der ANNOVAR Software [Wang *et al.*, 2010] und verwendet die daraus resultierenden Annotationen pro Variante zur automatisierten Interpretation und Klassifizierung nach 18 der 28 ACMG-AMP Kriterien. Für die 10 anderen Kriterien, wie zum Beispiel Informationen zur Segregation, ist eine automatisierte Erfassung nicht möglich. Sie können aber, sofern vorhanden, in einer Datei an das Programm übergeben werden [Li und Wang, 2017]. In dieser Arbeit wird InterVar in der Version 0.1.7 20170608 verwendet.

IV.XI H3M2

Längere Regionen mit homozygotem Genotyp in einem diploiden Genom werden als „Regions of Homozygosity“ (ROH) bezeichnet und erstrecken sich meist über eine Länge von dutzenden Kilobasen bis zu einigen Megabasen. Eine geographische Isolation, Selektionsvorteile oder kulturelle Gewohnheiten können zur Bildung einer ROH beitragen. Lange ROHs über mehrere Megabasen entstehen meist durch elterliche Verwandtschaft (Konsanguinität), weil die beiden Allele an einem Ort von demselben gemeinsamen Vorfahren stammen und damit übereinstimmen (engl. Identity by Descent, IDB). Es ist bekannt, dass ROHs, die aus IBD hervorgehen, möglicherweise hoch penetrante krankheitsverursachende rezessive Mutationen enthalten. Sie spielen also eine große Rolle in der Prädisposition von seltenen sowie häufigen rezessiven Erkrankungen.

Die Detektion von ROHs wird meist unter Verwendung von SNP-Mikroarrays

durchgeführt, bei denen der Abstand zwischen zwei SNPs etwa 3 kbp beträgt. Das 1000-Genom-Projekt identifizierte und genotypisierte etwa 38 Millionen SNPs mit einer durchschnittlichen Distanz von 73 bp im menschlichen Genom, wobei 98% dieser gefundenen SNPs eine Allelfrequenz von mehr als 1% aufweisen. Werden nur die exonischen Marker betrachtet, ergeben sich Distanzen von 1 bp bis 26 Mbp zwischen zwei benachbarten SNPs (Durchschnitt: 500bp). Dadurch könnten große ROHs durch nur einige wenige ungleichmäßig verteilte SNP-Marker abgedeckt sein. Kleine isolierte ROHs hingegen weisen eine hohe Markerdichte auf. Mit dem heterogenen versteckten Markow-Modell für Homozygotien (engl. homozygosity heterogeneous hidden Markow model, H^3M^2) stellen Magi *et al.* einen Algorithmus vor, welcher die Distanz zweier aufeinanderfolgender SNP-Marker in die Berechnung des heterogenen Markow-Modells einbezieht und dadurch ROHs jeder Größe mit hoher Spezifität und Sensitivität in den ungleichmäßig verteilten SNP-Markern von WES-Daten detektieren kann [Magi *et al.*, 2014].

Das Softwarepaket ist auf der Sourceforge Seite[5] des H^3M^2-Projekts herunterladbar. Es beinhaltet die beiden Shell Skripte `H3M2BamParsing.sh` und `H3M2Analyze.sh` sowie einen Datensatz mit Positionen zu exonischen SNPs aus dem 1000-Genom-Projekt. Da die beiden Programme `R-Skripte` ausführen, muss die `R` Umgebung installiert sein. Im ersten Schritt wird mit dem `H3M2BamParsing.sh`-Skript die Frequenz des nicht-Referenz-Allels, die sogenannte B-Allel-Frequenz (BAF), berechnet. Diese nutzt das Skript `H3M2Analyze.sh` im zweiten Schritt um ROHs zu detektieren. Beide Skripte werden in Version 1 mit den Standardparametern angewendet.

IV.XII Variant Effect Predictor

Ein typisches diploides menschliches Genom weist in Bezug auf die Genomreferenzsequenz etwa 3,5 Millionen Einzelnukleotidvarianten (SNVs) und 1000 Kopienzahlveränderungen auf. Etwa 20.000 – 25.000 dieser Varianten sind proteinkodierend, wobei 10.000 eine Aminosäure verändern. Nur 50 – 100 Varianten hingegen sind proteintrunkierend oder führen zu einem Verlust der Proteinfunktion [Gonzaga-Jauregui *et al.*, 2012]. Die manuelle Überprüfung einer so großen Anzahl an Varianten ist unpraktisch und kostet viel Zeit. Zudem stellen fehlende funktionale Informationen und die Interpretation mehrerer Varianten innerhalb eines Haplotyps weitere Schwierigkeiten dar.

Die Ensembl Variant Effect Predictor (VEP)-Software bietet Tools und Methoden für eine automatisierte Annotation und Priorisierung von Varianten, sowohl in großen als auch in kleinen Sequenzierungsprojekten. Die Annotation in einer standardisierten Art und Weise verkürzt den Zeitaufwand für die manuelle Überprüfung, wodurch viele der allgemeinen Herausforderungen, die mit der Analyse von Varianten verbunden sind, bewältigt werden können. Der Variant Effect Predictor unterstützt die Annotation von SNVs, Insertionen, De-

[5] `https://sourceforge.net/projects/h3m2/`

letionen, Multinukleotidaustauschen, Veränderungen in Tandem Repeats und Mikrosatelliten sowie größerer struktureller Variationen, wie zum Beispiel Kopienzahlveränderungen. Dabei liefert die Software für jede Variante detaillierte Informationen über ihren Effekt auf Transkript- und Proteinebene oder ihre Auswirkung auf regulatorische Regionen.

Zur Annotation der Transkriptinformationen wird für das menschliche Genom der GENCODE Gen Datensatz verwendet. Dieser stellt eine vollständige Zusammenfassung von ENSEMBLs evidenzbasierten Transkriptvorhersagen mit manuellen Annotationen dar und bildet somit den umfangreichsten Satz von Transkriptisoformen für diese Spezies. Da eine Variante mehrere verschiedene Transkripte oder regulatorische Regionen „überlappen“ kann, verwendet VEP Informationen über das Transcript Support Level [ENSEMBL, 2020b] und APPRIS [Rodriguez *et al.*, 2013], um nur eine Annotation pro variiertem Allel für das Haupttranskript auszugeben. Neben der Angabe der Veränderung auf Proteinebene in der HGVS Nomenklatur [Rodriguez *et al.*, 2013], liefert der Algorithmus Hinweise auf die Auswirkung der Aminosäureveränderung unter Verwendung der biophysikalischen Eigenschaften des Proteins. Dazu werden vorberechnete Werte von Pathogenitätsvorhersage-Tools, wie zum Beispiel SIFT [Vaser *et al.*, 2016], PolyPhen-2 [Adzhubei *et al.*, 2010] oder MutationTaster [Schwarz *et al.*, 2014] für jeden möglichen Aminosäureaustausch verwendet. Weitere Pathogenitätsvorhersage-Tools, wie zum Beispiel CADD [Kircher *et al.*, 2014] lassen sich über Plugins einbinden. Varianten in nicht kodierenden Bereichen können einen Effekt auf die Genregulation haben. Die Annotation der Variante mit einem Konservierungswerts wie GERP [Davydov *et al.*, 2010] hilft diese Auswirkungen besser einzuschätzen. Außerdem inkludiert VEP Überlappungen mit bekannten Varianten aus der dbSNP (siehe Anhang III.II), der HGMD (siehe Anhang III.I) und COSMIC [Tate *et al.*, 2019] Datenbank, gibt Allelfrequenzen bekannter Varianten aus dem 1000-Genom-Projekt [Auton *et al.*, 2015], dem Exome Sequencing Projekt [NHLBI, 2014] und dem gnomAD Projekt (siehe Anhang III.VI) an und fügt den bereits publizierten Varianten PubMed IDs der entsprechenden Publikationen hinzu. Die klinische Signifikanz aus der ClinVar Datenbank (siehe Anhang III.III) sowie mit dieser Variante assoziierte Erkrankungen werden mit einem OMIM (siehe Anhang III.IV) oder Orphanet (siehe Anhang III.V) Identifer annotiert.

