Masterarbeit

Genomische Ansätze zur Identifizierung seltener autosomal dominanter/rezessiver Krankheitsgene

eingereicht von

Christian Grabner MSc

zur Erlangung des akademischen Grades

Master of Science

(MSc)

an der

Medizinischen Universität Graz

ausgeführt am

Institut für Humangenetik

unter der Anleitung von Betreuer/in

Assoz.-Prof. Mag. Dr. Christian Windpassinger

Sankt Johann im Pongau, 12. August 2019

Eidesstattliche Erklärung

Ich erkläre ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst habe, andere als die angegebenen Quellen nicht verwendet habe und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe

Sankt Johann im Pongau am 12.08.2019 *Christian Grabner*

Inhaltsverzeichnis

I. Abkürzungen

ASIC	application-spezific integrated circuit
ATAC	assay for Transposon-accessible chromatin
ATP	Adenosintriphosphat
BAC	bacterial artificial chromosome
BWA	Burrows-Wheeler Alignment
CCD	charged coupled device
CCS	circular consenus sequence
cDNA	complementary DNA
ChIP-Seq	Chomatin-Immunoprecipitation
CLR	continuous long read
CNV	copy number variation
Contigs	contiguous stretches
DAPI	4',6'-Diamidin-2-phenylindol
DEG	differential expressed genes
dNTP	Desoxyribonukleosidtriphopshat
ddNT	Didesoxyribonukleosidtriposphat
DNA	deoxic ribonucleic acid
E.coli	Escherichia coli
ELAND	efficiant large-scale alignment of nucleotide databases
emPCR	emulsion polymerase chain reaction
FFPE	formalin-fixed paraffin-embedded
Indel	insertion-deletion
ISFET	ion sensitive field effect transistor
NGS	next generation sequencing
MAQ	mapping and assembly with quality
MASP1	mannan-binding lectin serine protease 1
MeDIP	methylated immunoprecipitation
MLL2	histone-lysine N-methyltransferase
MRE	methylation sensitive restriction enzyme digestion
mRNA	messenger RNA
miRNA	microRNA
NHLBI	National Lung Heart and Blood Instute
lncRNA	long noncoding RNA
PACP1	artificial chromosome
PCR	polymerase chain reaction
pRNA	packaging RNA
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
rRNA	ribosomal RNA
SBR	single base resolution
SGS	second generation sequencing
SMRT	single molecule real-time
SNP	single nucleotide polymorphism

SNV	single nucleotide variant
STS	sequence tagged sites
ssDNA	single stranded DNA
TF	transcription factor
TGS	third generation sequencing
TOPMed	Trans-Omics for Precision Medicine
VCF	variant call format
WES	whole exome sequencing
WGS	whole genome sequencing
WGBS	whole genome bisulfite sequencing
YAC	yeast artificial chromosome
ZMW	zero-mode wave guide

II. Datenbanken

OMIM® - Online Mendelian Inheritance in Men
HGMD® - The Human Gene Mutation Database
dbSNP - Single Nucleotide Polymorphism Database
ExAC - Exome Aggregation Consortium
GEO - Gene Expression Omnibus Database
CADD - Combined Annotation Depletion
GERP - Genomic Evolutionary Rate Profiling
gnomAD - Genome Aggregation Database

III. Abbildungsverzeichnis

IV. Zusammenfassung

Die Anzahl an seltenen monogenen Erkrankungen wird derzeit auf >5000 geschätzt, wobei bei der Hälfte der molekulare Mechanismus noch immer nicht bekannt ist. Der traditionelle Ansatz zur Identifizierung von Krankheitsgenen erfolgte bisher über die Auswahl von Genen aufgrund von Ähnlichkeiten mit vergleichbaren Krankheiten, der möglichen Relevanz für die Physiologie einer Krankheit oder durch Bestimmung der chromosomalen Position durch verschiedene Genkartierungsansätze.

Die Sanger-Sequenzierung bildet dabei den Abschluss dieser zweischrittigen Analyse. Aufgrund der Begrenzung bezüglich der Readlänge und des Probendurchsatzes ist die Sanger-Sequenzierung sehr aufwendig und kostspielig und so für eine Hochdurchsatzsequenzierung nicht geeignet.

Die Entwicklung von hocheffizienten Sequenziertechnologien erlaubt die Parallelisierung der Sequenzierreaktion und die Erhöhung der Gesamtzahl an Reads, welche zu einer massiven Reduktion der Sequenzierkosten führt. Seit der Einführung des Roche 454 GS FLX haben verschiedene Hersteller Plattformen auf den Markt gebracht, die für eine massive parallele Sequenzierung geeignet sind, deren zugrundeliegende Technologie auf unterschiedlichen Detektionsmechanismen basiert, die je nach Applikation verschiedene Vor-und Nachteile aufweisen.

Diesen als Second Generation Sequencing bezeichneten Technologien ist gemein, dass sie für die Detektion des Signals eine Amplifikation durch die Polymerasekettenreaktion benötigen, welche zu Basensequenzfehler und andern Bias führen können.

Die Einführung von Third Generation Sequencing-Plattformen erlaubt die Sequenzierung einzelner DNA-Moleküle auf Basis unterschiedlicher zugrundeliegender Technologien unter Entstehung sehr langer Reads, wobei die Sequenzierung durch eine biologische, synthetische oder Hybridnanopore eine vielversprechende Alternative für die Zukunft darstellt. Mit der NGS-Technologie sind verschiedenste spezifische Applikationen zur Untersuchung von Genom, Transkriptom und Epigenom möglich, deren wahre Herausforderung die bioinformatische Analyse der großen Menge an produzierten Daten sein wird.

Gene Panels und Whole Exome Sequencing, für das es einige Strategien zur Genfindung gibt, haben sich als Standardmethode für die Identifizierung von Krankheitsge-

nen etabliert, werden aber in Zukunft vom Whole Genome Sequencing abgelöst werden, da hier keine Amplifikation durch Polymerasekettenreaktion erforderlich ist und auch nicht-codierende Regionen untersucht werden können.

Die Herausforderung wird in Zukunft nicht mehr die Entdeckung sondern Interpretation von Varianten sein und für das Verständnis von Genotyp-Phänotyp-Korrelation wird es nötig sein, die Daten aus allen unterschiedlichen Disziplinen wie genetische Sequenzen, Expressionsprofile, Methylierungs- sowie Proteinmengendaten zu kombinieren und daraus globale Profile zu erstellen, welche in Form von Geninteraktionsnetzwerke auf Basis der Systembiologie neue Einblicke in zelluläre Mechanismen unter normalen und abnormalen Bedingungen, wie Krebs oder immunologische Störungen geben werden.

V. Abstract

The amount of rare monogenic diseases is estimated to be >5000 and for the half of these the molecular mechanism is still unknown. The traditionel approach for the identification of disease-causing genes took place on the selection of genes due to similarities of related diseases, the possible relevance for the physiology of a disease or the determination of the chromosomal position by various mapping approaches. Sanger sequencing provide the conclusion of this two-step approach.

Because of limitations in terms of read length Sanger sequencing is highly costly and expensive and therefore not suitable for high-throughput sequencing. The development of high efficient sequencing technologies allows the parallelization of the sequencing reaction and the increase in total number of reads, which leads to a massive reduction of sequencing costs.

Since the introduction of the Roche 454 GS FLX, various manufacterer have introduced platforms onto the market, which are suitable for massively parallel sequencing, whose underlying technology is based on different detection mechanims, which show depending on application various advantages and disadvantages. These technologies referred to as second generation sequencing have in common, that they need amplification via polymerase chain reaction for the detection of the signal, which can lead to base sequence errors and other biases.

The introduction of third generation sequencing platforms allow the sequencing of single dna molecules on the basis of various underlying technologies under develoment of very long reads, whereby sequencing through a biological, synthetic or hybrid nanopore constitute a promising alternative in the future.

With the next generation technology divers specific applications are possible for examination of the genome, transcriptome and epigenome, whose true challenge the bioinformatic analysis of the large amounts of produced data is.

Gene panels and whole exome sequencing, for which several strategies for gene finding exist, has established as standard methods for the identification of disease-causing genes, but are replaced by whole genome sequencing in the future, since here no amplification through polymerase chain reaction is necessary and also non-coding regions can be examined. The challenge in the future will no longer be the

discovery but rather the interpretation of variants and for the understanding of the genotype-phenotype correlation it will be necessary to combine all data of various disciplines such as genetic sequences, expression profiles, methylation and protein abundance data and to generate global profiles, which in the form of gene interaction networks on basis of system biology will give new insights in cellular mechanisms under normal and anormal conditions like cancer or immunological disorders.

1. Einleitung

Vor mehr als hundert Jahren hat Gregor Mendel entdeckt, dass vererbte Merkmale von zellulären Einheiten kontrolliert werden, die später als Gene bekannt sind. In den letzten Jahren hat sich das Verständnis über diese Gene außerordentlich, durch das Wissen über die Molekularbiologie der DNA, gesteigert. Es ist inzwischen möglich, eine genaue Beschreibung von Genen und der DNA zu erhalten, seit es durch die Entwicklung von Techniken möglich ist, durch Mapping (Zuordnung) die Gene in der DNA zu kartieren und danach jede DNA-Einheit, als Nukleotid bekannt, zu sequenzieren (National Research Council, 1988).

Diese Kartierungsansätze weisen jedoch Anwendungsgrenzen auf, da es schwierig oder unmöglich ist vorherzusagen, ob eine Krankheit durch eine einzelne Nukleotid-Mutation oder strukturelle genomische Variationen verursacht wird oder ob sie ohne Information zur Familie dominant oder rezessiv vererbt wird. Des Weiteren reduzieren diese Kartierungsansätze häufig die Kandidatengene nicht genügend für die nachfolgende Sanger-Sequenzierung, wenn der Krankheitslocus zu groß bleibt (Gilissen et al., 2012) und dies sich als limitierender Faktor auf die Sequenzierkosten auswirkt. Die genetische Heterogenität, die Tatsache, dass ein ähnlicher Krankheitsphänotyp durch unterschiedliche mutierte Gene oder Allele verursacht werden kann, macht die Korrelation zwischen Genotyp und Phänotyp so schwierig (Kayman Kurekci and Dincer, 2015).

Die Anzahl an seltenen monogenen Erkrankungen wird derzeit auf >5000 geschätzt, wobei bei der Hälfte die zugrundeliegenden Gene nicht bekannt sind (Stand 2012). Die Identifizierung von Genen, die für diese Krankheiten verantwortlich sind, erlauben eine molekulare Diagnose bei Patienten, die Untersuchung von Anlageträgern sowie pränatale Testungen. Die Genidentifizierung bildet dabei den ersten Schritt zum besseren Verständnis der physiologischen Rolle des vorliegenden Proteins im Krankheitsverlaufs und ermöglicht als Startpunkt die Entwicklung von therapeutischen Behandlungen. Die NGS-Technologie ist im Moment dabei die medizinische Genomforschung zu revolutionieren und erlaubt, ohne vorhergehende Priorisierung von Kandidatengenen, umfangreiche Applikationen in der Erforschung und Diagnostik von Mendelschen Krankheiten (Gilissen et al., 2012).

2. Genfindung in der Prä-NGS-Ära mittels Genkartierung und Sequenzierung

Die traditionelle Identifizierung von Krankheitsgenen beruht auf der Sanger-Sequenzierung von Kandidatengenen. Die Auswahl dieser Kandidatengene erfolgt (1) aufgrund von Ähnlichkeiten mit Genen die mit vergleichbaren Krankheiten assoziiert sind, (2) weil die vorhergesagte Proteinfunktion eine Relevanz für die Physiologie der Krankheit zu besitzen scheint oder (3) weil Mapping-Ansätze auf diese Gene in einer genomischen Region hinweisen (Gilissen et al., 2012). Diese letztgenannten Ansätze basieren auf der Identifikation der chromosomalen Position des Gens das wahrscheinlich für die Krankheit verantwortlich ist. Dazu wird die chromosomale Region des Kandidatengens so eng wie möglich begrenzt und die Gene in dieser Region nach Mutationen gescreent. Durch genomweite Analysen von molekularen Markern wie zum Beispiel Einzelnukleotid-Polymorphismen (SNPs), deren genaue chromosomale Position für jeden einzelnen SNP im humanen Genom definiert ist, werden sie dazu verwendet Karten zu erzeugen, die das gesamte Genom abdecken. Der Genkartierungsansatz für Einzelgenerkrankungen ist die Kopplungsanalyse (Linkage Analysis). Die Kopplungsanalyse basiert auf der Berechnung der Wahrscheinlichkeit, dass ein mutiertes Krankheitsallel zusammen mit verschiedenen genetischen Markern, auf Basis genetischer Informationen aus Familienstammbäumen, vererbt wird und das durch die Verwendung der Positionen dieser genetischen Marker gekoppelte oder zusammenhängende Loci detektiert werden können (Kayman Kurekci and Dincer, 2015).

2.1. Genetische Kopplungskartierung

Kopplungskarten beschreiben das Arrangement von DNA-Markern und Genen auf der Basis ihrer Vererbungsmuster. Gene die dazu neigen gemeinsam vererbt zu werden, liegen auf derartigen Karten nahe beieinander und jene die unabhängig voneinander vererbt werden sind entfernt liegend. Gene auf demselben Chromosom können fest, locker oder ungekoppelt sein, wiedergespiegelt in der Wahrscheinlichkeit mit der sie während der Meiose im Zuge der Gametogenese getrennt werden. Durch sogenanntes Crossing-Over können Gene getrennt werden, indem es, mithilfe eines

Chromosomenbruchs, zum Austausch von Teilen mit dem anderen Mitglied des Chromosomenpaares kommt. Je weiter zwei Gene auf einem Chromosom auseinander liegen, desto höher ist die Frequenz, mit welcher so ein Austausch zwischen ihnen stattfindet. Um den Grad des Austausches zwischen zwei Genen zu messen, muss die Frequenz der gemeinsamen Vererbung von parentalen Allel-Kombinationen aus statistisch signifikanten Proben gemessen werden. Vom praktischen Standpunkt aus gesehen benötigt man für die Detektion einer Kopplung, die Messung von Allel-Kombinationen, weitergegeben von einer Generation zur nächsten, von mindestens 10 Spermien-und Eizellen. Die Kartierungseinheit der Distanz wird dabei in centiMorgan (cM) angegeben. Laut Definition haben zwei Gene oder Loci, die 1 cM voneinander entfernt sind, eine 1% Wahrscheinlichkeit durch einen Austausch während der Spermien- und Eiproduktion voneinander getrennt zu werden. Auf das gesamte Genom betrachtet entsprechen 1 cM auf der „Genetic Linkage Map“ ungefähr 1 Million Nukleotidpaaren. (National Research Council, 1988).

Es gibt eine Vielzahl an polymorphen Markern (RFLP, SNPs, Mikrosatelliten), wobei der Restriktionsfragmentlängenpolymorphismus (RFLP) der erste in Verwendung war. RFLPs stellen Regionen von Sequenzen da, in der sich die Allele in der Verteilung bestimmter Restriktionsstellen unterscheiden. Durch Verdau der genomischen DNA mit Restriktionsenzymen und anschließender Elektrophorese können die RFLPs detektiert werden. Durch Southern-Blot-Hybridisierung mit radioaktiv-markierter komplementärer DNA zur Region von Interesse erscheinen unterschiedliche Allele als Banden unterschiedlicher Größe (Dear, 2001).

2.2. Physikalische Genkartierung

Physikalische Genkarten legen die Distanzen zwischen Messpunkten entlang des Chromosoms fest. Im Idealfall werden die Distanzen in Nukleotide gemessen, sodass die Karte eine direkte Beschreibung eines DNA-Moleküls bereitstellt. Die wichtigsten Messpunkte von physikalischen Genkarten sind die Spaltstellen von Restriktionsenzymen (National Research Council, 1988).

Physikalische Genkarten beinhalten das Finden von einer zusammenhängenden Reihe (Contigs) von geklonten DNA-Fragmenten, die überlappende Abschnitte des

Genoms beinhalten. Die Überlappungen definieren dabei die Positionen der Klone relativ zueinander. Wenn zumindest einige der Klone Marker beinhalten, die unabhängig gemapped worden sind, dann ist die Position des gesamten Contigs im Genom bekannt. Der Ausgangspunkt für die physikalische Genkartierung ist das Erstellen einer Library von geklonten genomischen Fragmenten entweder durch mechanische Fragmentierung oder partiellem Restriktionsverdau. Die Fragmente werden danach üblicherweise in bakterielle Wirte wie *E.coli* durch Bakteriophagen, Plasmide, Cosmide oder andere Vektorsysteme transformiert. Für große Genome stehen auch PACs, BACs oder YACs zur Verfügung. Wenn die Library etabliert ist, können die überlappenden Klone mithilfe verschiedener Ansätze identifiziert werden. Entweder durch Screening auf Sequence Tagged Sites (STS) mittels PCR oder durch Restriktionsverdau mit einem oder mehreren Enzymen und anschließender Elektrophorese. Klone, die mehrere Fragmentgrößen gemeinsam haben, müssen überlappende Anteile des Genoms darstellen (Dear, 2001). Die physikalische Genkarte, als Vorgänger der Sequenzierung bezeichnet, ist ein Hybrid aus „Restriction Map“ und „Contig Map“. Restriktionskarten zeigen die Reihenfolge und Abstände zwischen Spaltstellen von standortspezifischen Restriktionsendonukleasen. „Contig Maps“ repräsentieren die Struktur von angrenzenden Regionen des Genoms durch Bestimmung der Anordnungsbeziehung unter einem Set von Klonen und sind abhängig von der kontinuierlichen Existenz einer bestimmten Klonsammlung. Die Herstellung von reinen „Restriction Maps“ ist schwierig, da die Restriktionsstellen für die geeigneten Enzyme nichtzufällig verteilt und manchmal durch Methylierungs-Systeme blockiert sind, welche die DNA in vivo kovalent modifizieren. Zusätzlich besteht für die Anwender nicht immer ein leichter Zugang zu DNA-Klonen. Reine „Contig Maps“ sind schwierig zu erstellen, da die Kontinuität an jedem Punkt verloren gehen kann, wo Klone nicht verfügbar oder die Anordnungsbeziehungen unklar sind. Hochrechnungen vergangener Experimente deuten darauf hin, dass eine „Contig Map“ eines menschlichen Chromosoms mit einer durchschnittlichen Größe wahrscheinlich 200 bis 1000 Lücken aufweist (Olson et al., 1989).

2.3. Fluoreszenz-in-situ-Hybridisierung

Im Gegenzug dazu stellt die Fluoreszenz-in-situ-Hybridisierung eine direkte und elegante Methode dar um physikalische Arrangements von Markern entlang des Chromosoms zu beobachten. Die Proben, normalerweise genomische DNA oder cDNA werden durch Biotin oder Digoxygenin via Nick-Translation gelabelt und können nach Denaturierung zu ihren korrespondierenden Sequenzen an den Metaphasechromosomen hybridisieren. Zur Detektion der Proben werden die Slides mit fluoreszenzmarkierten Proteinen (Avidin oder Antidigoxygenin-Antikörper) inkubiert, die an des Hapten binden und die Probe so für die Fluoreszenzmikroskopie sichtbar machen. Die Chromosomen selbst werden mittels DAPI (4',6'-Diamidin-2-phenylindol) gegengefärbt, einem anderen Fluorophor, um die Proben zu detektieren zu können. Die resultierenden Bilder zeigen deutlich die Lokalisation der gebundenen Probe, typischerweise zwei benachbarte Punkte zu den zwei korrespondierenden Chromatiden (Dear, 2001).

2.4. First-Generation Sequencing

Der größte Durchbruch, der den Fortschritt in der Technologie der DNA-Sequenzierung in 1977 für immer verändert hat, war die Entwicklung von Sanger's Kettenabbruch- oder Didesoxy-Methode. Das Mischen von radioaktivmarkierten chemischen Analoga (ddNTPs) der Desoxyribonukleotide anteilsmäßig in die DNA-Verlängerungsreaktion von Standard-dNTPs resultiert in DNA-Stränge jeder Größe, da die Didesoxynukleotide zufällig beim Verlängern des Stranges eingebaut werden und so die Reaktion anhalten. Dieser Kettenabbruch entsteht, da die ddNTPs aufgrund der fehlenden 3'Hydroxylgruppe, die für die Verlängerung der DNA-Kette benötigt wird, keine Bindung mit dem 5' Phosphat des nächsten dNTPs eingehen können. Nach der Ausführung von 4 parallelen Reaktionen, jede mit einer individuellen ddNTP-Base, und dem Auftragen auf ein Polyacrylamidgel, kann man auf die Nukleotidsequenz, nach Sichtbarmachen der Banden, in der Originalvorlage rückschließen (Heather and Chain, 2016).

Die ursprüngliche Verwendung von Autoradiographie zum Sichtbarmachen der Banden mit Phospho- oder Tritrium-Radiomarkierung wurde in den darauffolgenden Jah-

ren infolge von Verbesserungen der Methode durch eine fluorometrische Detektion ersetzt (Heather and Chain, 2016).
Die zweite wichtige und richtungsgebende Methode aus dieser Zeit ist die, von Allan Maxam und Walter Gilbert, entwickelte basenspezifische chemische Spaltung. Hierbei muss die DNA zuerst durch radioaktives P^{32} an der 5'Phosphatgruppe markiert werden. Durch unterschiedliche chemische Behandlungen werden danach selektiv Basen von DNA-Stellen entfernt. Hydrazin entfernt Basen von Pyrimidinen (Cytosin und Thymin) und in hoher Salzkonzentration nur von Cytosin. Die Basen von Purinen (Adenin und Guanin) können durch Säure entfernt werden und Guanin allein mithilfe von Dimethylsulfat. Das Phosphodiester-Rückgrat wird danach mit Piperidin gespalten und die erhaltenen Fragmente werden danach wie bei der Sanger Sequenzierung mittels Polyacrylamidgel visualisiert (Heather and Chain, 2016).

