medgen 2019 · 31:212–221

<https://doi.org/10.1007/s11825-019-0246-2>

Online publiziert: 15. Juli 2019

© Der/die Autor(en) 2019

Kerstin U. Ludwig² · Franziska Degenhardt¹ · Markus M. Nöthen¹

¹ Institut für Humangenetik, Universitätsklinikum Bonn, Bonn, Deutschland

² Emmy-Noether-Gruppe „Kraniofaziale Genomik“, Institut für Humangenetik, Universitätsklinikum Bonn, Bonn, Deutschland

Die Rolle seltener Varianten bei häufigen Krankheiten

Hinführung zum Thema

Die Aufklärung der genetischen Faktoren bei den in der Bevölkerung häufigen Krankheiten, den sog. Volkskrankheiten, hat in den letzten Jahren rasante Fortschritte gemacht. Während krankheitsbeitragende, hochfrequente genetische Varianten (engl. „common variants“) mittlerweile in großer Zahl identifiziert sind, gelang die Identifizierung seltener Varianten (engl. „rare variants“) in der Regel nur bei monogen vererbten Subtypen bzw. für Varianten mit hoher Penetranz. Neue Sequenzieretechnologien ermöglichen seit Kurzem auch die exom- bzw. genomweite Sequenzierung großer Patienten- und Kontrollkollektive. Somit kann der Beitrag seltener Varianten zur Ätiologie von Volkskrankheiten zukünftig systematisch erfasst werden, erste entsprechende Untersuchungen werden bereits durchgeführt. In diesem Artikel betrachten wir die Bedeutung seltener Varianten bei der Aufklärung der genetischen Ätiologie häufiger Erkrankungen insbesondere unter folgenden Aspekten: Bei welchen häufigen Krankheiten ist ein Beitrag seltener Varianten zu erwarten, wie können diese Varianten identifiziert werden, und welches Potenzial bieten seltene Varianten für das Verständnis biologischer Prozesse bzw. für die Translation in die klinische Praxis?

Die genetische Architektur häufiger Krankheiten

In der Bevölkerung häufige Krankheiten, wie z. B. schizophrene Störungen, entzündliche Darmerkrankungen, koronare Herzkrankheit oder Alzheimer De-

menz, entstehen bei den meisten Patienten aus dem Zusammenspiel von genetischen Faktoren und Umgebungseinflüssen. Darüber hinaus nimmt man an, dass unter einer Diagnose unterschiedliche Krankheitsätiologien zusammengefasst werden (Abb. 1). Der genetische Anteil an der Krankheitsentstehung kann im Prinzip hochfrequente Varianten mit niedriger Penetranz ebenso umfassen wie seltene Varianten entlang des gesamten Penetranzspektrums. Sowohl in Simulationsstudien als auch durch Untersuchungen realer Daten konnte gezeigt werden, dass eine Vielzahl an Kombinationen von Varianten des gesamten allelischen Spektrums als Szenarien für die genetische Architektur von Volkskrankheiten infrage kommt [2, 5]. Dabei ist es naheliegend, dass mit zunehmender Polygenität, d. h. je mehr genetische Varianten beim einzelnen Patienten an der Entstehung der Krankheit beteiligt sind, sowohl seltene als auch häufige Varianten im Schnitt geringere Effektstärken aufweisen werden [2].

Schätzungen des Anteils häufiger/seltener Varianten an der Heritabilität

Der relative Beitrag häufiger Varianten an der Gesamterblichkeit (-heritabilität) lässt sich mit statistischen Programmen auf der Basis systematischer SNP- („single nucleotide polymorphism“)-Daten, meist aus genomweiten Assoziationsstudien (GWAS), abschätzen (sog. „SNP-basierte Heritabilität“, z. B. GCTA [61], LD-score-Regression [7]). Die für viele Volkskrankheiten beobachtete Differenz zwischen Gesamtheritabilität

und SNP-Heritabilität (oft als „missing heritability“ bezeichnet) kann zumindest zum Teil durch seltene Varianten bedingt sein, sicher quantifizieren lässt sich der Beitrag seltener Varianten bisher in der Regel nicht. Es gibt auch Stimmen, die die „missing heritability“ für ein überschätztes Phänomen halten, da die Gesamtheritabilität, wenn sie als sog. „broad-sense heritability“ geschätzt wurde, auch aus Interaktionen zwischen Genen bzw. Signalwegen sowie Gen-Umwelt-Wechselwirkungen resultieren kann (epistatische Effekte) [63]. Damit ist sie nicht direkt vergleichbar mit der SNP-Heritabilität, welche ausschließlich additive Geneffekte misst. Aber auch wenn die Schätzung der Gesamtheritabilität auf additive Effekte beschränkt ist (sog. „narrow sense heritability“), verbleibt bei den Volkskrankheiten in der Regel eine erhebliche „missing heritability“. Basierend auf Gesamtgenomsequenzdaten einer großen populationsbasierten Kohorte, konnte in einer kürzlich auf bioRxiv erschienenen Studie zumindest für die Merkmale Körpergröße und BMI (Body-Mass-Index) gezeigt werden, dass die „missing heritability“ vollständig durch seltene Varianten erklärt wird [57].

Einfluss der Reproduktionsrate auf die genetische Architektur

Der Beitrag seltener Varianten an der Krankheitsentstehung wird bei verschiedenen Volkskrankheiten unterschiedlich groß sein. Bei denjenigen Volkskrankheiten, die bei den Betroffenen zu einer schwerwiegenden Beeinträchtigung mit einhergehender verminderter Repro-

Abb. 1 ▲ Genetische Architektur häufiger Erkrankungen. Volkskrankheiten sind in der Regel polygen und genetisch heterogen. Verschiedene Varianten des allelischen Spektrums sind durch *Symbole* und *Farben* gekennzeichnet, unterschiedliche Positionen der Symbole stehen für verschiedene krankheitsbeitragende Gene. Die Größe der Symbole korreliert mit der jeweiligen Effektstärke. **a** Verschiedene Betroffene weisen eine individuelle Krankheitsarchitektur auf, zu der Varianten des gesamten allelischen Spektrums sowie Umgebungseinflüsse beitragen können. **b** Für häufige Erkrankungen ursächliche Varianten sind in der Regel nicht vollständig penetrant und kommen auch bei nicht erkrankten Personen der Allgemeinbevölkerung vor

duktion führen, ist zu erwarten, dass zumindest ein Teil der beitragenden Varianten erst kürzlich entstanden und damit das allelische Spektrum in Richtung der seltenen Varianten verschoben ist. *De novo* Mutationen sind dann auch eine mögliche Erklärung für die gleichbleibende Häufigkeit dieser Erkrankungen in der Bevölkerung. Zur Gruppe der Volkskrankheiten mit verminderten Reproduktionsraten gehören z.B. schizophrene Störungen [39], für die der Beitrag seltener Varianten bereits durch seltene Kopienzahlveränderungen (engl. „copy number variants“, CNVs) überzeugend demonstriert werden konnte [27, 52].

