CharitéCentrum für Frauen-, Kinder- und Jugendmedizin mit
Perinatalzentrum und Humangenetik (CC 17)

Klinik für Pädiatrie m. S. Neurologie

Direktor: Prof. Dr. Christoph Hübner

und

Institut für Medizinische Genetik

Direktor: Prof. Dr. Stefan Mundlos

Habilitationsschrift

Computer-unterstützte Suche nach krankheitsverursachenden DNA-Mutationen

zur Erlangung der Lehrbefähigung
für das Fach Experimentelle Genetik

vorgelegt dem Fakultätsrat der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Dr. rer. medic. Dominik Seelow
geboren in Berlin

Eingereicht:	November 2014
Dekanin:	Professor Dr. med. A. Grüters-Kieslich
1. Gutachter/in:	Prof. Dr. Andreas W. Kuß
2. Gutachter/in:	Prof. Dr. Thomas Meitinger

Inhaltsverzeichnis

1 Verzeichnis der Abkürzungen

1000G	*1000 Genomes Project* (1000-Genom-Projekt zur Ermittlung häufiger Polymorphismen)
CGH	*comparative genomic hybridisation* (vergleichende genomische Hybridisierung zur Suche nach Variationen der Kopienzahl genomischer Regionen)
CDS	*coding sequence* (protein-kodierende Sequenz eines Gens)
CNV	*copy number variant* (Variation der Kopienzahl eines Gens oder einer chromosomalen Region)
DNA	*deoxyribonucleic acid* (Desoxyribonukleinsäure oder DNS)
GOF	*gain of function* (Mutation, die zu einer neuen Proteinfunktion führt)
HPO	*Human Phenotype Ontology* (Ontologie, in der menschliche (Krankheits-)Symptome hierarchisch strukturiert sind)
LOF	*loss of function* (Mutation, die zu einem Funktionsverlust führt)
NGS	*Next Generation Sequencing* (Hochdurchsatzsequenzierung)
SNP	*single nucleotide polymorphism* (Einzelnukleotidpolymorpohismus)
VCF	*Variant Call Format* Standardformat für Genotypen aus Hochdurchsatzsequenzierungen
WES	*Whole Exome Sequencing* (Sequenzierung sämtlicher kodierenden Sequenzen im Genom)
WGS	*Whole Genome Sequencing* (Sequenzierung des gesamten Genoms)

2 Einleitung

Die Erforschung monogener Krankheiten befindet sich zur Zeit im Umbruch. In der Vergangenheit wurden Krankheitsmutationen meist durch eine Genkartierung (Abschnitt 2.1 *Suche nach Krankheitsmutationen mittels Genkartierung*) gefunden. Im Verlauf der Genkartierung wurden durch Kopplungsanalysen in Familien chromosomale Regionen identifiziert, die gemeinsam mit der Krankheit, dem Phänotyp, vererbt wurden. In diesen wurden gezielt einzelne Kandidatengene sequenziert, deren Funktion den Phänotyp erklären würde – zum Beispiel Kanalproteine für neurologische Krankheiten wie Muskelschwäche.

Die Entwicklung von Hochdurchsatzverfahren zur DNA-Sequenzierung (*Next Generation Sequencing*, NGS, auch *Deep Sequencing* genannt) revolutioniert derzeit die Aufklärung der molekularen Ursachen genetischer Erkrankungen. Durch die Möglichkeit, die kodierende Sequenz (CDS, *coding sequence*) sämtlicher Gene eines Menschen auf einmal und für nur etwa 1.000 Euro zu analysieren (Exomsequenzierung oder *Whole Exome Sequencing* - WES), können nun auch die molekularen Ursachen sehr seltener genetischer Krankheiten bestimmt werden.

Allerdings tritt hier ein neues Problem auf: die hohe Variabilität des menschlichen Genoms. Bei der kompletten Sequenzierung des menschlichen Exoms werden in der Regel mehrere tausend Abweichungen von der Referenzsequenz des Menschen gefunden. Jede dieser Varianten könnte die Krankheitsursache sein – die experimentelle Validierung des Krankheitspotentials durch funktionelle Untersuchungen oder Tiermodelle ist jedoch aus Zeit- und Kostengründen praktisch ausgeschlossen. Um die Zahl der in Frage kommenden Varianten auf eine handhabbare Zahl einzuschränken, sind bioinformatische '*in silico*' Verfahren unerlässlich. Der Arbeitsschwerpunkt der Erforschung monogener Erkrankungen verschiebt sich deshalb immer mehr von der Laborarbeit hin zu bioinformatischen Analysen.

In dieser Habilitationsschrift werde ich die Entwicklung verschiedener computerbasierter Verfahren vorstellen, die die Suche nach krankheitsverursachenden DNA-Mutationen erleichtern. Diese können in verschiedenen Strategien zur Aufklärung der molekularen Ursachen genetischer Krankheiten eingesetzt werden, die ich im Folgenden kurz erläutern werde. Allen gemein ist, dass sie web-basiert sind und mittels eines normalen Internetbrowsers verwendet werden können, so dass keine Installation von Software durch die Anwender erforderlich ist. Außerdem bieten alle Lösungen leicht zu benutzende Benutzerschnittstellen und können somit auch von Forschern oder Klinikern benutzt werden, die nur geringe Computerkenntnisse besitzen. Die Programme bieten darüber hinaus meist umfangreiche Möglichkeiten, eigenes Hintergrundwissen über die zu erforschende Krankheit einzubringen und erlauben es über die übersichtliche und umfassende Ausgabe der Ergebnisse, die Resultate direkt zu beurteilen oder gegebenenfalls die eigenen Vorgaben anzupassen. Die direkte Benutzung der Software durch die Experten, die mit einer Krankheit vertraut sind, erlaubt es ihnen, ihr eigenes Wissen über die Krankheit unmittelbar einzubringen - ohne dass Informationen durch die Auslagerung der Computerauswertung an Bioinformatiker verloren gehen.

Die hier vorgestellten Verfahren können Forscherinnen und Forschern dabei helfen, die molekularen Ursachen genetischer Krankheiten möglichst schnell, bequem und unter einem minimalen Einsatz von Arbeitszeit und finanzieller Mittel aufzuklären.

2.1 Suche nach Krankheitsmutationen mittels Genkartierung

Bei der klassischen Genkartierung monogener Krankheiten (siehe Abbildung 1) wird zuerst eine genomweite Genotypisierung mit genetischen Markern durchgeführt. In der Vergangenheit wurden dazu meist hochpolymorphe Mikrosatelliten eingesetzt, inzwischen werden aus Kosten- und

Zeitgründen in der Regel SNP-Chips zur simultanen Genotypisierung mehrerer zehn- oder hunderttausender Einzelnukleotidpolymorphismen (*single nucleotide polymorphisms* - SNPs) verwendet.

Mittels einer Kopplungsanalyse werden genomische Regionen identifiziert, deren Vererbung mit der Krankheit gekoppelt ist. Prinzipiell sind alle Gene in einer so gefundenen genomischen Region potentielle Kandidatengene für die Erkrankung ('positionelle Kandidaten'). Die Größe dieser Regionen und damit die Zahl der positionellen Kandidatengene kann mit Hilfe einer optionalen Feinkartierung häufig noch weiter eingeschränkt werden.

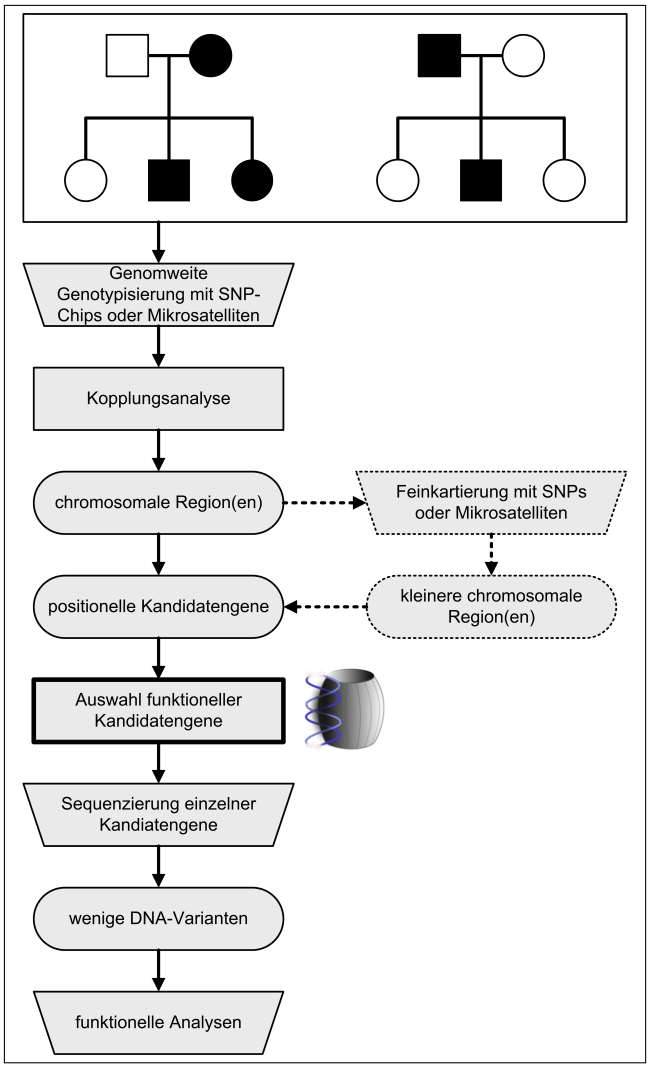


Abb. 1: **Klassischer Weg der Genkartierung**

Hier wird der Verlauf einer klassischen Suche nach Krankheitsmutationen über eine Genkartierung dargestellt. Arbeitsschritte im Labor werden als Trapez wiedergegeben, bioinformatische oder Denkarbeitsschritte als Rechtecke und Ergebnisse als abgerundete Rechtecke. Schritte, für die von mir entwickelte Software eingesetzt werden kann, sind fett umrandet.
Die gestrichelt eingezeichnete Feinkartierung mit zusätzlichen genetischen Markern ist optional.
Die von mir entwickelte Software **GeneDistiller** (Abschnitt 3.1.1) hilft bei der Auswahl funktioneller Kandidatengene (fett umrandet).
Die kodierenden Sequenzen der ausgewählten Gene werden dann mittels Sanger-Sequenzierung auf Abweichungen von der Referenzsequenz überprüft. Häufig wird die Sequenzierung beendet, sobald die erste Variante mit schwerwiegenden Auswirkungen auf das Protein gefunden wird.

Unter den positionellen Kandidaten werden nun 'funktionelle Kandidatengene' ausgewählt, deren Funktion oder Expression den Phänotyp erklären könnte. Dies kann im Beispiel einer genetisch bedingten Hautkrankheit bedeuten, dass gezielt Gene gesucht werden, die in der Haut exprimiert werden. Im Falle eines Stoffwechseldefekts wären beispielsweise Gene, deren Proteine im entsprechenden Stoffwechselweg beteiligt sind, aussichtsreiche Kandidaten. Die Auswahl der jeweiligen Gene hängt daher einerseits sehr stark von den zur Verfügung stehenden Informationen über die Gene, andererseits aber noch stärker vom Wissen der Forscher oder Kliniker über die entsprechende Krankheit. Um diese Arbeit zu vereinfachen, habe ich während meiner Promotion die Software GeneDistiller[1] entwickelt, die im Abschnitt 3.1.1 näher beschrieben wird. Die dabei verwendeten Verfahren, die auch zur Beurteilung der Relevanz von potentiellen Krankheitsmutationen wichtig sind, erläutere ich weiter unten im Abschnitt 2.3 *Auswahl von Kandidatengenen.*

In der Regel werden nun nacheinander die kodierenden Bereiche der wahrscheinlichsten Krankheitsgene sequenziert. Wird dabei eine DNA-Variante entdeckt, die zu einer schwerwiegenden Veränderung im Protein führt und die nicht (oder bei rezessiven Erkrankungen nur selten und nur heterozygot) in gesunden Verwandten oder Kontrollen aus der selben Population gefunden wird, so können sich nun funktionelle Analysen zur Bestimmung des Krankheitspotentials der Variante anschließen (Abbildung 1).

Allerdings erfordert eine Kopplungsanalyse ausreichend viele 'informative' Meiosen, in denen die Vererbung der Allele genetischer Marker eindeutig mit der Vererbung der Krankheit in Zusammenhang gebracht werden kann, um chromosomale Regionen mit einer ausreichenden statistischen Sicherheit identifizieren zu können. Selbst bei der Analyse hoch polymorpher Mikrosatelliten oder sehr vieler benachbarter Einzelnukleotidpolymorpohismen (SNPs) sind dazu mindestens 10 informative Meiosen erforderlich – dies bedeutet, dass entweder mehrere Familien mit der selben Krankheit oder aber große Familien mit mehreren Betroffenen gefunden und in die Analyse eingeschlossen werden müssen.

2.2 Homozygotiekartierung

Im Falle konsanguiner Familien verringert sich das oben beschriebene Problem drastisch, da sich bei der Vererbung eines Krankheitsallels über zwei blutsverwandte Eltern die Zahl der Meiosen, in denen das Krankheitsallel übertragen wird, in der konsanguinen 'Schleife' verdoppelt. Die sogenannte Homozygotiekartierung[2] erlaubt es, die mit der Krankheit gekoppelte Genregion lediglich durch die Genotypisierung weniger Betroffener durchzuführen.

In der Vergangenheit wurde für die Homozygotiekartierung zumeist eine Mehrpunkt-Kopplungsanalyse eingesetzt, in der vor allem zwei verschiedene Algorithmen zum Einsatz kamen:

1. Dies ist zum einen der *Lander-Green-Algorithmus*[3], der linear mit der Anzahl der eingesetzten Marker skaliert und deshalb prinzipiell gut für den Einsatz in genomweiten Analysen geeignet ist, sowohl mit relativ wenigen Mikrosatelliten als auch mit einer mittleren Anzahl von SNPs. Da in einer Kopplungsanalyse im Gegensatz zu einer Assoziationsanalyse oder auch einer Feinkartierung initial nur die Vererbung chromosomaler Abschnitte mit der der Krankheit verglichen wird, reichen hier etwa 10.000 informative SNPs aus, um die Krankheitsregion identifizieren zu können. Um Zeit zu sparen, werden deshalb in der Regel nicht alle SNPs in die Analyse eingeschlossen, sondern diejenigen ausgewählt, in denen die Verteilung der beiden Genotypen in der jeweiligen Population möglichst ausgeglichen ist. Die Verwendung weiterer Marker würde keinen signifikanten Informationsgewinn bringen, den Zeitaufwand aber beträchtlich erhöhen. In der nachfolgenden Feinkartierung werden dann selbstverständlich alle SNPs in den potentiellen Krankeitsregionen verwendet.

   Dieser Algorithmus hat aber einen gravierenden Nachteil: er skaliert exponentiell zur Zahl der Personen bzw. Meiosen. Sehr große konsanguine Familien, oder solche mit mehreren blutsverwandten Eltern, können deshalb nicht komplett analysiert werden. Zudem kann das Vorkommen mehrerer konsanguiner 'Schleifen' zu nicht mehr akzeptablen Laufzeiten führen; zum Beispiel erforderte die Analyse sämtlicher 50.000 vorhandener SNPs in einem Beispiel aus meiner Dissertation[4] eine Laufzeit von etwa 2.000 Stunden, also 12 Wochen.

2. Alternativ kann der *Elston-Stewart-Algorithmus*[5] eingesetzt werden, der linear zur Anzahl der betrachteten Personen bzw. Meiosen skaliert. Er wäre somit für konsanguine Familien deutlich besser geeignet, allerdings skaliert er exponentiell zur Zahl der verwendeten Marker. Er ist somit vor allem für Zweipunkt-Analysen geeignet, Mehrpunktanalysen werden sehr stark verlangsamt beziehungsweise, beim Einsatz von mehr als 8 Markern zur Betrachtung von Haplotypen, praktisch unmöglich. Dies wird insbesondere bei der Verwendung von

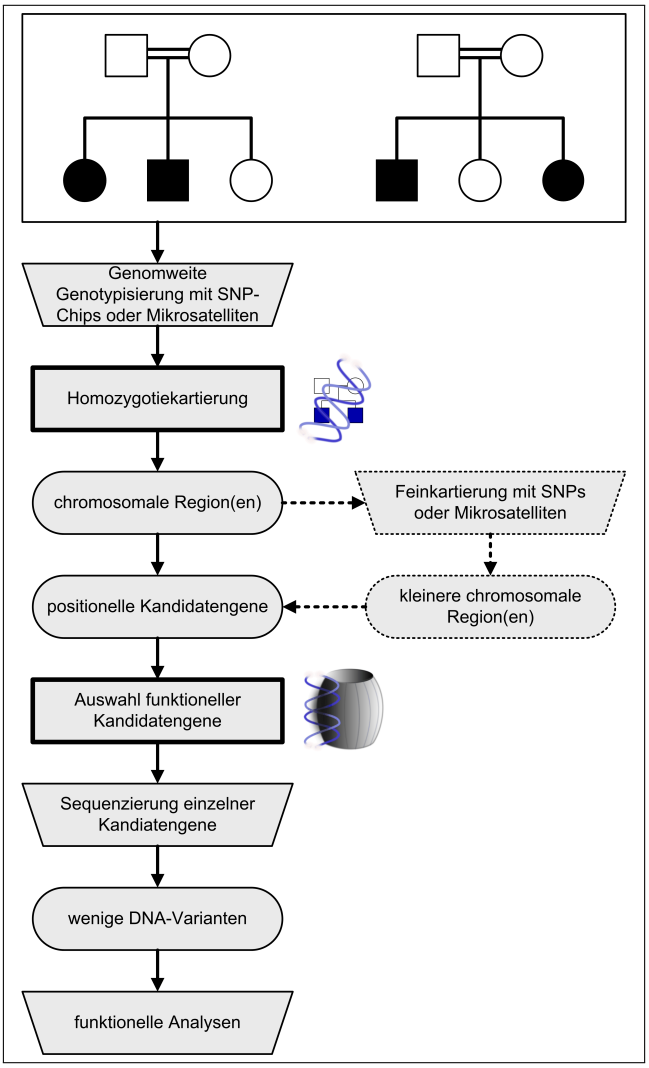


Abb. 2: **Homozygotiekartierung**

Die Suche nach krankheitsverursachenden DNA-Mutationen in konsanguinen Familien unterscheidet sich lediglich in der Analyse der Genotypen zur Identifizierung der mit der Krankheit gekoppelten Genregion vom klassischen Verfahren (Abbildung 1).
Für diese Teilaufgabe, die Homozygotiekartierung (fett umrandet), kann die Software **HomozygosityMapper** (siehe Abschnitte 3.2.1 und 4.1.1) eingesetzt werden.

SNP-Markern zum Problem, da aufgrund deren geringer Informativität Mehrpunktanalysen in kleinen Familien zwingend erforderlich werden.

Beide Algorithmen sind daher nur eingeschränkt für genomweite Genkartierungsprojekte mit großen konsanguinen Familien mittels SNP-Chips geeignet, da hier im Gegensatz zu etwa 400 sehr informativen Mikrosatelliten mehrere tausend wenig informative SNP-Marker betrachtet werden müssen.

Allerdings ist eine initiale Kopplungsanalyse zur Suche nach Krankheitshaplotypen gar nicht erforderlich, wenn ausreichend viele betroffene Personen genotypisiert wurden: Grundlage der Homozygotiekartierung ist das Vorhandensein eines homozygoten Haplotyps um den Krankheitslokus herum in den betroffenen Familienmitgliedern. Insbesondere dann, wenn lediglich Betroffene eingeschlossen wurden, ist die zeitaufwendige Kopplungsanalyse - in der auch für die nicht genotypisierten Familienmitgliedern 'wahrscheinliche' Genotypen errechnet werden müssen - vollkommen überflüssig. Es genügt, zuerst eine Suche nach homozygoten Abschnitten durchzuführen, die bei allen Betroffenen vorhanden sind. Nicht betroffene Familienmitglieder können dann durch eine anschließende Kopplungsanalyse dieser Bereiche oder die Betrachtung der Haplotypen zum Ausschluss von Regionen einbezogen werden.

Eine schnelle Homozygotiekartierung, die lediglich gemeinsame homozygote Regionen in den betroffenen Personen detektiert, wird durch die in den Abschnitten 3.2.1 und 4.1.1 vorgestellte

Software HomozygosityMapper[6,7] erreicht, für das oben angegebenene Beispiel verringert sich die Laufzeit so um das 24.000-fache von 2.000 Stunden auf etwa 5 Minuten.

Der Ablauf der Suche nach Krankheitsmutationen ist, bis auf die Durchführung einer Homozygotiekartierung, der selbe wie bei einer klassischen Kopplungsanalyse und wird in Abbildung 2 dargestellt.