Die Sanger-Sequenzierung wurde über die Jahre weiterentwickelt und hat so der Methode von Maxam und Gilbert den Rang abgelaufen. Der Einsatz der Kapillarelektrophorese zur verbesserten Auftrennung der DNA anstelle von Gelplatten hat zu einer höheren Zuverlässigkeit der Base Calls geführt. Die Anwendung von Fluoreszenz- anstatt radioaktivmarkierter Terminatoren hat die Detektion der Sequenz robuster und das Handling im Labor sicherer gemacht. Durch die Möglichkeit der Automatisierung ist die menschliche Beteiligung reduziert und die Effizienz erhöht worden. Durch diese Weiterentwicklungen einzelner Bereiche der Sanger-Sequenzierung wurde diese für das Human Genom Project und für die weiteren Jahrzehnte als Methode der Wahl angenommen und ist auch heute noch der Goldstandard der DNA-Sequenzierung (Wang, 2016, S.55).

Aufgrund der Sequenzierlänge von maximal einer Kilobase (Kb) und der limitierten Anzahl der Kapillaren kann die Sanger-Methode keinen hohen Durchsatz erreichen, der aber benötigt wird um die Sequenzierkosten zu senken. Das Sequencing-by-Synthesis Prinzip wurde daher die Basis für mehrere Next Generation Sequencing Technologien, die alle durch einen hohen Durchsatz gekennzeichnet sind. Die Verwendung von Nukleotiden mit reversiblen Terminatoren oder anderen spaltbaren

chemischen Modifikationen oder regulären nicht modifizierten Nukleotiden führt dazu, dass der neu synthetisierte Strang nicht dauerhaft beendet ist sondern überwacht werden kann, wenn oder nachdem eine Base eingebaut wurde. Diese Entwicklung gemeinsam mit weiteren Verbesserungen in anderen relevanten Bereichen macht es möglich die Sequenzierung von Millionen DNA-Fragmente gleichzeitig auszuführen. Mit dem Aufkommen des Next Generation Sequencing hat sich für die Sanger-Sequenzierung das First Generation Sequencing als Synonym etabliert (Wang, 2016, S. 55-57). Das Sequenzieren auf in dieser Zeit entwickelten neuen Plattformen wird auch als Second Generation Sequencing bezeichnet.

3. Next (Second) Generation Sequencing

Die Fähigkeit, zügig und genau Wissen über die DNA-Zusammensetzung zu erlangen ist für viele biologische Wissenschaften essentiell. In einer Zeit, wo man sich in Richtung synthetisches Genom hin bewegt verändern hocheffiziente Sequenziertechnologien den wissenschaftlichen Horizont und versprechen eine Ära der personalisierten Medizin für eine erhöhte menschliche Gesundheit (Pettersson et al., 2009). Ein beeindruckender Fortschritt wurde auf dem Gebiet des Next Generation Sequencing durch die Weiterentwicklung in den Bereichen der Molekularbiologie und des technischen Ingenieurwesens gemacht. Durch die Parallelisierung der Sequenzierreaktion kann die Gesamtzahl der produzierten Reads pro Lauf gesteigert werden und dadurch die Kosten massiv gesenkt werden. Schrittweise geht es in die Richtung der Anwendung von Next Generation Sequencing Technologien als Standardmethode, denn obwohl Replikation, Transkription, Translation, Methylierung oder Kernfaltung komplett unterschiedliche Prozesse darstellen, können sie alle durch die Nutzung der Sequenzierung studiert werden. Aktuelle Plattformen erlauben einen mühelosen Einblick in komplexe Mischungen aus RNA- und DNA-Proben (Buermans and den Dunnen, 2014) und durch die Miniaturisierung und Parallelisierung der Plattformen, die zu einem geringeren Sample- und Template-Verbrauch und zu einer gesteigerten Geschwindigkeit führen, bieten sie einen signifikant höheren Durchsatz im Vergleich zur Sanger-Methode. Diese State-of-the-Art Systeme faszinieren durch die Möglich-

keit, immer größere Bereiche des humanen Genoms resequenzieren zu können und dadurch phänotypische Varianten zu erklären und so das Verständnis für Krankheitsanfälligkeit und Pharmakogenomik zu erweitern (Pettersson et al., 2009).

3.1. Sequence Library Preparation

Die unterschiedlichen Plattformen benötigen eine Präprozessierung der genomischen DNA oder RNA-Transkripte in eine Bibliothek (Library) passend für das darauffolgende Sequenzieren. Die DNA oder RNA wird dabei in Plattform-spezifische Größenbereiche fragmentiert gefolgt von Modifikationen der Enden um Fragmente mit Blunt-Ends oder Überhänge zu erhalten und Adapter anhängen zu können. Die Fragmentierung kann durch verschiedene Techniken wie Sonifikation, akustisches Zerkleinern oder enzymatische Behandlung erfolgen (Quail et al., 2012; Wang, 2016, S. 58). Jede Sequenzierplattform verwendet ein unterschiedliches Set an einzigartigen Adaptersequenzen die für eine funktionierende Library an das 3' und 5' Ende angehängt werden, um mit den weiteren Schritten des Sequenzierprozesses fortführen zu können. Für manche Technologien ist ein Verfahren namens Nick Translation notwendig um funktionale Moleküle zu bekommen, für andere ist die Probe unmittelbar nach der Ligation fertig zum Laden. Abhängig vom System können diese Libraries direkt zum Sequenzieren verwendet werden oder benötigen vorher noch einen Prä-Amplifikationsschritt (Buermans and den Dunnen, 2014).

3.2. Pyrosequenzierung - Roche 454 GS FLX

Der Genome Sequencer 454 FLX von Roche verwendet eine sogenannte Emulsions-PCR in einer Öl-Wasser-Emulsion für die klonale Amplifikation, gefolgt von einer parallelen und individuellen Pyrosequenzierung der klonal amplifizierten Kügelchen (Pettersson et al., 2009). Die hergestellten Libraries werden zunächst mithilfe der Adaptersequenzen an Kügelchen (Beads) gebunden und in einer Wasser-in-Öl-Emulsion-PCR (emPCR), wo idealerweise ein DNA-Fragment an ein Kügelchen gebunden hat, im eigenen Emulsionströpfchen amplifiziert. Durch dieses Set Up können Reads mit einer Länge von ungefähr 400-500 Basenpaaren produziert werden. Diese Methode unterscheidet sich deutlich von den vorher existenten Methoden wie die Sanger-Sequenzierung und markiert die erste Welle an Second Generation Se-

quencern. Die Visualisierung der Nukleotide erfolgt nicht über radio- oder fluoreszenzmarkierte dNTPs mittels (Kapillar)-Elektrophorese, sondern durch eine lumineszierte Methode durch Messung der Pyrophosphat-Synthese in einem Prozess mit zwei Enzymen. Die ATP-Sulfurylase wird dazu verwendet um Pyrophosphat in ATP zu konvertieren. Danach wird das ATP als Substrat für die Luciferase verwendet, die abhängig von der Menge an Pyrophosphat direkt proportional Licht produziert (Heather and Chain, 2016).

Die Emission der Photonen beim Einbau der Nukleotide durch die DNA-Polymerase setzt eine Diphosphatgruppe (PPi) frei welche durch die oben erwähnte ATP-Sulfurylase unter Anwendung von Adenosinphosphosulfat in ATP katalysiert wird. Das Enzym Luciferase gemeinsam mit D-Luciferin und Sauerstoff kann das neu produzierte ATP dazu verwenden um Licht zu emittieren. Die Degradierung von nicht eingebautem dNTPs sowie der Abbau von ATP wird durch die Apyrase durchgeführt. Diese Enzyme sind auf kleineren Kügelchen immobilisiert, welche die großen Amplikon-tragenden Kügelchen umgeben. Alle weiteren Reagenzien inklusive der 4 Nukleotide werden bereitgestellt und fließen zu den Kügelchen mit den Templates die sich in einer Picotiterplatte befinden, wobei in jedes Well nur ein Kügelchen passt. Im Idealfall produzieren die Polymerase und ein dNTP ein Photon pro Zyklus, welches mit Hilfe einer CCD (Charged Couple Device) Kamera detektiert wird. Danach wird die Apyrase durchgespült um die überschüssigen Nukleotide abzubauen (Pettersson et al., 2009).

Die Pyrosequenzierung wird auch als Sequencing-by-Synthesis genannt, da sie eine DNA-Polymerase benötigt, um diesen Output zu produzieren. Die Vorteile der Pyrosequenzierungs-Technik gegenüber der Sanger-Methode sind die Verwendung von natürlichen Nukleotiden anstelle der stark modifizierten dNTPs , die den Kettenabruch initialisieren, die Beobachtung des Prozesses in real-time anstatt langwieriger Elektrophorese, das mögliche Anfügen der DNA an paramagnetische Kügelchen sowie die Möglichkeit des enzymatischen Abbaus von nicht eingebauten dNTPs und die dadurch nicht mehr notwendigen Waschschritte (Heather and Chain, 2016).

3.3. Sequenzierung durch reversible Termination - Illumina/Solexa

Illumina Geräte dominieren den High Throughput Sequencing (HTS) Markt seit der Einführung des Genome Analyzer II in 2006. Der Sequenzierprozess erfolgt nicht auf Basis einer Emulsions-PCR wie beim 454 GS FLX von Roche, sondern auf einer klonalen Amplifikation, auch „Bridge Amplification“ genannt, von adapter-ligierten DNA-Fragmenten auf der Oberfläche einer Flow Cell (Reuter et al., 2015). Zu Beginn der Bridge Amplification werden Forward- und Reverse-Oligonukleotide (mit einer spaltbaren Stelle), die komplementär zu den Adaptersequenzen, welche während der Herstellung der Sequence Library eingesetzt worden sind, an die innere Oberfläche der Spuren der Flow Cell gebunden. Je nach spezifischer Illumina-Plattform kann die Flow Cell in 1 (MiSeq), 2 (HiSeq 2500), oder 8 (HiSeq 2000, HiSeq 2500) getrennte Spuren (Lanes) unterteilt werden. Die Denaturierung der doppelsträngigen DNA-Fragmente in einzelsträngige Moleküle ist der erste Schritt um die Library auf die Flow Cell zu laden. Auf dieser hybridisieren die DNA-Fragmente mit den zuvor an der Oberfläche immobilisierten Oligonukleotide, die als Primer fungieren, und so wird von jedem individuellen Template-Molekül eine Kopie erstellt (Buermans and den Dunnen, 2014).

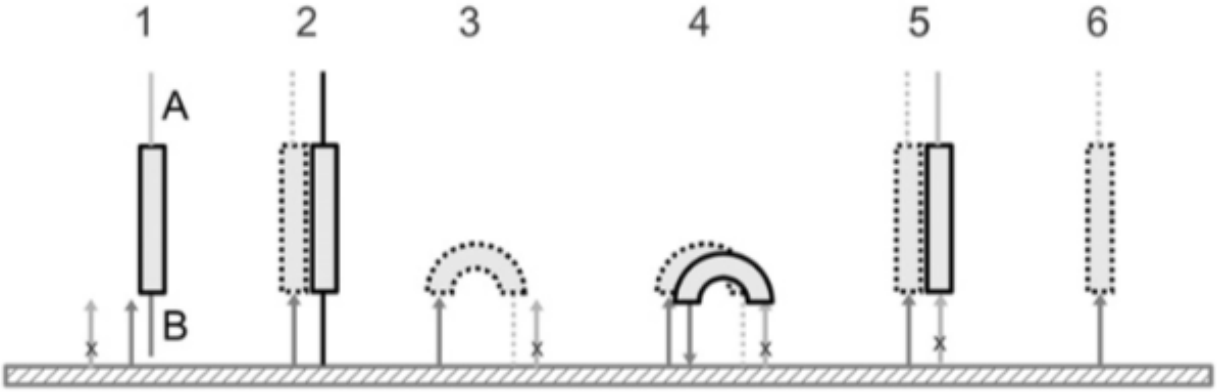

Abbildung 1: Ablauf der einzelnen Schritte der Bridge Amplification. (1) Bindung der einzelsträngigen DNA-Fragmente an die an der Flow Cell immobilisierten Primer. (2) Bildung einer Kopie des Templates, Entfernen des ursprünglichen Library-Moleküls. (3-5) Ausbildung einer Brückenstruktur und Bildung von identen Clustermolekülen durch isothermale Amplifikation (Buermans and den Dunnen, 2014).

Nach Entfernung der ursprünglichen Library-Moleküle bilden die an die Flow Cell gebundenen kopierten DNA-Fragmente durch isothermale Amplifikation lokale Cluster von identischen Template-Molekülen. Durch cyclische Abwechslung von Denaturierung, Annealing und Extension bei 60°C können die 3'Enden der kopierten Library-Moleküle mit den komplementären Oligos auf der Flow Cell hybridisieren, eine brückenartige Struktur ausbilden wodurch für jedes gebundene Fragment eine Kopie erstellt wird. Als finaler Schritt wird ein Strang des doppelsträngigen DNA-Fragments durch die spaltbare Stelle im Oligonukleotid entfernt und das 3'Ende mit ddNTP geblockt um zu verhindern, dass das offene 3'Ende als Sequenzier-Primerstelle für benachbarte Library-Moleküle dient (Buermans and den Dunnen, 2014).

Die danach stattfindende Sequenzierung erfolgt auf Illumina-Plattformen durch reversible Termination. Nach der klonalen Amplifikation der DNA-Fragmente mittels Bridge-PCR wird die Sequenzierung durch Nutzung von reversiblen Terminatornukleotiden (RT-Nucleotides) durchgeführt. Durch den Zusatz dieser Nukleotide, die fluoreszenzmarkiert und an der 3'OH-Gruppe durch 2-Cyano-Ethyl geschützt sind, zur Flowcell, kann die DNA-Polymerase diese modifizierten Nukleotide einbauen und den neuen DNA-Strang synthetisieren (Ambardar et al., 2016).

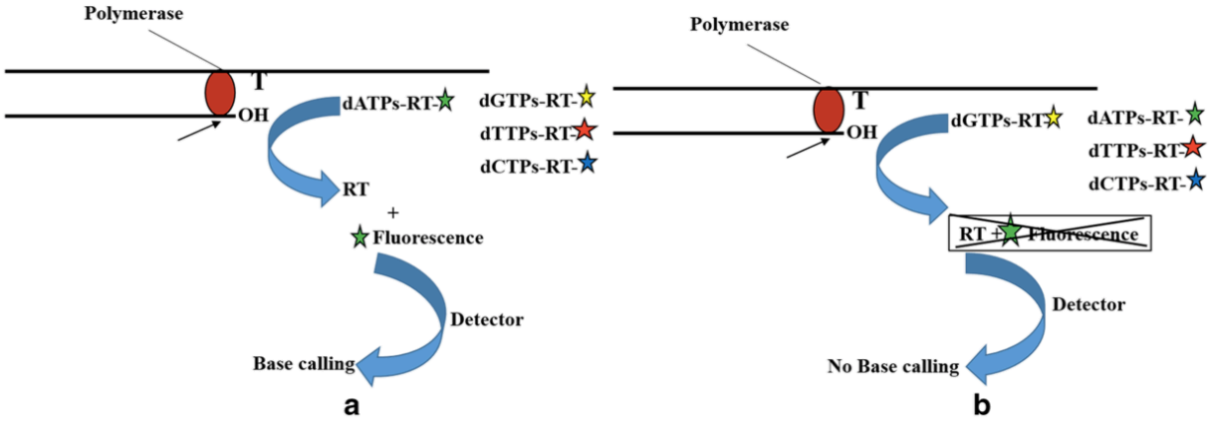

Abbildung 2: Schematische Darstellung der Sequenzierung durch reversible Termination (a) während des Einbaus eines komplementären Nukleotids (b) wenn Nukleotid nicht eingebaut wird (Ambardar et al., 2016).

Dieser Ablauf beinhaltet drei aufeinanderfolgende Schritte, (1) Einbau der komplementären RT-Nukleotide durch eine mutierte DNA-Polymerase zu dem an der Flowcell gebundenen DNA-Strang, (2) Detektion der unterschiedlichen Fluoreszenzen für

die 4 Basen und (3) Wiederherstellung der freien 3'OH-Gruppe durch Spaltung der funktionellen Gruppe und des Reporter-Moleküls. Durch die Repetition dieses Zyklus wird der vorliegende Template-Strang nach und nach sequenziert. RT-Nukleotide werden eingebaut und abgebildet gefolgt von der Entfernung der Fluorophore und Aktivierung der terminierenden Basen durch De-Protektion der freien 3'OH-Gruppe, welche eine neue Runde des Nukleotideinbaus erlaubt. Der Vorteil gegenüber der Pyrosequenzierung besteht in der Bias-Reduktion durch die Präsenz aller vier Nukleotide anstelle eines Typs in jedem Zyklus und die Vermeidung von Homopolymerfehler, da für jeden Einbau einer neuen Base erst der reversible Terminator abgespalten werden muss (Ambardar et al., 2016).

3.4. Sequenzierung durch Detektion von H^+-Ionen – Ion Torrent (Thermofisher)

Der Ion PGM (Personal Genome Machine) Sequencer von Ion Torrent wurde Ende 2010 mit der Halbleitersequenziertechnologie auf den Markt gebracht und ist wie der MiSeq von Illumina ein in seiner Größe kleines Gerät mit schnellen Turn-Over Raten und einer ursprünglichen Readlänge von 100bp für klinische Applikationen (Liu et al., 2012). Diese von Jonathan Rothenburg nach dem Verlassen von 454 Life Sciences entwickelte sogenannte „Post-Light" Sequenziertechnologie verwendet weder Fluoreszenz noch Lumineszenz zur (optischen) Detektion des Einbaus von Nukleotiden, sondern die Änderung im pH-Wert durch das Freisetzen von Protonen. Analog zur Pyrosequenzierung erfolgt die Amplifikation der DNA-Fragmente durch eine Emulsions-PCR (emPCR), die danach über eine Picotiter-Platte gespült werden, gefolgt von den Nukleotiden (Heather and Chain, 2016). Beim Einbau eines Nukleotids durch die DNA-Polymerase wird ein Proton frei, welches zu einer Änderung im pH-Wert führt und so vom Ion PGM erkannt wird, ob eine Base eingebaut wurde oder nicht. Die Nukleotide werden nacheinander auf den Chip gespült, wobei eine Spannung nur beim Einbau des korrekten Nukleotids erfolgt und beim Hinzufügen zweier doppelt so groß ist. (Liu et al., 2012).

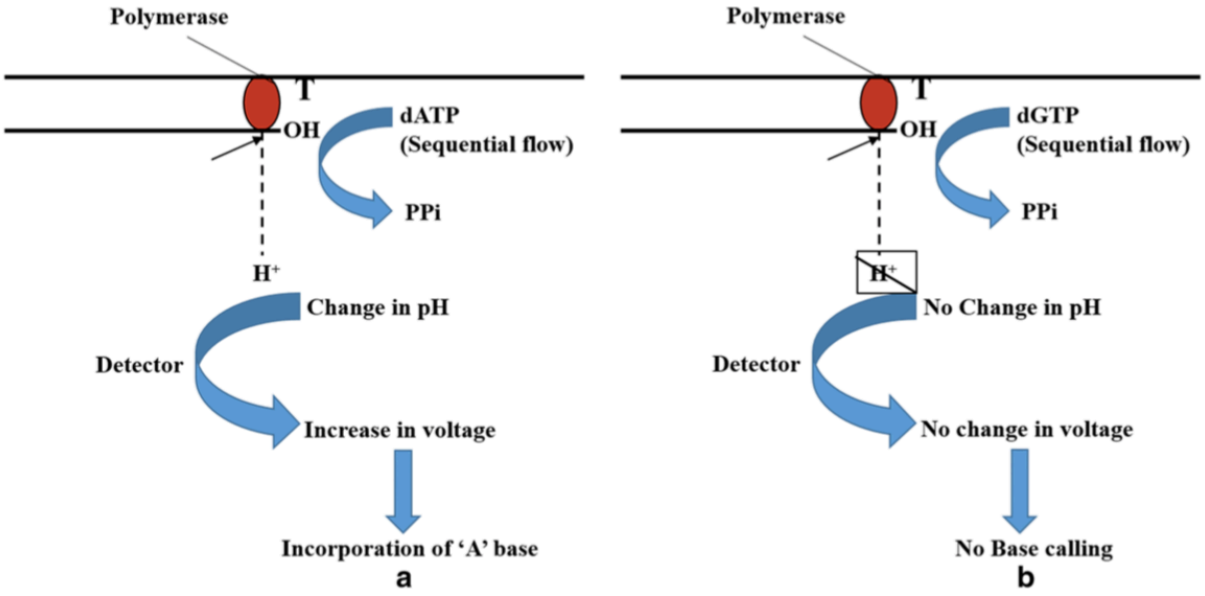

Abbildung 3: Schematische Darstellung der Beteiligung der pH-Wertänderung während der Sequenzierung durch Detektion von H^+-Ionen (a) wenn komplementäres Nukleotid eingebaut (b) nicht eingebaut wird (Ambardar et al., 2016).