Dagegen stehen Erkrankungen mit spätem Manifestationsalter (z. B. koronare Herzerkrankung, Alzheimer Demenz) oder allenfalls mäßiger bzw. fehlender klinischer Beeinträchtigung (z. B. androgenetische Alopezie), bei denen zunächst von einem geringen Beitrag seltener Varianten auszugehen ist [37, 62]. So wurde für die androgenetische Alopezie erst kürzlich durch Analyse der Daten der UK-Biobank ein relevanter Beitrag sel-

tener Varianten nahezu ausgeschlossen [62]. Grundsätzlich können aber auch bei Erkrankungen ohne starkem Selektionsdruck Subtypen vorliegen, die einer eher monogenen Vererbung unterliegen. So sind z. B. bei der Alzheimer Demenz seltene, hoch penetrante Mutationen in Familien mit frühem Krankheitsbeginn, u. a. in *APP*, *PSEN1*, *PSEN2* [60] und *SORL1* [41], berichtet. Auch bei der koronaren Herzkrankheit wurden kausale Gene in eher monogenen Formen identifiziert, u. a. *PCSK9* in familiärer Hypercholesterinämie [1].

Warum lohnt es sich, seltene Varianten bei häufigen Erkrankungen zu identifizieren?

Unabhängig davon, welchen Anteil seltene Varianten an der Gesamtheritabilität einzelner Volkskrankheiten erklären, lohnt sich deren Identifizierung aus mehreren Gründen:

Identifizierung neuer krankheitsbeitragender Gene

Seltene Varianten können zur Identifizierung neuer Gene beitragen. Dies kann durch die Beobachtung eines *de novo* Auftretens, von Kosegregation in mehrfach betroffenen Familien oder der Häufung in einem Patientenkollektiv im Vergleich zu Kontrollen geschehen. Dadurch können sowohl Krankheitsgene an bereits z. B. im Rahmen von GWAS-Untersuchungen beschriebenen Loci als auch bisher noch völlig unbekannte Gene identifiziert werden.

Priorisierung von Kandidatengen an bekannten Loci

In den GWAS-Loci findet sich das stärkste Assoziationssignal meist in nicht-kodierenden Bereichen, daher ist das kausale Krankheitsgen häufig nicht direkt evident. Als Teil des allelischen Spektrums können jedoch auch seltene Varianten vorliegen und dadurch hinweisgebend auf das kausale Krankheitsgen sein. Dieser Ansatz war bereits bei verschiedenen Volkskrankheiten erfolgreich, z. B. bei der Identifizierung des Gens *GLP1R* in Diabetes Typ 2 [59], oder mehrerer Gene an GWAS-Loci für chronisch-entzündliche Darmerkrankungen [45] (eine Übersicht findet sich in [6]). Analog kann die Identifizierung seltener Varianten auch zur Priorisierung von Krankheitsgenen in krankheitsassoziierten CNV-Loci beitragen, wenn in der betroffenen Region mehrere Gene liegen.

In der Literatur wird auch die Möglichkeit diskutiert, dass seltene Varianten die tatsächlich ursächlichen Varianten an einem GWAS-Locus darstellen können. Bei dieser Theorie der „synthetischen Assoziationen“ treten mehrere seltene Varianten unabhängig voneinander auf dem mit der Erkrankung assoziierten Haplotyp auf. Das Assoziationssignal einer den Haplotyp markierenden häufigen Variante bildet somit den gemeinsamen Effekt der seltenen Varianten ab [15]. Für die tatsächliche Existenz dieser synthetischen Assoziationen gibt es jedoch bisher nur wenige Beispiele (z. B. *HOXB13*-Locus bei Prostatakrebs [49], *NOD2*-Locus bei chronisch-entzündlichen Darmerkrankungen [3]).

Identifizierung von Krankheitsgenen in genomischen Regionen ohne vorbestehende genetische Evidenz

Es ist anzunehmen, dass ein Teil der Krankheitsgene in genomischen Regionen liegt, die nicht durch GWAS oder CNV-Befunde impliziert worden sind (z. B. *SETD1A* in Schizophrenie [50]). Dies kann zum einen daran liegen, dass die bisher untersuchten GWAS-Kohorten nicht zur Detektion einer tatsächlich vorhandenen Assoziation ausreichen (Powerproblem), dass eine klinische Subgruppe an Patienten (mit z. B. einer spezifischen molekularen Signatur) in den bisherigen Untersuchungen unterrepräsentiert ist, oder die bisher verwendeten Arrays diese Regionen nicht ausreichend abdecken (einschließlich der Möglichkeit mangelnden Kopplungsungleichgewichts zwischen genotypisierten und kausalen Varianten). Auf der anderen Seite wird es auch genomische Loci geben, an denen es keine funktionell relevanten häufigen Varianten gibt – diese Loci werden durch den GWAS-Ansatz grundsätzlich nicht erfasst. Bisher gibt es noch keine systematischen Abschätzungen darüber, auf wie viele Krankheitsgene diese Hypothese zutrifft. Die aktuellen Daten des GTEx-Konsortiums [11] zeigen zwar für über 80 % der proteinkodierenden Gene in mindestens einem humanen Zellsystem einen *cis*-eQTL-Effekt durch häufige Varianten. Für die Identifizierung in einer GWAS ist aber entscheidend, ob die spezifische regulatorische Sequenz, die für die Expression des Krankheitsgens im pathophysiologischen Kontext genutzt wird, eine häufige Variante mit eQTL-Effekt aufweist. Dieser Anteil ist möglicherweise geringer. Mit der Identifizierung seltener Varianten besteht also auch eine gute Chance, völlig neue, nicht durch GWAS abgreifbare Krankheitsgene zu finden.