Die aktuelle Version HomozygosityMapper2012[7] erlaubt es darüber hinaus, die Genotypen gesunder Familienmitglieder zu berücksichtigen und zum Ausschluss von Regionen zu verwenden, in denen auch Gesunde für die gleichen Allele wie die Betroffenen homozygot sind. Es bietet zudem eine Schnittstelle, um eine auf die möglichen Krankheitsregionen begrenzte Kopplungsanalyse durchführen zu können.

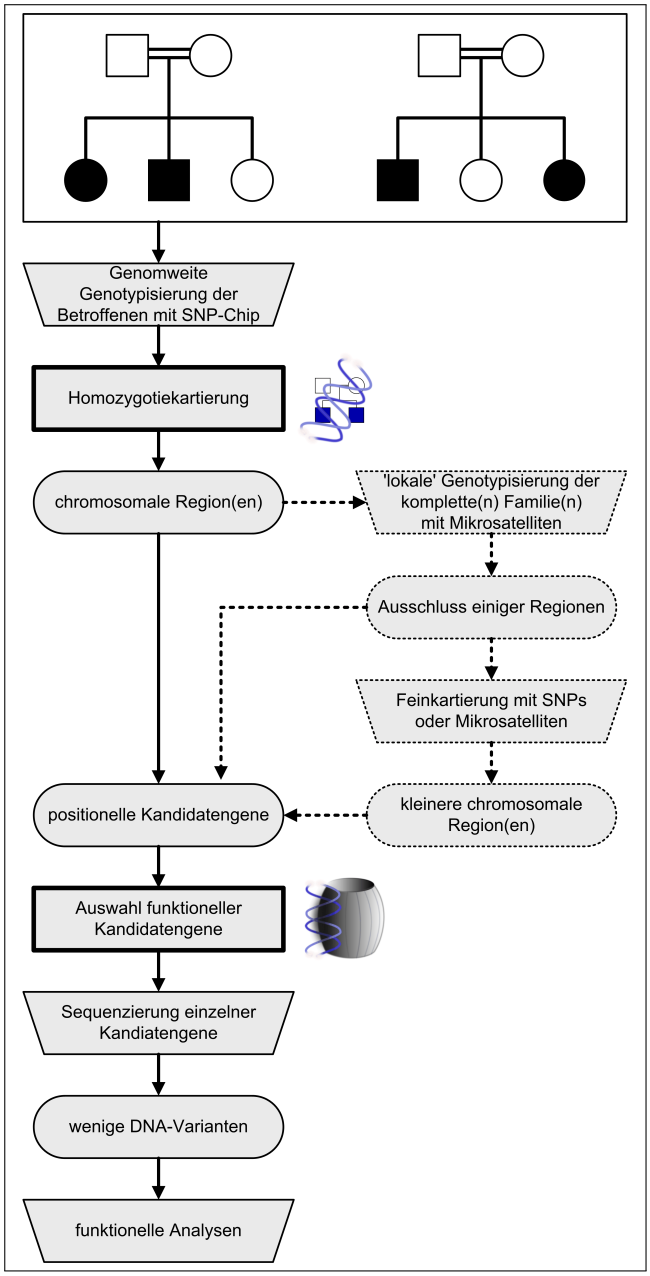


Abb. 3: **Initiale Homozygotiekartierung der Betroffenen**

Hier wird eine kostensparende Variante der Homozygotiekartierung dargestellt: Sind ausreichend viele (2-3) betroffene Personen eingeschlossen, genügt es, lediglich diese zu genotypisieren. Um mögliche Krankheitsregionen zu bestätigen oder auszuschließen, kann dann mit wenigen Mikrosatellitenmarkern überprüft werden, ob in den gefundenen Bereichen eine Kopplung zwischen der Vererbung der Krankheit und der Allele der genetischen Markern existiert. Gegenbefalls kann durch den Einsatz weiterer Marker eine Feinkartierung durchgeführt werden, um den Krankheitslokus weiter einzugrenzen - bei aktuellen SNP-Chips mit mehreren hunderttausenden SNPs oder einer Genotypisierung mittels einer Exomsequenzierung ist dies aber in der Regel nicht erforderlich.

Konnten weitere gesunde Familienmitglieder rekrutiert werden, kann es eine kostensparende Möglichkeit sein, diese - falls ausreichend viele Betroffene eingeschlossen werden konnten - nicht initial zu genotypisieren. In diesem Fall können die gefundenen homozygoten möglichen Krankheitsre-

gionen durch den Einsatz weniger Mikrosatellitenmarker in den gesamten Familien auf ihren Erbgang überprüft, zur Feinkartierung verwendet und gegebenenfalls als Krankheitslokus ausgeschlossen werden. Dieses Verfahren bietet sich insbesondere in Laboren an, in denen passende Mikrosatellitenmarker ohnehin zur Verfügung stehen (Abbildung 3).

Aufgrund der sinkenden Preise für Hochdurchsatzsequenzierungen kann eine weitere Kostensenkung dadurch erreicht werden, dass die Exome einiger betroffener Personen sequenziert werden. Mit Hilfe von HomozygosityMapper2012, das die Analyse von NGS-Genotypen im VCF-Format anbietet, können so homozygote Regionen direkt aus den Datensätzen einer Hochdurchsatzsequenzierung identifiziert werden. Dabei werden nicht nur die möglichen Krankheitsregionen ermittelt sondern in der Regel auch gleich die Krankheitsmutation - sofern sich diese in der kodierenden Sequenz befindet. In Frage kommende DNA-Varianten, die sich sowohl durch ihren Effekt auf das Protein (siehe Abschnitt 2.6 *Variantenbewertung*) als auch durch die Genfunktion (siehe Abschnitt *2.3* weiter unten) als Krankheitsursache anbieten, können dann mittels Sanger-Sequenzierung oder Restriktionanalyse in allen Familienmitgliedern studiert werden.

2.3 Auswahl von Kandidatengenen

Das Ergebnis von Genkartierungen über Kopplungsanalysen oder Homozygotiekartierungen ist nicht ein einzelnes Gen sondern eine chromosomale Region, innerhalb derer sich das Krankheitsgen befindet. Das Vorkommen von *crossing overs* in den Meiosen führt dazu, dass in der Regel keine kompletten Chromosomen von der Mutter oder dem Vater geerbt werden, sondern lange Segmente der beiden homologen Chromosomen jedes Elternteils rekombiniert werden. Mit jeder Meiose nimmt die Zahl dieser Rekombinationen zu, wodurch sich die gemeinsam mit einer Krankheit vererbten Haplotypen immer weiter verkleinern. Können nur wenige Personen in eine Kopplungsanalyse eingeschlossen werden, ist es deshalb nicht nur schwierig oder gar unmöglich, eine einzelne mögliche Krankheitsregion zu identifizieren; die gefundenen möglichen Regionen sind zudem sehr groß.

Kopplungsregionen können wenige Gene enthalten, sind aber bei der Kartierung kleinerer Familien in der Regel mehrere Megabasen groß und enthalten deshalb oft mehr als 100 verschiedene Gene (zum Beispiel 216 Gene in der initialen Genotypisierung eines Lokus' für Schizophrenie[8]). All diese Gene mittels Sanger-Sequenzierung nach krankheitsverursachenden Mutationen zu durchsuchen, wäre sehr kosten- und zeitaufwändig. Inzwischen bieten Hochdurchsatzverfahren zur DNA-Sequenzierung zwar eine günstigere Alternative (siehe Abschnitt 2.4 *target-enrichment Strategien*); diese resultieren aber in einer Vielzahl von DNA-Varianten, unter denen sich die krankheitsverursachende verbirgt. Eine Hochdurchsatzsequenzierung ist zudem erheblich teurer als die konventionelle Sequenzierung eines einzelnen Gens.

Die Zahl der Kandidatengene (oder der in einer Hochdurchsatzsequenzierung gefundenen Genen mit DNA-Varianten) kann deutlich verringert werden, indem das bestehende Wissen über den Phänotyp genutzt wird. Als erster Schritt bietet es sich an, in der Literatur bzw. in Krankheitsdatenbanken wie OMIM[9] oder HGMD[10] nach schon bekannten Krankheitsgenen zu suchen. Wird ein Gen gefunden, in dem Mutationen die studierte Krankheit oder einen ähnlichen Phänotyp auslösen, so avanciert dieses Gen natürlich zu einem sehr aussichtsreichen Kandidatengen.

Alternativ kann zum Beispiel anhand der betroffenen Organe oder Gewebe postuliert werden, dass das Krankheitsgen in diesen auch aktiv sein muss. Hier können Expressionsdatenbanken helfen, dieses Wissen bei der Auswahl der Gene zu berücksichtigen. Auch ist es möglich, in Datenbanken zu recherchieren, ob für eines der positionellen Kandidatengene Tiermodelle existieren, die zu einem ähnlichen Phänotyp führen. Für die Suche nach Genen, die Mitochondriopathien auslösen, kann die subzelluläre Lokalisation des Proteins im Mitochondrium herangezogen werden. Diese Beispiele zeigen, dass die Strategie zur Auswahl geeigneter 'funktioneller' Kandi-

datengene sehr stark vom Wissen über die Krankheit abhängt.

Ein bequemes Verfahren, Wissen über die positionellen Kandidatengene zusammenzutragen, stellt GeneCards dar, eine Website, die verschiedenartige Informationen über ein Gen zusammenträgt. Allerdings bot GeneCards in der Vergangenheit nur die Möglichkeit, Daten zu einzelnen Genen wiederzugeben, so dass für eine Vielzahl von positionellen Kandidaten zahlreiche Anfragen nötig waren. Eine Alternative sind automatische Priorisierungsverfahren, die über verschiedene Algorithmen (zum Beispiel Proteininteraktionsdaten) nach Verbindungen oder Ähnlichkeiten zwischen den positionellen Kandidaten und bekannten Krankheitsgenen suchen und die positionellen Kandidatengene nach diesen sortieren. Allerdings wird hierbei das meist umfangreiche Wissen der Kliniker oder Forscher über die Krankheit vernachlässigt, so zum Beispiel die geforderte Expression in bestimmten Geweben.

Die von uns entwickelte Software GeneDistiller[11], die auch in meiner Dissertation[12] beschrieben wird, verbindet deshalb beide Ansätze. Sie wird im Abschnitt *GeneDistiller* (3.1.1) vorgestellt.

2.4 *target-enrichment* Strategien

Bei einer Genkartierung mit nur wenigen Individuen werden häufig mehrere und bzw. oder sehr große mögliche Krankheitsregionen gefunden (siehe oben). Dies bedeutet, dass mehrere hundert Gene als positionelle Kandidaten in Frage kommen. Auch die Einschränkung auf funktionelle Kandidatengene führt in solchen Fällen häufig zu einer Vielzahl in Frage kommender Gene - deren Sequenzierung mit dem 'klassischen' Sanger-Verfahren war und ist aus Zeit- und Kostengründen nicht sinnvoll.

Die Entwicklung von Hochdurchsatzverfahren zur DNA-Sequenzierung (*Next Generation Sequencing*, NGS, auch *Deep Sequencing*) führte vor wenigen Jahren zu einer neuen Möglichkeit der Suche nach krankheitsverursachenden DNA-Mutationen: Während die Sequenzierung kompletter Genome oder Exome mit ausreichender Abdeckung anfangs noch extrem zeit- und kostenaufwändig war, so konnten durch die gezielte Anreicherung bestimmter Sequenzen doch zumindest die für die Krankheit in Frage kommenden chromosomalen Regionen (oder alternativ auch ausgewählte Kandidatengene) mit Hilfe der Hochdurchsatzsequenzierung auf Varianten untersucht werden. In diesem *target-enrichment* Verfahren[13] werden für die zu untersuchenden DNA-Abschnitte komplementäre DNA-Sonden erstellt, mit deren Hilfe die gewünschten Sequenzen durch ein Hybridisierungsverfahren angereichert, amplifiziert und schließlich auf Hochdurchsatzsequenziergeräten sequenziert werden können.

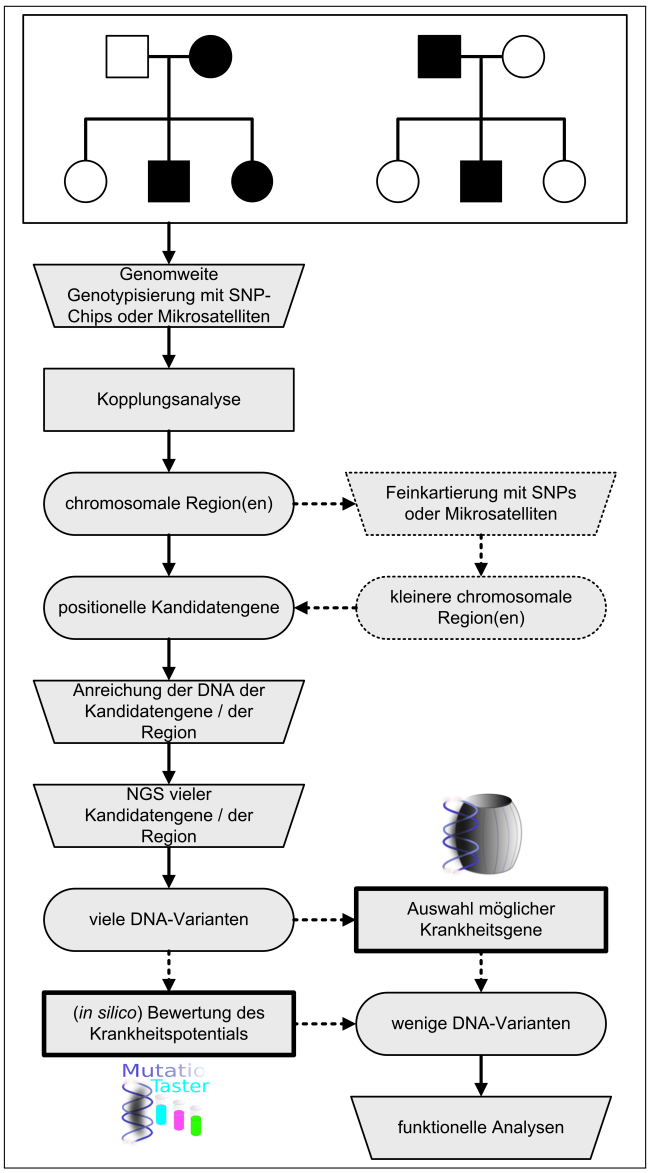


Abb. 4: **Mutationssuche via *target-enrichment***

Die Suche nach Krankheitsmutationen über das *target-enrichment* Verfahren beginnt mit einer klassischen Gen- oder Homozygotiekartierung. Allerdings werden nach der Bestimmung möglicher Krankheitsregionen im Genom nicht einzelne Kandidatengene ausgewählt und gezielt sequenziert sondern gleich alle Gene in einer Region (oder deren kodierende Bereiche) auf einmal. Dabei werden in der Regel mehrere hundert Abweichungen von der Referenzsequenz gefunden. Abhängig von der Kapazität der Hochdurchsatzsequenzierung bzw. der Methode zur DNA-Anreicherung kann eine vorhergehende Feinkartierung sinnvoll sein, um die Gesamtlänge der zu sequenzierenden DNA zu verringern. Die durch die Sequenzierung gefundenen Varianten können umgekehrt aber auch genutzt werden, um den Krankheitshaplotyp weiter einzugrenzen (hier nicht dargestellt).
Als nächster Schritt muss eine Einteilung dieser Varianten in wahrscheinlich harmlose und möglicherweise krankheitsverursachende unternommen werden. Dies kann einerseits durch eine Auswahl funktioneller Kandidatengene und die Vernachlässigung der Varianten außerhalb dieser geschehen. Eine weitere Reduzierung der Zahl möglicher Krankheitsmutationen kann durch die Bewertung des Krankheitspotentials der Variante selbst erzielt werden (fett umrandet). Dieses Verfahren wird durch die Software **MutationTaster** erleichtert, die im Abschnitt 4.2.1 vorgestellt wird.

Allerdings werden so, abhängig von der Größe der angereicherten Regionen bzw. der Gesamtlänge der angereicherten Gensequenzen, hunderte oder sogar tausende von Abweichungen von der Referenzsequenz gefunden. Mit Hilfe bioinformatischer Verfahren muss deshalb eine Auswahl der Varianten erfolgen, die das höchste Potential besitzen, die Krankheit auszulösen.

Dies kann zum einen dadurch geschehen, dass eine Auswahl anhand der Varianten beinhaltenden Gene durchgeführt wird; das heißt, dass Varianten in vielversprechenden funktionellen Kandidaten eher als mögliche Krankheitsursache angenommen werden. Eine weitere Reduzierung der Zahl möglicher Krankheitsmutationen kann durch die Bewertung des Krankheitspotentials der Variante selbst erzielt werden. Dies ist unter anderem mit der in dieser Arbeit vorgestellten Software MutationTaster[14,15] möglich - diese und andere Verfahren zur Ermittlung des Krankheitspotentials von DNA-Varianten werden in Abschnitt 2.6 *Variantenbewertung* vorgestellt. Abbildung 4 zeigt eine Übersicht über die Genkartierung mit Hilfe der gezielten Anreicherung chromosomaler Regionen.

2.5 Genom- oder Exomsequenzierung

Die meisten monogenen Krankheiten gehören zu den seltenen Erkrankungen, die nach der Definition der Europäischen Union eine Prävalenz von unter 5:10.000 haben*. Während es bei den 'häufigen' seltenen Erkrankungen möglich ist, ausreichend viele bzw. große Familien zu rekrutieren – und hier in vielen Fällen die krankheitsverursachenden Mutationen identifiziert werden konnten - ist dies bei den 'selteneren' seltenen Erkrankungen, an denen weltweit nur wenige Menschen leiden, nicht ohne Weiteres möglich, da für aussagefähige Kopplungsanalysen außerhalb konsanguiner Familien oft schlichtweg nicht genügend betroffene Menschen existieren oder rekrutiert werden können.