Die DNA-Fragmente werden an sogenannte „Ion Sphere Particles“ in den Microwells gebunden, wobei idealerweise jedes ein einzelnes Partikel beinhaltet. Nach Zugabe eines bestimmten Typs an dNTP zum Microwell wird der Einbau, durch die Schwankung des H^+-Ions durch Triggern eines ISFET- (Ion Sensitive Field Effect Transistor) Ionensensor am Boden jedes Wells, detektiert (Garrido-Cardenas et al., 2017; Heather and Chain, 2016). Das chemische Signal wird nach Detektion durch Veränderung des pH-Werts mithilfe einer Sensorschicht in der Mikrotiterplatte in Sekunden in ein digitales Signal umgewandelt. Die Detektion erfolgt direkt ohne die Verwendung von Laser-Scanner oder CCD-Kameras sowie nichtmodifizierten Basen und macht diese Methode deshalb extrem schnell sowie kostengünstig (Ambardar et al., 2016). Der Ion Torrent hat jedoch genau so wie der 454 FLX und alle anderen auf der Pyrosequenzierung basierenden Technologien das Problem, dass es durch den Einbau von mehreren passenden Nukleotiden hintereinander, sogenannter Homopolymere, zu einem Verlust des Signals und deswegen zu Schwierigkeiten bei der Interpretation der Sequenzen kommt (Heather and Chain, 2016).

3.5. Sequenzierung durch Hybridisierung und Ligation – SOLiD (Applied Biosystems)

Die Sequenzierung durch Ligation ist auch als SOLiD- (Support Oligonucleotide Ligation Detection) Sequenzier-Plattform bekannt. Das Sequencing-by-Ligation erfolgt aufgrund der Spezifität des Enzyms DNA-Ligase für Basenpaarung-Mismatches anstelle der Verlängerungsfähigkeit der DNA-Polymerase wie im Sequencing-by-Synthesis Prinzip. Die klonale Amplifikation der DNA-Fragmente, die über Adapter an Beads gekoppelt sind, erfolgt mittels Emulsions-PCR. Die Beads mit der amplifizierten DNA werden danach kovalent an eine Glasplatte gebunden. Für die Sequenzierung werden Sonden verwendet, die 8-9 Basen lang sind und eine oder zwei Basen beinhalten, gefolgt von drei degenerierten und drei Universalbasen, die fluoreszenzmarkiert sind. Nach der Zugabe von kurzen Primern zum Pool aus fluoreszenzmarkierten Sonden, die an die komplementäre Ziel-DNA hybridisieren, können diese Primer mithilfe der DNA-Ligase an die Sonden ligiert und die Fluoreszenz detektiert werden. Die Oligonukleotid-Sonden besitzen Cleavage-Sites um die Fluoreszenzmarkierung abzuspalten und das System für eine neue Runde vorzubereiten (Ambardar et al., 2016).

Die Aufnahme des Fluoreszenzsignals erfolgt während die Sonden zum Template komplementär gebunden haben und verschwindet nach Abspaltung der letzten 3 Basen. Am Ende eines Zyklus ist die Sequenz nur an einigen Positionen bekannt (mit Ausnahme der degenerierten Proben) und deswegen müssen die Zyklen mit Primern die ein oder zwei Basen kürzer sind, wiederholt werden, um die übersprungenen Positionen auch zu sequenzieren. Mit dieser Methode lässt sich die Sequenz des Fragments nach 5 Runden der Sequenzierung unter Verwendung sogenannter Ladder Primer-Sets ableiten (Liu et al., 2012). Aufgrund der geringen Readlänge von 35bp (die mittlerweile auf 85bp erhöht werden konnte), die zu Ungenauigkeit führen, wird es schwierig die Reads zu assemblieren (Ambardar et al., 2016) aber obwohl die SOLiD-Plattform nicht in der Lage ist die Readlänge und Tiefe von Illumina-Geräten zu produzieren, bleibt sie trotzdem aufgrund der geringen Kosten pro Base konkurrenzfähig (Heather and Chain, 2016) und findet in einer Variation dieser Methode durch die Complete Genomics (GC) Plattform, entwickelt in 2006, neue Anwendungsmöglichkeiten (Garrido-Cardenas et al., 2017). Zu den Applikationen des SO-

LiD gehören Whole Genome Resequencing, Targeted Resequencing, Gene Expression Profiling, Small RNA Analysis oder ChIP-Seq (Liu et al., 2012).

4. Third Generation Sequencing - Single Molecule Real Time (SMRT)

Second Generation Sequencing Technologien erreichen einen viel höheren Durchsatz durch die parallele Sequenzierung einer großen Anzahl von DNA-Molekülen. Dabei werden zigtausende idente Stränge an einem bestimmten Standort verankert, um danach in einem Prozess, bestehend aus aufeinanderfolgenden Wasch- und Scanschritten, gelesen zu werden. Dieser „Wasch und Scan"- Sequenzierprozess beinhaltet das Überschwemmen mit Reagenzien wie markierten Nukleotide, der Einbau der Nukleotide in den DNA-Strang, das Stoppen der Inkorporation, das Auswaschen des überschüssigen Reagenz, das Scannen um die eingebauten Basen zu identifizieren und das Behandeln der kürzlich eingebauten Basen, um die DNA-Templates für den nächsten „Wasch und Scan"-Schritt vorzubereiten. Der Zyklus wird so lange wiederholt bis die Reaktion nicht mehr durchführbar ist. Die Anordnung der DNA-Fragmente kann eine sehr hohe Dichte aufweisen, die zu einem extremen Gesamtdurchsatz und daraus resultierenden geringen Kosten pro identifizierter Base führt, wenn solche Geräte mit hoher Kapazität laufen. Der HiSeq 2000 von Illumina zum Beispiel kann mehr als 300 Gigabasen in einem einzelnen Lauf generieren (Schadt et al., 2010).

Das Prinzip der Amplifikation der DNA-Fragmente mittels Emulsions-PCR, um das Lichtsignal stark genug für eine zuverlässige Basendetektion durch CCD-Kameras zu machen hat die DNA-Analyse revolutioniert, bringt jedoch auch das Problem mit sich, dass Basensequenzfehler eingeführt oder bestimmte Sequenzen gegenüber anderen bevorzugt werden und so die relative Frequenz und Häufigkeit verschiedener DNA-Fragmente verändert wird, die vor der Amplifikation vorliegen (Pareek et al., 2011).

Diese Einschränkung führt zu Fehler in der Template-Sequenz sowie Amplifikations-Biases.

Das zweite Problem dieser Second Generation Sequencing (SGS)-Methoden ist die Zeitspanne bis zum Erhalt eines Resultats, welche generell sehr lang ist und mehrere

Tage dauern kann. Grund dafür sind die große Anzahl an Wasch- und Scanzyklen die benötigt werden. Des Weiteren liegt die Ausbeute für das Hinzufügen jeder Base bei <100% und das führt dazu, das eine Population von Molekülen immer mehr und mehr asynchron bei jeder weiteren Base wird. Das Resultat dieses Verlustes an Synchronizität, welcher auch als Dephasing bezeichnet wird, ist ein Ansteigen des Rauschens und der Sequenzierfehler beim Verlängern des Reads, wodurch die Readlänge, produziert von den meisten SGS-Systemen, signifikant geringer ist als die durchschnittlich erreichte Readlänge der Sanger-Sequenzierung. Des Weiteren steigt die Komplexität und Zeitdauer, die mit der Probenvorbereitung verbunden ist, sowie die Anforderung an die Speicherung der Unmengen an hoch informativen Daten und die Anforderung an das Alignment und die Assemblierung insbesondere in Anbetracht der kurzen Reads (Schadt et al., 2010).

Eine mögliche Lösung für diese Probleme ist die direkte Bestimmung einer Sequenz von einem einzelnen DNA-Molekül, durch Miniaturisierung in den Nanobereich, unter minimalen Einsatz von Biomolekülen, ohne vorangegangener PCR-Amplifikation mit ihrem Potential an Verfälschung. Das Sequenzieren eines einzelnen DNA-Moleküls wird auch als „Third Generation of High-Throughput Next Generation Sequencing Technology“ bezeichnet (Pareek et al., 2011).

Diese Single-Molecule-Sequencing-Technologien können je nach zugrunde liegendem Prinzip grob in drei unterschiedliche Kategorien unterteilt werden. (1) Sequencing-by-Synthesis Strategien, bei denen ein einzelnes Enzym an DNA-Polymerase beobachtet wird, wie es ein DNA-Molekül synthetisiert, (2) durch Nanopore-Sequencing, bei der einzelne DNA-Moleküle durch eine Nanopore gefädelt oder in der Nähe einer Nanopore positioniert werden und eine Detektion erfolgt, wenn individuelle Basen diese Pore passieren und (3) Direct Imaging (direkte Darstellung) von individuellen DNA-Molekülen durch die Verwendung von hochentwickelten Mikroskopiertechniken. Jede dieser Technologien bietet eine neue Herangehensweise um die DNA zu sequenzieren, mit Vor-und Nachteilen in Bezug auf verschiedene Applikationen (Schadt et al., 2010). Sequencer der dritten Generation werden kommerziell unter anderem von Pacific Biosystems (PacBio) zum Single-Molecule-Real-Time Sequencing und Oxford Nanopore Techniques (ONT) zum Nanopore-Sequencing, ver-

trieben, wobei Helicos Biosciences die erste Plattform zur Sequenzierung von einzelnen DNA-Molekülen darstellt. Neben dem Verzicht auf den Amplifikationsschritt während der Library-Preparation sind die zu erwartenden langen Reads die Stärke dieser Sequenziergeräte gegenüber der Second Generation Sequencing-Technologien, die mit einer durchschnittlichen Länge von 6-8kbp und einer maximalen Readlänge von 30-150kbp und darüber hinausgehend, um einiges größer sind als jener der Plattformen der 2.Generation (siehe Abbildung 4) (Bleidorn, 2016).

Platform	Sequencer	Costs sequencing platform	Reads per run/lane	Output per run/lane	Maximal read lengths[1]	Average run duration
Sanger	ABI 3730xl	$100,000	96	100 kbp	1000 bp	2–3 hours
454	GS FLX	$450,000	1,000,000	700 mpb	1000 bp	24 hours
Illumina	HiSeq 3000	$750,000	300,000,000[2]	150 gbp[3]	250 bp	4 days
Illumina	NextSeq500	$250,000	400,000,000	120 gbp[3]	150 bp	30 hours
Illumina	MiSeq	$100,000	25,000,000	15 gbp[3]	300 bp	24 hours
Ion Torrent	Proton II	$224,000	330,000,000	66 gbp	200 bp	4 hours
Ion Torrent	PGM 318	$50,000	5,000,000	2 gbp	400 bp	7 hours
PacBio	RS II	$700,000	50,000	400 mbp	54 kbp	3 hours
Nanopore	MinION	$1,000	80,000[4]	490 mbp[4]	150 kbp	n.a.[4]

[1]Estimated for high quality reads, individual reads could be longer.
[2]Single read run on one lane, capable of paired-end runs.
[3]Output for paired-end runs (in case of HiSeq a single lane).
[4]Machine run time is usually adjusted to need of sequencing depth, example given for an 48 hours run.

Abbildung 4: Gegenüberstellung des Outputs von ausgewählten Sequenzier-Plattformen (Bleidorn, 2016).

4.1. Single Molecule Real-Time Sequencing (SMRT) – Pacific Biosciences

Das Prinzip des Single Molecule Real-Time Sequencing von PacBio basiert auf dem Monitoring der Aktivität der DNA-Polymerase während sie unterschiedlich markierte Nukleotide in den DNA-Strang einbaut. Die Fluoreszenzmarkierungen befinden sich auf der Phosphatgruppe der verschiedenen Nukleotide und werden durch den Einbau der Basen freigesetzt. Die Detektion der eingebauten Nukleotide erfolgt durch Real-Time-Imaging, also dem Filmen der Inkorporation durch die Polymerase während der Strangsynthese. Der gesamte Prozess findet in einem sogenannten Zero-Mode Wave Guide (ZMW), einer von Aluminiumwänden umgebenen kleinen Vertiefung (Well) statt (Bleidorn, 2016). Der Zero-Mode Wave Guide ist eine mit einem Durchmesser von 70 nm und einer Länge von 100 nm, mit einer Metallschicht hinterlegten, Vertiefung auf einem Glasträger und verhindert aufgrund seiner geringen Größe, dass

sichtbares Laserlicht mit einer Wellenlänge von 600 nm vollständig durchgeht. Es kommt zu einem exponentiellen Abbau des Lichts wenn es in den Wellenleiter eintritt. Durch die Laserbeleuchtung aufwärts durch das Glas in den ZMW werden nur die unteren 30 nm des ZMW illuminiert (Schadt et al., 2010). In jedem der 150.000 ZMWs, die sich auf sogenannten SMRT-Zellen befinden, ist eine einzelne DNA-Polymerase am Boden der Glasoberfläche mittels Biotin-Streptavidin-Interaktion verankert (Rhoads and Au, 2015).

Als nächstes wandern die mit unterschiedlichen Fluorophoren gekoppelten Nukleotide in der benötigten Konzentration durch Diffusion im Nanobereich in Mikrosekunden in den ZMW hinunter und umgeben die DNA-Polymerase. Wenn kein korrektes Nukleotid dabei ist diffundiert es wieder nach oben und verlässt die Öffnung. Da der Laser nicht durch die Öffnung nach oben vordringt und die Fluoreszenzmarkierungen der Nukleotide anregt, tragen die markierten Nukleotide oberhalb des ZMW nicht zum Messsignal bei. Nur wenn die korrekten Basen durch die Unterseite der 30 nm des ZMW diffundieren, fluoreszieren sie. Die Detektion und der Einbau in den wachsenden Strang durch die DNA-Polymerase erfolgt nur, wenn es sich um das korrekte Nukleotid handelt. Dieser Prozess dauert Millisekunden und ist im Ausmaß ungefähr dreimal so lang wie die einfache Diffusion und dieser Unterschied in der Zeit resultiert in einer höheren Signalintensität von eingebauten gegenüber nicht eingebauten Nukleotiden und so zu einer höheren Signal-to-Noise Ratio (Bleidorn, 2016; Garrido-Cardenas et al., 2017; Schadt et al., 2010).

Die Bewegung der Nukleotide in die ZMWs, aus diesen heraus und ihre Präsenz währenddessen wird als Hintergrundrauschen (Background Noise) gemessen. Wenn die DNA-Polymerase ein markiertes Nukleotid in ihrem aktiven Zentrum hält, wird ein Fluoreszenzimpuls für den entsprechenden Farbstoff aufgezeichnet. Nach dem Einbau kommt es zu einer Abspaltung des Fluoreszenzfarbstoffes durch die normale Aktivität der Polymerase, welche zu einer Diffusion des Farbstoffs in den Hintergrund und zu einem Abfall des Emissionssignals führt. Danach beginnt der Ablauf von neuem und so werden etwa 2-4 Nukleotide pro Sekunde synthetisiert. Die Emissionsspektren bieten neben der Sequenzierung der Template-Moleküle die Möglichkeit eventuelle epigenetische Modifizierungen durch Unterschiede in den Emissionsmus-

tern aufzudecken. Pro SMRT-Zelle ist der Sequenzierprozess sehr schnell und dauert ungefähr 4 Stunden (Bleidorn, 2016; Schadt et al., 2010).
Der Erhalt der Sequenzinformation erfolgt während des Replikationsprozesses des Targetmoleküls, welches als sogenannte SMRTbell bezeichnet wird. Diese SMRTbell ist eine geschlossene einzelsträngige zirkuläre DNA, die durch Ligation von Hairpin-Adapter an beide Enden eines doppelsträngigen (dsDNA) Template-Moleküls, erzeugt wird (Rhoads and Au, 2015).

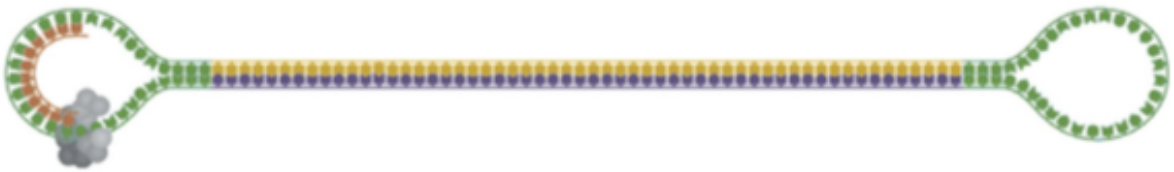

Abbildung 5: Darstellung eines SMRTbell-Templates. Ligation der Hairpin-Adapter (grün) an beide Seiten der doppelsträngigen DNA (gelb & violett). Die Polymerase (grau) verlängert den Strang (orange) durch den Einbau von Nukleotiden (Rhoads and Au, 2015).

Wenn eine SMRTbell auf den als SMRT-Zelle bezeichneten Chip geladen wird, diffundiert sie in einen der Zero-Mode Wave Guides und kann von der immobilisierten DNA-Polymerase an eine der beiden Hairpin-Adapter gebunden und die Replikation gestartet werden. Die Aufnahme des Replikationsprozesses aller ZMWs einer SMRT-Zelle erfolgt als Film von Lichtimpulsen und diese, jedem einzelnen ZMW zugehörigen Impulse, können danach zu einer Sequenz an Basen, einem sogenannten Continuous Long Read (CLR) interpretiert werden. Aufgrund der zirkulären Struktur der SMRTbell repliziert die Polymerase einen Strang der doppelsträngigen DNA, führt den Einbau über die Basen des Adapters fort, inkorporiert danach den anderen Strang. Diese Fähigkeit der Polymerase wird auch als Strand Displacement Capacity bezeichnet. Für Templates die kurz genug sind, kann die Polymerase mehrere Male um das Template zirkulieren. Dadurch können beide Stränge, wenn die Lebensspanne dieser spezialisierten Polymerase lang genug ist, gegenwärtig 360 min, mehrere Male („Passes“) in einem CLR sequenziert werden und dieser danach durch Erkennen und Ausschneiden der Adaptersequenzen in Multiple Reads (Subreads) aufgeteilt werden. Diese Konsensussequenz aus mehreren Subreads in einem ZMW erge-

ben einen sogenannten Circular Consensus Sequence (CCS) Read mit einer höheren Genauigkeit (99,999%) und einer effektiven Reduktion von zufälligen Fehlern. Wenn eine Zielsequenz zu lang ist um mehrere Male in einem CLR sequenziert zu werden, wird kein CCS Read generiert sondern ein einzelner Subread. Ein wichtiger Vorteil der PacBio-Sequenzierung ist die mögliche Readlänge. Der PacBio RS II weist mit der derzeitigen C4-Chemie durchschnittliche Readlängen von über 10 kb und maximale Readlängen von über 60 kb (höchster Rekord 92,7 kb) auf. Im Vergleich dazu liegt die maximale Readlänge des Illumina HiSeq 2500 durch Paired-End-Sequenzierung bei 250bp. Dadurch sind SGS-Technologien nicht in der Lage repetitive Sequenzen mit mindestens einer einzigartigen flankierten Sequenz zu umfassen. Die Herkunft des Reads kann nicht exakt bestimmt werden und die daraus resultierenden Alignments und Misalignments führen zu Problemen in der Downstream-Analyse, inklusive der Häufigkeitsabschätzung und dem Structural Varation (SV) Calling (Nakano et al., 2017; Rhoads and Au, 2015).

Durch die viel längere Readlänge ist es mit dem PacBio-System möglich, die genaue Lokalisation und Sequenz von repetitiven Regionen durch Bestimmung einmaliger Regionen in einem einzelnen Read zu lösen. Gemeinsam mit den drei anderen wesentlichen Vorteilen gegenüber dem Second-Generation-Sequencing, der hohen Konsensus-Genauigkeit, der niedrige Grad an Bias und die gleichzeitige Fähigkeit der epigenetischen Charakterisierung macht das PacBio-System die Sequenzierung von genomischen Regionen mit hohem/niedrigem G+C, Tandem Repeats und Interspersed Repeat Regions möglich und ist daher ideal für Whole Genome Sequencing, Targeted Sequencing, Complex Population Analysis, RNA-Sequencing oder Charakterisierung von epigenetischen Modifikationen (Nakano et al., 2017; Rhoads and Au, 2015).

4.2. Nanopore-based Sequencing – Oxford Nanopore Technologies (ONT)

In 2014 fand die Revolution der TGS-Plattformen durch die Entwicklung eines kommerziellen Sequenziergeräts dem MinION™, welcher die Nanopore-Technologie verwendet, seinen Fortschritt. Der MinION™ erkennt Nukleotide durch das Messen von Änderungen in der elektrischen Leitfähigkeit, die beim Durchtritt des DNA-Stranges

durch eine biologische Pore hervorgerufen werden. Dabei werden mit wenigen Hunderttausend Basenpaaren ähnliche Readlängen erzielt wie bei Pacific Biosciences. Der MinION™ ist mit 10 x 3 x 2cm und 90g das kleinste Sequenziergerät, das aktuell verfügbar ist. Als Ausgangsmaterial sollte hochqualitative genomische DNA bestehend aus langen Fragmenten (>30 kb) verwendet werden, die mit Standardextraktions- und Aufreinigungsmethoden erzielt werden kann (Lu et al., 2016).