Untersuchungen von Pathomechanismen

Bei seltenen Varianten sind im Vergleich zu häufigen Varianten im Mittel größere Effektstärken zu erwarten. Diese genetischen Effekte korrelieren mit den Effek-

medgen 2019 · 31:212–221 <https://doi.org/10.1007/s11825-019-0246-2>
© Der/die Autor(en) 2019

K. U. Ludwig · F. Degenhardt · M. M. Nöthen

Die Rolle seltener Varianten bei häufigen Krankheiten

Zusammenfassung

Häufige Krankheiten, die sog. Volkskrankheiten, sind in der Regel multifaktoriell verursacht, d. h. zu ihrer Entwicklung tragen sowohl genetische Faktoren als auch nicht-genetische Umgebungseinflüsse bei. Die geschätzte Gesamterblichkeit (-heritabilität) reicht von moderat bis vergleichsweise hoch. Die genetische Architektur ist komplex und kann das gesamte allelische Spektrum, von häufigen Varianten mit niedriger Penetranz bis hin zu seltenen Varianten mit höherer Penetranz, sowie alle möglichen Kombinationen umfassen. Während häufige Varianten seit mehreren Jahren mit großem Erfolg durch genomweite Assoziationsstudien (GWAS) identifiziert werden, war bisher die Identifizierung seltener Varianten, insbesondere aufgrund der großen Zahl beitragender Gene, nur begrenzt erfolgreich. Dies ändert sich derzeit dank der Anwendung von Hochdurchsatz-Sequenzierungstechnologien

(„next-generation sequencing“, NGS) und der daraus resultierenden zunehmenden Verfügbarkeit von exom- und genomweiten Sequenzdaten großer Kollektive. In diesem Artikel geben wir einen Überblick über die Bedeutung seltener Varianten bei häufigen Erkrankungen sowie den aktuellen Stand in Bezug auf deren Identifizierung mittels NGS. Wir betrachten insbesondere die folgenden Fragen: Bei welchen häufigen Krankheiten ist ein Beitrag seltener Varianten zu erwarten, wie können diese Varianten identifiziert werden, und welches Potenzial bieten seltene Varianten für das Verständnis biologischer Prozesse bzw. für die Translation in die klinische Praxis?

Schlüsselwörter

Volkskrankheiten · Ätiologie · Penetranz · Polygenie · Next-generation sequencing

The role of rare variants in common diseases

Abstract

In general, common diseases have a multifactorial etiology, i.e., both genetic and non-genetic environmental factors contribute to disease onset and progression. Heritability estimates for these disorders range from moderate to relatively high. The genetic architecture of common diseases is complex and may encompass the entire allelic spectrum, including common variants with lower penetrance and rare variants with higher penetrance, and all possible combinations thereof. Over recent years, a successful approach to the investigation of common variants has been the genome-wide association study (GWAS) approach. In contrast, elucidation to date of the role of rare variants in common diseases has been limited, in particular due to the large number of contributing genes. This challenge has now been overcome by the introduction of

high-throughput sequencing technologies (next-generation sequencing, NGS), and the increasing availability of exome- and genome-wide sequencing data from large collectives. In this article, we provide an overview of the importance of rare variants in common diseases, and the current state of their identification by NGS. In particular, we address three key questions: Which common diseases are likely to involve a contribution of rare variants; how can these variants be identified; and what potential do rare variants offer in terms of both our understanding of the underlying biological processes and the translation of genetic findings into clinical practice?

Keywords

Multifactorial · Etiology · Penetrance · Polygenicity · Next-generation sequencing

ten auf funktioneller Ebene [42], was wiederum zu eindeutigeren Ergebnissen bei Untersuchungen in den für die Krankheit relevanten Modellsystemen führt, z. B. in Tiermodellen oder in stammzellbasierten Experimenten. Seltene Varianten eignen sich dadurch ganz besonders für die Modellierung von Pathomechanis-

men. Auch wenn derartige molekulare Studien im Bereich der Volkskrankheiten erst begonnen haben, finden sich bereits erste Beispiele bei den chronisch-entzündlichen Darmerkrankungen [46] oder bei Diabetes Typ 2 [25].

Abb. 2 ▲ Das allelische Spektrum und Möglichkeiten der Erfassung. Zur Identifizierung häufiger Varianten werden kostengünstige Genotypisierungs-Arrays (SNP-Arrays) eingesetzt, die bei ausreichender Kohortengröße auch niedrig frequente Varianten detektieren können. Arrays sind nur bedingt geeignet für den Nachweis von seltenen Varianten (z. B. mittels des Exom-Kontents) und sind ungeeignet zur Detektion privater Mutationen oder verschiedener struktureller Varianten (z. B. kleinere „copy number variants“, CNVs). Technologien des „next-generation sequencing“ (NGS) ermöglichen grundsätzlich die Identifizierung von Varianten des gesamten allelischen Spektrums. Aufgrund noch hoher Kosten und der Herausforderungen in der Datenanalyse kommt NGS derzeit in erster Linie zur Detektion von privaten und seltenen Varianten zum Einsatz, aber auch zur Detektion häufiger, kleiner CNVs. Mittelfristig wird NGS auch für die Erhebung von Gesamtgenominformationen eingesetzt werden inkl. der Detektion häufiger Varianten (engl. „genotyping by sequencing“)

Translation

Bei Krankheiten, für die seltene Varianten mit hoher Penetranz identifiziert worden sind (z. B. familiäre Alzheimer Demenz oder koronare Herzkrankheit), finden diese zügig Eingang in die bestätigende und prädiktive Diagnostik. Bei reduzierter Penetranz ist die Aussagekraft der einzelnen Variante dagegen begrenzt und eine direkte Translation in die Diagnostik nicht möglich. Allerdings ist es für häufige, niedrig penetrante Varianten auf Basis von GWAS-Daten gelungen, die Information multipler Loci in sog. polygenen Risiko-Scores zusammenzufassen und damit in den Bereich einer klinisch relevanten Risiko-Einschätzung zu gelangen [22]. Es ist anzunehmen, dass die Informationen von seltenen, niedrig penetranten Varianten in diese Risiko-Scores integrierbar sind und dort zu einer Erhöhung der Vorhersagegenauigkeit führen werden [57]. Der tatsächliche Zugewinn in der Aussagekraft muss jedoch im Einzelfall gezeigt werden.

Genetische Befunde können bei Volkskrankheiten durch Favorisierung von „drug targets“ die Effizienz der Medikamentenentwicklung verbessern [33]. Dieser für GWAS-Befunde demonstrierte Effekt gilt ebenso für sel-

tene Varianten. Grundsätzlich kann die Identifizierung hochpenetranter seltener Varianten, analog zu den monogenen Krankheiten, auch bei Volkskrankheiten zur Entwicklung gentherapeutischer Ansätze führen [8], davon würde aber wahrscheinlich nur eine kleine Gruppe von Patienten profitieren. Vielversprechend für die Entwicklung neuer Medikamente ist auch die Identifizierung seltener protektiver Varianten und die nachfolgende Charakterisierung des funktionellen Mechanismus. Dass durch einen pharmakologischen Ansatz ein solcher Mechanismus erfolgreich imitiert werden kann, konnte z. B. für die koronare Herzkrankheit und das *ANPTLA*-Gen gezeigt werden [14].