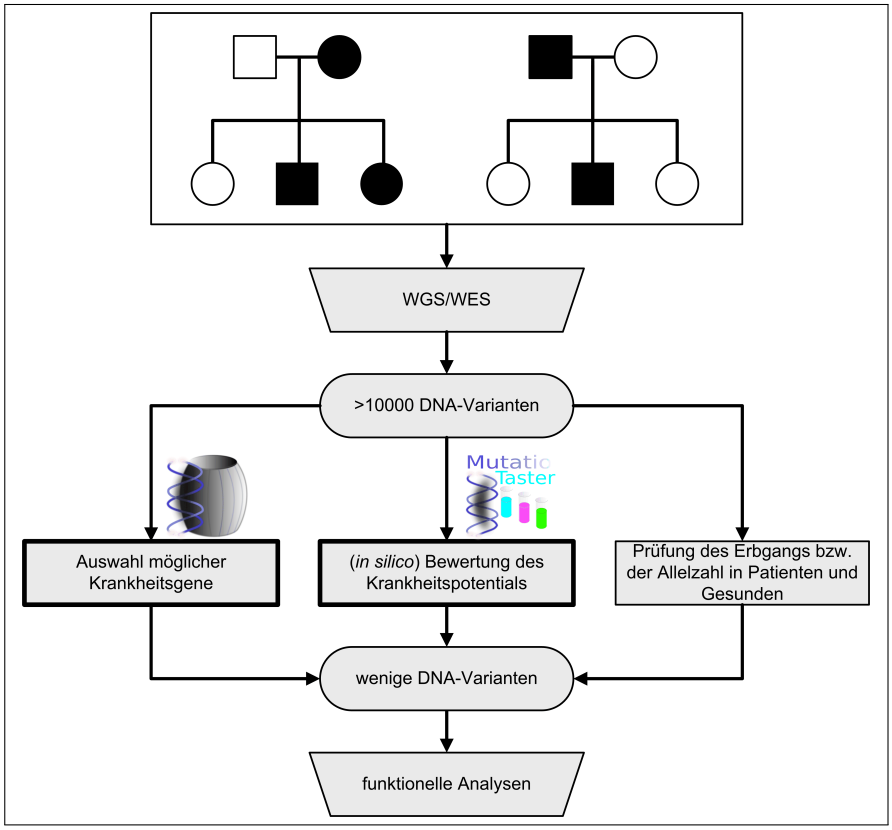


Abb. 5: **Genom- oder Exomsequenzierung**
Die Sequenzierung kompletter Genome (WGS) oder Exome (WES) liefert mehrere tausend Varianten. Um die wahrscheinlich krankheitsverursachende Variante zu ermitteln, können verschiedene Strategien eingesetzt und miteinander kombiniert werden:

1. Auswahl von Genen, die den Phänotyp erklären könnten
2. Bewertung des Krankheitspotentials der gefundenen Varianten (anhand des Effekts auf das Protein oder über den Abgleich mit Datenbanken, in denen Polymorphismen gespeichert sind)
3. Beschränkung auf Varianten, deren Erbgang oder Allelzahl dem Vererbungsmodell der Krankheit entspricht

Die inzwischen mit etwa 1.000 Euro pro Exom relativ kostengünstige Sequenzierung vollständiger Exome (*Whole Exome Sequencing*, WES) bietet die Möglichkeit, auch ohne die vorherige Einschränkung auf bestimmte Gene oder chromosomale Abschnitte die kodierenden Bereiche aller Gene auf potentielle Krankheitsmutationen hin zu untersuchen. Sie kann deshalb prinzipiell auch ohne die für eine Kopplungsanalyse erforderliche große Zahl von Meiosen, die entweder durch große Familien mit mehreren Betroffenen oder durch Betroffene aus verschiedenen Familien

*http://www.bmg.bund.de/praevention/gesundheitsgefahren/seltene-erkrankungen.html

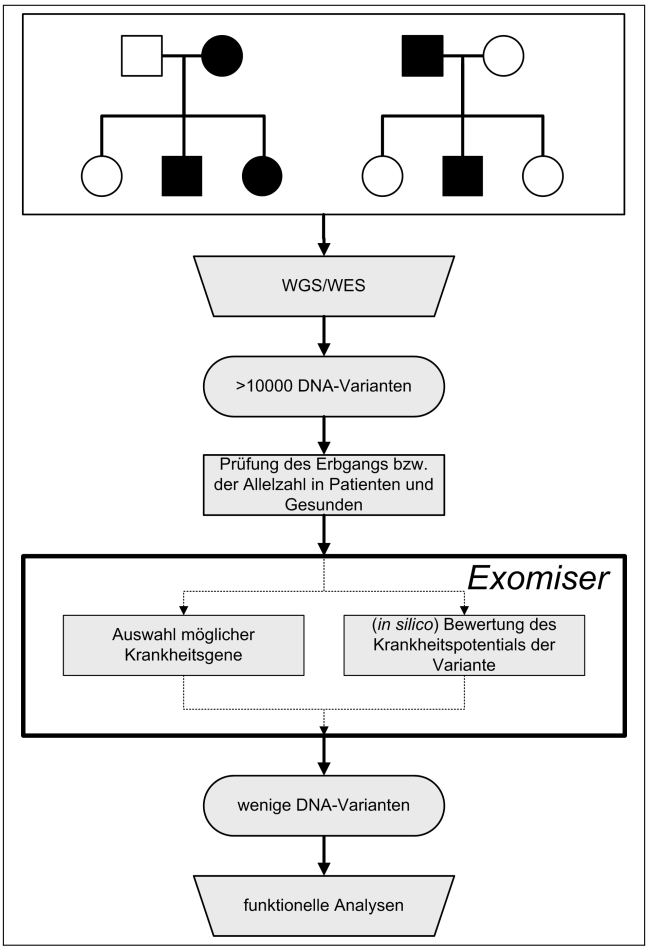


Abb. 6: **Genom- oder Exomsequenzierung**

Einige aktuelle Computerprogramme - wie der in dieser Arbeit vorgestellte **Exomiser** (Abschnitt 4.2.3) fassen sie Suche nach potentiellen Kandidatengenen und die Bewertung des Potentials von Varianten, die Gen- bzw. Proteinfunktion zu stören, zusammen.
Dabei wird unter anderem die Bewertung von Varianten durch MutationTaster verwendet.

erreicht werden kann, erfolgreich sein.

Im Jahr 2010 wurde zum ersten Mal der Einsatz einer Exomsequenzierung zur Aufklärung der Ursache einer monogenen Erkrankung beschrieben[16]. Das Verfahren wird in Abbildung 5 dargestellt. Seitdem konnten durch Hochdurchsatzsequenzierungen kompletter Exome oder Genome die genetischen Ursachen von mehr als 70 monogenen Erkrankungen aufgedeckt werden[17].

Wie in der Einleitung kurz beschrieben, verschiebt sich hier der Arbeitsaufwand weg von der Laborarbeit hin zur bioinformatischen Analyse der Ergebnisse: In einer Sequenzierung des kompletten Genoms (*Whole Genome Sequencing*, WGS) werden häufig mehrere Millionen DNA-Varianten detektiert, das Genom von Erzbischof Desmond Tutu weist beispielsweise mehr als 3,6 Millionen Abweichungen von der Referenzsequenz auf[18].

Um eine bequemere Einschränkung der Varianten auf solche mit hohem Krankheitspotential zu erreichen, wurden in der Zwischenzeit Computerprogramme entwickelt, die die Bewertung des Krankheitspotentials der Varianten mit der Bewertung des Krankheitspotentials der Gene zusammenfassen (siehe Abschnitt 2.6 und Abbildung 6). Dazu gehört der in dieser Arbeit in Abschnitt 4.2.3 vorgestellte Exomiser.

Die in einer Genom- oder Exomsequenzierung anfallenden DNA-Varianten können auch als genetische Marker in einer Kopplungsanalyse oder einer Homozygotiekartierung eingesetzt werden (Abbildung 7), um so ohne eine vorherige genomweite Typisierung von SNPs eine Suche nach Krankheitsregionen und Krankheitsmutationen durchzuführen. Die aktuelle Version der Software HomozygosityMapper (HomozygosityMapper 2012 - siehe Abschnitt 3.2.1) kann deshalb auch Genotypen aus NGS-Projekten zur Ermittlung der Krankheitsregionen verwenden.

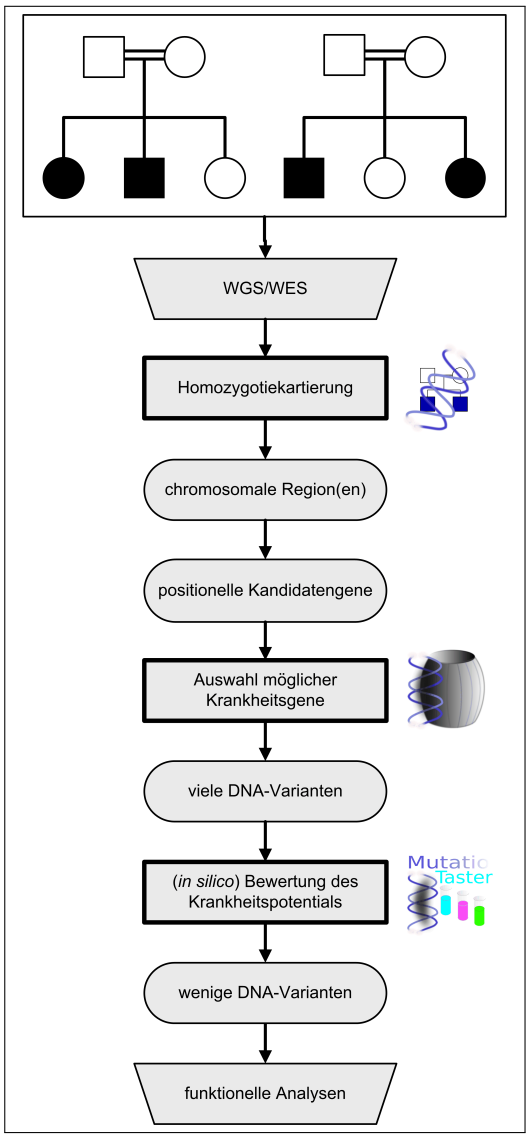


Abb. 7: **Genom- oder Exomsequenzierung und Homozygotiekartierung**

Diese Abbildung stellt den Ablauf einer Genkartierung mit einer initialen Hochdurchsatzsequenzierung dar. Die in einer oder mehreren Familien gefundenen Varianten können als genetische Marker verwendet und in einer Kopplungsanalyse oder, wie hier fett umrandet dargestellt, einer Homozygotiekartierung eingesetzt werden. Unsere Software **HomozygosityMapper** bietet die Möglichkeit, diese mit den Genotypdateien aus NGS-Projekten direkt vorzunehmen.
Im Anschluss können die in den vorhergehenden Abschnitten beschriebenen Techniken zur Variantenauswahl eingesetzt werden. Hier dargestellt ist das 'klassische Verfahren', bei dem zunächst aus den positionellen Kandidatengenen mögliche Krankheitsgene ausgewählt werden und die in ihnen gefundenen Varianten als potentielle Krankheitsmutationen betrachtet werden. Diese werden in einem weiteren Schritt auf ihr Krankheitspotential hin untersucht.
Diese Reihenfolge kann auch umgedreht werden, dies ist allerdings zeitaufwendiger, weil so mehr Varianten bewertet werden müssen.
Werden für beide Aufgaben vollständig automatisierte Computerprogramme verwendet, so ist natürlich auch der parallele Einsatz beider Methoden möglich, in diesem Fall sind alle DNA-Veränderungen aussichtsreiche Kandidaten, die sich in der Schnittmenge aus den Varianten in potentiellen Krankheitsgenen und den Varianten mit schwerwiegendem Effekt auf das Protein befinden. Alternativ können Programme eingesetzt werden, die automatisch die Bewertung des Krankheitspotentials von Gen und Variante zusammenfassen (siehe Abbildung 6).

2.6 Variantenbewertung

Um krankheitsverursachende DNA-Mutationen sicher zu identifizieren, muss deren Auswirkung nachgewiesen werden.
Der 'Goldstandard' zum Nachweise des Krankheitspotentials von DNA-Mutationen ist ein Tiermodell, in dem die in Frage kommende Mutation in ein Tier eingebracht wird - bei Funktionsverlust-Mutationen muss das Tier das Krankheitsallel homozygot, beziehungsweise bei compound-heterozygoten Fällen beide Krankheitsallele, tragen. Weisen die Versuchstiere einen dem Menschen entsprechenden (oder zumindest sehr ähnlichen) Phänotyp auf, bestätigt dies die Hypothese, dass ebendiese Mutation (bzw. Mutationen) krankheitsverursachend ist. Allerdings erfordert die Etablierung entsprechender Tiermodelle hohe Kosten und ist zudem sehr zeitaufwendig. Darüber hinaus stellen sich hier natürlich auch ethische Fragen. Aus diesen Gründen werden häufig andere Verfahren eingesetzt, die die Auswirkungen von Mutationen in Zellkulturen oder *ex vivo* untersuchen: Sprechen die Befunde beispielsweise dafür, dass die Krankheit durch einen Enzymdefekt ausgelöst wird, so kann die Untersuchung der katalytischen Aktivität des veränderten Proteins Aufschluss darauf geben, ob die Mutation tatsächlich zu einer Verminderung der enzymatischen Aktivität führt. In anderen Fällen kann mit Hilfe von Antikörpern

gegen das betroffene Protein beispielsweise in Immunfluoreszenzexperimenten in Zellen gezeigt wird, dass sich das mutierte Protein nicht mehr an der gewohnten Stelle befindet.

Die funktionelle Charakterisierung der tausenden (Exomsequenzierung) bis Millionen (Genomsequenzierung) von DNA-Varianten, die in genomweiten Hochdurchsatzsequenzierungen gefunden werden, im Labor oder gar im Tiermodell würde derzeit allerdings jeden Kosten- und Zeitrahmen sprengen. Es muss also eine sinnvolle Auswahl unter den Varianten getroffen werden, für die sich weitere Untersuchungen lohnen.

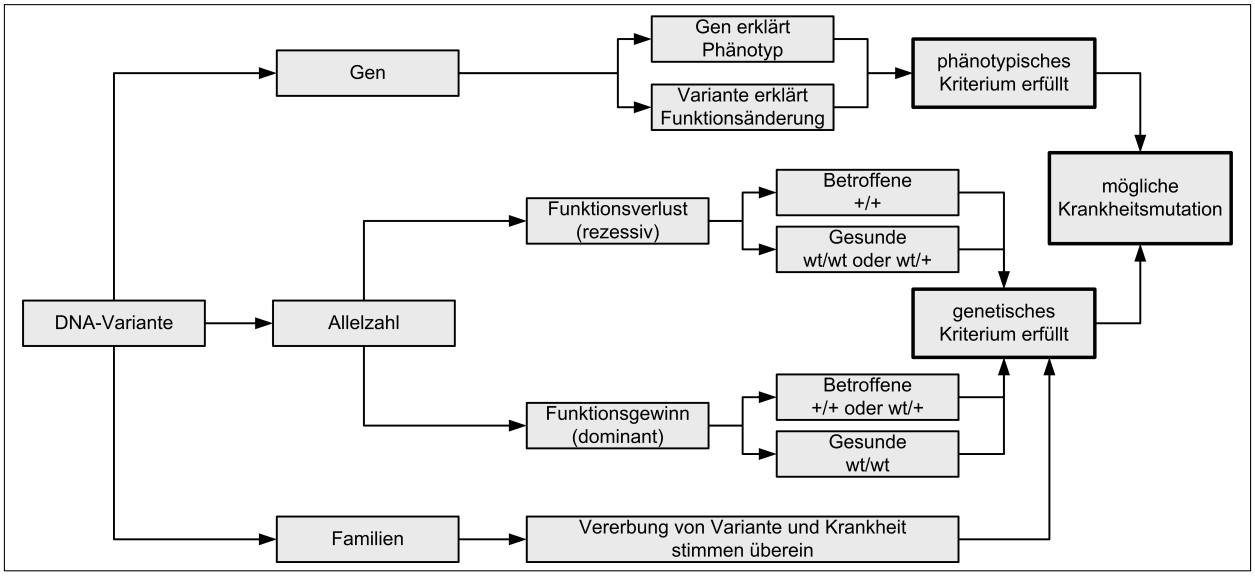


Abb. 8: Anforderungen an eine Krankheitsmutation

Diese Abbildung zeigt die Kriterien, die eine krankheitsverursachende Mutation erfüllen muss.
Dabei müssen sowohl der Phänotyp durch das betroffene Gen erklärbar sein, als auch eine signifikante Störung der Proteinfunktion, -lokalisation oder -expression durch das variante Allel ausgelöst werden können.
Neben diesen phänotypischen Erfordernissen muss auch die Vererbung der Krankheitsallele in betroffenen Familien mit dem Erbgang des Phänotyps übereinstimmen. Auch muss die Allelzahl in Betroffenen und Gesunden dem Vererbungsmodell der Krankheit entsprechen - dieser Test ist auch ohne die Untersuchung weiterer Familienmitglieder möglich.
Dargestellt sind hier autosomal dominant oder rezessiv vererbte monogene Krankheiten mit einer vollständigen Penetranz; compound-heterozgyote Mutationen erfodern selbstverständlich eine andere Allelverteilung.
wt: Wildtyp; +: mutmaßliche Krankheitsmutation

Damit eine Variante eine Krankheit auslösen kann, sind drei Grundvoraussetzungen erforderlich (Abbildung 8):

1. Der Erbgang der Mutation beziehungsweise die Allelzahl in Betroffenen und Gesunden entspricht dem Erbgang der Krankheit.

2. Das veränderte Protein führt zur Krankheit

3. Die Mutation verändert die Funktion des Proteins.

Alle drei Punkte werden durch von mir entwickelten Computerprogrammen adressiert:

1. Mit HomozygosityMapper[6,7] (Abschnitte 3.2.1 und 4.1.1) lassen sich bei rezessiv vererbten Krankheiten in konsanguinen Familien die chromosomalen Regionen identifizieren, in denen sich die Krankheitsmutation befinden muss - alle Varianten außerhalb dieser Regionen können vernachlässigt werden.

   Der einfache Abgleich der Allelzahlen ist in einer in unserer Arbeitsgruppe entwickelten internen Analysepipeline implementiert, die allerdings nicht publiziert wurde, da in diesem

Bereich genügend Lösungen existieren, die zum Teil noch darüber hinaus gehen - beispielsweise snpActs[†], welches sogar eine Kopplungsanalyse ermöglicht.

2. GeneDistiller[1] (Abschnitt 3.1.1) erlaubt es, unter Einbeziehung des Hintergrundwissens über den Phänotyp Kriterien zu formulieren, denen mögliche Krankheitsgene entsprechen müssen.

3. MutationTaster[14,15] (Abschnitt 4.2.1) analysiert das Krankheitspotential von Varianten anhand ihrer Auswirkungen auf das resultierende Protein.

Eine weitere Möglichkeit zur Verringerung der zu untersuchenden Varianten ist der Abgleich der gefundenen Varianten mit bekannten Polymorphismen. Die aktuelle Version des 1000-Genom-Projekts (1000G)[19] umfasst 79 Millionen Varianten[‡]. Da die im Rahmen dieses Vorhabens sequenzierten Personen nicht an schweren monogenen Erkrankungen leiden (zumindest nicht an solchen mit hoher Penetranz und frühem Krankheitsbeginn), können Allele, die im 1000G vorkommen, als Ursache dominanter Erkrankungen ausgeschlossen werden. Allerdings ist es aufgrund möglicher Sequenzierungsfehler sinnvoll, diesen Ausschluss erst dann vorzunehmen, wenn das betreffende Allel mehrfach auftritt. Im Falle rezessiver Erkrankungen ist ein solcher Ausschluss nur dann ratsam, wenn das Allel im 1000G-Kollektiv mehrfach *homozygot* vorkommt - heterozygote Träger des Allels sind schließlich nicht erkrankt und sind im Falle häufigerer monogener Krankheiten in einem Kollektiv von 1.000 Personen durchaus zu erwarten: Beispielsweise beziffert ein WHO-Report zur Mukoviszidose die Heterozygotenfrequenz für Krankheitsallele im *CFTR*-Gen im südlichen Afrika auf $1/42$[20].

Der Abgleich gefundener Varianten mit bekannten Polymorphismen unter der Berücksichtigung der Genotypenhäufigkeit ist in die aktuelle Version von MutationTaster integriert.

Es liegt auf der Hand, dass die Kombination der vorgeschlagenen Wege zu Bestimmung der tatsächlichen Krankheitsmutation in einer einzelnen Anwendung benutzerfreundlicher wäre. Zu diesem Zwecke wurden verschiedene Computerprogramme entwickelt, die die Kombination der möglichen Auswirkungen des Proteins auf den Phänotyp einerseits und der möglichen Auswirkungen der Variante auf das Protein andererseits erlauben, zum Beispiel eXtasy[21] und Exomiser[22], an dessen Entwicklung ich beteiligt war (Abschnitt 4.2.3), oder auch dessen Weiterentwicklung für diagnostische Zwecke PhenIX[23]. In allen drei Programmen wird der Phänotyp durch die Eingabe der bei Patienten vorhandenen Symptome aus der Human Phenotype Ontology (HPO)[24] definiert. Anhand verschiedener Algorithmen, die beispielsweise Phänotypen aus Mausmodellen (Exomiser) oder die Interaktion mit bekannten Krankheitsgenen (eXtasy) verwenden, wird jedem Gen ein Potential zugeordnet, die angegebenen Symptome auszulösen. Gleichzeitig wird der Schweregrad der Proteinveränderung durch eine Kombination der Vorhersagen verschiedener Variantenbewertungsprogramme bestimmt. In allen hier genannten Programmen ist MutationTaster eines der verwendeten Programme; die Autoren von eXtasy erwägen sogar, die anderen Variantenbewertungsprogramme nicht weiter zu berücksichtigen, da ihr Beitrag zum Gesamtergebnis einer Variante deutlich geringer ist (Yves Moreau, persönliche Kommunikation). Allerdings benutzen sowohl eXtasy als auch Exomiser die in frühen Versionen der Datenbank dbNSFP[25] gespeicherten Bewertungen von Varianten, die noch keine Werte aus MutationTaster2 enthalten. Ein generelles Problem ist, dass dbNSFP nur bekannte und zugleich nicht-synonyme Varianten enthält.

[†] http://snpacts.ikmb.uni-kiel.de/

[‡] ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/

2.7 Copy-Number Varianten

Neben punktuellen Mutationen in der DNA (Einzelbasenaustausche oder kleine Insertionen und/oder Deletionen) können auch Verluste oder Gewinne größerer Bereiche des Erbguts oder ganzer Chromosomen Krankheiten auslösen. Diese Variationen in der Kopienzahl (*copy number variants* - CNV) führen dazu, dass eine genomische Region gar nicht mehr vorhanden, hemizygot oder dupliziert ist. Neben den durch zusätzliche oder fehlende Chromosomen, zum Beispiel in Trisomie 23 (OMIM #190685) oder dem Turner-Syndrom (kein OMIM-Eintrag), verursachten Krankheiten, kann auch der Verlust oder Gewinn eines kleineren DNA-Segments zu einer Erkrankung führen: Das DiGeorge-Syndrom ('CATCH-22', OMIM #188400) wird durch eine Mikrodeletion hervorgerufen, die mehrere Gene umfasst. Auch die Veränderung der Kopienzahl einzelner Gene kann eine medizinische Relevanz haben, beispielsweise steigt durch eine erhöhte Kopienzahl des Proto-Onkogens ERBB2 (HER-2) das Krebsrisiko signifikant an[26]. Auch die Veränderung der Kopienzahl eines intragenischen Bereichs kann pathogen sein[27].