Das Variant Effect Predictor Programm ist plattformunabhängig als Online-Tool, Perl Skript oder über den REST (engl. Representational State Transfer, REST) Webservice verfügbar. Da das lokal zu installierende und auszuführende Perl-Skript die leistungsfähigste sowie flexibelste Methode für die Nutzung des VEP darstellt, findet es in dieser Arbeit Anwendung. Das VEP-Skript liest Daten im VCF-Format (siehe Anhang II.IV) ein und gibt die Varianten mit ihren entsprechenden Annotationen im `INFO` Feld der VCF-Datei wieder als VCF-Datei aus. Werden die zur Annotation verwendeten Dateien, die sogenannten „Cache“-Dateien, auf den lokalen Server heruntergeladen und das Skript im „offline“ Modus betrieben, arbeitet es am effektivsten. Es ist zudem möglich

mit dem Befehl `-custom` eigene Datensätze in die Annotation einzubringen. Zur schnelleren Bearbeitung der Daten können die Berechnungen des VEP-Skript unter Angabe des ganzzahligen Werts für den `-t` Parameter auf mehrere Prozessorkerne verteilt werden.

Die VEP-Software sowie die Referenzdatensätze werden ebenso wie die ENSEMBL Datenbank (siehe Anhang III.VII) quartalsweise vier mal im Jahr aktualisiert. In dieser Arbeit wurden die Versionen 89 bis 100 mit den in Tabelle IV.19 gelisteten Parametern zur Datenanalyse verwendet.

Tabelle IV.19: `VEP` *Parameter*

Parameter	**Erläuterung**
`--offline`	VEP arbeitet im Offline-Modus
`--dir`	Pfad zum Verzeichnis mit Cache und Plugin Dateien
`--fasta`	Pfad zur Referenzsequenz im FASTA-Format
`--everything`	Schaltet die Annotation mit allen zur Verfügung stehenden Daten ein
`--assembly`	Gibt an welcher Genombuild verwendet werden soll
`-i`	Pfad zur Eingabedatei im VCF-Format
`-o`	Pfad zur Ausgabedatei
`--vcf`	Schreibt die Ausgabedatei als VCF-Datei
`--plugin`	Verwendet das angegebene Plugin (mehrfache Anwendung möglich)
`-custom`	Fügt den angegebenen eigenen Datensatz zur Annotation hinzu (mehrfache Anwendung möglich)
`--buffer_size`	Adjustiert den Arbeitsspeicherverbrauch
`--pick`	Gibt nur eine Auswirkung pro Variante aus
`--pick_order`	Gibt die Reihenfolge der Kriterien zur Auswahl der auszugebenden Auswirkung an
`--force_overwrite`	Gibt an, dass die Ausgabedatei überschrieben werden soll, falls bereits vorhanden

IV.XIII Primer3

Primer3 ist ein weit verbreitetes Programm zum Design von PCR-Primer, Hybridisierungssonden und Sequenzierungsprimer welches im Jahr 2000 erstmals erschienen ist [Rozen und Skaletsky, 2000] und 2012 durch ein neues thermodynamisches Modell unter anderem zur Vorhersage von Schmelz- und Bindungstemperaturen erweitert wurde [Untergasser *et al.*, 2012]. Das Programm wird als Web-Service[6] sowie als kommandozeilenbasierte Software, zur Integration in bioinformatische Pipelines von Untergasser *et al.* bereitgestellt. Das `primer3_core`-Modul ist das Hauptprogramm zur Suche von Primerpaaren für

[6] `http://primer3.ut.ee/`

eine Sequenz unter Berücksichtigung der vom Benutzer angegebenen Parameter. Durch weitere intern aufgerufene Module wird zum Beispiel die Schmelztemperatur, die Wahrscheinlichkeit für die Bildung eines Primerdimers oder der Ausbildung von Sekundärstrukturen eines Primers berechnet. Da mehr als 150 Parameter für den Aufruf der `primer3_core`-Software existieren, erfolgt das Einlesen dieser über eine Datei (`--p3_settings_file`). Die Datei enthält Parameter als Schlüssel-Wert-Paare im Format `Schlüssel=Wert`. In Abbildung IV.7 ist die in dieser Arbeit verwendete Konfigurationsdatei dargestellt. Tabelle IV.20 beschreibt die darin enthaltenen Parameter.

```
1  Primer3 File - http://primer3.org
2  P3_FILE_TYPE=settings
3
4  PRIMER_MAX_NS_ACCEPTED=0
5  PRIMER_TASK=generic
6  PRIMER_PICK_LEFT_PRIMER=1
7  PRIMER_PICK_INTERNAL_OLIGO=0
8  PRIMER_PICK_RIGHT_PRIMER=1
9  PRIMER_OPT_SIZE=20
10 PRIMER_MIN_SIZE=18
11 PRIMER_MAX_SIZE=22
12 PRIMER_PRODUCT_SIZE_RANGE=100-400
13 PRIMER_EXPLAIN_FLAG=1
14 =
```

Abbildung IV.7: *Primer3 Settings Datei.*
Die Datei enthält Parameter als Schlüssel-Wert-Paare im Format `Schlüssel=Wert`*. Durch diese Parameter kann zum Beispiel die optimale Primerlänge und die Produktgröße definiert werden.*

Eine Eingabedatei kann mehrere Datenblöcke umfassen. Sie sind durch eine Zeile, welche nur ein =-Zeichen beinhaltet, getrennt. Die im Datenblock enthaltenen Angaben liegen wie in der Konfigurationsdatei als Schlüssel-Wert-Paare im Format `Schlüssel=Wert` vor. In dieser Arbeit finden die Schlüssel `SEQUENCE_ID`, `SEQUENCE_TEMPLATE` und `SEQUENCE_TARGET` Anwendung. `SEQUENCE_ID` bezeichnet dabei einen eindeutigen Identifier für die unter `SEQUENCE_TEMPLATE` angegebene DNA-Sequenz. Der zu vervielfältigende Bereich wird durch `SEQUENCE_TARGET` aufgeführt. Ein Wert von `200,201` bedeutet zum Beispiel, dass ein durch das Primerpaar entstehendes Amplikon bei Base 200 der gegebenen Sequenz beginnen und 201bp lang sein soll.