Strategien zur Identifizierung seltener Varianten bei Volkskrankheiten

Zur systematischen Detektion seltener Varianten werden als Technologie hauptsächlich SNP-Arrays und die Hochdurchsatzsequenzierung (engl. „next-generation sequencing“, NGS) eingesetzt (Abb. 2). SNP-Arrays waren zunächst primär für die Darstellung häufiger Varianten ausgelegt, die entweder direkt genotypisiert oder mittels statistischer

Vorhersage (sog. Imputation) indirekt erhoben werden. In Abhängigkeit von Kollektivgröße und Ethnizität können jedoch auch seltenere SNPs durch Imputation vorhergesagt werden. Gängige SNP-Arrays ermöglichen derzeit die Darstellung von Varianten bis hin zu einer minoren Allelfrequenz (MAF) von ca. 1 %, zukünftig wird dies sogar bis etwa 0,1 % MAF möglich sein [12, 24]. Im Prinzip können seltene Varianten auch direkt auf einen SNP-Array eingebracht werden, dies wurde z. B. für seltene Varianten der proteinkodierenden Bereiche im sog. Exom-Kontent von SNP-Arrays realisiert (Abb. 2). Damit sind allerdings keine nicht-kodierenden Bereiche abgedeckt, und damit diejenigen genomischen Regionen, die über ihre regulatorischen Effekte auf die Genexpression erheblich zur Ätiologie von Volkskrankheiten beitragen [10, 19]. Auch die sehr seltenen und privaten, in einzelnen Familien auftretenden Varianten werden mit den SNP-Arrays nicht dargestellt, wie auch bestimmte strukturelle Varianten (z. B. kleinere CNVs, Inversionen). Die Untersuchung dieser Teile des allelischen Spektrums wurde erst durch die technologischen Entwicklungen im Bereich des NGS möglich.

Abb. 3 ▲ Strategie zur Untersuchung seltener und privater Varianten bei Volkskrankheiten. Ausgehend von der klinischen Diagnose, werden im ersten Schritt systematisch Kandidatengene identifiziert, in denen seltene Varianten vermutlich einen Krankheitsbeitrag leisten. Im zweiten Schritt wird weitere genetische Evidenz durch Vergrößerung der Kohorte und gezielte Resequenzierung erhoben. Während identische Mutationen eher selten (aber nicht ausgeschlossen) sind, können Mutationen in rekurrenten Genen, Signalwegen oder molekularen Netzwerken statistisch mit der Erkrankung assoziiert werden. Durch Genotyp-Phänotyp-Korrelation können klinische und molekulare Subgruppen identifiziert werden. Dieser Zwei-Schritt-Ansatz wird derzeit als „Übergangsstrategie“ eingesetzt, bis WGS für große Kohorten finanzierbar ist. (Abbildung modifiziert nach [54], mit Genehmigung). GWAS genomweite Assoziationsstudie, CNV „copy number variants“ (Kopienzahlveränderungen), WGS „whole genome sequencing“

Zur Identifizierung seltener Varianten bei Volkskrankheiten kommen grundsätzlich die gleichen NGS-Technologien wie bei der Untersuchung monogener Erkrankungen zum Einsatz. Diese beruhen standardmäßig auf der Sequenzierung kurzer Abschnitte des Genoms (engl. „short read sequencing“), welche als Gesamtgenomsequenzierung („whole genome sequencing“, WGS), Exomsequenzierung („whole exome sequencing“, WES) und als Sequenzierung ausgewählter genomischer Regionen (meist Kandidatengene) in sog. Genpanels durchgeführt wird. In Zukunft werden auch verstärkt Sequenzierungen langer Einzelstrangsequenzen (engl. „long read sequencing“) zum Einsatz

kommen, mit denen zusätzliche, durch „short read sequencing“ nicht darstellbare Varianten (z. B. Repeat-Sequenzen [4]) und auch Haplotypinformationen [47] erfasst werden können. Aufgrund der genetischen Heterogenität sowie der biologischen Komplexität und dem dadurch entstehenden großen „mutational target“ ist allerdings die Verfügbarkeit großer Patientenkollektive sowie genomweiter Sequenzdaten von großen Kontrollkohorten eine Grundvoraussetzung für den Erfolg von NGS bei Volkskrankheiten. Dies verstärkt das Problem der Datenauswertung aufgrund der enormen Zahl seltener Varianten in individuellen Genomen [18] sowie die Herausforderungen an die Rechen- und

Datenspeicherkapazitäten (siehe „Datenanalyse“).

Außerhalb Deutschlands werden bereits großflächig sowohl WES als auch WGS durchgeführt, meist finanziert von öffentlichen Programmen. Diese Studien umfassen zum einen WES/WGS der Allgemeinbevölkerung, für die teilweise umfangreiche Phänotyp-Daten verfügbar sind (z. B. Estonia Genome Project (Estland) oder die UK-Biobank (UK)). Es werden aber auch systematisch Kollektive spezifischer Erkrankungen mittels NGS untersucht, z. B. im Gabriella Miller Kids First Programm (USA) für Erkrankungen im Kindesalter, oder im 100.000 Genomes Project (UK), u. a. für häufige Krebsformen.

In Deutschland wird WES/WGS zur Charakterisierung von populationsbasierten Kohorten und Patientenkollektiven mit Volkskrankheiten nicht oder nur sehr begrenzt gefördert. Notgedrungen werden von den Wissenschaftlern alternative, kostengünstige Strategien verwendet, welche in erster Linie dem Abgreifen der „low hanging fruits“ dienen. Diese Studiendesigns gehen meistens zweischrittig vor (Abb. 3): Zunächst werden Kandidatengene in einem limitierten Kollektiv mittels WES oder WGS identifiziert (alternativ werden die Gene z. B. basierend auf GWAS-Daten ausgewählt). Im zweiten Schritt wird dann über eine gezielte Resequenzierung unabhängige genetische Evidenz in einem größeren Kollektiv erhoben. Im Folgenden wird das Vorgehen näher erläutert und anhand ausgewählter Beispiele von Volkskrankheiten oder häufigen Merkmalen mit Krankheitsrelevanz illustriert.

Anreicherung von Patienten mit seltenen Varianten

Um die Chancen für die Identifizierung von Krankheitsgenen mittels WES/WGS zu erhöhen, können zunächst gezielt Patienten untersucht werden, bei denen angenommen wird, dass ihr Genom für krankheitsbeitragende, seltene Varianten angereichert ist (Abb. 4).