Die Untersuchung der Variationen der Kopienzahl durch die vergleichende genomische Hybridisierung (Array CGH, *comparative genomic hybridisation*) ist inzwischen zu einer Routineuntersuchung bei Kindern mit einer geistigen Behinderung oder Entwicklungsstörungen avanciert[28]. Darüber hinaus werden CNVs auch als Risikofaktor für komplexe Krankheiten betrachtet[29]. Allerdings konnte in diesem Fall bisher keine klare Beziehungen zwischen Phänotyp und Genotyp (CNV) gefunden werden.

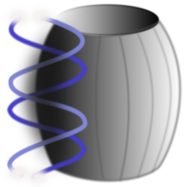
Die Suche nach krankheitsverursachenden oder -begünstigenden CNVs wird dadurch erschwert, dass einerseits auch bei gesunden Menschen ein beträchtliches Maß an Variationen der Kopienzahl genomischer Regionen existiert[30] und andererseits verschiedene Methoden zur Bestimmung der Kopienzahl unterschiedliche Ergebnisse liefern können[31].

In den vergangenen Jahren sind durch die umfangreichen Untersuchungen von CNVs im Rahmen von Studien komplexer Krankheiten zahlreiche Datenbanken entstanden, die die gefundenen CNVs in Gesunden und Betroffenen sowie die zur Analyse verwendeten Plattformen enthalten. Um das Krankheitspotential 'neuer' CNVs leichter ermitteln zu können, haben wir die webbasierte Software CNVinspector[32] entwickelt, mit deren Hilfe sowohl Forscher als auch klinisch tätige Ärzte diese gegen schon bekannte 'polymorphe' CNVs filtern und die in ihnen enthaltenen Gene studieren können. CNVinspector kann dabei sowohl mit CNVs eines einzelnen Patienten als auch mit Kohorten, die in Assoziationsstudien komplexer Krankheiten untersucht werden, verwendet werden. Die Software wird im Abschnitt 4.3.1 *CNVinspector* vorgestellt.

3 Vorarbeiten aus meiner Promotion

Meine Promotion beschäftigte sich mit verschiedenen Wegen, die für genetische Erkrankungen verantwortlichen Gene zu identifizieren. Da die Erkennung potentieller Krankheitsgene eine wichtige Grundlage für das Thema dieser Habilitationsschrift, die Identifizierung der ursächlichen Genmutationen, ist, habe ich die folgenden zwei Arbeiten aus meiner kumulativen Dissertation[12] in diese Habilitationsschrift aufgenommen.

3.1 Auswahl von Kandidatengenen



Um die Auswahl von Genen zu erleichtern, die einen bestimmten Phänotyp erklären, haben wir die Software GeneDistiller[1] entwickelt. GeneDistiller bietet darüber hinaus auch die Möglichkeit, verschiedene Informationen über ein oder mehrere Gen(e) auswählen und anzeigen.
http://www.genedistiller.org/

3.1.1 GeneDistiller

Zur Suche nach Kandidatengenen können zwei Ansätze verfolgt werden: Der klassische Weg besteht aus einer manuellen Suche nach Informationen über die positionellen Kandidaten in der Literatur oder in Internet-Datenbanken. Alternativ können Computerprogramme eingesetzt werden, die die positionellen Kandidaten beispielsweise aufgrund von Interaktionen oder Ähnlichkeit mit bekannten Krankheitsgenen priorisieren. GeneDistiller verbindet beide Ansätze:

1. Zum Einen bietet es durch die Einbindung umfangreicher genspezifischer Informationen aus vielen öffentlichen Datenbanken die Möglichkeit für die Forscher, sich detailliert über die positionellen Kandidatengene zu informieren. Im Gegensatz zur alten Version von GeneCards erlaubt es, eine chromosomale Region anzugeben und die gewünschten Informationen zu *allen* in ihr enthaltenen Genen anzuzeigen. Das Interface gestattet die Auswahl der jeweils relevanten Geninformation aus einer Vielzahl von Datenquellen, um so nicht 'in der Datenflut zu ertrinken'. Dabei stellt GeneDistiller weitere Optionen zur Verfügung: So können beispielsweise Gene nach bestimmten Kriterien gefiltert werden oder alternativ Schlüsselwörter in Genbeschreibungen oder genspezifischen Eigenschaften in verschiedenen Datenquellen wie zum Beispiel OMIM[9] oder der GeneOntology[33] hervorgehoben werden.

2. Zusätzlich haben die Benutzer die Wahl, Gene nach ihren Übereinstimmungen mit dem vorgegebenen Krankheitsmodell zu priorisieren, so dass beispielsweise Gene im gleichen Stoffwechsel- oder Signaltransduktionsweg wie bekannte Krankheitsgene zuerst und mitsamt der ausgewählten genspezifischen Daten aufgelistet werden.

Das Interface erlaubt es, beide Wege zu verbinden, also beispielsweise die Priorisierung auf Gene zu beschränken, die in einem bestimmten Gewebe oder Organ exprimiert werden. Durch diese Definition eines 'Krankheitsmodells' kann das Hintergrundwissen der Kliniker oder Forscher für die Gensuche herangezogen werden, ohne die Anwender durch die umfangreiche Suche nach den für sie relevanten Informationen im Internet oder der Literatur unnötig zu belasten. Da eine Abfrage innerhalb weniger Sekunden fertig gestellt wird und die Gründe für die Bewertung des Krankheitspotentials angezeigt werden, lässt sich das der Suche zugrunde gelegte Krankheitsmodell sehr schnell weiter verfeinern.

Kandidatengene für eine genetische Erkrankung können nicht nur durch Kopplungsanalysen oder Homozygotiekartierungen bestimmt werden. Auch im Rahmen genomweiter Assoziationsstudien (GWAS), die insbesondere bei der Suche nach den Ursachen komplexer Krankheiten eingesetzt werden, fallen mögliche Kandidatengene an. Das gleiche gilt für die direkte Hochdurchsatzsequenzierung eines Exoms oder Genoms, bei denen viele Gene mit möglichen Krankheitsmutationen detektiert werden. GeneDistiller erlaubt es deshalb, anstelle einer genomischen Region auch eine Liste von Genen anzugeben, unter denen dann wie oben beschrieben der beste funktionelle Kandidat bestimmt werden kann.

GeneDistiller wird kontinuierlich weiterentwickelt; wesentliche neue Möglichkeiten, die nicht in der initialen Publikation (die auf den folgenden Seiten wiedergegeben wird) aufgeführt sind, werden im folgenden aufgelistet:

- **Genomweite Suchen**: Da die Software inzwischen auf einem deutlich leistungsstärkeren Server als in der Vergangenheit läuft, ist es nun möglich, das gesamte Genom nach funktionellen Kandidaten zu durchsuchen.
- **Human Phenotype Ontology**: Die Human Phenotype Ontology[24] (HPO) bietet detaillierte Informationen über die klinischen Symptome, die durch Mutationen eines Gens verursacht werden können. GeneDistiller erlaubt es, diese zum Filtern und zum Priorisieren von Genen einzusetzen. Auch die Suche nach gemeinsamen Symptomen mit bekannten Krankheitsgenen ist möglich.
- **STRING-Interaktionsdatenbank**: Zusätzlich zu den initial vorhanden Proteininteraktionsdaten aus UniHI[34] konnte in der Zwischenzeit auch die deutlich umfangreichere STRING-Datenbank[35] integriert werden.
- **API und Ausgabe von Tabellen**: Während die ursprüngliche Version lediglich HTML-Seiten als Ergebnis lieferte, ist es nun möglich, die Informationen auch tabellarisch anzuzeigen. Dabei kann zwischen einer formatierten Textdatei oder Dateien im Microsoft-Excel-Format gewählt werden. GeneDistiller kann dabei über ein Programminterface (*application programming interface* - API) angesteuert werden, so dass es direkt aus anderen Applikationen benutzt werden kann. Hierfür sind alle Programmoptionen des Web-Interfaces verfügbar.
- **Primerdesign**: Um die Sequenzierung der kodierenden Bereiche eines Gens zu erleichtern, bietet GeneDistiller die Möglichkeit, automatisch geeignete Primer zu entwerfen. Dabei werden sämtliche Exons aller Transkripte des Gens einbezogen. Zur eigentlichen Primererstellung wird Primer3[36] verwendet.
- **Unterschiedliche Genomversionen**: Die aktuelle Version von GeneDistiller bietet die Möglichkeit, aus den Genomversionen 36 (hg18) und 37 (hg19) auszuwählen. Dies ist insbesondere für Hochdurchsatzsequenzierungsprojekte relevant, da hier sowohl für eine Anreicherung vor der Sequenzierung (*target-enrichment*) als auch in den Genotypdateien physikalische Positionen verwendet werden. Die Integration der aktuellen Genomversion 38 ist derzeit in Entwicklung.
- **Erzeugung von BED-Dateien**: Für die Anreicherung genomischer Regionen oder einzelner Genen für Hochdurchsatzsequenzierungen sind BED-Dateien erforderlich, die die anzureichernden Bereiche über deren genomische Positionen spezifizieren. GeneDistiller kann derartige Dateien für einzelne Gene oder komplette Kopplungsregionen erstellen. Dabei kann ausgewählt werden, ob lediglich die kodierenden Bereiche (samt flankierender Basen) oder die kompletten Gensequenzen enthalten sein sollen. Auch die durch den Benutzer festgelegten Filter zum Ausschluss von Genen werden berücksichtigt.

GeneDistiller kann unter der URL http://www.genedistiller.org/ kostenlos verwendet werden. Die Originalarbeit wird im Anhang zu dieser Habilitationsschrift wiedergegeben.

3.2 Homozygotiekartierung

3.2.1 HomozygosityMapper

HomozygosityMapper dient der Genkartierung in konsanguinen Familien. Das Verfahren der Homozygotiekartierung wird im Abschnitt 2.2 *Homozygotiekartierung* umfangreich beschrieben. Die Web-basierte Software erlaubt es, die von mehreren Betroffenen geteilten homozygoten Regionen auf einfache Art und Weise zu identifizieren. Dazu können Benutzer Genotypen auf einen Server hochladen, und, nach der Angabe der betroffenen und der gesunden Familienmitglieder, analysieren. Die Datenausgabe erfolgt sowohl als Text als auch visuell, dabei werden besonders lange gemeinsame homozygote Bereiche optisch hervorgehoben. Zur manuellen Eingrenzung oder Erweiterung der möglichen Krankheitsregionen ist eine Visualisierung der Genotypen implementiert, wobei besonders lange homozygote Segmente herausgehoben dargestellt werden. Zur Suche nach Kandidatengene können die Gene in den so gefundenen Region direkt in GeneDistiller studiert werden.

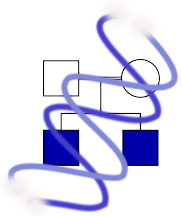
HomozygosityMapper bietet diverse Analyseoptionen, so dass auch bei genetischer Heterogenität - wenn mehrere Familien analysiert werden - Krankheitsregionen identifiziert werden können. Ebenso ist es möglich, auch in 'Inzucht-Populationen' Krankheitshaplotypen zu identifizieren, wenn ausreichend viele Betroffene eingeschlossen wurden. Bei einer engen Konsanguinität (Cousine-Cousin ersten Grades) genügen schon 2-3 betroffene Familienmitglieder zur Bestimmung möglicher Genorte.

Zum gemeinsamen Bearbeiten von Projekten erlaubt HomozygosityMapper den gemeinsamen Zugriff auf ein Projekt von verschiedenen Benutzerkonten. Es ist aber auch möglich, Daten anonym zu analysieren. Eine weitere Option ist, Projekte öffentlich zugänglich zu machen.

Von dieser Software existieren zwei Versionen und Publikationen. Das Manuskript zur initialen Version[6], die lediglich die Suche nach homozygoten Regionen in den Genotypen von SNP-Chips gestattet, befindet sich im Anhang zu dieser Habilitationsschrift. Auf den folgenden Seiten wird die aktuelle Version beschrieben, die im Rahmen von Hochdurchsatzsequenzierungen eingesetzt werden kann und weitaus umfassendere Analysemöglichkeiten bietet.

4 Eigene Arbeiten

4.1 Homozygotiekartierung



Unsere Software HomozygosityMapper erlaubt die schnelle und bequeme Durchführung von Homozygotiekartierungen im Internet.
Während die initiale Version[6] (siehe Abschnitt 3.2.1) auf Menschen und SNP-Chips beschränkt war, erlaubt die aktuelle Fassung auch die Analyse weiterer Spezies und die Nutzung von NGS-Genotypen.
http://www.homozygositymapper.org/

4.1.1 HomozygosityMapper2012

Seelow, D. & Schuelke, M.
HomozygosityMapper2012 - bridging the gap between homozygosity mapping and deep sequencing.
Nucleic Acids Research, July 2012, W516–520, 40

Die Durchführung einer genomweiten Hochdurchsatzsequenzierung (WES oder WGS) als initiale Analyse[16] stellt die Forscher vor das Problem, unter vielen tausenden DNA-Varianten die krankheitsverursachende finden zu müssen. Durch *in silico* Verfahren zur Auswahl von Varianten nach ihrem Krankheitspotential wie z.B. MutationTaster[15] lässt sich zwar eine deutliche Einschränkung erreichen, allerdings können - falls mehrere Personen sequenziert wurden - durch die Einbindung der genetischen Informationen große Teile des Genoms ausgeschlossen werden (siehe Abbildung 8).

Dazu kann die Allelverteilung betrachtet und mit dem Krankheitsmodell verglichen werden: Bei einer dominanten Erkrankung mit hoher Penetranz müssen alle Betroffenen einer Familie das Krankheitsallel besitzen, gesunde Kontrollen dürfen es nicht tragen. Bei einer rezessiven Erkrankung müssen alle Betroffenen zwei Krankheitsallele besitzen bzw. homozygot sein; ihre Eltern müssen jeweils eines der beiden Allele heterozygot tragen. Gesunde Personen dürfen, außer in Krankheiten mit unvollständiger Penetranz, nicht zwei Krankheitsallele besitzen (siehe Abbildung 8 im Abschnitt 2.6 *Variantenbewertung*).

Eine deutlich weiter gehende Reduktion der zu untersuchende Varianten lässt sich erreichen, wenn nicht einzelne Positionen des Genoms für sich betrachtet werden, sondern die klassischen Verfahren zur Genkartierung eingesetzt werden: Alle bei einer genomweite Sequenzierung anfallenden Varianten lassen sich als genetische Marker in einer Kopplungsanalyse oder einer Homozygotiekartierung nutzen.

Wir haben die initiale Version unserer Software HomozygosityMapper[6] deshalb so erweitert, dass nun auch Genotypdateien aus Hochdurchsatzsequenzierungen anstelle von SNP-Genotypen eingelesen und analysiert werden können.

HomozygosityMapper2012[7] erlaubt es somit, auch ohne eine initiale Genotypisierung mit 'klassischen' genetischen Markern potentielle Krankheitsregionen zu identifizieren und das Auffinden der krankheitsverursachenden Varianten erheblich zu beschleunigen. Eine direkte Verknüpfung von HomozygosityMapper und MutationTaster zur Vorhersage des Krankheitspotentials aller Varianten in homozygoten Regionen ist geplant, konnte aber aus Zeitgründen bislang nicht realisiert werden.

Für eine zusätzliche Kopplungsanalyse bietet HomozygosityMapper2012 die Möglichkeit, die Genotypen in den möglichen Krankheitsregionen zu exportieren. Neben den Genotypen werden dabei auch die Positionen und Allelfrequenzen der genetischen Marker ausgegeben. Die Dateien können direkt von ALOHOMORA[37] eingelesen werden, einem Computerprogramm, das SNP-Genotypen in die Eingabeformate für verschiedene Programme wie zum Beispiel GENEHUNTER[38] und ALLEGRO[39] zur Kopplungsanalyse umwandelt.

Zum Zeitpunkt der Publikation der zweiten Version hatte HomozygosityMapper bereits mehr als 600 registrierte Nutzer, die mehr als 7 Milliarden Genotypen mit HomozygosityMapper analysiert hatten (September 2014: 1400 Nutzer, über 20 Milliarden Genotypen). In HomozygosityMapper2012 wurden viele Wünsche unserer Benutzer realisiert. Unter anderem war dies die Ausweitung auf weitere Spezies neben dem Menschen für die Aufklärung genetischer Erkrankungen in Modellorganismen[40] oder für die Zucht von Nutztieren[41] nach genetischen Merkmalen. Eine weitere wesentliche Änderung ist die Einbindung der genetischen Informationen gesunder Familienangehörige: HomozygosityMapper2012 erlaubt es, die Krankheitsregionen auf solche einzuschränken, in denen alle Betroffenen das gleiche Allel homozygot besitzen und in denen gesunde Kontrollpersonen nicht für dieses Allel homozygot sind. Dieses Verfahren erlaubt es, mögliche Krankheitsregionen in einer einzelnen Familie mit nur wenigen Betroffenen zielsicherer zu bestimmen als bislang.

HomozygosityMapper kann unter der URL http://www.homozygositymapper.org/ kostenlos verwendet werden.

Die Originalarbeit wurde 2012 in *Nucleic Acids Research* publiziert und wird hier nicht wiedergegeben.

Seelow, D. & Schuelke, M.
HomozygosityMapper2012 - bridging the gap between homozygosity mapping and deep sequencing.
Nucleic Acids Research, July 2012, W516–520, 40
https://doi.org/10.1093/nar/gks487

4.2 Variantenbewertung

Die von uns entwickelte Software MutationTaster dient der *in silico* Bewertung des Krankheitspotentials von DNA-Varianten. Im Gegensatz zu den bekannten Programmen SIFT[42] und PolyPhen-2[43] ist MutationTaster nicht auf den Austausch einzelner Aminosäuren beschränkt, zudem erfolgt die Analyse auch auf DNA Ebene, so dass auch regulatorische Effekte betrachtet und Insertion und Deletionen bewertet werden können.
http://www.mutationtaster.org/

4.2.1 MutationTaster

Schwarz, J. M., Rödelsperger, C., Schuelke, M. & **Seelow, D.**
MutationTaster evaluates disease-causing potential of sequence alterations.
Nature Methods, August 2010, 575–576, 7

Wie im Abschnitt Variantenbewertung (2.6) geschildert, ist es vor funktionellen Analysen zur Überprüfung der Kausalität einer potentiellen Krankheitsmutation sinnvoll, deren mögliche Auswirkungen auf den Organismus zuerst durch eine *in silico* Analyse zu überprüfen. Für diese Aufgabe standen in der Vergangenheit zwar eine Reihe von web-basierten Computerprogramme bereit (z.B. PolyPhen-1[44] oder SNAP[45]), alle diese Programme zeichneten sich aber durch eine relativ geringe Vorhersagegenauigkeit aus und waren zudem auf die Analyse der Auswirkungen des Austauschs einer einzelnen Aminosäure in einem Protein beschränkt. Neben der eingeschränkten Funktionalität hatte dies auch zur Folge, dass zunächst die geänderte Transkriptsequenz in eine Aminosäuresequenz umgewandelt werden musste, um eine Variante bewerten zu können. Zudem wurden so weitere Auswirkungen der Variante, zum Beispiel auf das Splicing, nicht betrachtet.

Nach meinen eigenen Erfahrungen gab es darüber hinaus in vielen Arbeitsgruppen kein einheitliches Vorgehen zur Bewertung von Varianten; viele hierfür sinnvolle Ressourcen wie beispielsweise dbSNP[46] oder HapMap[47] wurden von einigen Forscherinnen und Forschern gar nicht konsultiert und auch das Vorhandensein multipler Transkripte, in denen die jeweilige Veränderung unterschiedliche Auswirkungen haben könnte, wurde bisweilen ignoriert. Im Rahmen des Studiengangs 'Molecular Medicine' an der Charité entschloss ich mich deshalb, ein zweimonatiges Praktikum zur Entwicklung eines computerbasierten Schemas zur Variantenbewertung anzubieten, das nicht auf die Bewertung von nicht-synonymen Varianten beschränkt sein sollte. Dabei wurden die Benutzer durch einen Entscheidungsbaum geführt, in dem sie Fragen zur Variante beantworten mussten und gegebenenfalls aufgefordert wurden, auf die entsprechenden WWW-Ressourcen (wie zum Beispiel nnsplice[48] zur möglichen Veränderung von *splice sites*) zuzugreifen und die Ergebnisse der externen Datenquellen oder Programme einzutragen. Abhängig von den Antworten wurde dann die nächste Frage ausgewählt und am Ende anhand der Antworten sowie der Ergebnisse eine Bewertung des Krankheitspotentials vorgenommen (Abbildung 9). Zum Training und zur Validierung wurde nur eine kleine Datenreihe aus 50 bekannten Krankheitsmutationen und Polymorphismen eingesetzt; durch eine unterschiedliche Wichtung der Teilergebnisse konnte eine Genauigkeit von etwa 80% erzielt werden.