Die Ausgabedatei ist analog der Eingabedatei aufgebaut und beinhaltet bis zu vier mögliche Primerpaare pro Eingabesequenz, deren Eigenschaften mit jeweils 29 Schlüssel-Wert-Paaren beschrieben sind. Die für diese Arbeit wichtigen Eigenschaften eines Primerpaars sind die Sequenzen des linken und rechten Primers (`PRIMER_LEFT_0_SEQUENCE`; `PRIMER_RIGHT_0_SEQUENCE`), deren Annealing Temperatur (`PRIMER_LEFT_0_TM`; `PRIMER_RIGHT_0_TM`) und prozentualer GC-Gehalt (`PRIMER_LEFT_0_GC_PERCENT`; `PRIMER_RIGHT_0_GC_PERCENT`) sowie die Länge des entstehenden PCR-Produktes (`PRIMER_PAIR_0_PRODUCT_SIZE`).

Tabelle IV.20: `primer3_core` *Argumente*

Argument	Erläuterung
`PRIMER_MAX_NS_ACCEPTED`	Maximale Anzahl erlaubter unbekannter Basen (N) im Primer
`PRIMER_TASK`	Beschreibt die Art der des Primerdesigns; z.B. *generic*: ein einfaches Primerpaar designen
`PRIMER_PICK_LEFT_PRIMER`	Wenn 1, dann sucht Primer3 nach linkem Primer
`PRIMER_PICK_INTERNAL_OLIGO`	Wenn 1, dann sucht Primer3 nach einer Hybridisierungssonde innerhalb des Amplikons
`PRIMER_PICK_RIGHT_PRIMER`	Wenn 1, dann sucht Primer3 nach rechtem Primer
`PRIMER_OPT_SIZE`	Optimale Länge eines Primers
`PRIMER_MIN_SIZE`	Maximal akzeptierte Länge eines Primers
`PRIMER_MAX_SIZE`	Minimal akzeptierte Länge eines Primers
`PRIMER_PRODUCT_SIZE_RANGE`	Gewünschte Länge des Amplikons. Angegeben im Format MIN-MAX
`PRIMER_EXPLAIN_FLAG`	Wenn 1, dann gibt das Programm zusätzliche Informationen über die Anzahl der getesteten Primer aus.

In dieser Arbeit wird die Version 2.5.0 des Programms `primer3_core` mit der in Abbildung IV.7 gezeigten Konfigurationsdatei und den in Tabelle IV.21 dargelegten Parametern gestartet.

Tabelle IV.21: `primer3_core` *Parameter*

Parameter	Erläuterung
`--p3_settings_file`	Pfad zur Konfigurationsdatei
`--output`	Pfad zur Ausgabedatei
`<<Eingabedatei>>`	Pfad zur Eingabedatei

IV.XIV RNAstructure

Das `RNAstructure`-Softwarepaket beinhaltet Programme zur Vorhersage und Analyse von RNA- und DNA-Sekundärstrukturen und wurde erstmals 1998 in der Literatur erwähnt [Mathews *et al.*, 1998]. Eine 2D-Struktur definiert sich als Resultat der Basenpaarungen (AU, GC und GU) innerhalb eines RNA-Moleküls. Zur Berechnung verwendet die Software thermodynamische Methoden und Parameter, welche auf den sogenannten „nearest neighbor rules“ [Mathews *et al.*, 2004] basieren, die ihrerseits die Stabilität einer Struktur vorhersagen. Dazu gehören sowohl Parameter für die Änderung der freien Energie

bei 37 °C als auch Parameter für die Enthalpieänderung, um die Konformationsstabilität bei einer beliebigen Temperatur vorhersagen zu können. Die wahrscheinlichste Struktur ist jene mit der geringsten freien Energie. Dabei ist die Vorhersage der Sekundärstruktur einer einzelnen Sequenz recht genau. Zum Beispiel können die 2D-Strukturen von Sequenzen mit weniger als 700 Basen mit einer durchschnittlichen Genauigkeit von 73% prognostiziert werden [Mathews *et al.*, 2004] [Reuter und Mathews, 2010].

Zur Berechnung der Sekundärstruktur einer einzelsträngigen RNA-Sequenz wird das Programm `Fold` des Softwarepakets angewendet. Die Ergebnisse im CT-Format sind mittels `ct2dot` zur weiteren Verarbeitung in die Dot-Bracket-Notation umzuwandeln, welche Basenpaarungen als zusammengehörige runde Klammern und ungepaarte Basen als Punkte darstellt (siehe Abbildung IV.8).

```
1 >Test
2 AUGCAGGCGCTGCGGGGGGGGGGCUGCGCCUGCAU
3 ((((((((((.((.........)).)))))))))).
```

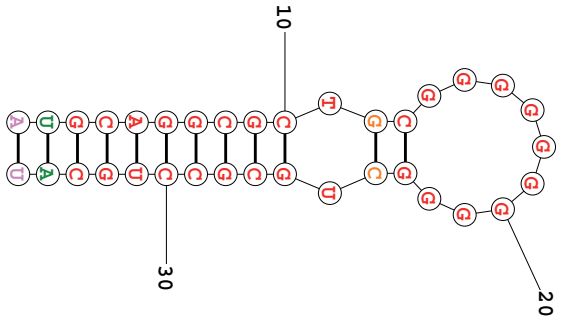


Abbildung IV.8: *Beispiel einer Dot-Bracket-Notation. Basenpaarungen werden als zusammengehörige runde Klammern und ungepaarte Basen als Punkte darstellt.*

Das Softwarepaket kann nach einer Registrierung kostenlos von der Homepage des Mathews lab[7] heruntergeladen werden. In dieser Arbeit findet Version 6.1 Anwendung. Die verwendeten Parameter der beiden Programme `Fold` und `ct2dot` sind in Tabelle IV.22 und Tabelle IV.23 erläutert.