Der Untersuchung mehrfach betroffener Familien (Abb. 4a) liegt die Annahme zugrunde, dass in diesen Familien seltene Varianten mit höherer Pene-

Abb. 4 ▲ Ansätze zur Anreicherung seltener Varianten bei Volkskrankheiten. Die Sterne repräsentieren seltene Varianten, unterschiedliche Farben stellen Mutationen in verschiedenen Genen dar. Eingekreiste Personen und -gruppen stellen zu sequenzierende Individuen dar. **a** Sequenzierung von Mitgliedern mehrfach betroffener Familien. **b** Trio-Design zur Untersuchung von *de novo* Mutationen. **c** Sequenzierung von Subgruppen basierend auf klinischen Daten (**oberer Teil**) oder aufgrund ihrer Extreme in normalverteilten, krankheitsrelevanten Merkmalen (**unterer Teil**). Bei diesem Ansatz können ggf. auch protektive Varianten identifiziert werden. **d** Untersuchung von Populationsisolaten. Nicht dargestellt sind Sequenzierungen umfangreicher Fall-Kontroll-Kohorten

tranz wahrscheinlich häufiger sind als bei Patienten mit sporadischem Auftreten. In diesem Ansatz werden innerhalb einer Familie Individuen von möglichst weit entfernten Ästen des Stammbaums sequenziert. Je niedriger der Verwandtschaftsgrad, desto weniger Varianten werden von diesen Familienmitgliedern geteilt – und desto höher ist die Chance, krankheitsrelevante Varianten zu identifizieren (z. B. [48]). Unter der Annahme vollständiger Penetranz werden zunächst diejenigen Varianten priorisiert, die mit der Erkrankung vollständig kosegregieren und nicht in gesunden Personen auftreten. Diese Annahme kann jedoch bei reduzierter Penetranz zu falsch-negativen Befunden führen. Daher werden in einem zweiten Schritt meist die Filterkriterien relaxiert und auch solche

Varianten betrachtet, die zusätzlich zu den Patienten in mindestens einem nicht betroffenen Familienmitglied vorkommen. Um einer möglichen genetischen Heterogenität in der betroffenen Familie Rechnung zu tragen, werden auch Varianten eingeschlossen, die nur bei einem (Groß-)Teil der Betroffenen vorliegen.

Die Sequenzierung eines betroffenen Kindes sowie seiner nicht betroffenen Eltern (sog. Trio, **Abb. 4b**) kommt besonders bei den Volkskrankheiten zum Einsatz, bei denen ein Beitrag von *de novo* Mutationen zu erwarten ist. Aufgrund des großen „mutational targets“ muss bei dem *de novo* Ansatz eine Vielzahl von Trios untersucht werden, um überzufällig häufig *de novo* Mutationen in einzelnen Genen, Signalwegen oder Netzwerken zu identifizieren. Untersuchungen zu

de novo Mutationen haben bereits erste Erfolge z. B. bei schizophrenen Störungen [17] oder auch bei bestimmten Formen von Autoimmunerkrankungen mit frühem Manifestationsalter [40] gezeigt.

Die Sequenzierung klinischer Subgruppen bzw. Individuen von beiden Enden der Merkmalsverteilung (**Abb. 4c**) beruht ebenfalls auf der Annahme, dass in diesen Kohorten ursächliche (oder auch protektive) Varianten angereichert sind. Ein Beispiel für die Anwendung dieser Strategie ist eine WGS-Analyse des HDL-Cholesterinspiegels (engl. „high density lipoprotein“, HDL-C; niedrige HDL-C-Level sind assoziiert mit einem erhöhten Risiko für kardiovaskuläre Erkrankungen). In Personen des oberen und unteren Spektrums der HDL-C-Verteilung wurden seltene Varianten mit

starken Effekten sowohl auf hohe als auch niedrige HDL-C Level identifiziert, darunter proteinkodierende Mutationen in *ABCA1* und *APOA1* [30].

Bei der Untersuchung von Populationsisolaten (■ Abb. 4d) wird die genetische Heterogenität der Patientenkollektive reduziert. Dadurch können krankheitsbeitragende Varianten, die in durchmischten Populationen selten sind, in höherer Frequenz auftreten und statistisch einfacher detektiert werden [43]. Dieser Ansatz führte z. B. zur Identifizierung mehrerer seltener Varianten, die in der Bevölkerung Sardiniens zur Regulation des Hämoglobinspiegels beitragen [13]. Auch in der isländischen Bevölkerung wurden seltene krankheitsassoziierte Varianten für verschiedene Volkskrankheiten mit dieser Strategie aufgedeckt (z. B. *LGR4* bei Osteoporose [55] oder Varianten in den Genen *CCDN2*, *PAM* und *PDX1* bei Diabetes Typ 2 [53]).

Priorisierung von Varianten

WES-Daten ermöglichen die Analyse seltener Varianten im proteinkodierenden Bereich. Auch die Auswertung von WGS-Daten fokussiert sich derzeit auf das Exom – zum einen, da dort der funktionelle Effekt leichter einzuschätzen ist, zum anderen ermöglicht WGS als „besseres Exom“ eine gleichmäßigere Abdeckung der proteinkodierenden Regionen [29]. Je nach Auswertestrategie bleibt eine Vielzahl an potenziellen Varianten übrig, die in Bezug auf einen Krankheitsbeitrag priorisiert werden müssen. Dies geschieht meist anhand funktioneller Annotationen (u. a. verschiedene *in silico* Prädiktions-Scores), Informationen über die Toleranz des Gens gegenüber Mutationen (z. B. RVIS-Score [38]), unabhängige Evidenz aus der Literatur und ggf. Ergebnissen funktioneller Untersuchungen. Die Frequenz der identifizierten Variante(n) in Referenzdaten gleicher Ethnizität ist bei häufigen Erkrankungen nur eingeschränkt aussagekräftig: Zum einen enthalten die meisten systematischen Referenzdatenbanken (z. B. gnomAD [23]) auch Patientenkollektive, zum anderen gehen seltene Varianten bei häufigen Erkrankungen oft mit reduzierter Penetranz

einher. Daher können Varianten bei Vorliegen in diesen Referenzdatenbanken nicht zwingend als krankheitsbeitragend ausgeschlossen werden. Gleichzeitig ist, ähnlich wie bei monogenen Erkrankungen, die Abwesenheit einer Variante in Referenzdaten auch keine hinreichende Evidenz für Kausalität.