Aufgrund dieser recht guten Werte wurde das Projekt zu einer Doktorarbeit ausgebaut[49]. Die so entwickelte Version, MutationTaster, integriert alle nötigen Datenquellen oder Computerprogramme und verfügt über ein einfach aufgebautes Interface, in das die Anwender die gefun-

dene DNA-Variante wahlweise über die Position oder im Sequenzkontext eingeben können und nach weniger als einer Sekunde eine Vorhersage erhalten. MutationTaster arbeitet sowohl auf der Protein- als auch auf der DNA-Ebene und ist somit in der Lage, auch synonyme Varianten zu bewerten. Die Bewertung erfolgt über einen Bayes-Klassifikator, der mit insgesamt etwa 600.000 Varianten mit bekanntem Krankheitspotential trainiert wurde. MutationTaster erreicht dabei eine Genauigkeit von insgesamt etwa 90%, die aber in einzelnen Bereichen, für die wenig Trainings- und Testdaten zur Verfügung standen (wie beispielsweise synonyme Krankheitsmutationen), deutlich geringer ist.

In einem direkten Vergleich mit diversen anderen Vorhersageprogrammen für Varianten, die einen einzelnen Aminosäureaustausch bewirken, schnitt MutationTaster mit 86% deutlich besser ab als alle anderen Programme. Weitere Informationen finden sich in der Originalarbeit, die auf den nächsten Seiten folgt.

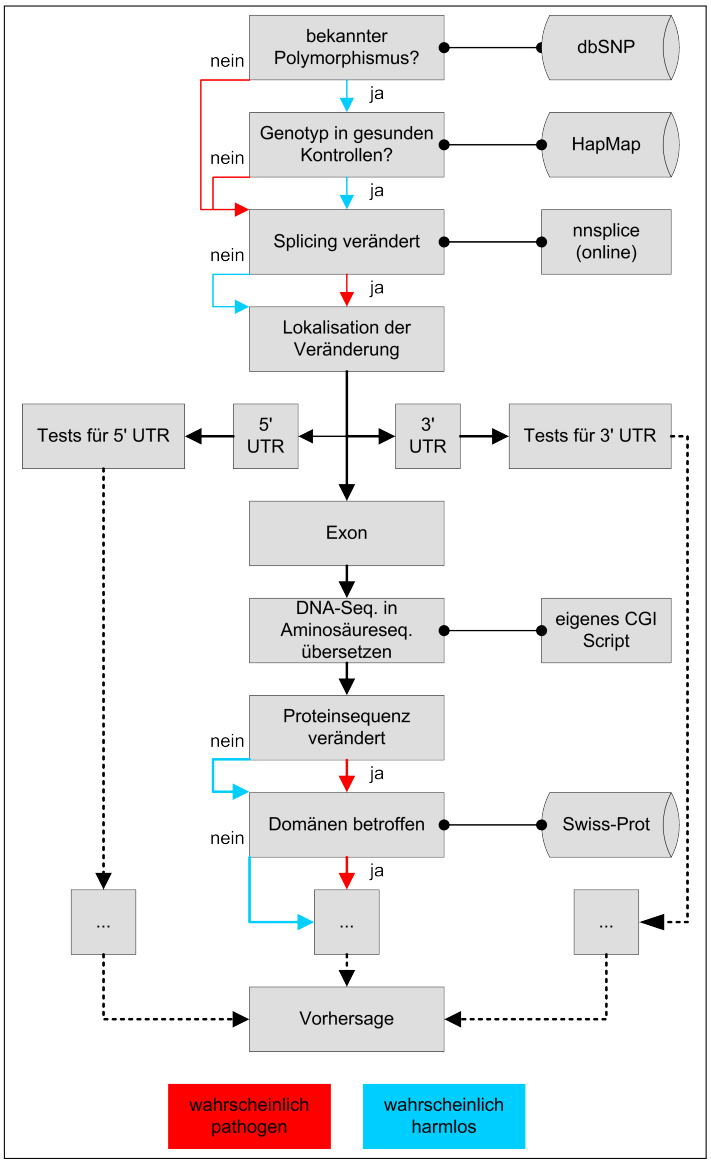


Abb. 9: **Variantenbewertung als Entscheidungsbaum**

Die Abbildung zeigt einen vereinfachten Teil des Entscheidungsbaums, der dem Vorläufer von MutationTaster zugrunde lag. Den Nutzern wurden von einem Web-Interface verschiedene Fragen gestellt, abhängig von der Antwort wurden sie zur nächsten Frage weitergeleitet. Zum Teil mussten die Fragen durch die Benutzung externer Programme bzw. Websites beantwortet werden. Abhängig von den Antworten oder der Ergebnisse der externen Datenquellen oder Programme steigt (rote Pfeile) oder sinkt (blaue Pfeile) die Wahrscheinlichkeit einer pathogenen Mutation.
Die Wichtung der einzelnen Faktoren wurde manuell anhand empirischer Daten festgelegt. Die gestrichelte Linien zeigen Teile des Entscheidungsbaums, die in dieser Abbildung nicht dargestellte werden.

Aufgrund der veränderten Anforderungen durch Hochdurchsatzsequenzierugen haben wir in der Folgezeit eine verbesserte Version entwickelt, MutationTaster2 wird im nächsten Abschnitt (4.2.2) vorgestellt. MutationTaster kann unter der URL http://www.mutationtaster.org/ kostenlos verwendet werden.

Die Originalarbeit wurde 2010 in *Nature Methods* publiziert und wird hier nicht wiedergegeben.

Schwarz, J. M., Rödelsperger, C., Schuelke, M. & **Seelow, D.**
MutationTaster evaluates disease-causing potential of sequence alterations.
Nature Methods, August 2010, 575–576, 7
https://doi.org/10.1038/nmeth0810-575

4.2.2 MutationTaster2

Schwarz, J. M., Cooper, D. N., Schuelke, M. & **Seelow, D.**
MutationTaster2: Mutation prediction for the deep-sequencing age.
Nature Methods, April 2014, 361–362, 11

Seit der ersten erfolgreichen Anwendung zur Aufklärung der Ursache einer monogenen Erkrankung im Jahr 2010[16], konnten durch Hochdurchsatzsequenzierungen kompletter Exome oder Genome viele weitere Krankheitsmutationen identifiziert werden. Bis ins Jahr 2013 wurden so die genetischen Ursachen von mehr als 70 monogenen Erkrankungen aufgedeckt[17].

Allerdings sind die meisten DNA-Varianten in Genen, die durch Exom- und insbesondere Genomsequenzierungen gefunden werden, entweder synonym oder liegen außerhalb der kodierenden Sequenz. Ihre Auswirkungen auf die Funktion des Proteins sind schwer vorherzusagen, da sie weniger offensichtlich sind als die von Aminosäureaustauschen, vorzeitigen Stopcodons oder gar Leserastermutationen (*frameshifts*). Die meisten Forscher - und auch die damals gängigen Computerprogramme zur Variantenbewertung (z.B. PolyPhen-2[43] oder SIFT[42]) - haben sich deshalb bislang auf die 'niedrig hängenden Früchte' der nicht-synonymen Varianten konzentriert.

Doch auch synonyme oder nicht-kodierende Varianten können Krankheiten hervorrufen, beispielsweise durch eine veränderte Expression oder Splicing. So führen nur etwa 55% der in der kommerziellen Version der Human Gene Mutation Database[10] (HGMD) gespeicherten Mutationen zu einem Aminosäureaustausch oder zu einem vorzeitigen Stopcodon.

Um dieses Problem zu adressieren, haben wir MutationTaster[14] stark verbessert. Die aktuelle Version, MutationTaster2[15], wurde gezielt dafür entwickelt, das Krankheitspotential solcher Varianten vorherzusagen. Neben neuen Tests wurden dazu die umfangreiche Sammlung an experimentell validierten funktionellen DNA-Elementen aus dem ENCODE-Projekt[50], wie zum Beispiel Transkriptionsfaktorbindestellen, integriert. Durch diese Erweiterungen und einen deutlich vergrößerte Satz an Trainingsfällen (mehr als 6 Millionen Polymorphismen aus dem 1000-Genom-Projekt (1000G)[19] und mehr als 100.000 bekannte Krankheitsmutationen aus der kommerziellen Version der HGMD) konnte die durchschnittliche Vorhersagegenauigkeit deutlich verbessert werden (Genauigkeit 90,5%, Sensitivität 90,5%, Spezifität 90,9%). Diese Werte sind über alle drei Vorhersagemodelle (synonyme/nicht-kodierende Varianten, nicht-synonyme Varianten, größere Auswirkungen auf die Aminosäuresequenz) konstant. Darüber hinaus konnten wir die Vorhersagegeschwindigkeit auf etwa 100 ms pro Analyse verringern. Um die Bewertung der vielen tausend Varianten zu erleichtern, die in einer Exomsequenzierung anfallen, haben wir ein automatisches Analysesystem entwickelt, das es den Anwendern erlaubt, ihre kompletten Genotypdateien im Standardformat VCF auf unseren Webserver hochzuladen und dort komfortabel auszuwerten. Varianten werden dabei parallel analysiert, die Kapazität des Systems liegt bei etwa 500.000 Varianten pro Stunde - die Auswertung einer kompletten Exomsequenzierung wird so in weniger als einer halben Stunde abgeschlossen.

Eine weitere Verbesserung stellt die Integration bekannter Polymorphismen aus dem 1000G sowie bekannter Krankheitsmutationen aus der nicht-kommerziellen Version der HGMD und NCBI ClinVar[51] dar. Bekannte Krankheitsmutationen werden automatisch als krankheitsverursachend eingestuft, MutationTaster2 zeigt zudem Informationen über die zugrundeliegende Krankheit an. Varianten, die in mehr als 4 gesunden Personen homozygot auftreten, werden automatisch als Polymorphismen erkannt - bei allen anderen Varianten aus dem 1000G werden die Genotyphäufigkeiten angezeigt. Zudem erlaubt das Abfragesystem auch den Ausschluss von Genotypen, die im 1000G mit benutzerdefinierten Häufigkeiten heterozygot und/oder homozygot vorkommen, um so die Zahl der möglichen Krankheitsvarianten drastisch zu verringern.

In einem Test mit dem Exom einer gesunden Person erreichte MutationTaster2 so eine Falsch-Positiv-Rate von nur 1% und war damit ähnlichen Programmen wie SIFT, PolyPhen-2 und PROVEAN[52] deutlich überlegen.

Bei einem Test mit nicht-synonymen DNA-Varianten mit bekanntem Effekt (harmloser Polymorphismus oder krankheitsverursachend) schnitt MutationTaster2 mit 88% um 2 Prozentpunkte besser als die Vorgängerversion und um 4 Prozentpunkte besser als die anderen Programme ab.

Die Originalarbeit mitsamt der zusätzlichen Informationen (*Supplement*) wird auf den folgenden Seiten wiedergegeben.

MutationTaster kann unter der URL http://www.mutationtaster.org/ kostenlos verwendet werden.

Die Originalarbeit wurde 2014 in *Nature Methods* publiziert und wird hier nicht wiedergegeben.

Schwarz, J. M., Cooper, D. N., Schuelke, M. & **Seelow, D.**
MutationTaster2: Mutation prediction for the deep-sequencing age.
Nature Methods, April 2014, 361–362, 11 4
https://doi.org/10.1038/nmeth.2890

4.2.3 Exomiser

Robinson, P. N., Köhler, S., Oellrich, A., Sanger Mouse Genetics Project, Wang, K., Mungall, C. J., Lewis, S. E., Washington, N., Bauer, S., **Seelow, D.**, Krawitz, P., Gilissen, C., Haendel, M. & Smedley, D.
Improved exome prioritization of disease genes through cross-species phenotype comparison.
Genome Research, February 2014, 340–348, 24

Der Exomiser ist eine Software zur Klassifizierung von Varianten anhand ihres vermuteten Effekts auf ein Gen bzw. Protein sowie der Wahrscheinlichkeit, dass das mutierte Gen den Phänotyp auslösen könnte. Dies geschieht über einen Algorithmus namens *PHenotypic Interpretation of Variants in Exomes* (PHIVE), der die Ähnlichkeit zwischen menschlichen Krankheiten und bekannten Mausphänotypen[33], die durch Genmutationen oder durch das gezielte Ausschalten von Genen (*gene knock-out*) ausgelöst werden, untersucht. Der Effekt auf das Protein wird durch die Kombination der Vorhersagen verschiedener Variantenbewertungsprogramme (MutationTaster1[14], PolyPhen-2[43] und SIFT[42]) errechnet. Einbezogen wird zudem das Vorkommen von Varianten im 1000-Genom-Projekt (1000G)[19].

Exomiser ist web-basiert und ermöglicht es Forschern, die Varianten aus vollständigen Exomsequenzierungen anhand der oben beschriebenen Kriterien zu priorisieren. Wie MutationTaster2 erlaubt auch diese Software, Varianten vorab nach bestimmten Kriterien, wie zum Beispiel der Abdeckung in der Sequenzierung oder dem Vorkommen im 1000G zu filtern. Die untersuchte Krankheit kann wahlweise durch ihren Titel in OMIM[9] oder über die Symptome der Patienten (über die Human Phenotype Ontology[24]) eingegeben werden.

Während der Exomiser natürlich nicht in der Lage ist, die krankheitsverursachende Mutation in allen Fällen zielsicher zu finden, so wurde diese doch in den meisten Testfällen an oberster Stelle platziert. Durch den Einsatz des PHIVE Algorithmus' kann so eine deutliche Verbesserung gegenüber einem nur auf den Variantenbewertungsprogramme basierenden Ansatz erreicht werden.

Der Exomiser kann unter der URL http://www.sanger.ac.uk/resources/databases/exomiser/ kostenlos verwendet werden.

Die Originalarbeit wurde 2014 in *Genome Research* publiziert und wird hier nicht wiedergegeben.

Robinson, P. N., Köhler, S., Oellrich, A., Sanger Mouse Genetics Project, Wang, K., Mungall, C. J., Lewis, S. E., Washington, N., Bauer, S., **Seelow, D.**, Krawitz, P., Gilissen, C., Haendel, M. & Smedley, D.

Improved exome prioritization of disease genes through cross-species phenotype comparison.

Genome Research, February 2014, 340–348, 24

https://doi.org/10.1101/gr.160325.113

4.3 Copy-Number Varianten

CNVinspector dient der Suche nach Variationen der Kopienzahl (CNV) chromosomaler Regionen, die in erkrankten Personen komplett deletiert, hemizygot oder dupliziert sein können. Das Programm erlaubt es dabei, wahlweise einzelne Patienten oder Kohorten zu betrachten. Die in den Patienten gefundenen CNV können sowohl gegen eigene Kontrollen als auch gegen öffentliche Daten gefiltert werden, um so diejenigen CNVs hervorzuheben, die in gesunden Personen nicht vorkommen und somit ein höheres Krankheitspotential aufweisen.
http://www.cnvinspector.org/

4.3.1 CNVinspector

Knierim, E., Schwarz, J. M., Schuelke, M. & **Seelow, D.**
CNVinspector: a web-based tool for the interactive evaluation of copy number variations in single patients and in cohorts.
Journal of Medical Genetics, August 2013, 529–533, 50

Wie in der Einleitung im Abschnitt 2.7 *Copy-Number Varianten* beschrieben, spielen Variationen der Kopienzahl (*copy number variants* - CNV) in der genetischen Diagnostik von Kindern mit Entwicklungsstörungen und geistiger Behinderung eine große Rolle, sie werden außerdem als Risikofaktor für komplexe Krankheiten betrachtet. Allerdings kommen CNVs auch in gesunden Personen häufig vor. Eine in einem Patienten gefundene Copy-Number Variation kann deshalb nicht automatisch als Krankheitsursache betrachtet werden.

Um CNVs in Patienten leichter auf ihr Krankheitspotential hin beurteilen zu können, haben wir die web-basierte Software CNVinspector[32] entwickelt. CNVinspector erlaubt es Ärzten, die CNVs ihrer Patienten mit bekannten CNVs zu vergleichen. Dabei können sowohl Übereinstimmungen mit bekannten krankheitsverursachenden CNVs gesucht werden als auch nur solche CNVs herausgefiltert werden, die nicht in gesunden Probanden gefunden wurden. Die Software verfügt deshalb über eine Datenbank, die sowohl die CNVs enthält, die in genomweiten Studien komplexer Krankheiten in gesunden und betroffenen Personen detektiert wurden, als auch die Daten aus DECIPHER[53], einer Datenbank, in der CNVs von Patienten und die damit direkt verbundenen Phänotypen erfasst werden.

Die Software ist jedoch nicht nur für die Analyse einzelner Patienten geeignet, sondern auch für den Vergleich zweier Kohorten, wie er beispielsweise in Studien komplexer Krankheiten erforderlich ist. Um die An- oder Abreicherung bestimmter CNVs als Krankheitsursache untersuchen zu können, ist es dabei möglich, die Maximal- und Mindesthäufigkeiten bzw. -frequenzen in den Fällen und Kontrollen anzugeben, und beispielsweise nur solche CNVs anzuzeigen, die in einer Kohorte deutlich häufiger sind als in einer anderen. Um einen möglichst umfassenden Vergleich zu gestatten, kann eine Kohorte (oder ein Einzelfall) auch gegen eine Vielzahl verschiedener anderer Kohorten gefiltert werden. Die Software bietet außerdem die Möglichkeit, den Vergleich auf solche CNVs zu begrenzen, die mit einer identischen Methode detektiert wurden, um so Unterschiede durch den Einsatz verschiedener Plattformen auszuschließen.

Da die Software auch von klinisch tätigen Ärzten ohne umfassende Computerkenntnisse benutzt werden soll, wurden die Interfaces so gestaltet, dass sie auch ohne informatische Kenntnisse

verwendet werden können. Die gefundenen CNVs werden sowohl als Liste als auch visuell dargestellt. Hierbei werden Verluste bzw. Gewinne genomischen Materials und die verschiedenen Kohorten farblich unterschieden. Zusätzlich werden die in den Regionen liegenden Gene angezeigt, um so schnell mögliche Krankheitsgene zu erkennen. CNVinspector bietet zudem eine direkte Schnittstelle zu GeneDistiller (siehe Abschnitt 3.1.1), um zusätzliche Informationen über die Gene nachschlagen zu können oder aber mit den im oben genannten Abschnitt beschriebenen Möglichkeiten nach dem wahrscheinlichsten Krankheitsgen zu suchen.

CNVinspector ist unter der URL http://www.cnvinspector.org/ kostenlos benutzbar.

Die Originalarbeit wurde 2013 im *Journal of Medical Genetics* publiziert und wird hier nicht wiedergegeben.

Knierim, E., Schwarz, J. M., Schuelke, M. & **Seelow, D.**
CNVinspector: a web-based tool for the interactive evaluation of copy number variations in single patients and in cohorts.
Journal of Medical Genetics, August 2013, 529–533, 50
https://doi.org/10.1136/jmedgenet-2012-101497

5 Diskussion

Während die hier vorgestellten computergestützten Verfahren zur Aufklärung der molekularen Ursachen genetischer Krankheiten von vielen Arbeitsgruppen erfolgreich verwendet werden, so besitzen sie natürlich auch einige Schwachstellen. Vielen der von mir (mit-)entwickelten Computerprogramme ist gemein, dass sie anfällig gegenüber eine unvollständiger Datenlage sind: MutationTaster kann beispielsweise eine Mutation, die eine metallbindende Aminosäure in einem Enzym verändert, besser als krankheitsverursachend identifizieren, wenn überhaupt bekannt ist, dass es sich um eine Metallbindestelle handelt. Fehlt diese Information, so kann zwar unter Umständen durch die Art des Austausches oder die Konservierung auf DNA- und Proteinebene dennoch eine richtige Vorhersage erfolgen, diese wäre jedoch mit einer erheblich größeren Unsicherheit behaftet. Diese Problematik tritt in GeneDistiller ebenfalls auf, wenn wichtige Informationen nicht vorliegen: so erfordert zum Beispiel die gezielte Suche nach mitochondrialen Proteinen natürlich, dass diese subzelluläre Lokalisation auch in unserer Datenbank verzeichnet ist.