Der Kern des Geräts besteht aus einer Flowcell auf der sich 2048 individuell adressierbare Nanoporen befinden, die in Gruppen von jeweils 512 Nanoporen durch einen sogenannten Application-Specific Integrated Circuit (ASIC) kontrolliert werden können (Jain et al., 2016). Dabei werden die aktuellen Werte der individuellen Nanoporen, wenn DNA-Moleküle passieren, vom Sequencer gestreamt und eine Analyse der aktuellen Stränge während des Sequenzierens ermöglicht. Vom MinION™ werden die Daten aller Kanäle simultan gestreamt, wobei jeder Kanal direkt ansprechbar ist und durch Umkehr der Spannung durch die Nanopore die Möglichkeit besteht den Read zu verwerfen. Moleküle, die vom Sequencer in Verwendung genommen wurden können entweder bis zur Vervollständigung sequenziert oder verworfen und ein alternatives Molekül probiert werden. Dieser Vorgang ermöglicht ein selektives Sequenzieren (Loose et al., 2016).

Vor der Benützung des MinION™s erfolgt ähnlich der SGS-Technologien eine Library Preparation um unterschiedliche Applikationen durchführen zu können. Es wird eine lange doppelsträngige DNA verwendet, damit das Sequenzieren von beiden Strängen ausgeführt werden kann. Die Konstruktion der Library umfasst mehrere Schritte, die in folgender Reihenfolge durchgeführt werden: (1) Fragmentierung der genomischen DNA mithilfe eines g-TUBE zum Beispiel von Covaris (2) ein optionaler „PreCR"-Schritt zum Reparieren von geschädigter DNA (3) End-Repair zur Erzeugung von Blunt-Ends in den DNA-Fragmenten (4) A-Tailing um eine Adenin-Base an das 3'Ende des Fragments zu hängen (5) Adapter Ligation und (6) His-Bead Purification zum Entfernen von Enzymen und Nukleotiden (Lu et al., 2016).

Vor Beginn der Sequenzierung müssen an beide Enden der genomischen DNA oder cDNA-Fragmente Adapter ligiert werden, welche das Erfassen des DNA-Stranges und das Binden von Enzymen am 5'Ende eines Stranges ermöglichen. Das Enzym

stellt dabei eine unidirektionale Einzelnukleotid-Verschiebung in Millisekunden entlang des Stranges sicher. Eine weitere Aufgabe der Adapter ist das Aufkonzentrieren des DNA-Substrates an der Membranoberfläche in der Nähe der Nanopore, um die Erfassungsrate der DNA um das mehrere tausendfache anzukurbeln (Jain et al., 2016). Es werden jeweils zwei Varianten von Adapter an beide Enden der doppelsträngigen DNA der Library-Fragmente angehängt, die als Leader-und Hairpin-Adapter bezeichnet werden. Der Leader-Adapter wird auch als Y-Adapter bezeichnet, da er eine ypsilonförmige Struktur besitzt, wohingegen der Hairpin-Adapter auch „HP“-Adapter genannt wird.

Der Ablauf der Sequenzierung beginnt am einzelsträngigen 5’Ende des Y-Adapters gefolgt vom Template-Strang, dem Hairpin-Adapter und dem komplementären Strang. Das Öffnen des doppelsträngigen DNA-Moleküls erfolgt durch ein Motorprotein, welches startet, wenn es sich dem Wendepunkt der Y-Adapter komplementären Region annähert. In dieser Phase passiert der erste Strang (Template) die Nanopore, dessen Geschwindigkeit durch das Motorprotein kontrolliert wird. Nach Erreichen des HP-Adapters wird der Durchtritt des komplementären Strangs durch die Pore in ähnlicher Art und Weise durch ein Protein, dem sogenannten Hairpin-Protein, kontrolliert. Nach der Sequenzierung kann die Information von einem Strang genutzt werden, diese wird als 1-directional (1D) Base Calling bezeichnet. 2D Base Calling wird verwendet, wenn die Information beider Stränge vorhanden ist und resultiert in einer höheren Basenqualität (Jain et al., 2016; Lu et al., 2016).

Beim Durchtritt der DNA durch die Pore, kann die Veränderung im Ionenstrom, verursacht durch die Differenz in den sich verschiebenden Nukleotidesequenzen, mithilfe eines Sensors mehrere tausendmal in der Sekunde detektiert und dem ASIC-Mikrochip übermittelt werden. Unterschiedliche Basen erzeugen verschiedene Signale und diese Unterschiede im Ionenstrom werden nach ihrer assoziierten Dauer, Amplitude und Streuung in diskrete Events eingeteilt. Danach erfolgt eine rechnerische Interpretation dieser 3-6 Nukleotide langen sogenannten Kmers durch graphische Modelle (Jain et al., 2016; Lu et al., 2016). Aufgrund der Geschwindigkeit des Prozesses und der Länge des Nanopore-Tunnels ist immer mehr als ein Nukleotid in der Pore. Folglich wird normalerweise das Signal von überlappenden 5-mers aufge-

zeichnet und die Cloud-Based Base-Calling Software MinKNOW muss danach zwischen 4^5 (1024) mögliche Ionenzustände für alle möglichen 5-mers unterscheiden, um eine Raw-Sequenz zu generieren. Deswegen liegt die sehr hohe Error-Rate für alle Reads, produziert durch diese Technik bei 25-40% (Bleidorn, 2016).

4.3. Nanopore-Technologie – Biologische/Solid State/ Hybrid-Nanoporen

Bei der Nanopore-Technologie werden, unter Anwendung von externer Spannung, Partikel die etwas kleiner als die Pore sind, durch diese hindurchtransportiert. Diese nanometergroßen Poren sind entweder in biologische Membranen eingebettet oder in fester Form (Solid-State) vorhanden und unterteilen die Reservoirs in *cis* und *trans* Kompartimente, die mit leitenden Elektrolyten gefüllt sind. Mithilfe von Elektroden können die Elektrolytionen unter Anlegen einer Spannung in Lösung elektrophoretisch durch die Pore wandern und dabei ein Ionenstrom-Signal produzieren. Durch Blockierung der Pore mit einem Analyt, wie das negativ geladene DNA-Molekül in der *cis*-Kammer, kommt es zu einem Blockieren dieses Stromflusses durch die Nanopore und zu einer Unterbrechung des Signals. Durch diese transiente Blockade können die chemischen und physikalischen Eigenschaften des Ziel-Moleküls durch statistische Analyse der Amplitude und Zeitdauer des Translokationsvorgangs kalkuliert werden. Nanoporen sind Single-Molecule-Sensing-Technologien, die dazu verwendet werden können biologische und chemische Moleküle auf Einzelmoleküllevel zu detektieren und haben deshalb großes Anwendungspotential in mehreren Bereichen wie die Analyse von Ionen, Peptide, Proteine, DNA, RNA, Polymere, Makromoleküle und Medikamente (Feng et al., 2015).

Biologische Nanoporen werden in der Einzelmoleküldetektion, DNA-Sequenzierung und Krankheitsdiagnose verwendet, aber zusehends von Solid-State-Nanopore-Sensoren abgelöst, die gemeinsam mit einem Feldeffekttransistor als synthetische Nanoporen auf einen Schaltkreis integriert werden können und so das Potential für die Miniaturisierung und Entwicklung des tragbaren Sequenziergeräts wie den MinION™ ermöglicht haben. Hybrid-Nanoporen sollen sich die Vorteile der Eigenschaften beider, biologischer und fester Nanoporen, zu Nutze machen (Feng et al., 2015).

Biologische Nanoporen werden auch Transmembranproteinkanäle genannt und sind häufig in einem Liposom, einer flachen Lipiddoppelschicht (Lipid Bilayer) oder ande-

ren Polymerfilmen integriert und besitzen den Vorteil, dass sie eine äußerst reproduzierbare Größe und Struktur aufweisen und klar definiert sind. Des weiteren können sie durch moderne Molekularbiologietechniken leicht modifiziert werden wie zum Beispiel durch das Einbringen von Mutationen in die DNA-Sequenz um den Aminosäurerest an dieser spezifischen Position zu verändern. Die drei wichtigsten und gut studierten biologischen Nanoporen sind das □-Hämolysin, MspA und Phi29 (Feng et al., 2015; Haque et al., 2013; Wang et al., 2015).

Das □-Hämolysin (□-HL, □-Toxin) ist ein vom humanpathogenen Staphylococcus aureus sekretiertes porenbildendes Exotoxin mit einer pilzförmigen Heptamerstruktur und einer molekularen Masse von 232,4 kDa. Dieser Transmembrankanal besitzt eine mit einem Durchmesser von 3.6 nm große Capstruktur und ein mit einem Durchmesser von 2.6 nm (auf der trans-Seite) großes Transmembran-□-Fass (□-Barrel) sowie eine äußere Dimension von 10 nm x 10 nm. Das □-HL ist *in vivo* in der Lage sich selbst schnell in einen planaren Lipid-Bilayer zu integrieren und einen Nanokanal mit einer Breite von 1.4 nm an der engsten Stelle zu formen. Da der innere Durchmesser des □-HL-Kanals und das einzelsträngige DNA-Moleküls mit seiner Größe mit 1.3 nm sehr dicht beieinander liegen, kann die □-HL-Nanopore aufgrund des Ionenstroms zwischen einzelne Nukleotide unterscheiden. Dadurch ist das □-Hämolysin ein sehr vielversprechendes Tool um biomolekulare Interaktionen und Strukturen auf Einzel-Moleküllevel zu analysieren. Aufgrund seiner intrinsischen Nanopore-Struktur ist ein Erfassen von diversen Analyten wie DNA, RNA, kleinen organischen Moleküle, Metallionen oder Proteine möglich. Aufgrund der geringen Porengröße (□1.4 nm) kann nur ssDNA sequenziert werden und das □-Barrel der Nanopore ist zu lang um zwischen einzelnen Nukleotiden und einem einzelnen langen DNA-Molekül zu unterscheiden. Der Vorteil dieser biologischen Nanopore liegt darin, dass sie bis zu einer Temperatur von 100°C funktionell stabil bleibt und einem breiten pH-Wertbereich (pH 2-12) standhält. Die □-Barrel-Struktur ist zudem zugänglich für gentechnische und chemische Modifikationen, um spezifische Binding Elements einzuführen (Feng et al., 2015; Haque et al., 2013; Wang et al., 2015).

Das MspA ist ein von Mycobacterium smegmatis Porin A sekretiertes trichterförmiges Oktamer, welches eine Nanopore mit einem Ausgang von 1.2 nm im Durchmesser und 0.6 nm Länge bildet und den Transport von wasserlöslichen Molekülen über bakterielle Zellmembrane erlaubt. Die MspA-Nanopore hat aufgrund der Eigenschaft einer einzelnen Verengung einen Vorteil gegenüber dem □-HL, da sie durch standortspezifischer Mutagenese optimiert werden kann. So wurden die negativ geladenen Aminosäuren in der Engstelle der Pore durch neutrale Asparaginreste und jene beim Eingang der Pore durch positiv geladene alkalische Aminosäuren ersetzt.
Diese gentechnische Veränderung ermöglicht ein leichteres Erfassen und Entschleunigen der DNA während der Translokation durch die Pore (Wang et al., 2015).

Das veränderte MspA weist eine höhere Basenauflösung als das □-HL auf, da größere Signalunterschiede zwischen den Basen erzeugt werden. Für eine präzise Single Base Resolution (SBR) darf die Länge der Erkennungsregion einer Nanopore □0.5 nm (der Phosphor-Phosphor-Abstand eines Nukleotids im ssDNA-Strang) nicht überschreiten. Die Konstriktion von MspA ist 0.6 nm lang und führt daher zu Störungen im Signal durch benachbarte Nukleotide, da immer 4 Basen zusammen zur Blockierung des Gesamtstroms beitragen. Um Signalüberlappungen von unterschiedlichen Basen, insbesondere der Desoxynukleotide Adenin und Guanin ausschließen zu können, ist die Entwicklung neuer Methoden notwendig, um die Genauigkeit zu erhöhen. MspA-Kanäle können sich ebenso wie □-HL in eine planare Lipiddoppelschicht integrieren um ein Nanopore zu formen. MspA besitzt eine hohe Robustheit und behält seine Kanalbildungsaktivität bei einem pH-Wert von 0 – 14, durch Inkubation bei 80°C in 2% SDS oder nach Extraktion für 30min bei 100°C. Durch zielgerichtete Mutagenese lassen sich die mutierten Kanäle aufgrund der Kristallstruktur für die gewünschten Applikationen überarbeiten (Haque et al., 2013; Wang et al., 2015).

Der Bakteriophage Phi29 ist der erste Nanokanal der weder aus einem Ionenkanal eingefügt in einen Lipid Bilayer noch aus einem Membranprotein besteht sondern aus 12 Kopien des Konnektorproteins Gp10 zu einem Dodekamer geformt ist. 6 Kopien der ATP-Binding DNA-Packaging RNA (pRNA) und das ATPase Protein Gp16 liefern

die für die Translokation der DNA benötigte chemische Energie. Mit einer Länge von 7 nm und einem Durchmesser von 3.6 nm auf der *cis*-Seite und 6 nm auf der *trans*-Seite hat der Bakteriophage Phi29 einen größeren Durchmesser als das □-HL und das MspA und ist so in der Lage größere Moleküle wie eine dsDNA zu translozieren. Die Art des Einbringens und Verankerns des Konnektorkanals in die Lipiddoppelschicht erfolgt in einem zweistufigen Ansatz. Als erstes wird der Konnektor in Lipidvesikel verpackt und in einem zweiten Schritt mit der planaren Lipiddoppelschicht fusioniert. Dadurch ist die Leitfähigkeit jeder einzelnen Pore fast identisch und vollkommen linear in Bezug auf die angewendete Spannung. Der Konnektorkanal ist sehr robust gegenüber einem breiten Spektrum an experimentellen Konditionen wie hohen Salzkonzentrationen oder pH-Werten. Durch die größere Nanopore ist es einfacher den Kanal zu modifizieren, um eine bessere Detektionsregion zu erzielen oder chemische Gruppen für neue diagnostische Applikationen einzubringen oder zu konjugieren (Feng et al., 2015; Haque et al., 2013).
Um den Nachteil der biologischen Nanoporen in Bezug auf fehlende Stabilität, konstante Porengröße und Fragilität von traditionellen Lipidmembranen zu korrigieren, wurden verschiedene synthetische Nanoporen durch unterschiedliche Verfahren für die DNA- und RNA Analyse hergestellt. Solid-State Nanoporen besitzen eine höhere chemische, thermische und mechanische Stabilität und eine Verstellbarkeit der Porengröße und können im Halbleiterprozess massengefertigt werden. Die primären Techniken erlauben die Produktion von synthetischen Nanoporen aus Siliziumnitrit (Si_3N_4), Siliziumdioxid (SiO_2), Aluminiumoxid (Al_2O_3), Bornitrid (BN), Graphen, Polymermembranen oder Hybridmaterial (Feng et al., 2015).

Si_3N_4-und SiO_2-Nanoporen werden aufgrund ihres niedrigen mechanischen Stresses und ihrer hohen chemischen Stabilität als Substrate verwendet und zeigen eine gute Leistungsfähigkeit in Lösungen mit hohen Eletrolytkonzentrationen, wobei die Poren bei unter 4°C und über 5°C geöffnet und geschlossen werden können. Der Durchmesser der Nanoporen lässt sich je nach Sensing-System im Nanometer oder sogar Angströmbereich kontrollieren und ist bei der konventionellen SiN-Membran 30 nm dick. Der Nachteil von synthetischen Nanoporen liegt in strukturellen Unregelmäßig-

keiten, langer Porenlänge und schwache Wiederholbarkeit. Eine unregelmäßige Geometrie führt zu erhöhten Störgeräuschen und die kürzeste erreichte Membrandicke mit 1.4 nm ergibt eine Porenlänge von 0.47 nm und würde zwar unter den 0.5 nm liegen, die für eine präzise SBR nötig sind, aber die ideale Porenlänge bei dieser Art von Nanoporen liegt mit 0.25 nm bei der Hälfte des Base Spacing (Feng et al., 2015; Haque et al., 2013; Wang et al., 2015).

Al_2O_3-Nanoporen besitzen eine gesteigerte elektrische Leistung, eine höhere Signal-to-Noise Ratio und ein niedrigeres Störgeräusch während der DNA-Translokation. Durch Atomlagenabscheidung können Al_2O_3-Membranen mit extremer Dünne (typisch sind 45-60 nm) bis hin zu einzelner Atomstärke hergestellt werden. Ein weiterer Vorteil gegenüber Si_3N_4-und SiO_2-Nanoporen ist die niedrigere Translokationsgeschwindigkeit der DNA durch die Nanopore aufgrund der elektrostatischen Interaktion der positiv geladenen Oberfläche des Al_2O_3 und des negativ geladenen doppelsträngigen DNA-Moleküls. Graphen-Membranen sind atomare Einzelschichten aus Kohlenstoff mit besonderen mechanischen und elektrischen Eigenschaften sowie ihrer Sub-nm-Dicke von 0.335 nm und werden aufgrund der verbesserungswürdigen räumlichen und zeitlichen Auflösung zum Erhalt von strukturellen Informationen von Molekülen auf Single-Base-Level, als potentielle Alternativen zu den traditionellen Solid-State-Nanoporen gesehen (Feng et al., 2015; Haque et al., 2013; Wang et al., 2015).

Hybrid-Nanoporen stellen eine alternative zu Solid-State-Nanoporen dar, die Zielmoleküle derselben Größe mangelhaft chemisch unterscheiden können. Diese chemische Spezifität kann durch funktionalisierende Oberflächen oder das Anhängen von spezifischen Erkennungssequenzen und Rezeptoren an die Nanopore verbessert werden. Nanoporen können mit Hairpins oder Rezeptoren funktionalisiert werden, um das Potential der Nukleotiderkennung zu erhöhen (Feng et al., 2015).

Die synthetischen Nanoporen können mit einem flüssigen Lipid-Bilayer beschichtet werden, wobei die Dicke und Oberflächenchemie genau durch verschiedene Lipide gesteuert werden, um Proteintranslokationen zu kontrollieren. Ein weiteres Konzept ist das Integrieren von □-Hämolysin in eine SiN-Membran, um einerseits die Fragilität der Lipiddoppelschichtmembran zu umgehen und andererseits das Potential zur standortspezifischen Gentechnik und chemischen Modifikation anwenden zu können.

Dieses Hybridsystem ist für einige Tage stabil, führt danach aber zur Verformung der Poren wenn sie mit der Solid-State Pore in Verbindung kommen und verlieren ihre Fähigkeit zwischen einzelnen Nukleotiden zu unterscheiden und führen zu einem Auslaufen von Strömungen, da Ionen durch die Regionen zwischen der biologischen und der Solid-State Pore aufgrund mangelhaften Abdeckung durchfließen können. Diese Faktoren limitieren derzeit die Entwicklung von Hybrid-Nanoporen zur routinemäßigen Verwendung in der DNA-Sequenzierung (Feng et al., 2015; Haque et al., 2013).

5. Alignment-Variant Calling-Filtering/Annotation

Mit dem Aufkommen von immer neueren Next Generation Sequencing Technologien steigt auch die Menge an Daten an, die produziert wird. Nachdem immer größere Datensätze erschwingbar werden, wird die Computeranalyse in der Zukunft der limitierende Faktor in der genomischen Forschung sein und die Bioinformatik mit ihren verschiedensten Analyse-Tools die wahre Herausforderung darstellen (Dolled-Filhart et al., 2013). Das folgende Kapitel gibt eine grundlegende Übersicht über den bioinformatischen Workflow, der beginnt wenn die Sequenzierung beendet ist.
Die Next Generation Sequencing Datenanalyse besteht aus mehreren aufeinander folgenden Phasen:

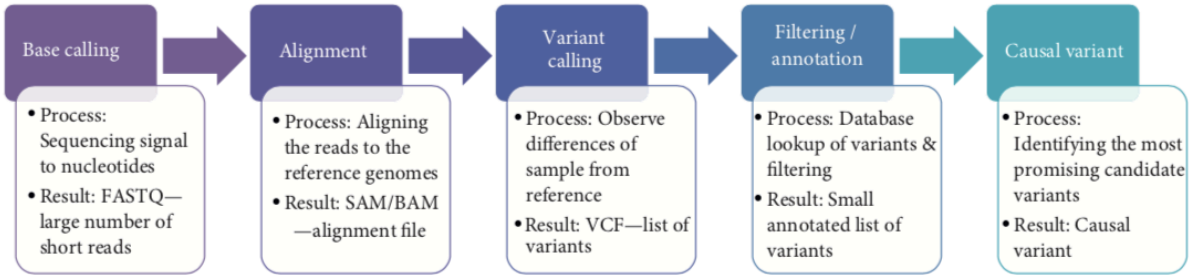

Abbildung 6: Ablauf der einzelnen Schritte des bioinformatischen Workflows nach der Sequenzierung (Dolled-Filhart et al., 2013).