Resequenzierung großer Patienten- und Kontrollkohorten

Die im ersten Schritt identifizierten Kandidatengene werden anschließend in umfangreichen Fall-Kontroll-Kohorten resequenziert mit dem Ziel, zusätzliche unabhängige Evidenz für die Relevanz einzelner Gene in der Krankheitsätiologie zu erhalten. Neben Genpanels kommerzieller Anbieter kommt besonders im Forschungsansatz zunehmend die amplikonbasierte Resequenzierung mittels „molecular inversion probes“ (MIPs) als Sequenzierungsmethodik zum Einsatz [32, 36, 46]. Hierfür werden an einen 30 bp-langen Backbone locuspezifische Primer angefügt, welche in einer Multiplexreaktion an die zu untersuchende DNA hybridisieren. Der Bereich zwischen den Primern wird komplementär zur Ziel-DNA aufgefüllt und der MIP zu einem Ring geschlossen. Unter Einbringung von Index-Barcodes und universellen Primern werden die Fragmente mittels PCR amplifiziert, anschließend können eine Vielzahl an Personen gemeinsam in einem Sequenzierlauf untersucht werden. Die MIPs wurden zunächst bei der Aufarbeitung von WES-Befunden, u. a. bei Autismus-Spektrum-Störungen [36] angewandt und werden seitdem sowohl bei monogenen als auch multifaktoriellen Erkrankungen eingesetzt [32, 46]. Gründe für den Einsatz von MIPs sind dabei vor allem Kosteneffizienz, Skalierbarkeit bei großen Kollektiven, Optimierungsmöglichkeiten im Labor sowie die Möglichkeit der projektübergreifenden Zusammensetzung der MIP-Pools.

Datenanalyse

Aufgrund der großen Zahl krankheitsrelevanter Gene bei Volkskrankheiten ist

die Datenanalyse eine besondere Herausforderung. Obwohl Gene mit seltenen, krankheitsbeitragenden Varianten prinzipiell in den zur Erkrankung beitragenden Signalwegen und molekularen Netzwerken angereichert sind [58], ist die Identifizierung dieser Gene mit statistischer Signifikanz schwierig, vor allem vor dem Hintergrund der Vielzahl seltener (und meist benignen) Varianten pro Genom. Mit zunehmend umfangreicheren Datensätzen wird dies einfacher werden, da zu erwarten ist, dass seltene Varianten auch in denselben Genen oder sogar identische Mutationen bei mehreren unverwandten Betroffenen identifiziert werden.

Seltene Varianten weisen teilweise erhebliche Unterschiede in ihrer Frequenz zwischen (Sub)-Bevölkerungen auf. Dadurch ist das Problem der Populationss-stratifikation bei seltenen Varianten deutlich stärker ausgeprägt als bei häufigen Varianten. Frequenzabschätzungen aus Referenzdatenbanken sind daher nur bedingt nutzbar. Eine statistische Korrektur der Stratifikation scheint grundsätzlich möglich, bedarf jedoch möglichst genomweiter Datensätze seltener Varianten [28].

Schließlich sind die statistischen Modelle, die zur Bewertung einer Krankheitsassoziation für häufige Varianten entwickelt wurden, nur eingeschränkt auf seltene Varianten übertragbar. Dies ist v. a. dadurch bedingt, dass aufgrund der niedrigen Allelfrequenz die statistische Aussagekraft bzgl. einer einzelnen, seltenen Variante beschränkt ist. Zudem basieren diese Tests in der Regel auf der Annahme einer asymptotischen Verteilung, welche bei seltenen Varianten nur bei sehr großen Kollektivgrößen gegeben ist. Darüber hinaus ist der Korrekturfaktor für multiples Testen bei seltenen Varianten aufgrund der vielfach höheren Anzahl im Vergleich zu häufigen Varianten deutlich erhöht. In Abwesenheit robuster Methoden zur statistischen Bewertung einzelner seltener Varianten kann es daher sinnvoll sein, die Gesamtheit seltener Varianten innerhalb einer genetischen Region („Mutationslast“) in sog. „burden“-Analysen zusammenzufassen [26]. Dabei werden typischerweise die minoren Al-

lele aller seltenen Varianten eines Gens und deren gemeinsame Assoziation mit der Erkrankung betrachtet, alternativ mehrere Gene eines Signalweges oder molekularen Netzwerks (siehe [6] für eine Übersicht verschiedener Methoden).

Es wird auch diskutiert, Varianten basierend auf *a priori* Evidenz zu filtern, z. B. aufgrund ihres erwarteten funktionellen Effektes. Dieser Ansatz reduziert zwar einerseits die Zahl eingeschlossener Varianten, zum anderen werden mögliche kausale Varianten angereichert [34]. Diese Strategie hat sich z. B. bei der Identifikation der Assoziation seltener Varianten im *PPARG*-Gen bei Diabetes Typ 2 als zielführend erwiesen [25]. Ein alternativer Ansatz kann hier die Integration von prädiktiven *in silico*- oder funktionellen Scores sein, mit denen Varianten in den „burden“ Analysen gewichtet werden. Dadurch würden potenziell funktionell relevante Varianten betont, aber keine (möglicherweise falsch-annotierten) Varianten aus der Analyse ausgeschlossen.

Herausforderungen

Die Zahl der systematischen NGS-basierten Analysen bei Volkskrankheiten hängt denen der monogenen Krankheiten deutlich hinterher. Angesichts der erforderlichen Kollektivgrößen sind die Kosten für die Generierung großer WGS-Datensätze derzeit noch sehr hoch und in der Regel nur in großen koordinierten Bemühungen zu stemmen. Die stetig sinkenden Kosten für NGS werden es aber in Zukunft ermöglichen, diese Datensätze für viele Volkskrankheiten zu generieren. Damit wird die systematische Identifizierung seltener Varianten mit hinreichender statistischer Aussagekraft möglich.

Auf Ebene der IT-Infrastruktur werden umfangreiche Rechen- und Speicherkapazitäten benötigt. Mit zunehmendem Umfang der erhobenen NGS-Daten (im Extremfall: WGS) steigen die Anforderungen an die rechnergestützte Datenspeicherung und -auswertung (inkl. des entsprechend ausgebildeten wissenschaftlichen Personals). Dieses Problem vervielfältigt sich durch die für die Analyse von Volkskrankheiten nötigen Kollektivgrößen. Entsprechende

Infrastruktur ist in Deutschland derzeit an nur wenigen Zentren vorhanden und alternative Ansätze wie „cloud storage“ und „cloud computing“ müssen in Betracht gezogen werden.

Einzelne Schritte der Auswertung von NGS-Daten müssen an die Charakteristika der Volkskrankheiten angepasst werden. Dies umfasst z. B. eine Optimierung des Varianten-Calling unter besonderer Berücksichtigung der Spezifität, da in großen Kohorten nicht mehr jede einzelne Variante im Labor experimentell bestätigt werden kann. Darüber hinaus sind die verfügbaren Methoden zur statistischen Bewertung seltener Varianten bisher noch unausgereift [56], auch die Kombination verschiedener Tests scheint die statistische Aussagekraft nicht zu verbessern [31]. Die Entwicklung anderer Ansätze zur *P*-Wert-Berechnung für seltene Varianten, z. B. mittels Permutationen [51], kann hier zielführend sein.