Tabelle IV.22: `Fold` *Parameter*

Parameter	**Erläuterung**
`<<FASTA Datei>>`	Pfad zur Eingabedatei im FASTA-Format
`<<CT Datei>>`	Pfad zur Ausgabedatei im CT-Format
`-m`	Maximale Anzahl an vorhergesagten Strukturen

[7] `https://rna.urmc.rochester.edu/RNAstructure.html`

Tabelle IV.23: `ct2dot` *Parameter*

Parameter	Erläuterung
`<<CT Datei>>`	Pfad zur Eingabedatei im CT-Format
`<<Zahl>>`	Anzahl der zu konvertierenden Strukturen; -1 = alle
`<<DB Datei>>`	Pfad zur Ausgabedatei im Dot-Bracket-Format
`-f`	Ausgabeformat der Dot-Bracket-Notation 1 = nur Struktur 2 = Struktur + Bezeichnung an rechter Seite 3 = Struktur + Bezeichnung in Titelzeile 4 = Struktur + Bezeichnung in Titelzeile und Sequenz

IV.XV Vienna RNA

`ViennaRNA` ist wie `RNAstructure` (siehe Anhang IV.XIV) ebenfalls ein Software-Paket zur Vorhersage und Analyse von Sekundundärstrukturen einzelsträngiger RNA-Moleküle [Lorenz *et al.*, 2011]. Zudem kann mit dem im Paket enthaltenen Programm `RNAforester` ein Alignment zum Vergleich von 2D-Strukturen in der Dot-Bracket-Notation durchgeführt werden. Höchsmann *et al.* erweiterten den „tree alignment"-Algorithmus von Jiang *et al.* [Jiang *et al.*, 1995], um lokale gleichartige Bereiche in RNA-Sekundärstrukturen ausfindig zu machen. Aus dem Ergebnis dieses Alignments werden relative „Similarity Scores" errechnet, welche die Ähnlichkeit zwischen den beiden alignierten RNA-Strukturen beschreiben [Höchsmann *et al.*, 2003].

Das `ViennaRNA`-Softwarepaket ist auf der Homepage der Theoretical Biochemistry Group der Universität Wien[8] öffentlich downloadbar. In dieser Arbeit wurde das Programm `RNAforester` aus Version 2.4.11 des `ViennaRNA`-Pakets mit den in Tabelle IV.24 gelisteten Parametern verwendet.

Tabelle IV.24: `RNAforester` *Parameter*

Parameter	Erläuterung
`< <<DB Datei>>`	Pfad zur Eingabedatei mit 2 zu vergleichenden Strukturen im DB Format
`-r`	Berechnung eines relativen Similarity Scores
`> <<Alignment Datei>>`	Pfad zur Ausgabedatei

[8] `https://www.tbi.univie.ac.at/RNA/#download`

Anhang V

Laborarbeit

Die in der NGS gefundenen Varianten werden mit einer zweiten unabhängigen Methode, der Sanger-Sequenzierung, validiert. Um *in silico* vorhergesagte Veränderungen der Genexpression *in vivo* nachzuweisen, kann ein Luziferase-Assay durchgeführt werden.

V.I PCR und Sanger-Sequenzierung

Die Polymerase-Kettenreaktion (engl.: Polymerase Chain Reaction, PCR) dient der *in vitro* Vervielfältigung von DNA Regionen. Dazu wird die DNA zu Einzelsträngen denaturiert. Anschließend folgt die Zugabe von regionsspezifischen Primern und die Synthese des jeweils komplementären Strangs durch das Enzym DNA-Polymerase unter Verwendung von hinzugefügten Nukleotiden. Dieses Vorgehen wird in mehreren Zyklen wiederholt, wobei die Produkte vorheriger Zyklen als Ausgangsstoff für den nächsten Zyklus dienen, was wiederum eine exponentielle Vervielfältigung ermöglicht [Neumeister *et al.*, 2018].

Die Sanger-Sequenzierung ist eine im Jahr 1977 von Frederick Sanger vorgestellte Methode zur DNA-Sequenzierung. Sie basiert auf dem selektiven Einbau von Didesoxynukleotiden, die eine weitere Synthese durch die DNA-Polymerase während einer *in vitro* DNA-Replikation verhindern [Sanger *et al.*, 1977]. Anschließend erfolgt die Beschreibung der durchzuführenden Laborarbeiten.

Die Primer werden mit Hilfe der beiliegenden Anleitung suspendiert und im Verhältnis 1:10 verdünnt. Eine Amplifikation der genomischen Bereiche mittels PCR findet mit einem 25 μl (Mikroliter) PCR-Ansatz (siehe Tabelle V.1) und dem PCR-Programm „RYR“ (siehe Tabelle V.2) statt.

Durch die PCR wird die DNA an spezifischen *Loci* amplifiziert. Zur Kontrolle sind die entstandenen Produkte anschließend neben einem Längenstandard auf ein Agarosegel aufzutragen und elektrophoretisch aufzutrennen. Dazu erfolgt die Herstellung eines 2 %igen Agarosegels, indem pro 100 ml (Milliliter) TAE-Puffer 2 g (Gramm) Agarose eingewogen und 2 μl Ethidiumbromid untergemischt werden.

Tabelle V.1: *PCR-Ansatz zur Amplifikation der Genomregionen.*

Volumen [μl]	**Reagenz**
12,5	ReadyMix Taq PCR Reaction Mix mit $MgCl_2$
2	Primer Forward und Reverse
1	genomische DNA (100 ng/μl)
9,5	steriles H_2O

Tabelle V.2: *Cycler-Programm „RYR60“ zur Durchführung der PCR-Reaktion.*

Schritt	**Temperatur** [°C]	**Zeit** [min]
1	94	5
2	94	0,5
3	60	0,5
4	72	0,75
5	*gehe zu Schritt 2*	*34 mal*
6	72	5
7	4	∞

Zu je 4 μl PCR-Produkt sind 1 μl Orange G-Ladepuffer und 5 μl Wasser hinzuzugeben. Diese resultierenden Lösungen werden in die Taschen des erstarrten Gels gefüllt, an das eine Spannung von 130 Volt angelegt ist. Nach etwa 15 Minuten ist die Laufstrecke der PCR-Produkte ausreichend lang, um das Gel unter ultraviolettem Licht auszuwerten.

Liegen PCR Produkte vor, werden diese mit Hilfe des EXO/SAP-Verdaus aufgereinigt. Hierbei beseitigt die Exonuklease I (EXO I) die noch vorhandenen Primer, während die Shrimp Alkaline Phosphatase (SAP) die übrigen Desoxyribonukleosidtriphosphate entfernt. Der Ansatz und das zugehörige Inkubationsprogramm können Tabelle V.3 und V.4 entnommen werden.

Tabelle V.3: *Ansatz für den EXO/SAP Verdau.*
Dieser enzymatische Verdau dient der Aufreinigung der PCR-Produkte

Volumen [μl]	**Reagenz**
20	PCR-Produkt
0,225	EXO I (20 U/μl)
0,9	SAP (1 U/μl)
4,875	steriles H_2O

Die Sequenzierreaktion erfolgt mit dem CEQ DTCS Quick Start Sequencing Kit (Beckman Coulter), der verschiedene fluoreszenzmarkierte Didesoxyribonukleosidtriphosphate (ddNTPs) enthält. Die Grundlage dieser Reaktion ist hier das von Sanger beschriebene Prinzip des Kettenabbruchs durch den Ein-

Tabelle V.4: *Cycler-Programm „Clean Up“.*
Dieses Programm stellt die optimalen Bedingungen für den EXO/SAP-Verdau bereit und führt somit die Aufreinigung der PCR-Produkte durch.