5.1 GeneDistiller

Neben fehlenden Geninformationen ist das größte Hindernis bei der Suche nach Krankheitsgenen mit GeneDistiller ein falsches Krankheitsmodell. Wird beispielsweise strikt nach einem mitochondrialen Protein gesucht, obwohl das vom Krankheitsgen kodierte Protein nicht mitochondrial ist, sondern über die Interaktion mit mitochondrialen Membranproteinen zu einer Mitochondriopathie führt, so kann es mit diesem Ansatz unmöglich entdeckt werden. Der Vorteil des hier vorgestellten Verfahrens gegenüber automatischen Lösungen, die Einbindung des Hintergrundwissens der Forscher und Kliniker, wird so zu einem Nachteil. Priorisierungsprogramme, die ohne ein spezifiziertes Krankheitsmodell sondern zum Beispiel nur über Proteininteraktionsdaten arbeiten, haben hier deutliche Vorteile. Allerdings unterliegen diese noch stärker dem Problem einer unvollständigen Datenlage - sind keine Interaktionsdaten vorhanden, so kann das Programm auch keine sinnvolle Vorhersage treffen. GeneDistiller bietet hier durch die Möglichkeit, verschiedene Modelle vorzugeben, eine gewisse Redundanz. Durch die Anzeige weiterer Informationen zu den vorgeschlagenen Kandidatengenen direkt in der Ausgabemaske kann das Modell gegebenenfalls auch leicht angepasst werden, um so weitere Datenquellen in die Priorisierung einzubeziehen oder stärker zu wichten. Darüber hinaus kann durch den Einsatz der Priorisierungsfunktion ohne das Filtern 'unpassender' Gene sichergestellt werden, dass kein Gen 'verloren geht', also aufgrund der Filteroptionen gar nicht erst angezeigt wird. Unter Umständen ergeben sich aus Informationen, die bei der Priorisierung nicht berücksichtigt wurden (zum Beispiel Phänotypen in Mausmodellen) weitere Hinweise für die Rolle eines Gens in der Krankheitsentstehung.

Allerdings ist dies bei bislang nicht näher charakterisierten Genen natürlich nicht möglich - für die Identifizierung von Genen, für die keine oder wenig Informationen vorliegen, ist dieses Verfahren deshalb nicht geeignet. Automatische Priorisierungsprogramme schneiden hier aber nur in den Fällen besser ab, in denen die für die Priorisierung verwendeten Daten vorliegen.

Eine Schwierigkeit im Betrieb von GeneDistiller liegt in der erforderlichen permanenten Aktualisierung der Daten. Hierfür wäre eine automatische Lösung wünschenswert, bei der die Datenquellen in regelmäßigen Abständen nach Neuerungen durchsucht und diese dann in die Datenbank eingespielt würden. Dies ist jedoch leider nicht möglich, da sich bei einigen der in GeneDistiller enthaltenen Daten das Format häufig ändert (dies betrifft beispielsweise Ensembl[54]) und die OMIM-Daten[9] nur auf Anfrage und darauf erfolgender Vergabe eines Zugangs zum FTP-Server heruntergeladen werden können. Die Aktualisierung von GeneDistiller muss daher manuell erfolgen, in den meisten Fällen sind dafür auch Anpassungen des Codes zum Einlesen der Daten

erforderlich. Aus diesen Gründen wird GeneDistiller nur etwa vierteljährlich aktualisiert. Zur Information der Benutzer über die Aktualität der Daten werden am Ende der Ausgabe der Ergebnisse die Zeitpunkte der letzten Aktualisierung der einzelnen Teildaten wiedergegeben.

Bereits realisierte, aber in der Originalarbeit noch nicht enthaltene, Erweiterungen von GeneDistiller werden im Abschnitt 3.1 vorgestellt.

Ausblick

In Gesprächen mit anderen Wissenschaftlern zeigt sich, dass viele von ihnen zwar vom Funktionsumfang von GeneDistiller beeindruckt sind, die Möglichkeiten zur Filterung und Priorisierung allerdings nur von wenigen Anwendern in vollem Umfang eingesetzt werden. Ein oft genannter Grund hierfür ist die Vielzahl der Optionen, die auf ungeübte Benutzer verwirrend wirken kann. GeneDistiller wird daher häufig als ein bloßes Werkzeug zur Wiedergabe von Geninformationen verwendet, wobei die weitergehenden Möglichkeiten (beispielsweise zum Ausschluss von Genen anhand der Expressionsdaten) ungenutzt bleiben.

Wir planen deshalb, das Interface von GeneDistiller so umzugestalten, dass die Anwender einfacher die für sie relevanten Optionen finden. Dazu werden wir eine neue Startseite erstellen, die häufige Modelle für die Kandidatengensuche (wie zum Beispiel die Suche nach in einem Gewebe exprimierten Genen oder den Vergleich aller Gene einer Region mit bekannten Krankheitsgenen) auflistet und in der Eingabemaske die hierfür nicht erforderlichen Optionen ausblendet.

Ein weiteres Vorhaben ist die Einbindung weiterer Spezies, insbesondere der Maus. Eine Entwicklungsversion zur Suche nach Genen in der Maus ist bereits lauffähig, allerdings noch nicht ausreichend getestet. Der zugrunde liegende Programmcode ist derart gestaltet, dass auch das Hinzufügen weiterer Spezies unproblematisch vonstatten gehen kann.

Vor einer erneuten Publikation werden wir zudem die Routine zur Erstellung von Primern für die Sequenzierung sämtlicher Exons verbessern, die derzeit häufig noch Vorschläge für Primerpaare liefert, die zu längeren Sequenzen führen würde als vorgegeben.

Geplant ist zudem die Integration von GeneDistiller in die Analysepipeline von MutationTaster (siehe Abschnitt 5.7).

5.2 HomozygosityMapper

Im Vergleich mit Computerprogrammen zur Kopplungsanalyse ist HomozygosityMapper um mehrere Größenordnungen schneller. Dieser Geschwindigkeitsvorteil wird jedoch durch Nachteile erkauft: in der ersten publizierten Version berücksichtigte HomozygosityMapper lediglich die genetischen Informationen der Betroffenen. Dies stellt in einem klassischen Fall der Homozygotiekartierung (mehrere Betroffene, oft 'sporadische' Krankheitsfälle aus verschiedenen konsanguinen Familien) kein Problem dar. In solchen Fällen ist kein längerer einheitlicher 'Krankheitshaplotyp' in den betroffenen Personen zu erwarten sondern lediglich eine homozygote Region in allen Betroffenen, inmitten derer der Krankheitslokus liegt. Werden aber einzelne Familien mit nur ein oder zwei betroffenen Personen analysiert, so ist es hilfreich, nach gemeinsamen homozygoten Haplotypen der Betroffenen zu suchen. Darüber hinaus dürfen, zumindest bei einer vollständigen Penetranz der Erkrankung, gesunde Familienmitglieder diesen Haplotyp nicht homozygot aufweisen. Die zweite Version von HomozygosityMapper erlaubt es, auch in derartigen Konstellationen Krankheitsloci bzw. -regionen zu identifizieren. Allerdings wird hierbei ein Algorithmus verwendet, der strikt auf die Minimierung von falsch-negativen-Ergebnissen (Fehler 2. Art) eingestellt ist, dabei werden falsch positive Bewertungen (Fehler 1. Art) in Kauf genommen. Der Grund für den Einsatz dieses Algorithmus ist, dass viele der auf SNP-Chips enthaltenen SNPs nicht informativ, also in allen Personen homozygot, sind. Es ist daher nicht ohne Weiteres zu

klären, ob es sich bei anscheinend gleichen Haplotypen in betroffenen und gesunden Familienmitgliedern wirklich um identische chromosomale Regionen oder nur um Abschnitte mit vielen nicht informativen SNPs handelt. Um sicherzustellen, dass keine potentielle Krankheitsregion übersehen wird, wurde die Software so entwickelt, dass zum Ausschluss von Krankheitsregionen die homozygoten Haplotypen in den Gesunden sehr lang sein müssen. Durch die Möglichkeit, die Genotypen optisch zu inspizieren, bietet HomozygosityMapper aber einen einfachen Weg, Fehler 1. Art leicht zu erkennen. Außerdem wird durch die Möglichkeit einer Kopplungsanalyse, die sich auf die Krankheitsloci in Frage kommenden Regionen des Genoms beschränkt, eine weitere Methode zur Einbeziehung der gesunden Personen angeboten.

In den ersten Jahren nach der Publikation der initialen Version zeigte sich eine weitere Schwierigkeit: Viele Forscherinnen und Forscher zögerten, ein Verfahren einzusetzen, das - im Gegensatz zu den LOD Scores in der Kopplungsanalyse - kein Maß der statistischen Sicherheit der identifizierten Regionen bietet. Durch das Aufkommen von NGS-Ansätzen hat hier aber eine erfreuliche Verschiebung der Kriterien stattgefunden: Während in der Vergangenheit häufig die Kombination aus einem Kopplungsbefund sowie einer durch einen Aminosäureaustausch vermuteten Beeinträchtigung der Proteinfunktion und des Ausschlusses einer Variante in hundert oder mehr gesunden Personen als 'Beweis' akzeptiert wurde (dies wird durch das häufige Auftreten angeblicher Krankheitsmutationen im 1000-Genom-Projekt[19] deutlich[55]), wird inzwischen normalerweise ein funktioneller Nachweis der Auswirkungen der Mutation gefordert. Die Existenz eines hohen LOD Scores ist dabei in den Hintergrund getreten, auch, weil dieser in Projekten mit nur einer Familie und wenigen Meiosen gar nicht erreicht werden kann. Die Gen-kartierung mit Hilfe von HomozygosityMapper ist deshalb inzwischen weithin akzeptiert, wie die weit mehr als 100 Publikation, die diese Software zitieren, zeigen.

Ausblick

Die Zunahme von Exomsequenzierungen als initiales Genotypisierungsverfahren wird die Verwendung von SNP-Genotypisierungschips marginalisieren. Bis auf die permanent erfolgende Einbindung neuer SNP-Chips in HomozygosityMapper sehen wir deshalb keine Notwendigkeit für wesentlichen Verbesserungen in diesem Bereich.

Die Verwendung von NGS-Genotypen in HomozygosityMapper ist hingegen verbesserungswürdig: HomozygosityMapper wurde für durch Genotypisierungschips ermittelte SNP-Genotypen entwickelt und speichert Genotypen daher entweder als 1) homozgyot für das Referenzallel, 2) heterozygot, 3) homozygot für das variante Allel oder 4) als nicht typisiert. Die eigentlichen Allele werden dabei nicht gespeichert - dies wäre aber erforderlich, um direkt aus HomozygosityMapper Varianten in möglichen Krankheitsregionen durch MutationTaster bewerten zu lassen. Zudem erlaubt die Verwendung von NGS-Genotypen auch die Ausnutzung von Varianten, die mehr als 2 verschiedene Allele aufweisen - also zum Beispiel Mikrosatelliten. Eine zukünftige Version von HomozygosityMapper wird deshalb ein vollkommen anderes Modell zur Speicherung der Allele aufweisen und zudem die oben genannte Möglichkeit bieten, alle Varianten der gefundenen homozygoten Regionen direkt durch MutationTaster analysieren zu lassen und als Ergebnis eine Liste der in Frage kommenden Varianten, sortiert nach ihrem Krankheitspotential, ausgeben. Dabei wird in einem späteren Schritt auch das Wahrscheinlichkeit, dass die betroffenen Gene den untersuchten Phänotyp hervorrufen, berücksichtigt werden (siehe Abschnitt 5.7).

In seltenen Fällen kommen auch in konsanguinen Familien compound-heterozygote Krankheitsmutationen vor. Während dies in Studien seltener Krankheiten unwahrscheinlich ist, steigt die Wahrscheinlichkeit mit der Häufigkeit der Krankheitsallele in der Bevölkerung, wie zum Beispiel bei der Mukoviszidose, deutlich an[56]. Ich entwickle zur Zeit deshalb eine Software, die Regionen mit gleichen Genotypen im Erbgut verschiedener Betroffener identifiziert und somit in der Lage

ist, sowohl homozygote als auch compound-heterozygote Krankheitsloci zu bestimmen. Gesunde Personen, die die selben Haplotypen tragen, werden dabei zum Ausschluss genutzt. Eine erste Version des Programms konnte bereits erfolgreich zur Bestimmung der krankheitsverursachenden Mutation in einem mitochondrialen DNA Depletionssyndrom eingesetzt werden[57].

5.3 MutationTaster

MutationTaster besitzt auch in der aktuellen Version vier gravierende Schwachstellen: die fehlende Unterscheidung zwischen Funktionsgewinn- und -verlustmutationen (*gain of function*, GOF bzw. *loss of function*, LOF), die Beschränkung auf monogene Krankheiten und intragenische Varianten sowie die Nichtberücksichtigung der Genfunktion in der Pathogenese.

Unterscheidung von GOF- und LOF-Mutationen

Die Unterscheidung von GOF- und LOF-Mutationen durch die Spezifikation des Erbgangs vor der Analyse wäre wünschenswert, ist allerdings aufgrund der Datenlage nicht ohne weiteres möglich: Weder HGMD[10] noch ClinVar[51] geben an, ob die gespeicherten Varianten die jeweilige Krankheit durch einen Funktionsgewinn (in den meisten Fällen dominant) oder -verlust (in den meisten Fällen rezessiv) auslösen. Durch den manuellen Abgleich mit anderen Datenquellen wie OMIM[9] oder der Primärliteratur ließen sich zwar die Vererbungsmodelle für die Krankheiten ermitteln, dies würde aber einen hohen Zeitaufwand erfordern, selbst wenn über Textmining eine Vorauswahl getroffen würde. Für einige Krankheitsgene stellt die HPO[24] Informationen über den Erbgang zur Verfügung, allerdings beruhen diese ebenfalls auf der Extraktion der Daten aus OMIM.
Erschwert würde eine Gruppierung der Varianten in GOF und LOF aber dadurch, dass einige Gene sowohl rezessiv als auch dominant vererbte Krankheiten hervorrufen können - abhängig von der jeweiligen Mutation. Für ein sinnvolles Training wäre aber eine klare Einstufung jeder Krankheitsmutation in GOF oder LOF erforderlich.

Komplexe Krankheiten

'Volkskrankheiten' wie Asthma oder Schizophrenie sind komplex, das heißt, das bei ihnen genetische und Umweltfaktoren zusammenwirken. Vermutet werden hier schwache Effekte einzelner, häufiger, genetischer Varianten, die das Krankheitsrisiko jeweils nur geringfügig erhöhen[58]. MutationTaster wurde darauf trainiert, das Krankheitsrisiko seltener Varianten mit einem großen Effekt richtig vorherzusagen, versagt jedoch bei häufigen Varianten mit einem geringen Effekt. Allerdings liegt, da bislang keine derartige Krankheit vollständig aufgeklärt werden konnte (bislang wurden lediglich für familiäre monogene Formen eindeutige Ursachen identifiziert[58]), auch kein experimentell verifiziertes Modell für komplexe Krankheiten vor.

Extragenische Varianten

MutationTaster arbeitet bislang Transkript-bezogen. Varianten, die außerhalb eines Gens, also zum Beispiel im Promotor, liegen, werden von MutationTaster deshalb derzeit nicht berücksichtigt. Wir haben inzwischen damit begonnen, diesen Schwachpunkt zu beheben (siehe weiter unten im Abschnitt 5.3).

Rolle der Gene in der Pathogenese

MutationTaster untersucht, wie auch andere Variantenbewertungsprogramme wie PolyPhen-2[43] oder SIFT[42], lediglich den Effekt einer Variante auf das Genprodukt. Dabei wird allerdings nicht berücksichtigt, ob das betreffende Gen überhaupt das Potential besitzt, die untersuchte Krankheit auszulösen. Die Einbeziehung der phänotypischen Informationen verbessert die Klassifizierung von DNA-Varianten erheblich, so wurde zum Beispiel in meisten Testfällen in der Exomiser-Publikation[22] die krankheitsverursachende Mutation mit dem höchsten Krankheitspotential bewertet. Auch in meinen Tests mit eXtasy[21] wurde die krankheitsverursachende Mutation normalerweise zumindest in den obersten 5% der nach dem Krankheitspotential sortierten Varianten platziert.

Obschon dieses Verfahren natürlich nicht garantiert, dass die ursächliche Mutation auch tatsächlich identifiziert werden kann, so können dennoch eine Vielzahl von Varianten mit relativ großer Sicherheit als Krankheitsursache ausgeschlossen werden. Allerdings unterliegt dieser Ansatz der weiter oben in der Diskussion von GeneDistiller beschriebenen Notwendigkeit entsprechender Daten für die Genpriorisierung. Dies kann insbesondere die Identifizierung von Krankheitsgenen, die bislang nur unzureichend charakterisiert wurden (für die beispielsweise keine Interaktionen beschrieben sind), erschweren.

Wir streben deshalb an, GeneDistiller und MutationTaster enger zu verzahnen, um so durch die Nutzung vieler Informationsquellen mit genspezifischen Daten, eine höhere Redundanz für die Bewertung des Krankheitspotentials von Varianten unter Berücksichtigung des Krankheitspotentials den Gens zu erreichen (siehe Abschnitt 5.7).

Nicht-proteinkodierende Gene

MutationTaster ist bislang auf proteinkodierende Gene beschränkt. Derzeit sind zwar nur wenige krankheitsverursachende Mutationen in anderen Genarten bekannt[59], dies kann jedoch auch der Tatsache geschuldet sein, dass bei der Suche nach Krankheitsmutationen normalerweise kodierende Sequenzen betrachtet werden und nicht-kodierende Varianten so schlichtweg nicht gefunden werden können.

Allerdings ist es uns derzeit nicht möglich, aussagekräftige Methoden zur Bewertung des Krankheitspotentials von Varianten in nicht proteinkodierenden Genen zu entwickeln, da aufgrund der geringen Anzahl bekannter Krankheitsmutationen kein sinnvolles Training möglich ist. Sollte die Zahl der bekannten Krankheitsmutationen stärker wachsen und somit ein größerer Trainings- und Testsatz zur Verfügung stehen, werden wir uns diesem Thema jedoch widmen.

Automatische Vorhersagen

Sind für eine Variante Polymorphismen oder Krankheitsmutationen beschrieben, erfolgt eine automatische Bewertung der Variante als 'harmlos' bzw. 'krankheitsverursachend'. Dies kann jedoch zu Fehleinschätzungen führen, da insbesondere die schon vor längerer Zeit beschriebenen 'Krankheitsmutationen' häufig nicht experimentell validiert sind und es sich in Wirklichkeit um seltene Polymorphismen ohne Krankheitspotential handelt. Ist eine Variante sowohl als Krankheitsmutation als auch im 1000G als Polymorphismus annotiert, wird sie deshalb als 'harmlos' eingestuft; in der Ausgabemaske wird aber auf die Beschreibung als Krankheitsmutation verwiesen.

Da einige relativ häufige Krankheitsallele heterozygot keine Krankheit auslösen (dies betrifft zum Beispiel die Mukoviszidose), werden nur solche Varianten automatisch als 'harmlos' klassifiziert, die in mindestens 4 Personen aus dem 1000G homozygot auftreten. Weil das Kollektiv keine Patienten mit schweren monogenen Erkrankungen umfasst, wird so in den meisten Fällen

sichergestellt, dass keine Krankheitsmutationen falsch-negativ als 'harmlos' betrachtet werden. Allerdings versagt dieses Verfahren bei Krankheitsmutationen mit einer niedrigen Penetranz oder einem Auftreten der Krankheit erst im hohen Lebensalter, wie beispielsweise im Falle der Chorea Huntington. Da derartige Krankheiten aber relativ selten auftreten, ist es jedoch unwahrscheinlich, dass das 1000G-Kollektiv mehr als vier Personen umfasst, die die gleiche Krankheitsmutation homozygot tragen. Dazu kommt, dass MutationTaster ohnehin für den Einsatz bei der Suche nach seltenen Varianten mit hoher Penetranz ausgerichtet ist und genutzt wird.

Das Vorkommen von (insbesondere rezessiven) Krankheitsallelen in Polymorphismusdatenbanken ist die Ursache dafür, dass MutationTaster derzeit weder dbSNP[46] noch den *Exome Variant Server** (EVS) zur automatischen Bewertung von Varianten nutzt. Während dbSNP einerseits auch Krankheitsallele enthält (seit Version 131 auch absichtlich) und andererseits häufig keine Genotypenhäufigkeiten zur Verfügung stehen, mit denen gezielt nach dem mehrfachen homozygoten Vorkommen der Variante in Kontrollen gesucht werden könnte, umfasst das Kollektiv des EVS vor allem kranke Personen. Auch wenn es sich bei diesen vor allem um Patienten mit komplexen Krankheiten handelt, kann nicht ausgeschlossen werden, dass sich unter den Varianten auch seltene monogene Krankheitsmutationen verbergen. Wir erwägen aber, die Daten des EVS zur automatischen Bewertung zu nutzen, um so auch seltenere Polymorphismen erkennen zu können, die im 1000G-Kollektiv nicht mindestens vier Mal homozygot vorkommen. Allerdings würde das oben beschriebene Vorkommen monogener Erkrankungen im EVS-Kollektiv eine deutlich höhere Schwelle für eine automatisch Klassifizierung erfordern, so dass sehr seltene Polymorphismen auch mit dieser Methode nicht sicher als solche erkannt würden.