Soll die Power der Analysen durch Zusammenführung von Datensätzen aus verschiedenen Quellen erhöht werden, besteht aufgrund technischer Unterschiede in verschiedenen Aspekten der NGS-Technologie und/oder der Datenanalyse das Risiko sowohl falsch-positiver als auch falsch-negativer Befunde. Eine gemeinsame Re-Analyse aller Rohdaten über bereitgestellte fastq-Files kann hier ein Teil des Lösungsansatzes sein und wurde z. B. für die ExAC-Daten durchgeführt [23].

Für die Bewertung der Funktionalität (und ggf. Kausalität) einzelner Varianten werden qualitativ hochwertige Annotationen benötigt. Diese *in silico* Tools sind im Bereich der proteinkodierenden Varianten bereits systematisch verfügbar und korrelieren mit Daten aus experimentellen Untersuchungen [16]. Schwieriger ist die Situation für die Bewertung nicht-kodierender Varianten, deren Zahl jedoch durch WGS deutlich steigen wird. Prädiktive Scores für funktionelle Effekte nicht-kodierender Varianten wurden kürzlich entwickelt (z. B. CADD [44], Eigen [21] oder LINSIGHT [20]), allerdings bleibt abzuwarten, inwieweit diese Scores die biologischen Effekte tatsächlich abbilden. Zweifel an der derzeitigen Aussagekraft bestehen besonders vor dem Hintergrund, dass der regulatorische Effekt

nicht-kodierender Varianten oft zelltyp-abhängig ist oder nur im pathophysiologischen Kontext (d. h. unter bestimmten Bedingungen) auftritt.

Nach erfolgreicher Identifizierung der genetischen Faktoren, die zu Volkskrankheiten beitragen, ist es für viele wissenschaftliche Fragestellungen (insbesondere für die Translation) notwendig, Varianten des gesamten allelischen Spektrums gemeinsam zu betrachten. Solche integrierten Analysen stehen derzeit allerdings noch am Anfang. Erste Studien wurden bei „monogenen“ Krankheiten durchgeführt, bei denen z. B. der Einfluss des polygenen Hintergrundes auf Ausprägung von Symptomen [35] oder Penetranz gezeigt werden konnte [9].

Schlussbetrachtung

Den NGS-Technologien kommt bei der systematischen Identifizierung seltener Varianten bei den multifaktoriell verursachten Volkskrankheiten zunehmende Bedeutung zu, dabei wird der Beitrag der seltenen Varianten bei den verschiedenen Krankheiten unterschiedlich groß sein. Die Generierung systematischer WES- und WGS-Daten in großen Patientenkollektiven hat gerade erst begonnen, dieser Ansatz wird aber mit sinkenden Kosten eine große Dynamik entwickeln. Die Anforderungen an Datenspeicherung und -analyse stellen die beteiligten Arbeitsgruppen vor große Herausforderungen, aber auch hier wird man von absehbaren Entwicklungen profitieren. Die Identifizierung des genetischen Beitrags über das gesamte allelische Spektrum hinweg wird zukünftig eine integrierte Betrachtung von seltenen und häufigen Varianten ermöglichen. Dies hat Implikationen für die Angabe von Wiederholungsrisiken, die Identifizierung molekularer Subtypen von Krankheiten und die Identifizierung von Risikopersonen. Außerdem wird ein wichtiger Beitrag zur Aufklärung kausaler pathophysiologischer Zusammenhänge bei Volkskrankheiten geleistet und die Identifikation neuer therapeutischer Targets ermöglicht.

Fazit für die Praxis

- Seltene Varianten tragen als Teil des allelischen Spektrums zu den Volkskrankheiten bei. Der Gesamtbeitrag seltener Varianten ist je nach Erkrankung unterschiedlich.
- In einzelnen Fällen können hochpenetrante seltene Varianten einen Großteil der individuellen Krankheitsätiologie erklären. Dies kann relevant für die genetische Beratung und ggf. die klinische Behandlung sein.
- Seltene Varianten mit größeren Effektstärken sind für funktionelle Untersuchungen besonders geeignet und können somit erheblich zum Verständnis der Pathophysiologie der Erkrankung beitragen.
- WGS zur Charakterisierung des gesamten allelischen Spektrums bei Volkskrankheiten ist technisch möglich, aber mit gegenwärtig noch hohen Kosten sowie umfangreichen Herausforderungen an IT-Infrastruktur, statistischen Methoden und Dateninterpretation verbunden.
- Eine integrierte Betrachtung von seltenen und häufigen Varianten wird bei den Volkskrankheiten in Zukunft zur Entwicklung einer personalisierten Medizin beitragen.

Korrespondenzadresse

Dr. Kerstin U. Ludwig

Emmy-Noether-Gruppe „Kraniofaziale Genomik“, Institut für Humangenetik, Universitätsklinikum Bonn
Venusberg-Campus 1, Gebäude 76,
53127 Bonn, Deutschland
kerstin.ludwig@uni-bonn.de

Danksagung. Die Autorinnen und der Autor danken Frau Dr. Stefanie Heilmann-Heimbach und Herrn Dr. Julian Hecker für die kritische Durchsicht des Manuskripts sowie Frau Christine Fischer für die Unterstützung bei der grafischen Illustration des Artikels.

Förderung. Diese Arbeit wurde gefördert durch die Deutsche Forschungsgemeinschaft (LU 1944/3-1).

Einhaltung ethischer Richtlinien

Interessenkonflikt. K.U. Ludwig, F. Degenhardt und M.M. Nöthen geben an, dass kein Interessenkonflikt besteht.

Für diesen Beitrag wurden von den Autoren keine Studien an Menschen oder Tieren durchgeführt. Für die aufgeführten Studien gelten die jeweils dort angegebenen ethischen Richtlinien.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz (<http://creativecommons.org/licenses/by/4.0/deed.de>) veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Literatur