Schritt	**Temperatur [°C]**	**Zeit [min]**
1	37	25
2	72	15
3	4	∞

Tabelle V.5: *SIQ-Beck Ansatz.*
PCR-Ansatz mit farbstoffmarkierten Nukleotiden für die anschließende Sequenzierung.

Volumen [μl]	**Reagenz**
0,5	10 μM M13-Primer Forward bzw. M13-Primer Reverse
2	CEQ DTCS Quick Start (Beckmann Coulter)
5,5	steriles H_2O
2	PCR-Produkt

bau von ddNTPs in die Sequenz während der PCR [Sanger *et al.*, 1977]. Der für die Sequenzierreaktion verwendete Ansatz (SIQ-Beck) und das PCR-Programm „SIQ-Beck“ sind in Tabelle V.5 und Tabelle V.6 beschrieben.

Anschließend wird das Produkt der Sequenzierreaktion mit Agencourt CleanSEQ magnetischen Beads (Beckman Coulter) mit dem Biomek3000 Roboter (Beckman Coulter) automatisch aufgereinigt. Die Analyse der Sequenzierreaktion erfolgt durch die Auftrennung mittels Kapillarpolyacrylamidgelelektrophorese im Sequenzierer CEQ^{TM} 8000 Genetic Analysis System (Beckman Coulter).

Tabelle V.6: *Cycler-Programm „SIQ-Beck“.*
PCR-Programm zur Erstellung und Vervielfältigung fluoreszierender PCR-Fragmente für die anschließende Sequenzierung.

Schritt	**Temperatur [°C]**	**Zeit [min]**
1	96	0,3
2	50	0,3
3	60	4
4	*gehe zu Schritt 1*	*34 mal*
5	4	∞

V.II Luziferase-Assay

Um zu testen, ob die im Patienten gefundene Variante des 3'UTRs einen Einfluss auf die Bindungsstelle von miRNAs hat, wird der mutierte sowie der wildtypische 3'UTR in jeweils einen Vektor kloniert. Dieser besitzt ein Luziferase-Gen stromaufwärts der 3'UTR Klonierungsstelle. Die klonierten Vektoren werden im Anschluss zusammen mit mikroRNA Mimics in HEK293 Zellen transfiziert und für ein Luziferase Assay verwendet.

V.II.I Oligonukleotid Design und Annealing

Zur Untersuchung der Variante c.*651C>T im *IL1RAPL1*-Gen (HG19 chrX:29974588C>T) werden 119bp lange Oligonukleotide für die wildtypische (WT) sowie mutierte (MU) Sequenz des 3'UTRs jeweils als Forward- und Reverse-Complement-Strang synthetisiert. Am jeweiligen 5‘ Ende des synthetisierten Oligos befindet sich das 3‘ Ende der Schnittstelle des XbaI Restriktionsenyzms (CTAGA) sowie am 3‘ Ende das 5‘ Ende der XbaI Schnittstelle (T). Zudem ist das 5' Ende des Oligonukleotids phosphoryliert (siehe Tabelle V.7). Die Oligonukleotide werden bei der Firma Sigma Aldrich bestellt.

Tabelle V.7: *Oligonukleotide für die Klonierung. gelb markiert: Sequenz der Restriktionsschnittstelle von XbaI; rot markiert: mutierte Base*

Wildtyp	Forward	5‘-[Phos] CTAGA CCACCTCATTCCCCACCTCTAC CTTTCTAATGGCGGCATGATGTGTAAACTCTGTG C A GGGGTGGGGGCGGGTCTAACTGTCTTAACATTCAAGT CACTGCTCTTCAGAATAC T -3‘
	Reverse-Complement	5‘-[Phos] CTAGA GTATTCTGAAGAGCAGTGACTT GAATGTTAAGACAGTTAGACCCGCCCCCACCCCT G C ACAGAGTTTACACATCATGCCGCCATTAGAAAGGTAG AGGTGGGGAATGAGGTGG T -3‘
Mutiert	Forward	5‘-[Phos] CTAGA CCACCTCATTCCCCACCTCTAC CTTTCTAATGGCGGCATGATGTGTAAACTCTGTG T A GGGGTGGGGGCGGGTCTAACTGTCTTAACATTCAAGT CACTGCTCTTCAGAATAC T -3‘
	Reverse-Complement	5‘-[Phos] CTAGA GTATTCTGAAGAGCAGTGACTT GAATGTTAAGACAGTTAGACCCGCCCCCACCCCT A C ACAGAGTTTACACATCATGCCGCCATTAGAAAGGTAG AGGTGGGGAATGAGGTGG T -3‘

Die jeweiligen Forward- und Reverse-Complement Oligonukleotide von wildtypischer und mutierter Sequenz hybridisieren im Annealing Schritt bei 95°C in 6 Minuten und in der anschließenden Abkühlphase von 30 Minuten bei Raumtemperatur in einem 10μl Ansatz (siehe Tabelle V.8) zu Doppelsträngen.

Tabelle V.8: *Annealing Ansatz.*

Menge	Wildtyp	Mutiert
$1\mu l$	WT_fwd	MT_fwd
$1\mu l$	WT_rev	MT_rev
$1\mu l$	Annealing Puffer	Annealing Puffer
$7\mu l$	H_2O	H_2O

V.II.II Restriktion und Ligation

Um die relative Genaktivität in einem Luziferase-Assay messen zu können, müssen die Oligonukleotide in einen entsprechenden Vektor, welcher ein Firefly Luziferase Gen stromaufwärts der Klonierungsstelle besitzt, eingefügt werden. Hierzu ist der pGL3-Promotor (pGL3P) Vektor und das Restriktionsenzym XbaI zu verwenden. Der $20\mu l$ Ansatz in Tabelle V.9 beschreibt die Linearisierung des Vektors in einem Thermocycler bei 37°C für insgesamt 2 Stunden. Nach einer Stunde erfolgt die Zugabe von $1\mu l$ Calf Intestine Phosphatase (CIP) zum Reaktionsansatz.