Ausblick

Während die Kosten für Exomsequenzierungen (WES) inzwischen so weit gesunken sind, dass diese als eine Routineuntersuchung in der Erforschung genetischer Krankheiten gelten dürfen, ist dies für komplette Genomsequenzierungen (WGS) bislang noch nicht der Fall. Allerdings ist auch hier damit zu rechnen, dass diese in Zukunft häufiger durchgeführt werden dürften. Neben der Erfassung extragenischer Bereiche bietet die WGS dadurch einen weiteren Vorteil gegenüber der WES, dass die der Sequenzierung vorangehende Anreicherung der Exons entfällt. Probleme in diesem Schritt führen dazu, dass einzelne Exons nicht oder nur unzureichend in die Hochdurchsatzsequenzierung eingeschlossen werden - dies resultiert in einer geringen (oder vollständig fehlenden) Abdeckung und kann so bewirken, dass Varianten in diesen Exons nicht oder nur mit einer sehr geringen Abdeckung, die beispielsweise das Erkennen heterozygoter Genotypen verhindert, erfasst werden. Der größte Gewinn gegenüber dem WES-Verfahren ist aber, dass so auch krankheitsverursachende Varianten außerhalb der kodierenden Bereichen detektiert werden können.

Ich gehe davon aus, dass viele monogene Erkrankungen, in denen bislang trotz NGS-Untersuchungen keine Krankheitsmutation identifiziert werden konnten, durch synonyme Veränderungen oder Mutationen außerhalb der kodierenden Bereiche ausgelöst werden. Zusätzlich zur Möglichkeit des Einsatzes von WGS, beispielsweise zur Ermittlung von (hemizygoten) Deletionen größerer DNA-Segmente (z.B. des Promotorbereiches), wird hier sicherlich auch eine genauere bioinformatische Analyse der nicht kodierenden Varianten erforderlich werden.

Eine Aufgabe in der Zukunft wird es sein, MutationTaster noch gründlicher auf diese Herausforderungen vorzubereiten. Das ENCODE-Projekt liefert bereits umfangreiche Daten zu funktionellen DNA-Elementen. Während diese derzeit schon in MutationTaster integriert sind, werden sie momentan noch nicht in einem zufriedenstellenden Maße genutzt: MutationTaster erkennt lediglich, ob sich eine Variante in einem derartigen Element befindet und nutzt diese Lokali-

*http://evs.gs.washington.edu/EVS/

sation über den Bayes-Klassifikator, um eine Aussage über das Krankheitspotential zu treffen. Dabei werden die Auswirkungen der Variante auf das betroffene Element nicht untersucht. Diesen Schwachpunkt der Analyse werden wir in der nahen Zukunft zu beheben versuchen - dazu werden beispielsweise Computerprogramme zur Detektion von Transkriptionsfaktorbindestellen zum Einsatz kommen. Im Rahmen einer von mir betreuten Masterarbeit wurden dafür bereits Vorleistungen erbracht[60]. Durch einen erfolgreichen DFG-Antrag konnte ich eine Postdoktorandinnenstelle für ebendieses Thema schaffen. Im Zuge dieses Projekts soll auch die Beschränkung auf intragenische Varianten entfallen.

Ein weiterer Punkt, der seit dem Oktober 2014 im Rahmen einer Doktorarbeit bearbeitet wird, ist eine bessere Unterscheidung der Auswirkungen von Varianten in verschiedenen Proteinfamilien. Wir gehen davon aus, dass die Beeinträchtigung einer bestimmten Proteindomäne (wie zum Beispiel einer Lipid-Bindestelle) durch veränderte Aminosäuren in unterschiedlichen Arten von Proteinen unterschiedliche Effekte auslösen dürfte. Für dieses Projekt werden wir die bekannten proteinkodierenden Gene in verschiedene Gruppen einteilen und dann dedizierte Modelle für die jeweiligen Gruppen erstellen und testen. Dabei sollen auch diejenigen Gene identifiziert werden, deren kompletter Funktionsverlust nicht zu einem schweren oder erkennbaren Phänotyp führt. Dies kann zum Beispiel durch die Untersuchung des 1000G nach homozgyoten Stop-Mutationen, die zu einem *nonsense-mediated decay*, also dem Abbauch der mRNA, führen, geschehen. Derartige Gene scheinen als Ursache schwerer rezessiver monogener Erkankungen auszuscheiden, da die im Rahmen des 1000-Genom-Projekts untersuchten Personen nicht an solchen Krankheiten leiden. Um eine mögliche Rolle dieser Gene in der Krankheitsentstehung durch Funktionsgewinnmutationen weitestgehend auszuschließen, muss gleichzeitig ein Abgleich mit bekannten Krankheitsgenen aus OMIM[9] und HGMD[10] erfolgen.

Insbesondere durch das zu erwartende Aufkommen von WGS-Projekten wird sich die Zahl der im Hinblick auf ihr Krankheitspotential zu bewertenden Varianten in der nahen Zukunft erheblich erhöhen. MutationTaster sollte deshalb noch besser als bisher im Stande sein, einerseits das Krankheitspotential von Varianten außerhalb der kodierenden Bereiche von Genen - und auch außerhalb der Gene selbst - zuverlässlich zu bewerten und andererseits eine noch schnellere Vorhersage treffen zu können. Für den letzten Punkt ist eine Lösung bereits in Entwicklung: Ab Ende des Jahres 2014 sollten wir insgesamt drei leistungsstarke Server für MutationTaster einsetzen können; wir rechnen damit, dadurch etwa 100 Analysen parallel durchführen zu können. Bei einer durchschnittlichen Analysezeit von unter 100 ms ließen sich so WGS-Projekte mit 3 Millionen Varianten in etwa 3.000 Sekunden, also 50 Minuten, bewerten. Die Analysezeit wird vermutlich sogar deutlich geringer sein, da Varianten ohne Aminosäureaustausch - die ja die Mehrzahl der genomweit detektiert Veränderungen ausmachen - erheblich schneller bewertet werden können. Durch eine vorhergehende Filterung gegen Veränderungen aus dem 1000G ließe sich die Zahl der Varianten, und somit die Laufzeit, zudem radikal verringern.

Eine höhere Qualität der Bewertung, kann neben den schon genannten Verbesserungen durch die Einbindung phänotypischer Informationen erreicht werden. In der Zukunft werden wir deshalb in einer neuen Applikation die Wahrscheinlichkeit einbeziehen, dass die durch Varianten veränderten Gene einen entsprechenden Phänotyp überhaupt verursachen könnten. Näheres dazu führe ich im Abschnitt 5.7 weiter unten aus.

5.4 Exomiser

Während die in Tests mit bekannten Krankheitsmutationen ermittelte Erkennungsrate der kausalen Varianten sehr hoch ist, schneidet die Software in der tatsächlichen Anwendung deutlich schlechter ab. Dies hat verschiedene Ursachen - eine wesentliche ist, dass Exomiser das Vorhandensein von Mausmodellen zu einem Gen für die Bewertung des Krankheitsrisikos durch

Veränderungen in diesem voraussetzt. Im Falle von Varianten in Genen, für die bislang keine Mausmodelle existieren, kann der Algorithmus zur Bewertung der Wahrscheinlichkeit der Beteiligung eines Gens an der Krankheitsentstehung nicht eingesetzt werden und es wird lediglich auf die Bewertung des Effekts der Variante auf das Genprodukt durch die integrierten Bewertungen von PolyPhen-2, SIFT und MutationTaster zurückgegriffen. Ein Vergleich mit HGMD[10] zeigte zwar, dass für fast 90% der in HGMD gespeicherten Krankheitsmutationen bzw. -gene auch Mausmodelle existieren, dies kann aber auch durch eine Stichprobenverzerrung erklärt werden: Wurde eine neue Mutation entdeckt, werden häufig Mausmodelle erstellt und gleichermaßen werden natürlich solche Gene bevorzugt auf Krankheitsmutationen hin untersucht, für die zum Phänotyp passende Mausmodelle existieren.

Die Erfahrungen im Einsatz der Software zeigen zudem, dass es zu falschen Priorisierungen der Varianten kommen kann, wenn die tatsächliche Krankheitsmutation von den Variantenbewertungsprogrammen nicht als pathogen eingestuft wird, gleichzeitig aber bekannte Krankheitsmutationen oder als pathogen vorhergesagte Varianten auftreten. In solchen Fällen kann die Bewertung des Effekts der Varianten auf das Genprodukt durch MutationTaster, PolyPhen und SIFT die Bewertung des Krankheitspotential der beteiligten Gene durch PHIVE übertreffen. Dies führt dazu, dass der Exomiser eine als nicht schädlich kategorisierte Variante im eigentlichen Krankheitsgen als weniger wahrscheinliche Krankheitsursache betrachtet als die als pathogen vorhergesagten Varianten in Genen, die für die Krankheit nicht relevant sind. Das Problem vergrößert sich natürlich, wenn sich diese Mutation in einem Gen befindet, für das kein Mausmodell existiert oder nur ein sehr unpräziser Phänotyp angegeben wurde.

Eine falsche oder unzureichende Angabe des Phänotyps stellt ein weiteres Problem dar - in diesem Falle werden entweder die falschen oder sehr viele Gene als potentielle Kandidaten betrachtet. Im ersten Fall wäre die Auswertung der Daten durch den Exomiser der durch MutationTaster oder ähnlichen Programmen unterlegen, im zweiten Fall brächte sie zumindest keinen Vorteil.

Eine Weiterentwicklung des Exomisers, an der ich nicht beteiligt war, ist die Software PhenIX[23]. Diese vermeidet das Problem nicht vorhandener phänotypischer Informationen zu den Genen dadurch, dass mit ihr lediglich bekannte Krankheitsgene ausgewertet werden können. Anstelle wie bei einer Exomsequenzierung vor der Hochdurchsatzsequenzierung die kompletten kodierenden Sequenzen anzureichern, kann so die Anreicherung auf die kodierenden Sequenzen der bekannten Krankheitsgene beschränkt und dadurch eine höhere Abdeckung erreicht werden. Dieser Ansatz eignet sich somit besser für diagnostische Zwecke, erlaubt es aber nicht, Mutationen in noch nicht als Krankheitsgen bekannten Genen zu identifizieren.

Ausblick

Für den Exomiser ist unter anderem eine bessere Verzahnung mit MutationTaster in Planung. Neben der Verwendung der Vorhersagen der aktuellen Version von MutationTaster, planen wir eine Lösung, mit der nicht in der Datenbank gespeicherte Varianten in Echtzeit an MutationTaster geschickt und analysiert werden. Dies ist insbesondere für Insertionen und Deletion relevant, die bislang nicht in dbNSFP annotiert sind.

5.5 CNVinspector

CNVinspector ist keine Software zur Analyse der Rohdaten aus den verschiedenen Plattformen für die Untersuchung von Varianten der Kopienzahl sondern dient der Visualisierung und dem Vergleich verschiedener CNVs, die mit beliebigen Methoden detektiert wurden.

Aufgrund der Vielzahl der Plattformen, Hersteller und Dateiformate wäre eine Schnittstelle für Rohdaten mit einem sehr hohen Implementierungs- und Wartungsaufwand verbunden gewe-

sen. Darüber hinaus würde ein solcher Ansatz den Transfer sehr großer Datenmengen über das Internet erfordern. Zur Prozessierung der Rohdaten existieren, neben den von den Herstellern der Auswertungsplattformen angebotenen Lösungen, ausreichend viele spezialisierte Computerprogramme wie beispielsweise CNV Workshop[61] oder arrayCGHbase[62]. Statt der Möglichkeit, Rohdaten direkt zu analysieren, bietet CNVinspector eine generische Schnittstelle zum Hochladen von Dateien an, die die einzelnen CNVs enthalten. Dies erlaubt es, Daten von verschiedenen Plattformen sehr leicht analysieren zu können - dabei müssen in den Ausgabedateien nur die Spaltennamen angepasst werden.

CNVinspector ist eine recht neue Software und wurde - trotz mehr als 300 registrierter Benutzerinnen und Benutzer - bislang erst einmal zitiert. Interessant ist dabei, dass mit CNVinspector nur etwas mehr als 100 verschiedene Projekte analysiert wurden. Daraus lässt sich ableiten, dass die Software momentan offenbar eher zum Studium schon publizierter Varianten der Kopienzahl eingesetzt wird, also beispielsweise, um eine Übersicht der bekannten Varianten, die mit einem Phänotyp assoziiert sind, zu erlangen. Da wir im Gegensatz zu den anderen Anwendungen auch nur sehr wenige Fragen oder gar Verbesserungsvorschläge zu CNVinspector erhalten, bleibt dies aber unklar. Über den tatsächlichen Nutzen der Software in der Aufklärung genetischer Krankheiten lässt sich daher im Moment noch keine Aussage treffen.

Für CNVinspector sind deshalb derzeit keine konkreten Verbesserungen geplant, möglich wäre aber die Einbindung in die avisierte neue Software (siehe Abschnitt 5.7), um die in WGS- und WES-Projekten detektierte Varianten der Kopienzahl visualisieren und bewerten zu können. Wir werden die Software aber in der nahen Zukunft aktualisieren, so dass zusätzlich zu den bestehenden Genomversionen GRCh37 und NCBI36 auch die aktuelle Version GRCh38 verwendet werden kann.

5.6 Einsatz der Verfahren

5.6.1 Studien unter Beteiligung unserer Arbeitsgruppe

Die folgenden Arbeiten zeigen exemplarisch den erfolgreichen Einsatz der hier vorgestellten Computerprogramme in Genkartierungsstudien, an denen ich beteiligt war:

Identifizierung von *PTRF*-Mutationen als Ursache autosomal-rezessiver Lipodystrophie

In diesem Projekt[63] wurde eine Genmutation identifiziert, die zu einer rezessiv vererbten Lipodystrophie (OMIM #613327) führt. Weil es sich um konsanguine Familien handelt, konnte der Genort mit HomozygosityMapper bestimmt werden. Zur Suche nach dem wahrscheinlichen Krankheitsgen unter den 74 positionellen Kandidatengenen wurde GeneDistiller eingesetzt, die Exons der so ermittelten Gene wurden mit dem Sanger-Verfahren sequenziert.

Da die Patienten aus verschiedenen Familien stammten und alle homozygote Mutationen im *PTRF*-Gen trugen, konnte sich eine direkte funktionelle Analysen, unter anderem mit elektronenmikroskopischen Untersuchungen, anschließen.

MutationTaster war zu diesem Zeitpunkt noch nicht publiziert und wurde deshalb in dieser Arbeit nicht zitiert - die Auswirkungen der Frameshift-Mutationen waren aber auch ohne dessen Einsatz klar.

Identifizierung von *ZC4H2*-Mutationen als Ursache des X-chromosomalen Wieacker-Wolff-Syndroms

Eine weitere Arbeit[64] war die Identifizierung von Mutationen im *ZC4H2*-Gen, die zum X-chromosomal vererbten Wieacker-Wolff-Syndrom (OMIM #314580) führen. Die betroffenen Familien wurden von verschiedenen Arbeitsgruppen unabhängig voneinander mit unterschiedlichen Methoden analysiert, dabei kamen sowohl eine Kopplungsanalyse mit nachfolgender Sequenzierung der Kopplungsregion (*target-enrichment*) als auch eine initiale Exomsequenzierung zum Einsatz. Die gefundenen Mutationen wurden mit MutationTaster auf ihr Krankheitspotential untersucht; das Gen *ZC4H2*, das sowohl anhand der genetischen Befunde, der Ergebnisse von MutationTaster als auch der Genbeschreibung aussichtsreich schien, wurde dann mit dem Sanger-Verfahren in weiteren Familienmitgliedern sequenziert.

Durch funktionelle Analysen, unter anderem *knockdown-rescue* Untersuchungen im Zebrafisch-Modell, konnte der Funktionsverlust des Proteins durch die von uns gefundenen Mutationen bestätigt werden.

Identifizierung von *ADAM9*-Mutationen als Ursache einer autosomal-rezessiven Form der progressiven Retinaatrophie in Hunden

Durch den Wunsch nach Hilfe bei der Genkartierung in Hunden aus einer Arbeitsgruppe der Ruhr-Universität in Bochum entstand eine weitere Arbeit[40], in der eine Genmutation identifiziert werden konnte, die für eine autosomal rezessive Form der progressiven Retinaatrophie in *Irish Glen Of Imaal* Terriern verantwortlich ist. Dabei wurden mit Hilfe der eigens dafür entwickelten Programmversion zur Analyse von caninen SNP-Genotypen in HomozygosityMapper homozygote Regionen in den betroffenen Hunden identifiziert. Die Kandidatenregionen wurden dann durch Mikrosatelliten weiter eingegrenzt. Die Sequenzierung des aussichtsreichen positionellen Kandidatengens *ADAM9* zeigte die homozygote Deletion zweier Exons, die durch cDNA-Analysen bestätigt werden konnte. Dadurch ergibt sich eine Leserasterverschiebung und ein vorzeitiges Stopcodon. Die Analysen zeigten fernerhin, dass die Genexpression in einem erkrankten Tier deutlich vermindert war.

Diese Deletion wurde nicht in anderen Hunderassen gefunden, wohl aber in nicht betroffenen Irish-Glen-Terriern. Eine spätere Untersuchung der initial als nicht betroffenen eingestuften Hunde zeigte, dass einige inzwischen eine Sehminderung entwickelt hatten. Dies legt eine unvollständige Penetranz bzw. die Einwirkung weiterer Gene oder Umweltfaktoren auf die Ausprägung des Phänotyps nahe.

5.6.2 Nutzung der Programme durch externe Gruppen

Die hier vorgestellten Programme können über das Internet kostenlos verwendet werden und werden weltweit genutzt.

MutationTaster

Insbesondere die Variantenbewertung durch MutationTaster wird häufig angewendet, um das Krankheitspotential von Varianten *in silico* zu bewerten. Die Datenbank dbNSFP[65] enthält für bekannte nicht-synonyme Varianten unter anderem vorberechnete Variantenbewertungen von PolyPhen-2[43], SIFT[42] und MutationTaster und wurde in viele Pipelines zur Hochdurchsatzsequenzierung integriert. Darüberhinaus werden die Daten aus dbNSFP auch in Computerprogrammen zur Bewertung von Varianten im Hinblick auf den Phänotyp wie eXtasy[21], Exomiser[22]

und PhenIX[23] verwendet. MutationTaster wurde so schon mehr als 600 Mal in anderen wissenschaftlichen Arbeiten zitiert.

HomozygosityMapper

HomozygosityMapper erreichte bislang deutlich mehr als 100 Zitate, und wurde von 1.400 angemeldeten Benutzerinnen und Benutzern zur Analyse von mehr als 7.000 verschiedenen Projekten eingesetzt. Dabei wurden weit mehr als 20 Milliarden Genotypen permanent in der Datenbank gespeichert - durch die Möglichkeit, Daten auch ohne Registrierung zu analysieren und anschließend zu löschen, liegt die tatsächliche Nutzung sogar weit höher. HomozygosityMapper hat eine rege Nutzergemeinschaft, die sich häufig durch Verbesserungsvorschläge an der Weiterentwicklung beteiligen.

GeneDistiller

Die Statistiken unseres Webservers zeigen, dass auch GeneDistiller mit durchschnittlichen 400 Zugriffen am Tag rege genutzt wird. Allerdings schlägt sich dies nicht in Zitaten nieder, so wurde GeneDistiller bislang erst etwa 50 Mal zitiert. Wir vermuten, dass GeneDistiller als frei verfügbare Informationsquelle von vielen Wissenschaftlern genutzt wird, ohne dies explizit in den Publikationen zu erwähnen. Dies ist allerdings für derartige web-basierte Angebote, wie auch beispielsweise Ensembl[54] oder NCBI Entrez[66], nicht ungewöhnlich.

Exomiser

Der Exomiser ist eine recht neue Entwicklung und wurdebislang erst weniger als 30 Mal zitiert. Im Institut für Medizinische Genetik der Charité wird inzwischen meist PhenIX verwendet, eine Weiterentwicklung für diagnostische Zwecke, das ich in der Diskussion zu Exomiser kurz vorgestellt habe.

CNVinspector

Wie weiter oben geschildert, wird das Programm CNVinspector zwar verwendet, wurde bislang aber erst einmal zitiert.

5.7 Zusammenführung der verschiedenen Programme

Während es bislang schon Interaktionen zwischen den einzelnen Programmen gibt (so können beispielsweise die von HomozygosityMapper bestimmten Krankheitsregionen oder die Gene, in denen MutationTaster krankheitsverursachende Mutationen vorhersagt, direkt in GeneDistiller studiert werden), sind die Programme bislang eigene, unabhängige Applikationen.