Im ersten Analyseschritt, dem Base Calling, wird das optische oder physikalisch-chemische Signal, das während des Sequenzierprozesses entsteht, durch plattformspezifische eigenentwickelte Algorithmen in eine Basenabfolge umgewandelt, wie zum Beispiel dem Algorithmus Bustard von Illumina. Unabhängig von der Art der Se-

quenziertechnologien, welche unterschiedliche Initial Raw Data (Bilddateien, Stromschwankungen) verwenden, werden die Base Call-Resultate für gewöhnlich im Standard FASTQ-Format gespeichert. Mithilfe des NGS QC Toolkit lassen sich andere NGS File-Formate wie FASTA, CFASTA, SFF und QUAL in das FASTQ-Format konvertieren. Eine charakteristische komprimierte FASTQ-Datei besitzt eine Größe von mehreren Gigabyte und beinhaltet mehr als 200 Millionen Reads, welche die Sequence Readouts der DNA-Fragmente aus der Sequenzier-Library darstellen. Das FASTQ-Format ist ein textbasiertes Format, das die Sequenz jedes Reads und für jede Base einen sogenannten Confidence (Quality) Score enthält. Dieser Phred Quality Score ist ein Maß für die Wahrscheinlichkeit, dass eine falsche Base eingebaut wurde und ist eine essentielle Komponente des FASTQ-Formats. Berechnet wird er durch folgende Formel:

$$Q = -10 \times \log_{10} P_{Err}$$

wobei P_{Err} die Wahrscheinlichkeit der Entstehung eines Base Call Errors ist. Bei einem Q-Score von 20 ist die Wahrscheinlichkeit 1% das eine inkorrekte Base erkannt wurde und bei Q30 1/1000, wobei High Quality Calls einen Q-Score zwischen 30 und 40 haben. Die Base Calls werden für eine bessere Visualisierung nach dem American Standard Code for Information Interchange (ASCII) verschlüsselt (Wang, 2016, S. 73-76).

Nach dem Base Calling erfolgt die Sequenzanalyse in drei Hauptschritten: (1) Alignment, (2), Variant Calling und (3) Variant Filtering and Annotation. Vor dem Alignment werden die FASTQ-Dateien einem Data Quality Control (QC) and Preprocessing-Schritt unterzogen um Reads mit geringer Qualität, Adaptersequenzen oder artifizielle Sequenzen wie PCR-Primer herauszufiltern (Dolled-Filhart et al., 2013).

5.1. Alignment

Im ersten Hauptschritt dem Alignment werden die kurzen Sequenzen zu Positionen im Referenzgenom zugeordnet und in Form einer Sequence Alignment/Map (SAM)- oder Binary Alignment/Map (BAM)-Datei abgespeichert (Dolled-Filhart et al., 2013). Als Quelle für das Referenzgenom dienen die UCSC (University of Santa Cruz) und das GRC (Genome Reference Consortium) (Wadapurkar and Vyas, 2018).

Aufgrund der Tatsache, dass jede der Millionen Reads mit jeder der 3 Milliarden mög-

lichen Positionen im humanen Genom verglichen werden muss, ist der rechnerische Schritt kein einfacher. Die Software muss dazu den Startpunkt für jeden Read im Referenzgenom festlegen, ein Umstand der durch die Variation in der Basenqualität, dem Mapping von einzigartigen gegen nicht-einzigartigen Reads sowie der Menge an kurzen Reads erschwert wird. Dieser kritische Prozess ist sehr rechnerisch- und zeitintensiv und auftretende Fehler im Alignment zum Referenzgenom ziehen sich durch die komplette restliche Analyse (Dolled-Filhart et al., 2013).

Die häufigste verwendete NGS Data QC Software umfasst das FastQC, FastX-Toolkit und NGS QC Toolkit. Mit diesen Toolkits lässt sich auf Per-Reads und Per-Base Q-Scores, Basenfrequenzverteilung, Readlängenverteilung und das Vorhandensein von duplizierten sowie artifiziellen Sequenzen untersuchen (Wang, 2016, S. 78). Für das Alignment der Sequenz zum Referenzgenom stehen verschiedene kommerziell sowie kostenlos erhältliche Softwareprogramme zu Verfügung, die sich alle jeweils in der Geschwindigkeit sowie Genauigkeit unterscheiden. Diese Alignment-Algorithmen können die potentielle Alignment-Lokation zügig innerhalb des Referenzgenoms unter Verwendung unterschiedlicher Herangehensweisen einschließlich Hash Tables, Spaced Seeds oder Contiguous Seeds identifizieren (Dolled-Filhart et al., 2013).

Die kurzen Reads, die während des NGS-Experiments erzeugt wurden, können entweder Single-End Reads oder Paired-End Reads einer Probe sein und zwischen Dutzenden und mehreren Hundert Basenpaaren schwanken, welche korrekt zur dazugehörigen Stelle im Genom aligniert werden müssen (Dolled-Filhart et al., 2013).

MAQ (Mapping and Assembly with Qualities) und ELAND (Efficient Large-scale Alignment of Nucleotide Databases) verwenden einen Hash-based Index, BWA Burrows-Wheeler Alignment), Bowtie oder SOAP2 einen BWT (Burrow-Wheeler Transform)-based Index, Novoalign und SOAP einen Genome-based Hash und SHRiMP einen Spaced-Seed-Ansatz. Des weiteren erlauben manche Algorithmen das Aufzeigen der „besten“ Matches unter Verwendung von heuristischen Ansätzen (BWA, Bowtie, MAQ) und andere alle möglichen Matches (SOAP3, SHRiMP). Manche Algorithmen (SARUMAN) können nur Single-End Reads verarbeiten, andere (BWA, Bowtie2) führen ein Gapped-Alignment oder (MAQ, Bowtie) Ungapped-Alignment aus. Manche fokussieren auf Geschwindigkeit (BWA, Bowtie) andere auf

Sensitivität (Novoalign) oder beides (Stampy). Für das Alignment von längeren Reads wie zum Beispiel der SMRT-Plattform von Pacific Biosciences empfiehlt sich die Verwendung von BLASR, LAST, LASTZ oder BWA-MEM (Dolled-Filhart et al., 2013; Wang, 2016, S.78-81).

Neben der Wahl des gewünschten Algorithmus ist die Wahl der Referenzsequenz, wenn mehrere vorhanden sind, ein wichtiger Faktor, da diese das Mapping-Ergebnis beeinflussen kann. Reads, die der gewählten Referenzsequenz ähnlicher sind, alignieren besser als jene die mehr von ihr abweichen. Wenn diese Abweichung groß genug ist wird der Read als Fehlpaarung verworfen. Diese durch die Verwendung mehrerer Referenzsequenzen entstehenden Verzerrungen werden auch als „Reference Bias“ bezeichnet und können durch das Abgleichen des Mapping-Ergebnisses mit verschiedenen Referenzsequenzen die passender sind oder die Verwendung von neueren Algorithmen wie den GenomeMapper, der die Fähigkeit besitzt, mehrere Genomsequenzen simultan als Referenzsequenz zu verwenden, behoben werden (Wang, 2016, S-81).

5.2. Variant Calling

Unter dem zweiten Hauptschritt, dem Variant Calling, versteht man den Prozess der Identifizierung von Single Nucleotide Variants (SNVs) aus NGS-Daten durch Verwendung von mathematischen und rechnerischen Tools (Chaitankar et al., 2016), wobei die Coverage einen der Hauptparameter darstellt, da sich identifizierte Mutationen in mehreren der Reads wiederfinden lassen sollten (Wadapurkar and Vyas, 2018). Nach dem Alignment der Reads mit dem Referenzgenom kann das Probengenom mit diesem verglichen und Mutationen identifiziert werden. Bei diesen Varianten kann es sich um genomisches Rauschen ohne funktionelle Effekte handeln oder doch um Varianten, die für Erkrankungen verantwortlich sind. Die Speicherung solcher Sequenzvariationen die SNPs, Insertionen-Deletionen (Indels), größere strukturelle Variationen und Annotationen beinhalten können, erfolgt standardgemäß im Variant Call Format (VCF) (Dolled-Filhart et al., 2013).

Die Anwendung einer Variant Calling Software zur Reduzierung von falsch positiven/negativen Varianten wird empfohlen (Chaitankar et al., 2016), da es für die Unterscheidung von „echten“ Varianten gegenüber Sequenzier- und Alignmentfehler mehrere Herausforderungen zu überwinden gibt (Day-Williams and Zeggini, 2011).

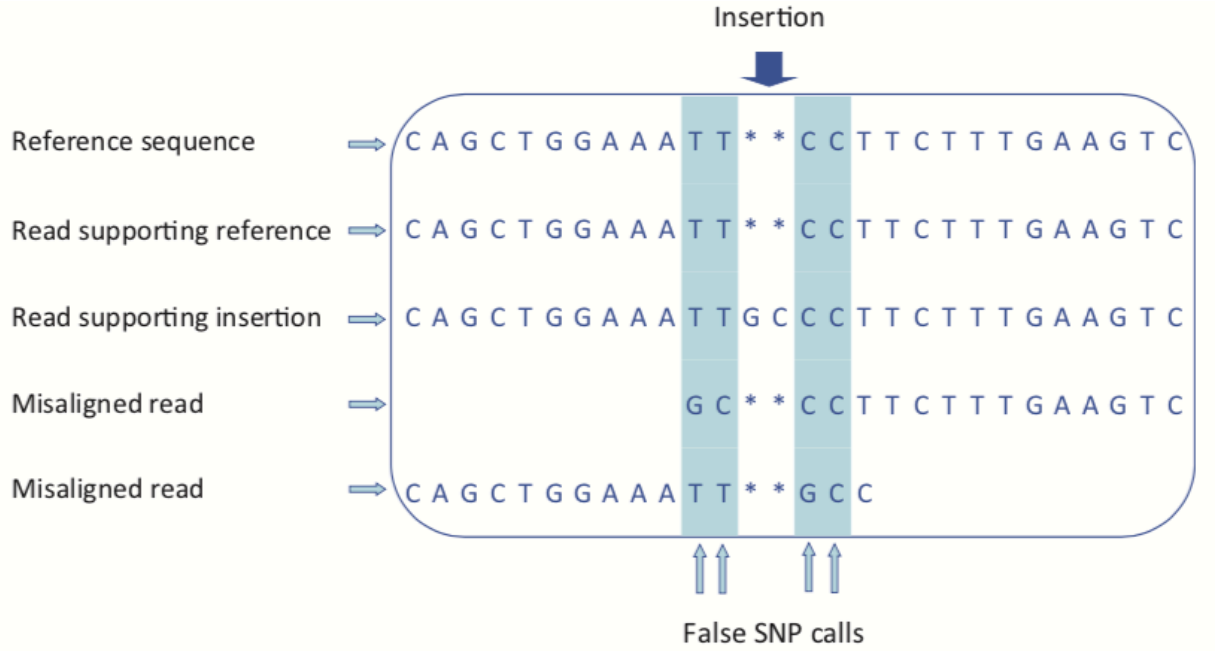

Abbildung 7: Konsequenzen eines Indel-Misalignments zu einem falsch positiven SNP Calling. Die eingefügten Sequenzen kommen am Beginn/Ende der Reads vor und führen dazu, dass die Insertion vom Alignment-Algorithmus schwer erkannt wird (Day-Williams and Zeggini, 2011).

Die drei Hauptquellen, welche die Identifikation von Varianten aus NGS-Daten erschweren sind: (1) Alignment-Artefakte durch die Anwesenheit von Indels (Day-Williams and Zeggini, 2011), welche den Hauptursprung der meisten falsch positiven SNP-Calls darstellen, vor allem wenn der verwendete Alignment-Algorithmus keine Gapped-Alignments durchführen kann, (2) Artefakte aus der Library-Preparation durch PCR-Artefakte und variabler GC-Gehalt in den kurzen Reads, wenn kein Paired-End Sequencing verwendet wurde und (3) Variable Q-Scores und höhere Error-Raten meist an Basen am Ende von Reads (Dolled-Filhart et al., 2013).
Die Konsequenz des Misalignments von Reads mit kleinen Indels führt dazu, dass es zu Fehlpaarungen rund um das Indel kommt, die von den Variant Calling- Algorithmen aber als High-Confident-SNP eingestuft werden, unterstützt durch multiple Reads mit einem potentiell guten Mapping-Score (Day-Williams and Zeggini, 2011).

Da die falsch positiven/negativen Calls von SNVs und Indels größere Sorgen bereiten, wird bei SNP-Calling-Algorithmen empfohlen, eine Rekalibration von Per-Base Q-Scores (GATK), Verwendung von Algorithmen mit großer Sensitivität (Novoalign), Likelihood-Ratio-Test und Linkage-Disequilibrium durchzuführen, um die SNP-Call-Genauigkeit zu erhöhen (Dolled-Filhart et al., 2013).

5.3. Variant Filtering and Annotation

Der dritte Hauptschritt nach dem Alignment und Variant Calling beinhaltet eine Kombination aus einer Filterung (Entfernen von Varianten, die einem bestimmten genetischen Model zuzuordnen sind oder in normalem Gewebe nicht vorkommen) sowie einer Annotation (funktionale Zuordnung von Varianten zu biologischen Prozessen) (Dolled-Filhart et al., 2013). Genaue Annotationen sind essentiell für die Lokalisierung von codierenden und nicht-codierenden Regionen von Genen sowie intergenischen Regionen im Genom und ihr funktionaler Zusammenhang mit Pathways im normalen und Krankheitszustand. Genaue Annotationen von Varianten in diesen Regionen sind wichtig, um herauszufinden, ob diese funktionellen Effekte krankheitserzeugend-pathogen oder zufällige genetische Drifts sind und ob die Varianten einen Einfluss auf Downstream-Produkte von Genen haben, die sie regulieren (Chakravorty and Hegde, 2017). Die Annotation von Varianten liefert eine biologische Signifikanz durch Identifizierung von krankheitsverursachenden Varianten (Wadapurkar and Vyas, 2018) und aufgrund der Tatsache, dass eine Exom-Analyse 20.000-30.000 Varianten aufzeigen kann, ist der erste Schritt die Bestimmung der Art der SNVs, ob es sich um eine synonyme, nicht-synonyme, Non-Sense Codon oder Consensus-Splice-Site-Veränderung handelt und der zweite Schritt das Abschätzen der Häufigkeit (Minor Allel Frequency, MAF) in der allgemeinen Bevölkerung (Chaitankar et al., 2016). Zu diesem Zweck haben groß angelegte Studien wie gnomAD (Karczewski et al., 2019), ExAC, ESP6500, 1000 Genome Project, oder dbSNP Sequenzvarianten von tausenden von Exomen und Genomen katalogisiert, um eine hochwertige Quelle für Allelfrequenzschätzungen darbieten zu können (Chaitankar et al., 2016).

Das Filtern lässt sich auf Basis eines genetischen Stammbaumes oder mit normalen und Krebsproben desselben Individuums durchführen. Im Falle des genetischen

Stammbaumes kann aufgrund der unterschiedlichen Vererbungsmuster gefiltert werden. Bei autosomal rezessiver Vererbungsmuster können zum Beispiel die Varianten, die heterozygot bei den Eltern und homozygot oder compound heterozygot im Kind sind, ausgewählt werden. Beim Beispiel Krebs ist eine gängige Methode jene Varianten wegzufiltern, die in normalen und Krebsproben vorhanden sind, sodass nur die somatischen Varianten übrigbleiben (Dolled-Filhart et al., 2013). Die Annotation von SNPs und Indels erfolgt mithilfe von computergestützten Annotation Tools, die mit Links zu speziellen Databases wie dbSNP verbunden sind (Wadapurkar and Vyas, 2018). Annotation Tools zum Klassifizieren von Varianten sind ANNOVAR in Kombination mit SIFT, PholyPhen2, Provean Annotation Scores, SnpEff, und der Variant Effect Predictor, wobei die Wahl der Software einen großen Einfluss auf die Varianteninterpretation hat (Chaitankar et al., 2016; Wadapurkar and Vyas, 2018).
Zusätzlich zur Bestimmung der Art und Inzidenz der Variante kann die Basenkonservierung und funktionelle Vorhersage mit der LJB23 Database, der Combined Annotation Depletion (CADD) Database und dem Genomic Evolutionary Rate Profiling (GERP) Score zugeordnet werden (Chaitankar et al., 2016; Wadapurkar and Vyas, 2018).

6. Ausgewählte NGS-spezifische Applikationen

Next Generation Sequencing Technologie-Ansätze sind heute in klinischen Studien sehr gut anwendbar und erlauben genomische, transkriptomische und epigenomische Auswertungen von Studien mithilfe einer Kombination aus Genom-, Exom-, mRNA und Bisulfit-Sequenzierung aufgrund von sinkenden Sequenzierkosten und dem Erhöhen von langen Reads durch Single Molecule Sequencing Technologien wie dem MinION™ Nanopore Sequencer. Diese NGS-Techniken in Verbindung mit verbesserten Analysealgorithmen sind in vielen Bereichen klinischer Studien wie chronische Erkrankungen, Krebs oder Neurobiologie nützlich und nach bereits verbesserter medizinischer Behandlung nach NGS-Diagnostik ist anzunehmen, dass sie die Medizin in der Klinik in Richtung Präzisionsmedizin verschieben (Vijay et al., 2016).

6.1. Gene Panels-Whole Exome/Genome Sequencing

Das Aufkommen der Next Generation Sequencing Technologie hat zu einer Veränderung des diagnostischen Workflows bei Patienten mit seltenen Krankheiten geführt, der sich meist über Jahre, unter Einbeziehung der Dokumentation von klinischen Manifestationen, Durchführung von biochemischen Tests und bildgebenden Verfahren bis hin zur Sanger Sequenzierung als genetische Analyse bei Verdacht auf eine genetische Störung, gezogen hat. NGS-basierende Tools können auf die Auswirkungen eines einzelnen Gens oder einer kleinen Gruppe von Genen hinweisen und so in einem beachtlichen Teil der Fälle helfen, eine schnelle Diagnose zu erstellen (Fernandez-Marmiesse et al., 2018).

Aus diesem Grund widmen sich viele Konsortien der Standardisierung der Hochdurchsatzsequenzierung, deren Schlüsselfaktoren für die Sequenziertechnologie in der klinischen Anwendung die Genauigkeit und Reproduzierbarkeit sind (Vijay et al., 2016).

Die Kombination aus Next Generation Sequencing und Segregationsanalyse kann zu einer Identifizierung einer pathogenen Mutation in einem Gen führen, welche bekanntermaßen eine Krankheit verursacht, die vorher mit einem differenten Phänotyp verknüpft war. Dieser Prozess wird als Reverse Phenotyping bezeichnet. Dadurch können bisher unerkannte zusätzliche Eigenschaften, in Patienten und Familienmitglieder, in einer retrospektiven klinischen Untersuchung aufgedeckt werden. So zeigen Ergebnisse vom Reverse Phenotyping, dass die Abfolge und Intensität mit derer Symptome in seltenen Erkrankungen auftreten von Patient zu Patient stark variieren und so eine Erklärung darstellen, warum es so schwierig ist eine Diagnose zu begründen (Fernandez-Marmiesse et al., 2018).

Der erste generalisierbare Ansatz zur Lokalisation von Risikogenen ohne vorherige Kenntnis basiert auf der Kopplungsanalyse gefolgt von einer positionellen Klonierung. Diese Kopplungsanalysen benötigen aber die Bestimmung einer ausreichenden Zahl an Probanden und deren Familien und sind deshalb für seltene Erkrankungen, wo nur ein oder ein paar Individuen betroffen sind, nicht geeignet. Des weiteren sind einfache Kopplungsanalysen zur Detektion von Co-Segregationen in Familien im Falle von Locus- und phänotypischer Heterogenität nicht sensitiv genug, im Gegensatz zu

NGS-Methoden, die das Potential besitzen, alle Arten von genetischen Variationen des humanen Genoms auf Basenpaarebene, in einem einzelnen Experiment, zu identifizieren (Wang et al., 2013).

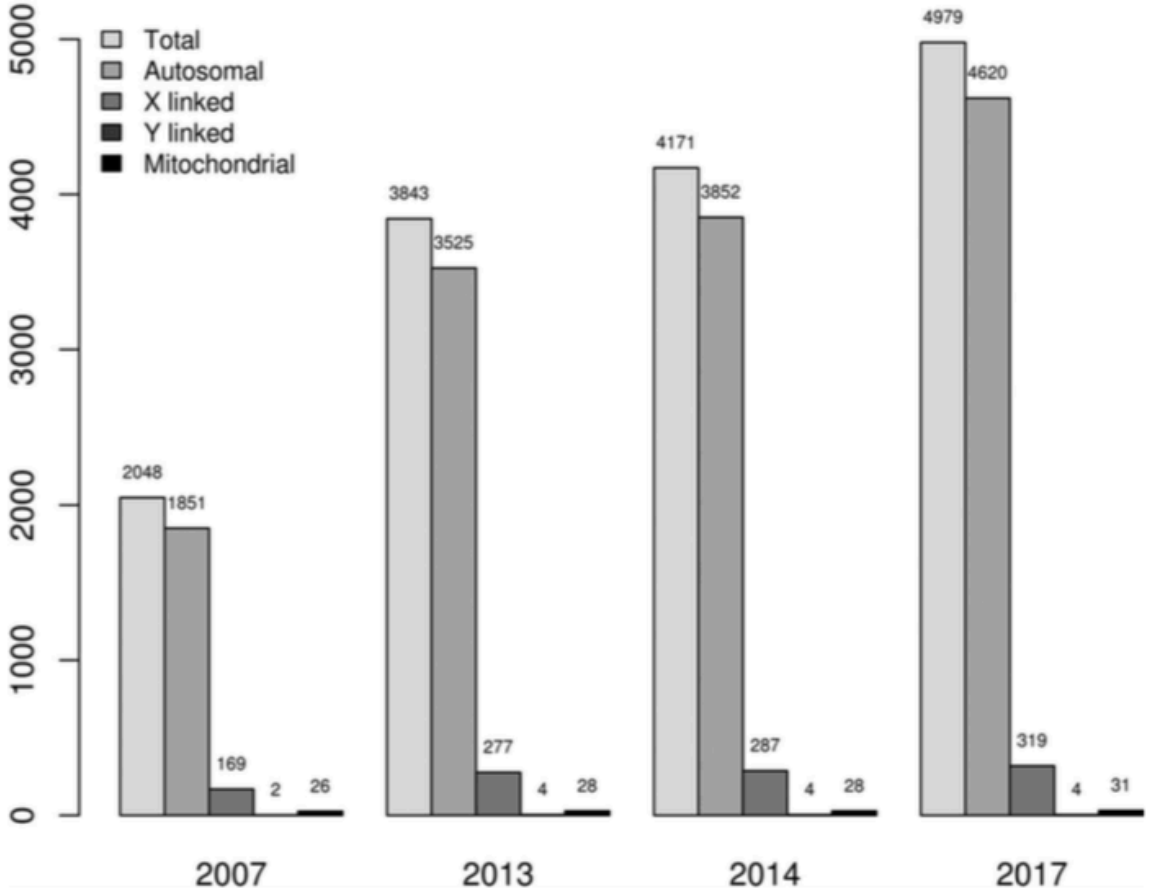

Abbildung 8: Steigende Anzahl der Einträge von entdeckten Genen in die OMIM-Datenbank, deren molekulare Grundlage mit einem bestimmten Phänotyp assoziiert wird, nach Einführung von Next Generation Sequencing (Fernandez-Marmiesse et al., 2018).