Tabelle V.9: *Restriktionsansatz zur Linearisierung des Vektors pGL3P*

Menge	Reagenz
$2\mu l$	pGL3P
$1\mu l$	XbaI
$2\mu l$	Cut Smart
$15\mu l$	H_2O
$1\mu l$	CIP (nach einer Stunde)

$4\mu l$ des Restriktionsproduktes werden mit $4\mu l$ H_2O und $2\mu l$ Ladepuffer auf ein 1,5%iges Agarosegel aufgetragen und bei 170V 20 Minuten aufgetrennt. War die Restriktion erfolgreich, sind die verbleibenden $16\mu l$ des Restriktionsproduktes mit Hilfe des DNA Aufreinigungs-Kits Nucleospin der Firma Macherey Nagel nach Herstellerprotokoll aufzureinigen. Die enstprechenden DNA Mengen zur Ligation des mutierten sowie wildtypischen doppelsträngigen Oligonukleotids in den linearisierten pGL3P Vektor werden mit Hilfe des Online-Tools „in silico ligation calculator“[1] bestimmt. Bei 100ng des 5010bp langen Vektors, einem 119bp langen Oligonukleotid und einer 1:6 (Vektor:Insert) Verteilung sind laut Online-Tool 14,25ng Insert und 100ng des Vektors zu verwenden. Die daraus resultierenden Ansätze für WT und MU ergeben sich wie in Tabelle V.10 dargestellt. Die Ligationsreaktion inkubiert bei 16°C über Nacht.

[1] `http://www.insilico.uni-duesseldorf.de/Lig_Input.html`

Tabelle V.10: *Ansätze zur Ligation des Vektors mit dem jeweiligen Insert.*

Reagenz	**Wildtyp Ansatz**	**Mutiert Ansatz**
T4 DNA Ligase	$1\mu l$	$1\mu l$
T4 10x Puffer	$2\mu l$	$2\mu l$
Vektor linearisiert pGL3P	$2{,}4\mu l$	$2{,}4\mu l$
WT Insert (1:100)	$1{,}89\mu l$	-
MU Insert (1:100)	-	$1{,}68\mu l$
H_2O	$12{,}71\mu l$	$12{,}92\mu l$

V.II.III Transformation

Zur Multiplizierung der Plasmide (pGL3P Vektor + Insert) werden die Ligationsprodukte für WT und MU in kompetente *E.coli* Bakterien transformiert. Dafür ist das komplette ($50\mu l$) Ligationsprodukt auf 1ml aufgetaute *E.coli* Bakterien zu gegeben und für 30 Minuten auf Eis zu inkubieren. Damit die Bakterien möglichst viel des Ligationsprodukts aufnehmen, findet die sogenannte „heat shock“ Methode Anwendung. Dabei wird der *E.coli*-Plasmid-Mix für exakt 90 Sekunden in ein 42°C heißes Wasserbad gestellt und anschließend direkt wieder für 2 Minuten auf Eis inkubiert. Zum Abschluss erfolgt die Zugabe von $250\mu l$ LB-Medium (engl. lysogeny broth) ohne Ampicillin zu jeder der beiden (WT + MU) *E.coli* Ansätze. Nach schüttelndem Inkubieren für 30 Minuten bei 37°C werden die Zellen in Petrischalen mit LB-Medium inklusive Ampicillin ausgestrichen und über Nacht bei 37°C inkubiert.

Nur jene Bakterien, bei denen eine erfolgreiche Transformation stattgefunden hat und die das richtige Plasmid mit dem Gen für Ampicillin Resistenz beinhalten, können auf diesem Medium wachsen.

V.II.IV Isolierung, Restriktion und Sequenzierung des Plasmids

Um zu überprüfen, ob die Plasmide das jeweilige Insert (WT oder MU) in korrekter Richtung und ohne Sequenzveränderung aufgenommen haben, werden von jeder Petrischale 12 Bakterienkolonien gepickt. Jede gepickte Kolonie ist in einem separaten 12ml Gefäß mit 3ml LB-Medium inklusive Ampicillin bei 37°C über Nacht schüttelnd zu inkubieren.

Zur Isolierung der Plasmid-DNA wird jeweils 1ml aus jeder 3ml *E.coli* Zellkultur in ein eigenes 1,5ml Eppendorfgefäß überführt. Eine 5 minütige Zentrifugation bei 6000 Umdrehungen pro Minute (engl. Rounds per Minute, rpm) ist notwendig, um anschließend vorsichtig den Überstand abzunehmen. Darauffolgend wird das Zellpellet in $100\mu l$ Resuspendierungspuffer P1 des Qiagen EndoFree Plasmid Maxi-Kits resuspendiert. Nach Zugabe von $100\mu l$ Lysepuffer P2 (Qiagen EndoFree Plasmid Maxi-Kit) und der Inkubation der Lösung für 5 Minuten bei Raumtemperatur, wird die Lysierungsreaktion durch Zugabe von $100\mu l$ gekühltem Puffer P3 (Qiagen EndoFree Plasmid Maxi-Kit) ge-

stoppt. Dieser P3 Puffer sorgt zudem für die Präzipitation der Plasmid-DNA. Nach kurzem Vortexen folgt eine weitere Inkubationsphase von 5 Minuten bei Raumtemperatur. Die Zentrifugation der Ansätze für 10 Minuten bei 13500 rpm entfernt die Zellfragmente. Der Überstand, welcher nun die Plasmid-DNA enthält, wird jeweils auf 1ml 100%igen Ethanol pipettiert, um die DNA ausfallen zu lassen. Hierzu ist es notwendig die Ansätze wiederrum 10 Minuten bei 13500 rpm zu zentrifugieren. Im Anschluss erfolgt ein Waschschritt mit 150μl 70%igem Ethanol und ein Zentrifugationsschritt bei 13500 rpm für 5 Minuten. Zur Gewinnung der DNA wird der Überstand entfernt, die Plasmid-DNA Pellets bei 50°C getrocknet und anschließend in 50μl Millipore Wasser resuspendiert.

Für 5μl jedes Ansatzes aus der MiniPrep ist ein Restriktionsverdau durch das Enzym XbaI (siehe Tabelle V.11) anzusetzten, um auf einem Agarosegel zu prüfen, ob das Insert die richtige Größe aufweist. Ist dies der Fall, so wird ein Reaktionsansatz (siehe Tabelle V.12) mit Primer RV4-rev (5‘-GACGATAGTCATGCCCC
GCG-3‘) hergestellt und zur Firma StarSeq[2] zur Sequenzierung geschickt. Anhand dieser lässt sich feststellen, ob das Insert die korrekte Orientierung und keine Sequenzfehler aufweist.

Tabelle V.11: *Restriktionsansatz 2 zur Linearisierung des Vektors pGL3P*

Reagenz	**Menge**
Plasmid DNA aus MiniPrep	5μl
XbaI	1μl
CutSmart	1μl
H_2O	3μl

Tabelle V.12: *Ansatz zur Sequenzierung des Inserts*

Reagenz	**Menge**
Plasmid DNA aus MiniPrep	1μl
Primer (RV4-rev; 10nm)	1μl
H_2O	5μl

V.II.V MaxiPrep

Für jeweils eine Kolonie jeder Bedingung (WT + MU) mit richtigem Insert erfolgt die Vorbereitung für eine MaxiPrep, indem die restlichen 2ml der Kultur in 20ml LB-Medium inklusive Ampicillin über Nacht bei 37°C schüttelnd kultiviert werden.