Insbesondere für die Verwendung phänotypischer Informationen für eine bessere Bewertung des Krankheitspotentials einer Variante ist eine direkte Verknüpfung von GeneDistiller und MutationTaster unerlässlich. Bislang beurteilt MutationTaster eine Variante lediglich im Hinblick auf ihren Effekt auf das Gen bzw. Protein - ob das Gen selbst überhaupt das Potential hat, die untersuchte Krankheit zu verursachen, wird dabei nicht berücksichtigt. Die nächste Version unserer Software wird deshalb GeneDistiller und MutationTaster zusammenführen. Anhand eines von den Anwendern eingegeben Phänotyps wird dabei zuerst eine Bewertung der Gene anhand ihres möglichen Krankheitspotentials für ebendiesen Phänotyp vorgenommen werden. Im Gegensatz zu den bestehenden Computerprogammen wie eXtasy[21], dem Exomiser[22] oder PhenIX[23] wird dabei die komplette Palette der Informationen genutzt werden, die GeneDistiller bereitstellt. Das

bedeutet, dass neben den 'Netzwerkinformationen', zum Beispiel Proteininteraktionsdaten, auch weitere genspezifische Daten eingesetzt werden. Hierbei sind beispielsweise Ähnlichkeiten in den funktionellen Beschreibungen, subzellulärer Lokalisationsdaten aus der GeneOntology[33] und der gewebsspezifischen Expression zwischen den Kandidatengenen und bekannten Krankheitsgenen, die ähnliche Phänotypen verursachen, zu nennen. Dabei wird es den Benutzern überlassen bleiben, ob sie Gene, die ihre Kriterien (wie beispielsweise die Expression in einem Gewebe) nicht erfüllen, komplett ausschließen oder lediglich niedriger priorisieren wollen. Die Applikation wird dann das errechnete Krankheitspotential der Gene mit der Bewertung des Effekts einer Variante auf das Protein (durch MutationTaster) zusammenbringen. Die Beurteilung des Krankheitspotentials der Gene aufgrund des von den Benutzern vorgegebenen Krankheitsmodells wird dabei nach erfolgter Analyse durch die Anwender wiederholt werden können, um das Modell gegebenenfalls anhand der Ergebnisse weiter verfeinern zu können. Eine Schnittstelle für die schnelle und bequeme Angabe komplexer Phänotypen habe ich bereits für die Software GrabBlur[67] entwickelt, die dem anonymisierten Austausch von NGS-Genotypen mitsamt der phänotypischen Beschreibungen der Patienten dient.

Eine weitere Möglichkeit zum Ausschluss von Varianten ist das Vererbungsmuster der Krankheit: In einer dominanten Erkrankung mit hoher Penetranz sollten erkrankte Personen Träger des Krankheitsallels sein, nicht betroffene Personen jedoch nicht. Im Falle (autosomal-) rezessiver Krankheiten müssen Erkrankte homozygot sein oder bei compound-heterozygoten Erbgängen zwei Krankheitsallele tragen. Gesunde dürfen diesen Genotyp nicht aufweisen (siehe Abbildung 8). Bislang musste eine derartige Analyse vor der Erstellung der VCF-Datei, die von MutationTaster analysiert werden soll, von den Benutzern selbst durchgeführt werden. Die zukünftige Software wird diesen Schritt integrieren, so dass die Forscher und Kliniker - nach Angabe des vermuteten Erbgangs - eine einzige Genotypendatei für die ganze Familie bzw. mehrere Familien mitsamt der Stammbäume mit der neuen Software analysieren können. Denkbar ist hier auch der Einsatz einer Kopplungsanalyse innerhalb der Applikation, um so große Bereiche des Genoms von vornherein ausschließen zu können. Informationen über den Erbgang können auch zur Suche nach Krankheitsgenen verwendet werden, da zumindest für einen großen Teil der bekannten Krankheitsgene derartige Informationen (Funktionsverlust/rezessiv bzw. Funktionsgewinn/dominant) vorliegen.

Auch die Integration von HomozygosityMapper verspricht Vorteile: Nach der Veränderung des Speichermodells für Genotypen (siehe weiter oben im Abschnitt 5.2) können die in den von HomozygosityMapper bestimmten Regionen vorhandenen DNA-Varianten direkt von der neuen Applikation analysiert werden und wie oben beschrieben unter Berücksichtigung des Phänotyps auf ihr Krankheitspotential hin bewertet werden.

6 Zusammenfassung

In dieser Habilitationsschrift beschreibe ich verschiedene Strategien und von mir entwickelte oder mitentwickelte Computerprogramme, die es Forschern und Klinikern erlauben, unter einem möglichst minimalen Einsatz von Zeit und Geld die wahrscheinlichsten Ursachen monogener Krankheiten zu bestimmen.

Alle hier vorgestellten Programme sind web-basiert, das heißt, dass sie direkt und ohne die Notwendigkeit, Software zu installieren, benutzt werden können. Sie zeichnen sich überdies durch eine sehr einfache Benutzbarkeit aus. Dies versetzt die mit einer Krankheit vertrauten Wissenschaftlern oder Ärzten in die Lage, sich selber stärker in die Aufklärung der Krankheitsursachen einzubringen, ohne die Datenauswertung komplett an Bioinformatiker angeben zu müssen, womit in der Regel Informationsverluste einhergehen, da diese weit weniger über die betreffende Krankheit wissen.

Die Software **GeneDistiller**, die schon in meiner Promotion entstanden ist, dient primär der Suche nach Kandidatengenen für monogene Krankheiten, kann aber darüber hinaus auch dazu verwendet werden, diverse Informationen über ein einzelnes oder eine Reihe von Genen komfortabel und schnell anzuzeigen oder die Gene nach selbst gewählten Eigenschaften zu durchzusuchen, zu filtern und nach ihrem erwarteten Krankheitspotential zu priorisieren. Neben dem Einsatz im Rahmen einer klassischen Genkartierung können mit GeneDistiller auch die Eigenschaften der Gene studiert werden, in denen in Exom- oder Genomsequenzierung potentiell krankheitsverursachende Varianten detektiert wurden. GeneDistiller kann auch in der Erforschung komplexer Krankheiten eingesetzt werden, um den möglichen Einfluss der den durch Assoziationsstudien bestimmten Loci benachbarten Gene auf die Krankheit zu beurteilen.

Eine weitere hier vorgestellte Software ermöglicht die Bestimmung von Krankheitsloci in konsanguinen Familien. Die Software **HomozygosityMapper** arbeitet erheblich schneller und ist deutlich einfacher zu benutzen als klassische Computerprogramme für die Kopplungsanalyse. Die erste Version der Software gestattete die direkte Analysen von Genotypen, die mittels SNP-Chips gewonnen worden. Die aktuelle Version erlaubt es darüber hinaus, Homozygotiekartierungen auch mit den durch Exom- oder Genomsequenzierungen gewonnenen Daten durchzuführen.

Zur raschen *in silico* Bewertung des Krankheitspotentials von DNA-Varianten haben wir ein weiteres Computerprogramm entwickelt, **MutationTaster**. Im Gegensatz zu allen bisherigen Variantenbewertungsprogrammen arbeitet MutationTaster auch auf DNA-Ebene und erlaubt es deshalb, neben nicht-synonymen auch synonyme und nicht kodierende DNA-Varianten zu analysieren. Auch von diesem Programm gibt es zwei publizierte Versionen; die zweite greift die in der Zwischenzeit verfügbar gewordene umfangreiche Sammlung an Polymorphismen aus dem 1000-Genom-Projekt und funktioneller genomischer Elemente aus dem ENCODE-Projekt auf und verfügt zudem über wesentlich mehr Tests und einen deutlich vergrößerten Satz an Trainingsdaten. Sie erlaubt es darüber hinaus, durch eine web-basierte Analysestrecke, die wahrscheinlichsten Krankheitsmutationen aus den tausenden oder Millionen von Varianten herauszufiltern, die in kompletten Exom- oder Genomsequenzierungen detektiert werden.

Die Variantenbewertung durch MutationTaster wird von einer weiteren Software genutzt, an deren Erstellung ich beteiligt war: der **Exomiser** verbindet die Wahrscheinlichkeit der Pathogenität einer Variante mit der Wahrscheinlichkeit, dass das betroffene Gen an der Pathogenese einer bestimmten Erkrankung beteiligt ist.

Das Programm **CNVinspector** dient der benutzerfreundlichen Identifizierung potentiell krankheitsverursachender Variationen in der Kopienzahl von Genen oder chromosomalen Regionen. Hierbei können die durch ArrayCGH-Experimente, SNP-Chips oder andere Methoden bestimmten duplizierten, deletierten oder hemizygoten Regionen in einzelnen Patienten oder Kohorten

identifiziert und mit gesunden Kontrollen verglichen werden. Dazu können wahlweise eigene Kontrollen oder aber öffentlich publizierte Studien herangezogen werden. Das Programm bietet ein Interface zu GeneDistiller, so dass eine schnelle Untersuchung und Bewertung der betroffenen Gene möglich wird.

7 Liste der einbezogenen eigenen Publikationen

Originalarbeiten in Zeitschriften mit *peer review*-Verfahren als Erst- bzw. Letztautor

	Abschnitt	Publikation	IF
1	4.1.1	**Seelow, D.** & Schuelke, M. HomozygosityMapper2012 - bridging the gap between homozygosity mapping and deep sequencing. *Nucleic Acids Research*, July 2012, W516–520, 40	8.278
2	4.2.1	Schwarz, J. M., Rödelsperger, C., Schuelke, M. & **Seelow, D.** MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods*, August 2010, 575–576, 7	20.721
3	4.2.2	Schwarz, J. M., Cooper, D. N., Schuelke, M. & **Seelow, D.** MutationTaster2: Mutation prediction for the deep-sequencing age. *Nature Methods*, April 2014, 361–362, 11	23.565
4	4.3.1	Knierim, E., Schwarz, J. M., Schuelke, M. & **Seelow, D.** CNVinspector: a web-based tool for the interactive evaluation of copy number variations in single patients and in cohorts. *Journal of Medical Genetics*, August 2013, 529–533, 50	5.703

Originalarbeiten in Zeitschriften mit *peer review*-Verfahren als Koautor

	Abschnitt	Publikation	IF
5	4.2.3	Robinson, P. N., Köhler, S., Oellrich, A., Sanger Mouse Genetics Project, Wang, K., Mungall, C. J., Lewis, S. E., Washington, N., Bauer, S., **Seelow, D.**, Krawitz, P., Gilissen, C., Haendel, M. & Smedley, D. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research*, February 2014, 340–348, 24	14.397

Impact Factors aus Thomson Reuters *ISI Web of Knowledge*. Stand kein *Impact Factor* für das betreffende Jahr zur Verfügung, so wurde das jeweils nächstliegende Jahr herangezogen.

8 Literaturangaben

1. Seelow, D., Schwarz, J. M. & Schuelke, M. GeneDistiller—Distilling Candidate Genes from Linkage Intervals. *PLoS ONE* **3,** e3874 (2008).
2. Lander, E. S. & Botstein, D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236,** 1567–1570 (1987).
3. Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. & Lander, E. S. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58,** 1347–1363 (1996).
4. Konrad, M. *et al.* Mutations in the tight-junction gene claudin 19 (CLDN19) are associated with renal magnesium wasting, renal failure, and severe ocular involvement. *American Journal of Human Genetics* **79,** 949–957 (2006).
5. Elston, R. C. & Stewart, J. A general model for the genetic analysis of pedigree data. *Hum. Hered.* **21,** 523–542 (1971).
6. Seelow, D., Schuelke, M., Hildebrandt, F. & Nürnberg, P. HomozygosityMapper - an interactive approach to homozygosity mapping. *Nucleic Acids Res.* **37,** W593–599 (2009).
7. Seelow, D. & Schuelke, M. HomozygosityMapper2012 - bridging the gap between homozygosity mapping and deep sequencing. *Nucleic Acids Res.* **40,** W516–520 (2012).
8. Stöber, G. *et al.* Periodic catatonia: confirmation of linkage to chromosome 15 and further evidence for genetic heterogeneity. *Hum. Genet.* **111,** 323–330 (2002).
9. Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.* **37,** D793–796 (2009).
10. Stenson, P. D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* (2013).
11. Seelow, D., Schwarz, J. M. & Schuelke, M. GeneDistiller–distilling candidate genes from linkage intervals. *PLoS ONE* **3,** e3874 (2008).
12. Seelow, Dominik. *Erstellung eines computergestützten Verfahrens zur Suche nach Kandidatengenen für rezessiv vererbte Krankheiten in konsanguinen Familien.* Diss. (Charité - Universitätsmedizin Berlin, 2009).
13. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **7,** 111–118 (2010).
14. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* **7,** 575–576 (2010).
15. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* **11,** 361–362 (2014).
16. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42,** 30–35 (2010).
17. Doherty, D. & Bamshad, M. J. Exome sequencing to find rare variants causing neurologic diseases. *Neurology* **79,** 396–397 (2012).
18. Schuster, S. C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463,** 943–947 (2010).
19. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

20. *Report of a joint meeting of WHO/ECFTN/ICF (M)A/ECFS. The molecular genetic epidemiology of cystic fibrosis* (World Health Organization, 2004).

21. Sifrim, A. *et al.* eXtasy: variant prioritization by genomic data fusion. *Nat. Methods* **10,** 1083–1084 (2013).

22. Robinson, P. N. *et al.* Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res.* **24,** 340–348 (2014).

23. Zemojtel, T. *et al.* Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci Transl Med* **6,** 252ra123 (2014).

24. Robinson, P. N. *et al.* The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *American Journal of Human Genetics* **83,** 610–615 (2008).

25. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* **32,** 894–899 (2011).

26. Peiró, G., Mayr, D., Hillemanns, P., Löhrs, U. & Diebold, J. Analysis of HER-2/neu amplification in endometrial carcinoma by chromogenic in situ hybridization. Correlation with fluorescence in situ hybridization, HER-2/neu, p53 and Ki-67 protein expression, and outcome. *Mod. Pathol.* **17,** 227–287 (2004).

27. Koenig, M. *et al.* Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. *Cell* **50,** 509–517 (1987).

28. Kearney, H. M. *et al.* American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Genet. Med.* **13,** 680–685 (2011).

29. Craddock, N. *et al.* Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464,** 713–720 (2010).

30. Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. & Pritchard, J. K. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38,** 75–81 (2006).

31. Zhang, Z.-F. *et al.* Detection of submicroscopic constitutional chromosome aberrations in clinical diagnostics: a validation of the practical performance of different array platforms. *Eur. J. Hum. Genet.* **16,** 786–792 (2008).

32. Knierim, E., Schwarz, J. M., Schuelke, M. & Seelow, D. CNVinspector: a web-based tool for the interactive evaluation of copy number variations in single patients and in cohorts. *J. Med. Genet.* **50,** 529–533 (2013).

33. Blake, J. A. *et al.* Gene Ontology annotations and resources. *Nucleic Acids Res.* **41,** D530–535 (2013).

34. Chaurasia, G. *et al.* UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res.* **35,** D590–594 (2007).

35. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41,** D808–815 (2013).

36. Untergasser, A. *et al.* Primer3–new capabilities and interfaces. *Nucleic Acids Res.* **40,** e115 (2012).

37. Rüschendorf, F. & Nürnberg, P. ALOHOMORA: a tool for linkage analysis using 10K SNP array data. *Bioinformatics* **21,** 2123–2125 (2005).

38. Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. & Lander, E. S. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58,** 1347–1363 (1996).

39. Gudbjartsson, D. F., Thorvaldsson, T., Kong, A., Gunnarsson, G. & Ingolfsdottir, A. Allegro version 2. *Nat. Genet.* **37,** 1015–1016 (2005).

40. Kropatsch, R. *et al.* Generalized progressive retinal atrophy in the Irish Glen of Imaal Terrier is associated with a deletion in the ADAM9 gene. *Mol. Cell. Probes* **24,** 357–363 (2010).

41. Lee, K.-T. *et al.* Whole-genome resequencing of Hanwoo (Korean cattle) and insight into regions of homozygosity. *BMC Genomics* **14,** 519 (2013).

42. Sim, N.-L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40,** W452–457 (2012).

43. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).

44. Sunyaev, S. *et al.* Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10,** 591–597 (2001).

45. Bromberg, Y., Yachdav, G. & Rost, B. SNAP predicts effect of mutations on protein function. *Bioinformatics* **24,** 2397–2398 (2008).

46. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).

47. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).

48. Reese, M. G., Eeckman, F. H., Kulp, D. & Haussler, D. Improved splice site detection in Genie. *J. Comput. Biol.* **4,** 311–323 (1997).

49. Jana Marie Schwarz. *MutationTaster - ein web-basiertes Computerprogramm zur Bewertung des Krankheitspotentials von DNA-Mutationen.* Diss. (Freie Universität Berlin, 2013).

50. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

51. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980–985 (2014).

52. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* **7,** e46688 (2012).

53. Firth, H. V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* **84,** 524–533 (2009).

54. Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Research* **40,** D84–D90 (2011).

55. Andreasen, C. *et al.* Mutations in genes encoding cardiac ion channels previously associated with sudden infant death syndrome (SIDS) are present with high frequency in new exome data. *Can J Cardiol* **29,** 1104–1109 (2013).

56. Ten Kate, L. P., Teeuw, M., Henneman, L. & Cornel, M. C. Autosomal recessive disease in children of consanguineous parents: inferences from the proportion of compound heterozygotes. *J Community Genet* **1,** 37–40 (2010).

57. Knierim, E., Seelow, D., Gill, E., von Moers, A. & Schuelke, M. Clinical application of whole exome sequencing reveals a novel compound heterozygous TK2-mutation in two brothers with rapidly progressive combined muscle-brain atrophy, axonal neuropathy, and status epilepticus. *Mitochondrion* **20,** 1–6 (2015).

58. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).

59. Maass, P. G., Luft, F. C. & Bähring, S. Long non-coding RNA in health and disease. *J. Mol. Med.* **92,** 337–346 (2014).

60. Daniela Hombach. *Application of in silico methods to predict the disease potential of DNA sequence variants in transcription factor-binding sites.* Masterarbeit (Charité - Universitätsmedizin Berlin, 2014).

61. Gai, X. *et al.* CNV Workshop: an integrated platform for high-throughput copy number variation discovery and clinical diagnostics. *BMC Bioinformatics* **11,** 74 (2010).

62. Menten, B. *et al.* arrayCGHbase: an analysis platform for comparative genomic hybridization microarrays. *BMC Bioinformatics* **6,** 124 (2005).

63. Rajab, A. *et al.* Fatal cardiac arrhythmia and long-QT syndrome in a new form of congenital generalized lipodystrophy with muscle rippling (CGL4) due to PTRF-CAVIN mutations. *PLoS Genet.* **6,** e1000874 (2010).

64. Hirata, H. *et al.* ZC4H2 mutations are associated with arthrogryposis multiplex congenita and intellectual disability through impairment of central and peripheral synaptic plasticity. *Am. J. Hum. Genet.* **92,** 681–695 (2013).

65. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.* **34,** E2393–2402 (2013).

66. Acland, A. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **42,** D7–17 (2014).

67. Stade, B., Seelow, D., Thomsen, I., Krawczak, M. & Franke, A. GrabBlur–a framework to facilitate the secure exchange of whole-exome and -genome SNV data using VCF files. *BMC Genomics* **15 Suppl 4,** S8 (2014).

9 Danksagung

Am Zustandekommen dieser kumulativen Habilitationsschrift haben naturgemäß viele Menschen einen Anteil.

Besonders bedanken möchte ich mich bei Markus Schülke für sein Vertrauen in mich und die großen Freiheiten, die ich in seiner Arbeitsgruppe genieße - und natürlich für die vielen fruchtbaren Diskussionen.

Einen großen Beitrag am Zustandekommen dieser Arbeit hat Jana Marie Schwarz, ohne die das Projekt MutationTaster sicherlich nie so weit gekommen wäre.

Mein Dank gilt außerdem Evelyn Seelow, die unzählige Texte und E-Mails von mir gelesen und Verbesserungsvorschläge gemacht hat.

Bedanken möchte ich mich außerdem bei allen Mitgliedern der Arbeitsgruppe von Markus Schülke für das Testen neuer Funktionen der Software - und natürlich für die angenehme Zusammenarbeit.

Nicht zuletzt danke ich allen Benutzerinnen und Benutzern unserer Software, die sich durch viele Fehlermeldungen und Verbesserungsvorschläge gewissermaßen aktiv an der Entwicklung beteiligt haben.

10 Erklärung

§ 4 Abs. 3 (k) der HabOMed der Charité

Hiermit erkläre ich, dass

- weder früher noch gleichzeitig ein Habilitationsverfahren durchgeführt oder angemeldet wurde,
- die vorgelegte Habilitationsschrift ohne fremde Hilfe verfasst, die beschriebenen Ergebnisse selbst gewonnen sowie die verwendeten Hilfsmittel, die Zusammenarbeit mit anderen Wissenschaftlern/Wissenschaftlerinnen und mit technischen Hilfskräften sowie die verwendete Literatur vollständig in der Habilitationsschrift angegeben wurden,
- mir die geltende Habilitationsordnung bekannt ist.

Ich erkläre ferner, dass mir die Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis bekannt ist und ich mich zur Einhaltung dieser Satzung verpflichte..

.....	.....
Datum	Unterschrift