Die Identifizierung neuer Gene, die mit humanen Erkrankungen oder bekannte Gene, die mit neuen Phänotypen assoziiert sind, bildet die biologische Grundlage für das Verständnis einer Erkrankung und ist wichtig für das Patientenmanagement und in manchen Fällen sogar für die therapeutische Behandlung (Fernandez-Marmiesse et al., 2018).

Die NGS-basierten Testmethoden für die Erforschung seltener Erkrankungen können nach Kosten, Leichtigkeit der Analyse und Umfang in 3 Kategorien eingeteilt werden: (1) Parallele Sequenzierung von codierenden Sequenzen (Exons) von Gruppen von

Genen verbunden durch ähnliche oder überlappende Phänotypen (**Gene Panels**), (2) **Whole Exome Sequencing** (WES), bei dem alle bekannten codierenden Regionen des humanen Genoms sequenziert werden und (3) **Whole Genome Sequencing** (WGS), wo das komplette humane Genom sequenziert wird (Fernandez-Marmiesse et al., 2018).

Für die Isolierung von krankheitsverursachenden Genen erfolgt entweder eine Analyse des Exoms durch WES oder Genoms durch WGS einer Gruppe von Patienten mit denselben klinischen Charakteristika, gefolgt von einer Filterung von Varianten, die einem häufigen Gen in manchen oder allen Mitgliedern der Gruppe zugeordnet werden können, oder die Analyse von einzelnen Patienten mit ihren Eltern und/oder aussagekräftigen Mitgliedern der Familie und Filterung der Varianten auf Basis der unterschiedlichen Vererbungsmodi (autosomal dominant, rezessiv, x-linked, oder *de novo*), um die Menge an Varianten soweit zu reduzieren, dass nur eine kleine Anzahl übrig bleibt, die eine Identifikation des kausalen Gens erlaubt (Fernandez-Marmiesse et al., 2018).

Whole Exome Sequencing wurde als Standardmethode in der genetischen Forschung angenommen (Lelieveld et al., 2015) und dominiert die neueren Jahre der Erforschung von seltenen Erkrankungen, denn obwohl es nur 1% (□30Mb) des humanen Genoms abdeckt, welches in Protein translatiert wird, stellt sie im Gegensatz zum WGS eine deutlich höhere in Bezug auf Kosten und Zeit effiziente Methode zum Sammeln und Analysieren von genomischen Daten dar (Fernandez-Marmiesse et al., 2018) und besitzt eine hohe Genauigkeit für die Detektion von SNVs und kurzen Indels (Vijay et al., 2016). Im Durchschnitt wurden 45000 SNVs durch WES erhalten, wobei 39% in codierenden, 4% in untranslatierten Regionen (UTR) und 56% in intronischen Regionen nahe UTRs gefunden wurden (Requena et al., 2017). Es wird geschätzt, dass das Exom 85% der Mutationen beherbergt, die einen großen Effekt auf krankheitsverursachende Merkmale haben. Der Großteil von monogenen Krankheiten wird durch exonische Mutationen verursacht, wobei 60% allein zu den Missense-und Nonsense-Mutationen gehören. Des weiteren wurden durch großangelegte Genom- und Exom- Sequenzierungsprojekte nicht nur essentielle Informationen zur Varianten-Frequenz in bestimmten Populationen gesammelt, sondern konnte auch gezeigt wer-

den, dass ein typisches humanes Genom ungefähr 100 echte Loss-of-Function Varianten beherbergt, die circa 20 Gene komplett inaktivieren. Der Fokus auf diese „Healthy Human Knockouts“ könnte in der Zukunft helfen, den wahren Effekt von Varianten zu detektieren, der als krankheitsverursachend angenommen wurde und das Phänomen der Belastbarkeit von Genen aufzuklären. Auch die pharmakogenomische Forschung wird zunehmend thematisiert, da es nicht nur möglich ist, genetische Ursachen zu detektieren, warum manche Patienten auf gewisse Medikamente nicht ansprechen, sondern auch der Versuch der Vorhersage des Erfolgs des Ansprechens eines Medikamentes basierend auf genetische Informationen (Petersen et al., 2017).

Aktuelles klinisches Next Generation Sequencing basiert auf der Verwendung von Gene Panels und Exomanalysen mit vorangegangenem selektivem Erfassen (Capture) der Zielregionen. Gene Panels, die einem eines Patienten bestimmten Phänotyps zugehörig sind, stellen die erste schnelle und kostengünstige Ebene der diagnostischen Testung dar (Meienberg et al., 2016). Gene Panels bieten viel schnellere Turnaround Times, weniger Zufallsbefunde und eine höhere Abdeckung und erhöhen so die Chancen der Detektion von CNVs und somatischen Mosaiken im Gegensatz zu WES, sind jedoch nicht in der Lage neue krankheitsverursachende Gene zu identifizieren (Fernandez-Marmiesse et al., 2018).

Wenn sie negativ sind, kann in einem zweiten Schritt ein all umfassendes WES durchgeführt werden, welches attraktiv für klinische Applikationen ist (Suwinski et al., 2019). Dieser Schritt des Erfassens der Zielregion führt zu Beschränkungen bezüglich ausreichender Abdeckung von codierenden Exons, insbesondere GC-reiche Regionen (Meienberg et al., 2016). Daraus folgt, dass die Umsetzung im klinischen Setting aufgrund der verringerten Sensitivität verglichen mit der Sanger-Sequenzierung viel langsamer passiert, da für klinische Applikationen der Anspruch in puncto Robustheit und Qualität viel höher ist als im forschenden Anwendungsgebiet und es für neue klinische Tests bezüglich Spezifität und Sensitivität erforderlich ist, dass sie gleich oder besser als bestehende Tests funktionieren. WES ist exomweit sehr sensitiv, besitzt aber an bestimmten Regionen aufgrund von locusspezifischen Eigenschaften und Sequenzier-Bias eine niedrige Sensitivität. Manche dieser falsch-negativen Ergebnisse können durch Feintuning und oder Verbesserungen von

Mapping und Variant Calling reduziert werden, der Hauptgrund für das mangelnde Auffinden von Varianten ist das Fehlen einer ausreichenden Sequenzabdeckung, die auch durch verbesserte Algorithmen nicht gelöst werden kann (Lelieveld et al., 2015). Whole Genome Sequencing erlaubt die Sequenzierung des gesamten humanen Genoms und bietet als PCR-freie Applikation eine bisher beispiellos vollständige Abdeckung der codierenden Regionen des Genoms (Low-Coverage durch Schwierigkeiten beim Capturing und Sequenzieren von extremen GC-Gehalten). Daher muss WGS aus klinischer und technischer Perspektive in der Zukunft als ernsthafte Alternative zum WES bei der genomischen Testung von Mendelsche Störungen angesehen werden. WGS führt zu einer genomweiten Read Coverage und erlaubt so eine vertrauenswürdige Detektion von CNVs, die maßgebend an der Krankheitslast beteiligt sind. Durch das Sinken von Kosten und der Reduzierung von Turnaround Times, einschließlich der Datenanalyse durch Verwendung von speziellen Tools (z.B. GENALICE MAP), auf ein paar Tage, kann die diagnostische Ausbeute, durch Verwendung von virtuellen Gene Panels in silico zur Vermeidung von unerwünschten Ergebnissen, auf 73% gesteigert werden. WGS übertrifft die konventionelle auf den Phänotyp gerichtete Einzelgenanalyse um eine Zehnerpotenz (Meienberg et al., 2016). Bei der Detektion von Varianten übertrifft das WGS das WES an Leistung in Bezug auf abgedeckte codierende Regionen. Bei einer hohen Sequenziertiefe (95x-160x) erfasst WGS 98% der codierenden Regionen mit einer 87fachen Abdeckung verglichen mit den 95% des WES mit der minimalsten 20fachen Abdeckung. Ein weiterer Vorteil des WGS ist, neben der höheren Coverage aufgrund der fehlenden Beteiligung von Capture-Methoden, die Sequenzierungs-Bias einführen können, die Fähigkeit auch nicht-codierende DNA zu sequenzieren und machen so das WGS zu einem scheinbar geeigneteren Test für klinische Applikationen. Um dieses Problem der fehlenden Sequenziertiefe bei schwierigen Regionen zu umgehen, werden bei heutigen Exom-Anreicherungsdesigns von WES Proben generiert, die nahe bei der interessierenden Region liegen und mithilfe von Paired-End Reads auch diese benachbarten Regionen der Zielregion sequenziert (Lelieveld et al., 2015).

6.2. Strategien zur Variantenfindung bei der Exomsequenzierung

Die anfängliche Priorisierung ist der erste Schritt der Variantenfindung und führt selten zur Identifizierung der pathogenen Variante an sich. Um die verursachende Mutation aus der Liste von ungefähr 150-500 verbliebenen Varianten zu finden, benötigt es zusätzliche Strategien, die vom Traditional Mapping und anderen häufigen Ansätzen auf die Exomsequenzierung adaptiert wurden (Gilissen et al., 2012).

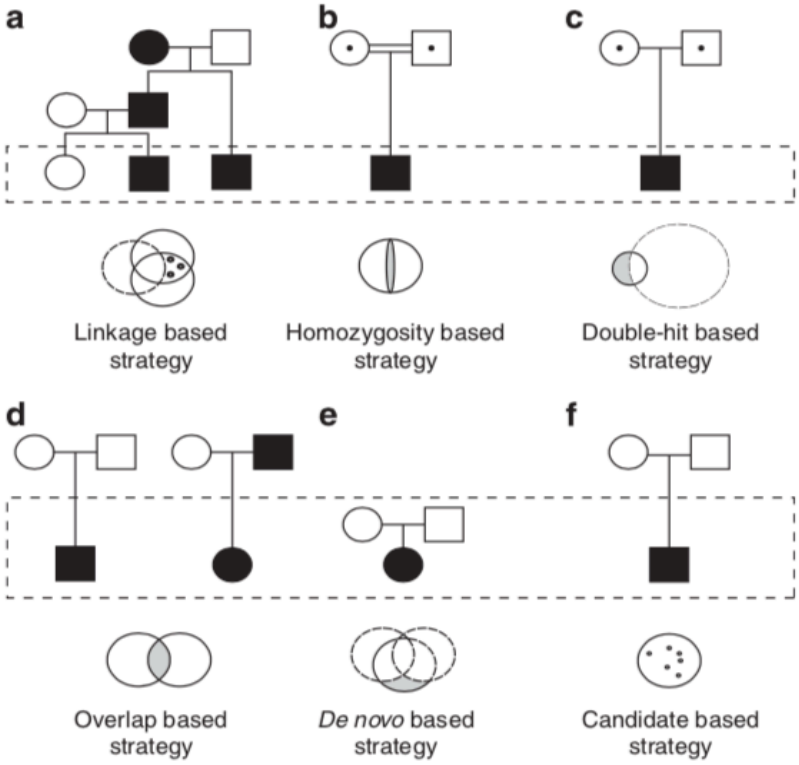

Abbildung 9: Strategien der Krankheitsgenfindung für die Exomsequenzierung (Gilissen et al., 2012).

Die spezifische Verwendung der einzelnen Strategien, für die Identifizierung eines Gens welches für Einzelgen-Defekte verantwortlich ist, erfolgt abhängig von der vorhandenen Information wie Vererbung, Stammbaum und genetische Heterogenität (Kayman Kurekci and Dincer, 2015).

Die *Kopplungsstrategie* beschreibt die Identifizierung von gemeinsamen Variationen in einer mehrfach mit einer monogenen vererbten Erkrankung betroffenen Familie durch Sequenzierung. Des Weiteren können nicht betroffene Familienmitglieder sequenziert werden um benigne Variationen auszuschließen (Gilissen et al., 2012), die

gesunden Individuen fungieren dabei als Kontrollen, um die Kandidaten zu verifizieren (Kayman Kurekci and Dincer, 2015). Die Auswahl der genetisch am weitesten voneinander betroffenen Familienmitglieder reduziert die Auswahl der gemeinsamen benignen Variationen auf ein Minimum. Das Kombinieren von Sequenzdaten zweier betroffener Geschwister, die 50% der DNA teilen, führt zu einer ähnlichen Reduktion der Varianten die in Betracht gezogen werden. So konnten Kravitz *et al.* den Haplotyp für alle Varianten von drei betroffenen Geschwistern bestimmen und so die Anzahl der Kandidatengene von 14 auf 2 reduzieren (Gilissen et al., 2012).

Die *Homozygotie-Kartierung* beruht auf der Annahme, dass bei einer rezessiv vererbten Erkrankung und vermuteten Konsanguinität, die Krankheit durch eine homozygote Variante verursacht wird, die von beiden Eltern vererbt wurde und auf einem großen Abschnitt einer homozygoten Region liegt (Gilissen et al., 2012). Die Detektion von homozygoten Regionen im Genom durch SNP-Array sind aufschlussreich und können die Anzahl an Varianten einer Familie mit rezessiver Vererbung reduzieren, da nur jene Varianten in den homozygoten Regionen für die Pathogenität verantwortlich sind. Gemeinsame homozygote Varianten werden dazu verwendet, die Kandidatenvariante zu finden. Sirmaci *et al.* konnten auf diese Weise die Ursache des Malpuech-Michels-Mingarelli-Carnevale Syndrom in zwei betroffenen Familien identifizieren. Die homozygote Region wurde dabei auf Chromosom 3q27 identifiziert und dabei eine *MASP1*-Mutation, die mit dem Phänotyp co-segregiert, mittels Exomsequenzierung bestätigt (Rabbani et al., 2012). In einer weiteren Studie konnten Becker *et al.* 17 der 318 nicht-synonymen Varianten durch Exomsequenzierung als autosomal homozygote Varianten identifizieren, wobei nur 3 davon in großen homozygoten Regionen lokalisiert waren und die verursachende Mutation in der größten Region entdeckt wurde (Gilissen et al., 2012).
Der Hauptunterschied der Homozygosity-Kartierung im Vergleich zur Linkage Strategy ist, dass Varianten, obwohl sie homozygot sind, nur dann detektiert werden, wenn sie sich in einem großen homozygoten Areal befinden. Durch diese Strategie wird die Anzahl der Varianten reduziert und sie erlaubt die Identifikation des Krankheitsgens in individuellen Fällen ohne zusätzliche Familienmitglieder, kann homozygote Loci aber

nur in Regionen detektieren, die eine ausreichende Dichte an informativen SNPs beinhalten (Gilissen et al., 2012).

Die *Double Hit Strategy* wird verwendet wenn einzelne Personen von einer rezessiven Erkrankung betroffen sind, aber keine ausreichende Anzahl an Familienmitglieder vorhanden ist. Hierbei werden homozygote und compound-heterozygote Varianten eben als Double Hit Strategy gesucht (Rabbani et al., 2012). Gilissen *et al.* konnte mit dieser Strategie die Anzahl von Kandidatengenen in zwei Patienten mit Sensenbrenner-Syndrom von 139 und 158 auf 3 beziehungsweise 4 und Pierce et al. bei einem Patienten mit Perrault-Syndrom von 207 nicht-synonymen Varianten auf 1 reduzieren. Durch diese Studien konnte somit gezeigt werden, dass diese Strategie sehr wirksam für die Identifizierung der genetischen Ursache einer Störung durch Sequenzierung eines einzelnen Individuums ist (Gilissen et al., 2012).

Die *Overlap Strategy* wird für nicht verwandte sporadische Individuen verwendet und führt zur Bestimmung der gemeinsamen geteilten mit der Krankheit assoziierten Variante (Kayman Kurekci and Dincer, 2015). Diese Strategie wird zur Suche nach Einzelgenmutationen in mehreren nichtverwandten Patienten mit ähnlichem Phänotyp in Abwesenheit von genetischer Heterogenität angewendet und ist wichtig für die Krankheitsgenfindung in dominanten Erkrankungen, da es viel mehr heterozygote nicht-synonyme als homozygote nicht-synonyme Varianten gibt. Durch das Kombinieren der Daten von einer zunehmenden Anzahl an Patienten verringert sich die Zahl an Genen mit Mutationen in mehreren betroffenen Patienten und resultiert in weniger Kandidatengene zur Weiterverfolgung. Hoischen *et al.* konnten mit dieser Strategie das Gen für das Bohring-Optitz-Syndrom durch Kombination der Daten von 3 Individuen identifizieren. In jedem dieser Patienten wurden zwischen 130 und 222 neue nicht-synonyme Varianten gefunden, die durch die Überlappung von je zwei beliebigen Patienten auf 26 Kandidatengene und durch Überlappung der Daten aller 3 Patienten auf 1 Einzelgen reduziert wurden. Ng *et al.* konnten im Falle dies Kabuki- Syndroms, einer seltenen weltweiten Erkrankung, die in den meisten Fällen sporadisch auftritt und dominant vererbt wird, die kausale Variante im *MLL2*-Gen in 7 von 10

Familien identifizieren und durch Sanger-Sequenzierung eine Frameshift Indel in weiteren 2 Familien aufdecken, welche von der Exomsequenzierung nicht erkannt wurde. Des weiteren können auch Translokation, Inversionen und CNVs mit der identifizierten Variante, in Patienten mit demselben Phänotyp, überlappen und dieselbe Krankheit auslösen, welche von der Sequenzierung unerkannt bleiben. So konnten Ng *et al.* in einer Studie mit einer Kohorte von 43 Patienten in nur 26 Patienten (60%) eine MLL2 Mutation nachweisen. Der Unterschied in der Prozentzahl an positiven MLL2 Mutationen bei der Entdeckung und nachfolgenden Kohortenstudie ergibt sich folglich aufgrund der klinischen und genetischen Heterogenität des Kabuki-Syndroms und unterstreicht einmal mehr die Wichtigkeit der genauen und einheitlichen klinischen Phänotypisierung für eine erfolgreiche Krankheitsgenfindung mittels Overlap Strategy.
Diese Strategie ist die erste für die Identifizierung eines mutierten Gens bei einer dominante Erkrankung durch Verwendung des Overlap-Ansatzes in nicht verwandten Personen (Gilissen et al., 2012; Rabbani et al., 2012).

Bei der *De Novo Strategy* werden die Exome eines Trios bestehend aus einem betroffenen Kind und seinen Eltern sequenziert und auf Varianten fokussiert, die beim Patienten aber nicht bei den Eltern detektiert werden. Eine beträchtliche Menge der *de novo* Mutationen treten sporadisch in Fällen auf, meistens Föten, die nicht überleben und diese letalen Mutationen scheiden aus der Population aus und werden deshalb nicht identifiziert. Die Mutationsrate wird auf $7{,}6 \times 10^{-6} – 2{,}2 \times 10^{-8}$ geschätzt (Kayman Kurekci and Dincer, 2015). Bei der Priorisierung werden zuerst die nichtpathogenen Varianten gefiltert und danach die Varianten, die in den Eltern vorhanden sind, ausgeschlossen (Rabbani et al., 2012).
Übrig bleiben eine begrenzte Anzahl an potentiell pathogenen Varianten, da das durchschnittliche Exom nur 0-3 *de novo* Mutationen aufweist. Besonders wenn eine Störung hauptsächlich sporadisch und mit einer reduzierten Fruchtbarkeit auftritt, wie es bei intellektueller Beeinträchtigung der Fall ist, könnte der Grund in *de novo* Mutationen liegen. Im Gegensatz zur Overlap Strategy, die nur bei seltenen Erkrankungen funktioniert, die hauptsächlich monogen sind, ist das Mutational Target, also die

Menge im Genom besetzt mit Genen die wenn sie mutieren zur Krankheit führen, zu groß um zwei Patienten mit Mutationen in denselben Gen zu finden. Diese großen Abschnitte führen auch zu einer ansteigenden Wahrscheinlichkeit, dass eine de novo Mutation während der Meiose in einen der Genen passiert und eine Krankheit verursacht (Gilissen et al., 2012). Aufgrund von Sequenzierfehlern und Mapping-Artefakte sollte eine Bestätigung durch Sanger-Sequenzierung mit hoher Genauigkeit erfolgen, wobei die Detektion der *de novo* Mutation nicht ausreicht um die Ursächlichkeit der Krankheit zu bestätigen. Weitere Wiederholungen und funktionelle Analysen sollten durchgeführt werden, um die kausale oder schädliche Variante zu bestimmen, da die Pathogenität einer Variante nicht nur vom Typ und der Lokalisation der Mutation, sondern auch von seinen funktionalen Effekten abhängt (Rabbani et al., 2012). Die Anreicherung und Sequenzierung aller Proben eines Trios sollte im selben Experiment erfolgen, wobei die *De Novo* Strategy ein relativ hohe Anzahl an Experimenten mit jeweils 3 Proben des Trios benötigt und deshalb nur angewendet wenn keine der anderen Strategien erfolgreich ist und elterliche Proben vorhanden sind (Gilissen et al., 2012). Die Trio-based Exomsequenzierung ist eine wirksame Methode, um neue verursachende Gene für sporadisches Autismus-Spektrum zu identifizieren. Iossifov *et al.* haben 343 Familien mit jeweils einer einzelnen von der Autismus-Spektrum-Störung betroffenen Person und zumindest einem nicht betroffenen Bruder oder Schwester sequenziert und analysiert. Das Ergebnis dieser Studie waren Gendisruptionen, aber keine Missense Mutationen in den betroffenen Kindern und das Aufdecken von 350-400 Genen, die eine Suszeptibilität für Autismus aufweisen (Rabbani et al., 2012).