[2] `https://www.starseq.com/`

Das Qiagen EndoFree Plasmid Maxi Kit dient zur Extraktion der Plasmid-DNA. Der erste Schritt besteht aus der Zentrifugation der beiden 200ml Kulturen bei 6000 x g für 15 Minuten bei 4°C in einer Avanti J-26 XP-Zentrifuge. Anschließend wird der Überstand verworfen und die Zellpellets in 10ml Puffer P1 resuspendiert. Nach Zugabe von 10ml P2 Puffer und gründlichem Mischen, ist der Ansatz 5 Minuten bei Raumtemperatur zu inkubieren. Anschließend werden 10ml gekühlter P3 Puffer auf den Lyseansatz pipettiert, die Ansätze durch vorsichtiges Invertieren der Tubes 4-6-mal gemischt, auf eine QIAfilter-Kartusche gegeben und für 10 Minuten bei Raumtemperatur inkubiert. Durch Entfernen des Verschlusses an der Kartusche und unter Verwendung eines Spritzenkolbens ist das Lysat durch den Filter in ein 50ml Flacon zu überführen. Auf dieses Filtrat werden nun 2,5ml ER-Puffer gegeben, bevor es vorsichtig durch zehnfaches Invertieren gemischt und darauffolgend 30 Minuten auf Eis inkubiert wird. In der Zwischenzeit erfolgt die Vorbereitung von QIAGEN tips, indem 10ml QBT-Puffer die QIAGEN-Säulen passiert. Das Zelllysat fließt nach der Inkubationszeit durch die vorbereiteten QIAGEN-Säulen. Der Durchfluss ist zu verwerfen und die Säulen zwei Mal mit jeweils 30ml QC-Puffer zu waschen. Anschließend wird die DNA mit 15ml QN Puffer eluiert und durch Zugabe von 10,5ml Isopropanol präzipitiert. Es folgt ein Zentrifugationsschritt der Ansätze bei 5000 x g und 4°C für 90 Minuten. Der Überstand ist danach zu verwerfen. Im letzten Schritt wird das DNA Pellet mit 5ml endotoxinfreiem 70%igem Ethanol gewaschen und wiederrum bei 5000 x g 30 Minuten zentrifugiert. Nach erneutem vorsichtigen Verwerfen des Überstandes ist es notwendig das resultierende Pellet für 5-10 Minuten luftzutrocknen, um den restlichen Ethanol zu entfernen und das Pellet anschließend in 250μl TE Puffer zu lösen.

V.II.VI Zellkultur

HEK293-Zellen sind unter sterilen Zellkulturbedingungen zu halten. Alle Arbeiten mit lebenden Zellen erfolgen in einer sterilen Werkbank. Die HEK293 Zelllinie wird in einem Medium aus Dulbecco's Modified Eagle Medium (DMEM) mit hohem Glukoseanteil (4500mg/L), 1% Penicillin/Streptomycin und 10% fetalem Rinder Serum in einem CO_2-Inkubator bei 37°C und einem CO_2-Gehalt von 5% kultiviert.

V.II.VII Transfektion

Um im Anschluss an die Transfektion die Luziferaseaktivität messen zu können, werden jeweils 75.000 HEK293 Zellen in 12 well Platten ausgesäht.

Einen Tag nach dem Aussähen der Zellen ist die Transfektion der beiden 3'UTR Konstrukte und der unterschiedlichen miRNAs mittels Lipofectamine2000 der Firma Thermo Fischer durchzuführen. Da alle Kombinationen der beiden 3'UTRs (WT und MU) mit den 4 miRNAs und einer siRNA als Negativkontrolle in Tripletts hergestellt werden (siehe Tabelle V.13), erfolgt die Berechnung des Ansatzes für 30 Proben. Hierzu wird ein Mastermix mit 225μl

Opti-MEM sowie 8μl Lipofectamine pro Triplett, also insgesamt 2,25ml Opti-MEM und 80μl Lipofectamine, in einem 15ml Falcon angesetzt. In 10 weiteren 1,5ml Eppendorfgefäßen (eins pro Tripplet) werden 225μl Opti-MEM mit 3750ng der jeweiligen Plasmid-DNA, 15μl des jeweiligen 20µM miRNA Stocks und 150ng pRL angesetzt. Anschließend ist zu jedem der 10 Eppendorfgefäße 230μl des Mastermixes, bestehend aus Opti-MEM und Lipofectamine, hinzuzugegeben und für 5 Minuten bei Raumtemperatur zu inkubieren. Nach der Inkubationszeit erfolgt die Überführung von jeweils 150μl jedes Triplett Ansatzes in ein Well der ausgesähten Zellen, bevor die Zellen erneut im CO_2-Inkubator bei 37°C inkubiert werden. 24 Stunden später erfolgt die Entfernung des Transfektionsreagenz durch Waschen der Zellen mit Dulbeccos phosphatgepufferter Salzlösung (engl. Dulbecco's Phosphate-Buffered Saline, DPBS). Danach werden die Zellen mit neuem Kulturmedium versorgt. Nach weiteren 24 Stunden ist es schließlich möglich diese zu ernten.

Tabelle V.13: *Schema des Triplettansatzes für das Luziferaseassay*

	WT			MU		
	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3
Mir-4672						
Mir-5195-5p						
Mir-6801-3p						
Mir-6810-3p						
NC						

V.II.VIII Luziferase Assay

Luziferase Experimente dienen zur Untersuchung von posttranskriptionellen Effekten der Genexpression. Hierzu wird die Biolumineszenz des sogenannten Firefly Luziferase-Enzyms genutzt welches sich aus der meistverbreiteten Leuchtkäferart in Nordamerika, *Photinus pyralis*, extrahieren lässt. Zur Analyse der posttranskriptionalen Genregulation durch miRNAs ist es notwendig ein Teil des betreffenden 3'UTRs des zu untersuchenden Gens stromabwärts des Luziferase-Gens in einen Expressionsvektor (z.B. pGL3P) zu klonen. Dieser wird anschließend gemeinsam mit der entsprechenden miRNA in Zellen (z.B. HEK293) kotransfiziert. Diese Zellen überexprimieren den Vektor sowie die miRNA. 48 Stunden nach der Transfektion werden die Zellen lysiert und das Lysat mit einem Luminometer analysiert, um die enzymatische Aktivität der Luziferase zu messen. Durch den direkten Zusammenhang des Firefly Luziferase-Gens mit dem zu untersuchenden 3'UTR ist es möglich die gemessene Aktivität unmittelbar mit der Genexpression zu korrelieren.