Die *Candidate Strategy* wird bei einer einzelnen dominant betroffenen Person ohne vorliegen von Familienmitgliedern oder weiteren betroffenen Patienten verwendet. Aufgrund der Limitationen erfolgt die Priorisierung auf Basis der vorhergesagten Wirkung der Variante auf die Struktur und Funktion des Proteins. Dabei wird auf Stop- und Frameshift-Mutationen sowie kanonischen Splice-Site-Mutationen priorisiert. In Bezug auf Missense-Varianten wird entweder auf die Wirkung auf die Proteinstruktur (Grantham Score) oder evolutionäre Konservierung der Nukleotidvariante (PhyloP-

oder GERP Score) analysiert, da pathogene Varianten dazu tendieren, eine deutlich höhere Konservierung aufzuweisen, als benigne Varianten. Mithilfe der dbSNP-Datenbank oder HGMD können die Scores von Varianten verglichen werden, wobei der PhyloP-Score für eine pathogene Variante bei > 2,5 liegt (Gilissen et al., 2012).
Neuere Ansätze bedienen sich der Genfunktion in Beziehung mit dem Phänotyp und der bekannten Pathophysiologie zur Priorisierung unter Verwendung von Computional Predictors, also Vorhersageprogramme, beim fehlenden Vorliegen von eindeutigen Varianten (z.B. trunkierende Mutationen). Ehrlich *et al.* konnten mithilfe von bioinformatischen Tools (SUSPECTS, ToppGene und Endeavour) und traditionellem Mapping-Ansatz KIF1A als wahrscheinlichste pathogene Variante für hereditäre spastische Paraparese priorisieren. Die Variante selbst wurde unabhängig durch Verwendung von 3 verschiedenen Prediction Tools (MutationTaster, Polyphen, SIFT) identifiziert. Die Kombination aus Genlevel-und genomischer Varianteninformation besitzt demnach hohes Potential für die Priorisierung von pathogenen Varianten (Gilissen et al., 2012).
Praktisch wird in Studien eine Kombination aus verschiedenen Strategien verwendet deren finaler Schritt für die Validierung die traditionelle Sanger-Sequenzierung, als Goldstandard, für die Detektion der Mutation ist. Für einen definitiven Beweis der Pathogenität einer Variante bedarf es einer Validierung in weiteren unabhängigen Kohorten und/oder funktionellen Analysen (Gilissen et al., 2012).

6.3. Transkriptom-Analyse – RNA-Seq

Für das Verstehen von phänotypischen Variationen ist die genaue Identifizierung von differentiell exprimierten Genen (Differentially Expressed Genes, DEGs) unter spezifischen Bedingungen entscheidend. Die Methode erster Wahl für Genexpressionsstudien ist heutzutage das High-Throughput Transcriptome Sequencing (RNA-Seq) (Costa-Silva et al., 2017), das sich durch die hohe Sensitivität, Genauigkeit, Reproduzierbarkeit und Flexibilität als Goldstandard in der Transkriptomforschung etabliert hat. In den letzten 10 Jahren wurden dadurch über 53.700 Datensätze in der Gene Expression Omnibus (GEO) Datenbank hinterlegt, welche Informationen über das gesamte Transkriptom inklusive differentielle Expression von codierenden und nicht-

codierenden Genen, die Entdeckung von non-coding RNAs wie microRNAs, long non-coding RNAs (lncRNAs) Genfusionen und Splice- Varianten in unterschiedlichen experimentellen Zuständen beinhalten. Immer mehr Anzeichen sprechen dafür, dass Veränderungen im Transkriptom die Folge biologischer Veränderungen sind und das RNA-Seq die treibende Kraft hinter der Erforschung dieser regulatorischen Netzwerke in Zellen, Geweben, Organismen und Krankheiten ist (Chao et al., 2019), welches für mehrere spezifische Applikationen, wie Traditionel Transcriptome Profiling, Genfusions-Events, Identifikation von neuen Transkripten/ exprimierten SNPs oder alternatives Splicing, verwendet werden kann. Im Vergleich zu anderen auf Hybridisierung basierenden Technologien wie dem DNA-Microarray und der quantitativen reversen Transkription bietet das RNA-Seq eine konstante Quantifizierung (Zhao et al., 2014) und ist nicht von Genomannotationen und vordefinierten spezies-spezifischen Proben zur Transkriptmessung abhängig sondern erlaubt die Detektion von bekannten sowie neuen Transkripten einschließlich Varianten und seltenen Transkripten. Durch diese Technologie lassen sich zigtausende unterschiedlich exprimierte Gene und Isoformen sowie Mutationen und Keimbahnvariationen von genetisch exprimierten Varianten identifizieren (Palomares et al., 2019).

Am Beginn der RNA-Seq Technologie werden die RNA-Proben fragmentiert, in kleine komplementäre DNA-Sequenzen umgeschrieben (cDNA) und danach von einer High-Throughput Plattform sequenziert (Costa-Silva et al., 2017).

Obwohl die direkte Sequenzierung von RNA-Molekülen möglich ist, werden die meisten RNA-Seq-Experimente auf Geräten durchgeführt die DNA-Moleküle sequenzieren. Dies liegt daran, dass kommerzielle Geräte, die für die DNA-Sequenzierung entwickelt wurden, eine größere technische Reife besitzen. Die cDNA-Library ist ein essentieller Schritt des RNA-Seq, wobei jedes cDNA-Insert eine bestimmte Größe besitzt und von Adaptersequenzen flankiert ist, um die Amplifikation und Sequenzierung auf spezifischen Plattformen durchführen zu können. Die cDNA Library Preparation variiert je nach untersuchter RNA-Spezies und kann sich in Bezug auf Größe, Sequenz, strukturelle Eigenschaften und Häufigkeit unterscheiden (Hrdlickova et al., 2017).

Die mRNA muss für die Fragmentierung angereichert werden, da das gesamte RNA-Extrakt typischerweise zu 98% aus rRNA besteht. Die Anreicherung der Transkripte kann entweder durch die Poly(A)-Affinitätsmethode oder durch Depletion von ribosomaler RNA durch Verwendung von sequenzspezifischen Proben erfolgen (Lowe et al., 2017). Der Poly(A)-Tail ist an den meisten mRNAs und lncRNAs von eukaryotischen Organismen vorhanden und kann als technisches Hilfsmittel für die Anreicherung verwendet werden. Die Auslese der RNA mit Poly(A)-Tail wird mithilfe von magnetischen oder Zellulose-Kügelchen, die mit Oligo(dT)-Primer beschichtet sind, durchgeführt. Die Poly(A)-Aufreinigung ist die bevorzugte Variante außer wenn nur ein sehr niedriger RNA-Gehalt verfügbar ist (Hrdlickova et al., 2017). Wenn als Startmaterial Formalin-fixierte oder in Paraffin eingebettete Proben verwendet werden, ist die mRNA in vielen Fällen stark degradiert und aus diesem Grund gibt es mehrere rRNA-Depletion Protokolle. Die Ribo-Zero Methode entfernt rRNA durch Hybridisation Capture und nachfolgender Bindung an magnetische Kügelchen (Zhao et al., 2014).

Nach der Sequenzierung werden die kurzen generierten Sequenzen gegen ein Genom oder Transkriptom gemapped (Costa-Silva et al., 2017). Die Nukleotidsequenzen besitzen dabei ein Länge von ungefähr 100 bp, können aber auch zwischen 30 bp und 10000 bp schwanken. RNA-Seq erlaubt die rechnerische Rekonstruktion des originalen Transkripts aus vielen kleinen Fragmenten des Transkriptoms durch Alignment der Reads gegen das Referenzgenom oder gegeneinander (*de novo* Assembly). Die Menge an RNA die eingesetzt wird, ist für RNA-Seq um einiges geringer (Nanogrammbereich) als beim Microarray (Mikrogrammbereich) und dies erlaubt eine bessere Untersuchung von zellulären Strukturen bis hin zum Einzelzelllevel (Lowe et al., 2017).

Im nächsten Schritt wird das Expressionslevel für jedes Gen oder jede Isoform geschätzt (Costa-Silva et al., 2017), also identifiziert, welche Gene an einem bestimmten Zeitpunkt aktiv sind und die Read Counts dazu verwendet, um das relative Genexpressionslevel genau abzubilden. Die Fortschritte in der RNA Seq Library Preparation aufgrund der Suche nach Transkriptomdaten für Einzelzellen haben in einer erhöhten Sensitivität resultiert. So ist die Einzelzell-Transkriptomanalyse inzwischen sehr gut beschrieben und findet auch im in situ RNA-Seq ihre Anwendung, bei dem

die Transkriptome von individuellen Zellen direkt in fixierten Geweben untersucht werden (Lowe et al., 2017).

Die RNA-Sequenzierung von Einzelzellen erlaubt die Messung von biologischen Variationen in einer heterogenen zellulären Population und die Dissektion der Tranksriptomkomplexität, die in der Gesamtmessung der Genexpression maskiert ist. Mit der Entwicklung einer mikrofluidischen Methode, um cDNA aus Einzelzellen für das High-Throughput Whole Transcriptome-Sequencing vorzubereiten, ermöglicht diese Plattform die Manipulation von Einzelzellen, minimiert Kontaminationen und bietet eine erhöhte Sensitivität und Messgenauigkeit, welche wichtig für die Unterscheidung von biologischer Variabilität und technischem Rauschen ist (Streets et al., 2014).

RNA Expression Profiling kann zur Entdeckung von molekularen Markern für Krankheitsdiagnose, Prognose und Drug-Targeting führen (Li et al., 2014) und findet seine Anwendung auch in der Messung der Genexpression und Identifizierung von strukturellen Variationen wie Indels und Fusionstranskripte in somatischen Driver-Mutationen von Tumoren (Sun et al., 2016).

RNA-Seq Technologien werden auch zunehmend für klinische Applikationen verwendet. So unterstützen die neuesten Brustkrebs Guidelines die Verwendung von mRNA-basierenden prognostischen Assays, um der Behandlungsentscheidung neben anderen klinisch-pathologischen Faktoren zu assistieren. Diese Assays bieten des Weiteren einen besseren Blick auf alternative Transkripte die durch Splicing-Veränderungen oder strukturellen Varianten aufkommen und in einer Bandbreite humaner Erkrankungen wie Entwicklungsstörungen, neurodegenerative Erkrankungen und Krebs verwickelt sind. RNA-Seq wird in der Zukunft von einem Discovery Tool zu einem Diagnostic Tool mit klinischen Applikationen wie Patienten-Stratifizierung, Diagnose und personalisierte Behandlung (Palomares et al., 2019).

6.4. Epigenom-Analyse

Der Fortschritt von molekularbiologischen Techniken betrifft neben den Bereichen Whole Exome Sequencing, Whole Genome Sequencing und Expression Profiling auch die Untersuchung des Epigenoms durch eine Vielzahl von unterschiedlichen Sequenziertechnologien wie CHiP-Seq, DNase-Seq, ATAC-Seq, MeDIP-Seq, MRE-

Seq oder Bisulfite-Seq. Die Epigenetik kann als Mechanismus definiert werden, der zu erblichen Veränderungen in Genfunktionen führt, ohne dass die Sequenz des Genoms betroffen ist (Chen and Li, 2016). Einer der wichtigsten Aspekte der epigenetische Modifikation ist die DNA-Methylierung, die eine Schlüsselrolle in der Regulation der Genexpression einnimmt. Sie kommt am Häufigsten an der C5-Position von Cytosinen (5mC) innerhalb von CpG-Dinukleotiden, aber auch non-CpG-Cytosinen in embryonalen Stammzellen vor. DNA-Methylierung ist in einer Vielzahl von biologischen Prozessen wie embryonale Entwicklung, X-Chromosom-Inaktivierung, Genomic Imprinting und dem Silencing von transposablen Elementen involviert und führt aberrant zur Pathogenese und Progression von zahlreichen Krankheiten wie Krebs und immunologische Störungen. Darum ist die Detektion dieser 5mC-Stellen und das Einschätzen ihrer DNA-Methylierungslevel von großer Bedeutung und kann, durch Verstehen der Methylierungsmuster, zur Identifizierung von krankheitsassoziierten Genen und potentiellen Drug Targets führen (Jia et al., 2018).

Methylation dependant Immunoprecipitation-Sequencing (MeDIP-Seq) verwendet einen Anti-Methyl-Cytosin-Antikörper, um methylierte DNA-Fragmente anzureichern und danach zu sequenzieren wohingegen das Methylation sensitive Restriction Enzyme Digestion-Sequencing (MRE-Seq) auf einer Sammlung von Restriktionsenzymen beruht, die CpG-enthaltene Sequenzmotive erkennen, aber nur schneiden, wenn das CpG unmethyliert ist. Die verdauten DNA-Fragmente mit den unmethylierten CpGs an ihren Enden werden angereichert und die CpGs durch Sequenzierung aufgedeckt. Diese zwei Methoden bilden zusammen eine robuste effektive und erschwingliche Plattform, für die Erforschung genomweiter DNA-Methylierungen (Xing et al., 2018).

Der Goldstandard für die Detektion von Methylierungen ist das Whole Genome Bisulfite Sequencing (WGBS-Seq) aufgrund seiner Fähigkeit 5-modifizierte Cytosine mit hoher Auflösung zu mappen. Dabei werden unmethylierten Cytosine durch Bisulfitbehandlung in Uracil umgewandelt, so dass nur methylierte Cytosinreste detektiert werden. Es ist jedoch sehr laborintensiv und benötigt eine 30fache Coverage, die für die Routineanalyse vieler Proben, insbesondere solcher mit großen Methylomen (z.B. humane) noch immer sehr teuer ist (Li et al., 2015; Xing et al., 2018).

Chromatin Immunoprecipitation-Sequencing (ChIP-Seq) kann die Assoziation von DNA-wechselwirkenden Proteinen wie RNA-Polymerase, Nukleosomen oder Transkriptionsfaktoren quantifizieren und jeder Position im Genom zuordnen. Das Prinzip erfolgt wieder durch Immunopräzipitation (Leleu et al., 2010), bei der die DNA, an der spezifische Proteine gebunden haben, durch Beschallung in Fragmente mit 300-1000 bp Länge zerteilt werden. Die spezifischen Antikörper erkennen nur das Protein (z.B. Transkriptionsfaktor) von Interesse und der Antikörper, der an diesen TF gebunden ist, welcher wiederum an die DNA gebunden ist, erlaubt die selektive Gewinnung und Isolierung des Chromatinkomplexes. Bei diesem Vorgang wird die DNA vom TF freigesetzt danach aufgereinigt und das Resultat ist eine, der genomischen Region entsprechende, angereicherte DNA-Probe die danach sequenziert werden kann. Da das Experiment in tausenden Zellen zur selben Zeit abläuft, hat man eine genügende Menge an DNA für weitere Analysen aber auch genügend Anreicherung in der Probe, also Kopien der an den TF gebundenen DNA, um sie von experimentellen Störgeräuschen zu unterscheiden (Pavesi, 2016).

DNase-Seq wird dazu verwendet, um genomische Regionen mit regulatorischen Elementen wie Promotoren oder Enhancer zu identifizieren. Dazu spaltet die DNase I die DNA bevorzugt an „offenen“ Chromationregionen, die häufig als aktiv angesehen werden. Die Enden der durch DNase I –Verdau des Chromatins entstandenen Fragmente entsprechen den zugänglichen genomischen Regionen. Nach dem Verdau werden Adapter für die Library Preparation an die DNA-Fragmente ligiert und diese sequenziert. So ist es möglich genomische Regionen, die vor DNase I– Verdau durch den gebundenen Transkriptionsfaktor geschützt sind, also cis-Elemente von nahen Genen, zu identifizieren. Es ist auch möglich die Bindung von transkriptionsregulatorischen Proteinen durch die Suche nach Sequenzmotiven innerhalb von Transcription Factor Footprints vorherzusagen (Chaitankar et al., 2016). DNase-Seq wurde aber schnell durch das Assay for Transposon Accessible Chromatin-Sequencing (ATAC-Seq) ersetzt (Reuter et al., 2015). ATAC-Seq kann zur Identifizierung von Regionen von aktiven regulatorischen Regionen verwendet werden und ist aufgrund der benötigten 500-50.000 Zellen im Vergleich zu den 50 Millionen benötig-

ten Zellen des DNase-Seq eine schnellere und effizientere Alternative, die auch schon erfolgreich für das Single-Cell Open Chromatin-Profiling angewandt wurde (Chaitankar et al., 2016). Die Entdeckung von Biomarkern aus FFPE-Proben durch ATAC-Seq wäre wünschenswert, denn das würde es erlauben klinische Proben aus Biobanken in großangelegten ATAC-Seq Profiling-Studien einzubeziehen und obwohl die Verwendung von „Open“ Chromatin als epigenetischer Biomarker noch sehr selten ist, wird die Flexibilität und Leichtigkeit des ATAC-Seq die Verwendung von „Open“ Chromatin in der klinischen Forschung und Praxis ankurbeln (Dirks et al., 2016).

7. Diskussion und Zukunftsperspektiven

Das Hauptaugenmerk der Arbeit liegt im Vorstellen der einzelnen Plattformen und ihrer spezifischen Technologien dahinter, wobei bei den einzelnen Plattformen bewusst auf technische Spezifikationen mit Ausnahme von Abbildung 4 verzichtet wurde, um die Arbeit nicht mit Zahlen zu überladen, sowie dem Aufzeigen des breiten Spektrums an Möglichkeiten durch Next Generation Sequencing durch ausgewählte Applikationen in den Bereichen der Genom-, Transkriptom,- und Epigenom-Untersuchung zur Durchführung spezifischer Analysen. Des Weiteren wird in kompakten Zusammenfassungen einerseits auf den bioinformatischen Workflow nach der Sequenzierung und andererseits auf die verschiedenen Strategien der Variantenfindung bei der Exomsequenzierung eingegangen, die alleine oder in Kombination je nach Vorliegen spezifischer Informationen angewendet werden können. Die Second Generation Sequencing-Technologien, insbesondere Illumina und Ion Torrent haben die Sanger-Sequenzierung, durch die Möglichkeit der parallelen Sequenzierung von Millionen von Fragmenten, infolge der Erhöhung der Geschwindigkeit und Reduktion der Kosten pro Base abgelöst, wobei diese weiterhin als Goldstandard zur Verifizierung von Genen dient.

WES und Targeted Sequencing haben sich als Standardmethodik für die Identifizierung von genetischen Varianten/Mutationen in der klinischen Diagnostik etabliert (Chaitankar et al., 2016), werden aber zunehmend vom WGS als weitestgehend verwendete Applikation abgelöst, welche gewaltige Mengen an Genomsequenzen generiert, die zum Aufbau von Referenzgenomen dient. Das 1000 Genome Project hat

Genomdaten von 2504 Individuen, in 26 unterschiedlichen Populationen, innerhalb von 5 Jahren, generiert. Das neueste NGS-Instrument, der HiSeq X Ten von Illumina hat Forschern die Sequenzierung von 20.000 Genomen im Zuge des NHLBI TOPMed Programms in 2015 erlaubt und dieser aktuelle Trend in der humanen Genomforschung hat zur Initialisierung der Precision Medicine Initiative durch die US-Regierung und ähnlichen Projekten von Profit/ Nonprofit-Organisationen in unterschiedlichen Ländern geführt. Die Anhäufung von humanen genomischen Daten hat die Bedeutung eines geeigneten humanen Referenzgenoms als wichtigen Aspekt der personalisierten Medizin aufgezeigt und durch die Projekte 100.000 Genome Project und GenomeAsia 100K wird dieses Vorhaben in naher Zukunft beschleunigt. Des Weitern wurden Third Generation Sequencing Technologien eingeführt, die im Moment noch hauptsächlich auf Forschungslevel verwendet werden und den aktuellen NGS-Methoden helfend und ergänzend zur Verfügung stehen. Dennoch darf ein Paradigmenwechsel in der Genomforschung in der nahen Zukunft zu erwarten sein (Park and Kim, 2016.).