Nach dem Waschen der transfizierten Zellen mit DPBS wird 150μl 1x Lysis Puffer des Promega Renilla Luciferase Assay System Kits hinzugegeben und dieser Ansatz schüttelnd für 15 Minuten bei Raumtemperatur inkubiert. Anschließend erfolgt die Resuspendierung der Zellen und das Überführen dieser

in PCR Reaktionsgefäße. Zum Messen der Luziferaseaktivität werden die zuvor lysierten Zellen in eine luminometerkonforme 96-Well Platte pipettiert. Das Gerät detektiert die Luziferaseaktivität durch Hinzufügen eines speziellen Puffers und Substrats zu jeder entsprechenden Probe auf der 96-Well Platte. Der verwendete Renilla-Puffer ist Teil des Promega Renilla Luziferase Assay System Kits. Der Firefly Puffer wird im Labor selbst hergestellt (siehe Tabelle V.14).

Tabelle V.14: *Ansatz für den Firefly Puffer*

Reagenz
25mM Tris-P pH 7.8
10mM MgSO4
2mM ATP pH 7.5
50µM Luziferin
For a 10 ml mix add 6,6 ml H_2O

Vor dem Start der Messung ist es notwendig die beiden Injektoren des Luminometers mit H_2O zu waschen und mit dem Firefly beziehungsweise dem Renilla Puffer zu befüllen. Für alle Proben werden erst die Firefly und anschließend die Renillaaktivität gemessen.

Die resultierenden Firefly Aktivitätswerte sind durch den Renilla Aktivitätswert zu normalisieren. So lassen sich Unterschiede in der Transfektionseffizienz ausgleichen. Dabei gilt die Annahme, dass die Transfektionseffizienz für das Renillasowie Firefly-Konstrukt gleich ist. So werden die Firefly Werte durch die Renilla Werte dividiert, um normalisierte relative Luziferaseaktivitätswerte zu erhalten.

Publikationen und Kongressteilnahmen

Publikationen

Schröder J.C., Läßig A.K., Galetzka D., Peters A., Castle J.C., **Diederich S.**, Zechner U., Müller-Forell W., Keilmann A. und Bartsch O. (2013). *„A boy with homozygous microdeletion of NEUROG1 presents with a congenital cranialdysinnervation disorder [Moebius syndromevariant]“*. Behavioral and Brain Functions 9(7):2-9.

Reuter M.S., Musante L., Hu H., **Diederich S.**, Sticht H., Ekici A. B., Uebe S., Wienker T.F., Bartsch O., Zechner U., Oppitz C., Keleman K., Jamra A. R., Najmabadi H., Schweiger S., Reis A. und Kahrizi K. (2014). *„NDST1 missense mutations in autosomal recessive intellectual disability“*. American Journal of Medical Genetics Part A 164A(11):2753-63.

Komlosi K., **Diederich S.**, Fend-Guella D.L., Bartsch O., Winter J., Zechner U., Beck M., Meyer P. und Schweiger S. (2018) *„Targeted next-generation sequencing analysis in couples at increased risk for autosomal recessive disorders“*. Orphanet Journal of Rare Diseases. 13(1):23.

Fend-Guella D.L., von Kopylow K., Spiess A.N., Schulze W., Salzbrunn A., **Diederich S.**, El Hajj N., Haaf T., Zechner U. und Linke M. (2019) *„The DNA methylation profile of human spermatogonia at single-cell- and single-allele-resolution refutes its role in spermatogonial stem cell function and germ cell differentiation“*. Molecular Human Reproduction. 25(6):283-294.

Arash-Kaps L., Komlosi K., Seegräber M., **Diederich S.**, Paschke E., Amraoui Y., Beblo S., Dieckmann A., Smitka M. und Hennermann JB. *„The Clinical and Molecular Spectrum of GM1 Gangliosidosis“*. The Journal of Pediatrics. 215:152-157

Kongressteilnahmen

25. Jahrestagung der Deutschen Gesellschaft für Humangenetik in Essen
Poster: *„A novel homozygous truncating mutation in the ALS2 gene leading to juvenile primary lateral sclerosis“*. Diederich S., Poarangan C., Hao H., Ropers H.H., Zechner U., Schweiger S.
Abstract veröffentlicht in Medizinische Genetik-Berlin, Band 26, Heft 1, Seite 132 (2014)

26. Jahrestagung der Deutschen Gesellschaft für Humangenetik in Graz
Poster: *„Targeted next-generation sequencing of 1222 genes in routine diagnostics of patients with intellectual disability“*. Diederich S., Komlósi K., Fend-Guella D.L., Bartsch O., Jacob S., Hao H., Wienker T.F., Ropers H.H., Zechner U., Schweiger S.
Abstract veröffentlicht in Medizinische Genetik-Berlin, Band 27, Heft 1, Seite 179 (2015)

27. Jahrestagung der Deutschen Gesellschaft für Humangenetik in Lübeck
Poster: *„Targeted next-generation sequencing in search for monogeniccauses of intellectual disability in children“*. Diederich S.,Komlosi K., Fend-Guella DL., Bartsch O.,Hao H., Wienker TF., Ropers HH., Zechner U., Schweiger S.
Abstract veröffentlicht in Medizinische Genetik-Berlin, Band 28, Heft 1, Seite 147 (2016)

28. Jahrestagung der Deutschen Gesellschaft für Humangenetik in Bochum
Vortrag: *„Three years of experience with targeted next-generation sequencing of developmental delay“*. Diederich S., Komlósi K., Fend-Guella DL., Bartsch O., Hao H., Wienker TF., Ropers HH., Zechner U., Schweiger S.
Abstract veröffentlicht in Medizinische Genetik-Berlin, Band 29, Heft 1, Seite 105 (2017)

29. Jahrestagung der Deutschen Gesellschaft für Humangenetik in Münster
Poster: *„Comparison of whole exome sequencing to targeted next-generation sequencing in patients with developmental and neuromuscular disorders“*. Diederich S., Komlósi K., Bartsch O., Wienker TF., Ropers HH., Winter J., Zechner U., Schweiger S.
Abstract veröffentlicht in Medizinische Genetik-Berlin, Band 30, Heft 1, Seite 155 (2018)

30. Jahrestagung der Deutschen Gesellschaft für Humangenetik in Weimar
Poster: *„Our experience with whole exome next-generation sequencing“*. Die-

derich S., Dewi S., Selig M., Winter J., Bartsch O., Schweiger S.
Abstract veröffentlicht in Medizinische Genetik-Berlin, Band 31, Heft 1, Seite 118 (2019)

Selbstständigkeitserklärung

Hiermit erkläre ich, Stefan Diederich (geboren am 03.06.1989 in Adenau), die hier vorliegende Arbeit selbstständig und ohne unerlaubte Hilfe angefertigt zu haben und alle verwendeten Hilfsmittel und Inhalte aus anderen Quellen als solche kenntlich gemacht zu haben. Zudem versichere ich, dass die vorliegende Arbeit noch an keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat.

Mir ist der Inhalt der Promotionsordnung bekannt.