Trotz des aufkommenden Enthusiasmus für NGS und seinen Möglichkeiten gibt es dennoch Einschränkungen, denn obwohl die Methode über die nächsten Jahre noch verbessert werden wird, ist es mit der derzeitigen technologischen Basis unmöglich, eine 100% Abdeckung und Genauigkeit zu erreichen. Ebenfalls wird die korrekte Interpretation von detektierten Varianten eine Herausforderung darstellen, da es unmöglich ist alle genetischen Varianten, die zu einer bestimmten Krankheit oder einem bestimmten Merkmal beitragen, zu identifizieren. Hinzu kommt die Tatsache, dass nach einer genauen Filterung von seltenen und proteinverändernden Varianten 100-200 potentiell krankheitsverursachende Veränderungen in einem Exom übrigbleiben. Dazu kommen weitere 100 gesundheitsschädlichen Mutationen die Loss of Function in proteincodierenden Region verursachen, die jedes Individuum mit sich trägt, also auch pathogene Varianten, die nicht die Krankheit betreffen (Incidental Findings). Die Herausforderung besteht inzwischen daher eher in der Interpretation von Varianten als der Detektion von Varianten bei der Übersetzung von Sequenzinformationen in die klinische Praxis (Lohmann and Klein, 2014).

Daher ist die korrekte Annotation von Varianten in codierenden, nicht- codierenden und intergenischen Regionen von großer Bedeutung um deren funktionelle Bedeutung zu verstehen, ob es sich um krankheitsverursachende Varianten oder um einen zufälligen genetischen Drift handelt (Chakravorty and Hegde, 2017).

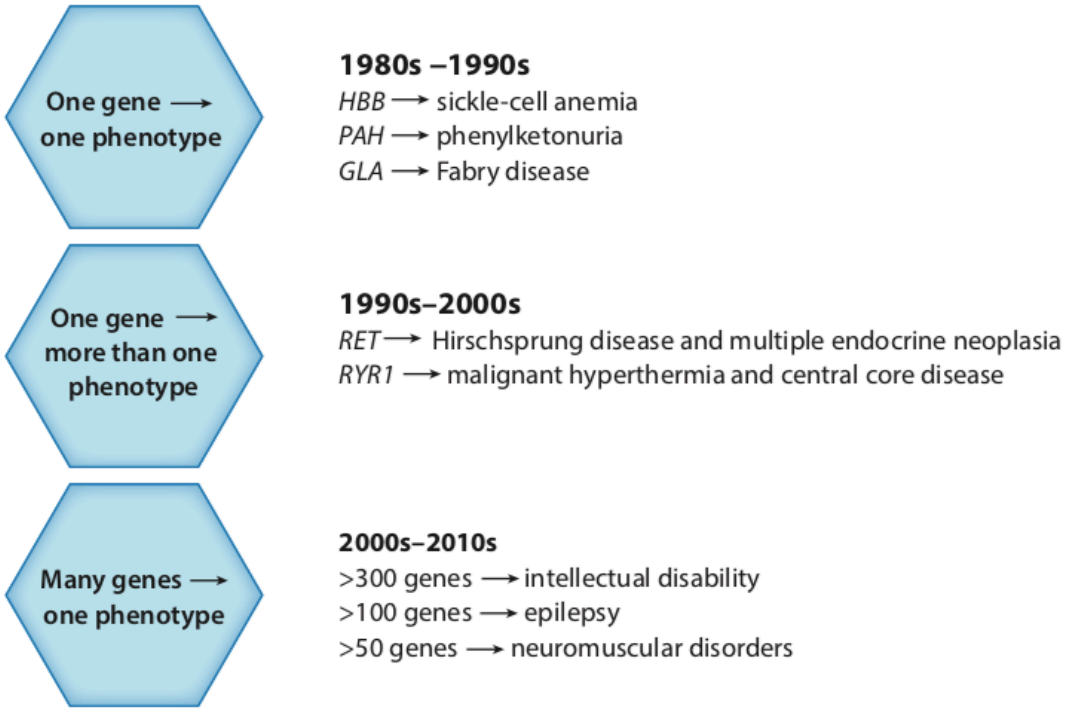

Abbildung 10: Entdeckung des Konzepts der Pleiotropie in Bezug auf Phänotyp-Genotyp-Assoziation (Chakravorty and Hegde, 2017).

Das Verständnis von Genotyp-Phänotyp-Korrelation und Gen-Krankheit-Assoziation hat sich in den letzten 30 Jahren extrem entwickelt, da die Forschung Beweise für die genetische Pleiotropie und multigene Effekte von Krankheiten gefunden hat und neue Belege darauf hindeuten, dass synergistische Effekte von heterozygoten Varianten in verschiedenen Genen den Krankheitsphänotyp beeinflussen können. So kann das *RET*-Gen einerseits für die Hirschsprung-Krankheit andererseits aber auch für die Multiple Endokrine Neoplasie verantwortlich sein. Es können aber auch ganze Sets von Genen mit verschiedenen Krankheiten assoziiert sein, die einen heterogenen Phänotyp aufweisen wie zum Beispiel die intellektuelle Beeinträchtigung mit 300 Genen, die Epilepsie mit 100 Genen oder neuromuskuläre Störungen mit 50 Genen (Chakravorty and Hegde, 2017).

Die Mehrheit der funktionellen Studien hat sich auf ein oder ein paar Gene fokussiert, welche wertvolle jedoch limitierte Einblicke in molekulare und zelluläre Pathways liefern, da Gene und deren Produkte (Proteine oder miRNAs) Teil eines kombinatorischen funktionellen Netzwerks sind und jede genetische (Sequenzvariation/Mutation) oder epigenetische Veränderung einen potentiellen Einfluss auf physiologische Prozesse hat, der zur zellulären/organismischen Anpassung oder Krankheit führt. Daher ist es unerlässlich globale Profile von Transkriptom und Epigenom zusammen mit lncRNAs, miRNAs und anderen nicht-transkribierten Sequenzen unter normalen (Kontrolle) und abnormalen (Krankheit) Bedingungen mit dem Fokus auf unterschiedliche Zelltypen zu erstellen und so eigene Gene Interaction Networks zu erzeugen (Chaitankar et al., 2016).

Diese Art von „Omics"-Daten beinhalten auch genetische Sequenzen (Genomics), genomweite Expressionsprofile (Transcriptomics), Daten zur Proteinhäufigkeit (Proteomics) sowie Methylierungsdaten und bieten neuartige Einblicke auf zelluläre Komponenten und deren zelluläre Mechanismen in biologischen Systemen. Rechnerische Systembiologie-Methoden können diese umfangreiche Akkumulation an heterogenen Daten aus den einzelnen Disziplinen kombinieren und so nützliche Tools liefern, um zusätzliche wertvolle auf die Systembiologie bezogene Einblicke, in zelluläre Mechanismen unter spezifischen biologischen Konditionen, wie Krebs oder infizierte Zellen, zu erlangen (Chen and Li, 2016).

NGS-Technologien verhelfen Forschern und Klinikern hierbei nicht nur zu neuen Informationen über das humane Genom, sondern gehen über die Genomforschung hinaus. Die Verwendung von ChIP-Seq, ATAC-Seq und Methyl-Seq in epigenetischen Studien erlaubt Forschern genomische regulatorische Mechanismen, Momentaufnahmen von DNA-Protein-Interaktionen, Methylmodifizierungen des Genoms, Enhancer-Regulationsmechanismen sowie Varianten all dieser Mechanismen im Krankheitszustand zu finden. Die Transkriptomforschung erlaubt Forschern die RNA bis hinunter zum Single-Transcript-Level zu sequenzieren, welches eine Relevanz in der klinischen Genomforschung besitzt, um so kryptische Splice Sites von Varianten, Insertion von Pseudoexonsequenzen, Downstream Frameshifts, die Entstehung von verfrühten Stop-Codons, allelspezifische Expression, Nonsense-mediated Decay,

differenzielle Exonanwendung oder Transkripthäufigkeiten zu identifizieren. Zu diesem Zweck werden neue Ansätze für Single-Cell RNA Sequencing entwickelt, um unterschiedliche Zellpopulationen zu charakterisieren und mithilfe von Long-Read Sequencing Technologien die spezifische Transkripthäufigkeit zu verstehen. Das funktionelle Ziel hier wie auch in der translationalen humanen Krebsforschung ist die Detektion von molekularen Biomarkern, welche die Verschmelzung von Forschung und klinischen Therapeutika erlauben, wobei in dieser genomischen Ära die NGS-Methode der Wahl von der Diagnose des Klinikers auf eine mögliche Krankheit beim Patienten abhängig sein wird und die beste Option für die molekulare Diagnose und genomische Medizin, die kombinierte Anstrengung von Forschern, medizinischen Genetikern und Klinikern erfordert (Chakravorty and Hegde, 2017).

8. Quellenverzeichnis

Ambardar, S., Gupta, R., Trakroo, D., Lal, R., Vakhlu, J., 2016. High Throughput Sequencing: An Overview of Sequencing Chemistry. Indian J. Microbiol. 56, 394–404. https://doi.org/10.1007/s12088-016-0606-4

Bleidorn, C., 2016. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. Syst. Biodivers. 14, 1–8. https://doi.org/10.1080/14772000.2015.1099575

Buermans, H.P.J., den Dunnen, J.T., 2014. Next generation sequencing technology: Advances and applications. Biochim. Biophys. Acta BBA - Mol. Basis Dis. 1842, 1932–1941. https://doi.org/10.1016/j.bbadis.2014.06.015

Chaitankar, V., Karakülah, G., Ratnapriya, R., Giuste, F.O., Brooks, M.J., Swaroop, A., 2016. Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research. Prog. Retin. Eye Res. 55, 1–31. https://doi.org/10.1016/j.preteyeres.2016.06.001

Chakravorty, S., Hegde, M., 2017. Gene and Variant Annotation for Mendelian Disorders in the Era of Advanced Sequencing Technologies. Annu. Rev. Genomics Hum. Genet. 18, 229–256. https://doi.org/10.1146/annurev-genom-083115-022545, CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

Chao, H.-P., Chen, Y., Takata, Y., Tomida, M.W., Lin, K., Kirk, J.S., Simper, M.S., Mikulec, C.D., Rundhaug, J.E., Fischer, S.M., Chen, T., Tang, D.G., Lu, Y., Shen, J., 2019. Systematic evaluation of RNA-Seq preparation protocol performance. BMC Genomics 20, 571. https://doi.org/10.1186/s12864-019-5953-1

Chen, B.-S., Li, C.-W., 2016. Constructing an integrated genetic and epigenetic cellular network for whole cellular mechanism using high-throughput next-generation sequencing data. BMC Syst. Biol. 10, 18. https://doi.org/10.1186/s12918-016-0256-5

Costa-Silva, J., Domingues, D., Lopes, F.M., 2017. RNA-Seq differential expression analysis: An extended review and a software tool. PLOS ONE 12, e0190152. https://doi.org/10.1371/journal.pone.0190152

Day-Williams, A.G., Zeggini, E., 2011. The effect of next-generation sequencing technology on complex trait research: NEXT-GENERATION SEQUENCING AND COMPLEX TRAIT RESEARCH. Eur. J. Clin. Invest. 41, 561–567. https://doi.org/10.1111/j.1365-2362.2010.02437.x

Dear, P.H., 2001. Genome Mapping, in: John Wiley & Sons, Ltd (Ed.), Encyclopedia of Life Sciences. John Wiley & Sons, Ltd, Chichester, UK, p. a0001467. https://doi.org/10.1038/npg.els.0001467
ling for biomarker discovery. Clin. Epigenetics 8, 122. https://doi.org/10.1186/s13148-

016-0284-4

Dolled-Filhart, M.P., Lee, M., Ou-yang, C., Haraksingh, R.R., Lin, J.C.-H., 2013. Computational and Bioinformatics Frameworks for Next-Generation Whole Exome and Genome Sequencing. Sci. World J. 2013, 1–10. https://doi.org/10.1155/2013/730210

Feng, Y., Zhang, Y., Ying, C., Wang, D., Du, C., 2015. Nanopore-based Fourth-generation DNA Sequencing Technology. Genomics Proteomics Bioinformatics 13, 4–16. https://doi.org/10.1016/j.gpb.2015.01.009

Fernandez-Marmiesse, A., Gouveia, S., Couce, M.L., 2018. NGS Technologies as a Turning Point in Rare Disease Research , Diagnosis and Treatment. Curr. Med. Chem. 25. https://doi.org/10.2174/0929867324666170718101946, CC BY-NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/legalcode)

Garrido-Cardenas, J., Garcia-Maroto, F., Alvarez-Bermejo, J., Manzano-Agugliaro, F., 2017. DNA Sequencing Sensors: An Overview. Sensors 17, 588. https://doi.org/10.3390/s17030588

Gilissen, C., Hoischen, A., Brunner, H.G., Veltman, J.A., 2012. Disease gene identification strategies for exome sequencing. Eur. J. Hum. Genet. 20, 490–497. https://doi.org/10.1038/ejhg.2011.258

Haque, F., Li, J., Wu, H.-C., Liang, X.-J., Guo, P., 2013. Solid-state and biological nanopore for real-time sensing of single chemical and sequencing of DNA. Nano Today 8, 56–74. https://doi.org/10.1016/j.nantod.2012.12.008

Heather, J.M., Chain, B., 2016. The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Hrdlickova, R., Toloue, M., Tian, B., 2017. RNA-Seq methods for transcriptome analysis: RNA-Seq. Wiley Interdiscip. Rev. RNA 8, e1364. https://doi.org/10.1002/wrna.1364

Jain, M., Olsen, H.E., Paten, B., Akeson, M., 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 17, 239. https://doi.org/10.1186/s13059-016-1103-0

Jia, Z., Shi, Y., Zhang, L., Ren, Y., Wang, T., Xing, L., Zhang, B., Gao, G., Bu, R., 2018. DNA methylome profiling at single-base resolution through bisulfite sequencing of 5mC-immunoprecipitated DNA. BMC Biotechnol. 18, 7. https://doi.org/10.1186/s12896-017-0409-7

Kayman Kurekci, G., Dincer, P., 2015. Exome Sequencing for The Identification of Mendelian Disease Genes. Erciyes Tıp DergisiErciyes Med. J. 36, 139–143. https://doi.org/10.5152/etd.2014.7804

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., 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., Vallabh Minikel, E., 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., Neale, B.M., Daly, M.J., MacArthur, D.G., 2019. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv 531210. https://doi.org/10.1101/531210

Leleu, M., Lefebvre, G., Rougemont, J., 2010. Processing and analyzing ChIP-seq data: from short reads to regulatory interactions. Brief. Funct. Genomics 9, 466–476. https://doi.org/10.1093/bfgp/elq022

Lelieveld, S.H., Spielmann, M., Mundlos, S., Veltman, J.A., Gilissen, C., 2015. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions: HUMAN MUTATION. Hum. Mutat. 36, 815–822. https://doi.org/10.1002/humu.22813

Li, D., Zhang, B., Xing, X., Wang, T., 2015. Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. Methods 72, 29–40. https://doi.org/10.1016/j.ymeth.2014.10.032

Li, P., Conley, A., Zhang, H., Kim, H.L., 2014. Whole-Transcriptome profiling of formalin-fixed, paraffin-embedded renal cell carcinoma by RNA-seq. BMC Genomics 15, 1087. https://doi.org/10.1186/1471-2164-15-1087

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., Law, M., 2012. Comparison of Next-Generation Sequencing Systems. J. Biomed. Biotechnol. 2012, 1–11. https://doi.org/10.1155/2012/251364

Lohmann, K., Klein, C., 2014. Next Generation Sequencing and the Future of Genetic Diagnosis. Neurotherapeutics 11, 699–707. https://doi.org/10.1007/s13311-014-0288-8

Loose, M., Malla, S., Stout, M., 2016. Real-time selective sequencing using nanopore technology. Nat. Methods 13, 751–754. https://doi.org/10.1038/nmeth.3930

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., Shafee, T., 2017. Transcriptomics technologies. PLOS Comput. Biol. 13, e1005457. https://doi.org/10.1371/journal.pcbi.1005457

Lu, H., Giordano, F., Ning, Z., 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics Proteomics Bioinformatics 14, 265–279. https://doi.org/10.1016/j.gpb.2016.05.004

Meienberg, J., Bruggmann, R., Oexle, K., Matyas, G., 2016. Clinical sequencing: is WGS the better WES? Hum. Genet. 135, 359–362. https://doi.org/10.1007/s00439-015-1631-9

Nakano, K., Shiroma, A., Shimoji, M., Tamotsu, H., Ashimine, N., Ohki, S., Shinzato, M., Minami, M., Nakanishi, T., Teruya, K., Satou, K., Hirano, T., 2017. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. Hum. Cell 30, 149–161. https://doi.org/10.1007/s13577-017-0168-8
National Research Council, Mapping and Sequencing the Human Genome (National Academy Press. Washington, DC, 1988)

Olson, M., Hood, L., Cantor, C., Botstein, D., 1989. A common language for physical mapping of the human genome. Science 245, 1434–1435. https://doi.org/10.1126/science.2781285

Palomares, M.-A., Dalmasso, C., Bonnet, E., Derbois, C., Brohard-Julien, S., Ambroise, C., Battail, C., Deleuze, J.-F., Olaso, R., 2019. Systematic analysis of TruSeq, SMARTer and SMARTer Ultra-Low RNA-seq kits for standard, low and ultra-low quantity samples. Sci. Rep. 9, 7550. https://doi.org/10.1038/s41598-019-43983-0

Pareek, C.S., Smoczynski, R., Tretyn, A., 2011. Sequencing technologies and genome sequencing. J. Appl. Genet. 52, 413–435. https://doi.org/10.1007/s13353-011-0057-x

Park, S.T., Kim, J., n.d. Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing 8.

Pavesi, G., 2016. ChIP-Seq Data Analysis to Define Transcriptional Regulatory Networks, in: Nookaew, I. (Ed.), Network Biology. Springer International Publishing, Cham, pp. 1–14. https://doi.org/10.1007/10_2016_43

Petersen, B.-S., Fredrich, B., Hoeppner, M.P., Ellinghaus, D., Franke, A., 2017. Opportunities and challenges of whole-genome and -exome sequencing. BMC Genet. 18, 14. https://doi.org/10.1186/s12863-017-0479-5

Pettersson, E., Lundeberg, J., Ahmadian, A., 2009. Generations of sequencing technologies. Genomics 93, 105–111. https://doi.org/10.1016/j.ygeno.2008.10.003

Quail, M., Smith, M.E., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., Gu, Y., 2012. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics 13, 341. https://doi.org/10.1186/1471-2164-13-341

Rabbani, B., Mahdieh, N., Hosomichi, K., Nakaoka, H., Inoue, I., 2012. Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. J. Hum. Genet. 57, 621–632. https://doi.org/10.1038/jhg.2012.91

Requena, T., Gallego-Martinez, A., Lopez-Escamez, J.A., 2017. A pipeline combining multiple strategies for prioritizing heterozygous variants for the identification of candidate genes in exome datasets. Hum. Genomics 11, 11. https://doi.org/10.1186/s40246-017-0107-5

Reuter, J.A., Spacek, D.V., Snyder, M.P., 2015. High-Throughput Sequencing Technologies. Mol. Cell 58, 586–597. https://doi.org/10.1016/j.molcel.2015.05.004

Rhoads, A., Au, K.F., 2015. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics 13, 278–289. https://doi.org/10.1016/j.gpb.2015.08.002, CC BY 4.0 (http://creativecommons.org/licenses/by/4.0/)

Schadt, E.E., Turner, S., Kasarskis, A., 2010. A window into third-generation sequencing. Hum. Mol. Genet. 19, R227–R240. https://doi.org/10.1093/hmg/ddq416

Streets, A.M., Zhang, X., Cao, C., Pang, Y., Wu, X., Xiong, L., Yang, L., Fu, Y., Zhao, L., Tang, F., Huang, Y., 2014. Microfluidic single-cell whole-transcriptome sequencing. Proc. Natl. Acad. Sci. 111, 7048–7053. https://doi.org/10.1073/pnas.1402030111

Sun, Z., Bhagwate, A., Prodduturi, N., Yang, P., Kocher, J.-P.A., 2016. Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations. Brief. Bioinform. bbw069. https://doi.org/10.1093/bib/bbw069

Suwinski, P., Ong, C., Ling, M.H.T., Poh, Y.M., Khan, A.M., Ong, H.S., 2019. Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics. Front. Genet. 10, 49. https://doi.org/10.3389/fgene.2019.00049

Vijay, P., McIntyre, A.B.R., Mason, C.E., Greenfield, J.P., Li, S., 2016. Clinical Genomics: Challenges and Opportunities. Crit. Rev. Eukaryot. Gene Expr. 26, 97–113. https://doi.org/10.1615/CritRevEukaryotGeneExpr.2016015724

Wadapurkar, R.M., Vyas, R., 2018. Computational analysis of next generation sequencing data and its applications in clinical oncology. Inform. Med. Unlocked 11, 75–82. https://doi.org/10.1016/j.imu.2018.05.003

Wang, X., 2016. Next-generation sequencing data analysis. Taylor & Francis, Boca Raton.

Wang, Y., Yang, Q., Wang, Z., 2015. The evolution of nanopore sequencing. Front. Genet. 5. https://doi.org/10.3389/fgene.2014.00449

Wang, Z., Liu, X., Yang, B.-Z., Gelernter, J., 2013. The Role and Challenges of Exome Sequencing in Studies of Human Diseases. Front. Genet. 4. https://doi.org/10.3389/fgene.2013.00160

Xing, X., Zhang, B., Li, D., Wang, T., 2018. Comprehensive Whole DNA Methylome Analysis by Integrating MeDIP-seq and MRE-seq, in: Tost, J. (Ed.), DNA Methylation Protocols. Springer New York, New York, NY, pp. 209–246. https://doi.org/10.1007/978-1-4939-7481-8_12

Zhao, W., He, X., Hoadley, K.A., Parker, J.S., Hayes, D., Perou, C.M., 2014. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. BMC Genomics 15, 419. https://doi.org/10.1186/1471-2164-15-419