1. Abifadel M, Varret M, Rabes JP et al (2003) Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat Genet* 34:154–156
2. Agarwala V, Flannick J, Sunyaev S et al (2013) Evaluating empirical bounds on complex disease genetic architecture. *Nat Genet* 45:1418–1427
3. Anderson CA, Soranzo N, Zeggini E et al (2011) Synthetic associations are unlikely to account for many common disease genome-wide association signals. *Plos Biol* 9:e1000580
4. Ardui S, Ameur A, Vermeesch JR et al (2018) Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res* 46:2159–2168
5. Astle WJ, Elding H, Jiang T et al (2016) The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* 167:1415–1429.e19
6. Bomba L, Walter K, Soranzo N (2017) The impact of rare and low-frequency genetic variants in common disease. *Genome Biol* 18:77
7. Bulik-Sullivan BK, Loh PR, Finucane HK et al (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 47:291–295
8. Carvalho M, Sepodes B, Martins AP (2017) Regulatory and scientific advancements in gene therapy: state-of-the-art of clinical applications and of the supporting European Regulatory Framework. *Front Med* 4:182
9. Castel SE, Cervera A, Mohammadi P et al (2018) Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat Genet* 50:1327–1334
10. Claussnitzer M, Dankel SN, Kim KH et al (2015) FTO obesity variant circuitry and adipocyte browning in humans. *N Engl J Med* 373:895–907
11. GTEx Consortium (2017) Genetic effects on gene expression across human tissues. *Nature* 550:204–213
12. Consortium UK, Walter K, Min JL et al (2015) The UK10K project identifies rare variants in health and disease. *Nature* 526:82–90
13. Danjou F, Zoledziwska M, Sidore C et al (2015) Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels. *Nat Genet* 47:1264–1271
14. Dewey FE, Gusarova V, O'dushlaine C et al (2016) Inactivating variants in ANGPTL4 and risk of coronary artery disease. *N Engl J Med* 374:1123–1133
15. Dickson SP, Wang K, Krantz I et al (2010) Rare variants create synthetic genome-wide associations. *Plos Biol* 8:e1000294
16. Findlay GM, Daza RM, Martin B et al (2018) Accurate classification of BRCA1 variants with saturation genome editing. *Nature* 562:217–222
17. Fromer M, Pocklington AJ, Kavanagh DH et al (2014) De novo mutations in schizophrenia implicate synaptic networks. *Nature* 506:179–184
18. Genomes Project C, Auton A, Brooks LD et al (2015) A global reference for human genetic variation. *Nature* 526:68–74
19. Gupta RM, Hadaya J, Trehan A et al (2017) A genetic variant associated with five vascular diseases is a distal regulator of Endothelin-1 gene expression. *Cell* 170:522–533.e15
20. Huang YF, Gulko B, Siepel A (2017) Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet* 49:618–624
21. Ionita-Laza I, McCallum K, Xu B et al (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 48:214–220
22. Khera AV, Chaffin M, Aragam KG et al (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 50:1219–1224
23. Lek M, Karczewski KJ, Minikel EV et al (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536:285–291
24. Loh PR, Danecek P, Palamara PF et al (2016) Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* 48:1443–1448
25. Majithia AR, Flannick J, Shahinian P et al (2014) Rare variants in PPARG with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. *Proc Natl Acad Sci USA* 111:13127–13132
26. Marouli E, Graff M, Medina-Gomez C et al (2017) Rare and low-frequency coding variants alter human adult height. *Nature* 542:186–190
27. Marshall CR, Howrigan DP, Merico D et al (2017) Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat Genet* 49:27–35
28. Mathieson I, McVean G (2012) Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 44:243–246
29. Meienberg J, Bruggmann R, Oexle K et al (2016) Clinical sequencing: is WGS the better WES? *Hum Genet* 135:359–362
30. Morrison AC, Voorman A, Johnson AD et al (2013) Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat Genet* 45:899–901
31. Moutsianas L, Agarwala V, Fuchsberger C et al (2015) The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *Plos Genet* 11:e1005165
32. Nejentsev S, Walker N, Riches D et al (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* 324:387–389
33. Nelson MR, Tipney H, Painter JL et al (2015) The support of human genetic evidence for approved drug indications. *Nat Genet* 47:856–860
34. Nicolae DL (2016) Association tests for rare variants. *Annu Rev Genomics Hum Genet* 17:117–130
35. Niemi MEK, Martin HC, Rice DL et al (2018) Common genetic variants contribute to risk of

- rare severe neurodevelopmental disorders. *Nature* 562:268–271
36. O’roak BJ, Vives L, Fu W et al (2012) Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* 338:1619–1622
 37. Park JH, Gail MH, Weinberg CR et al (2011) Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc Natl Acad Sci U S A* 108:18026–18031
 38. Petrovski S, Gussow AB, Wang Q et al (2015) The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. *Plos Genet* 11:e1005492
 39. Power RA, Kyaga S, Uher R et al (2013) Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA Psychiatry* 70:22–30
 40. Pullabhatla V, Roberts AL, Lewis MJ et al (2018) De novo mutations implicate novel genes in systemic lupus erythematosus. *Hum Mol Genet* 27:421–429
 41. Raghavan NS, Brickman AM, Andrews H et al (2018) Whole-exome sequencing in 20,197 persons for rare variants in Alzheimer’s disease. *Ann Clin Transl Neurol* 5:832–842
 42. Raychaudhuri S (2011) Mapping rare and common causal alleles for complex human diseases. *Cell* 147:57–69
 43. Reich DE, Lander ES (2001) On the allelic spectrum of human disease. *Trends Genet* 17:502–510
 44. Rentzsch P, Witten D, Cooper GM et al (2018) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* 47(D1):D886–D894
 45. Rivas MA, Beaudoin M, Gardet A et al (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* 43:1066–1073
 46. Rivas MA, Graham D, Sulem P et al (2016) A protein-truncating R179X variant in RNF186 confers protection against ulcerative colitis. *Nat Commun* 7:12342
 47. Roe D, Vierra-Green C, Pyo CW et al (2017) Revealing complete complex KIR haplotypes phased by long-read sequencing technology. *Genes Immun* 18:127–134
 48. Sasaki MM, Skol AD, Hungate EA et al (2016) Whole-exome sequence analysis implicates rare IL17REL variants in familial and sporadic inflammatory bowel disease. *Inflamm Bowel Dis* 22:20–27
 49. Saunders EJ, Dadaev T, Leongamornlert DA et al (2014) Fine-mapping the HOXB region detects common variants tagging a rare coding allele: evidence for synthetic association in prostate cancer. *Plos Genet* 10:e1004129
 50. Singh T, Kurki MI, Curtis D et al (2016) Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat Neurosci* 19:571–577
 51. Sondhi A, Rice KM (2018) Fast permutation tests and related methods, for association between rare variants and binary outcomes. *Ann Hum Genet* 82:93–101
 52. Stefansson H, Rujescu D, Cichon S et al (2008) Large recurrent microdeletions associated with schizophrenia. *Nature* 455:232–236
 53. Steinthorsdottir V, Thorleifsson G, Sulem P et al (2014) Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet* 46:294–298
 54. Stessman HA, Bernier R, Eichler EE (2014) A genotype-first approach to defining the subtypes of a complex disease. *Cell* 156:872–877
 55. Styrkarsdottir U, Thorleifsson G, Sulem P et al (2013) Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. *Nature* 497:517–520
 56. Timpson NJ, Greenwood CMT, Soranzo N et al (2018) Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat Rev Genet* 19:110–124
 57. Wainschein P, Jain DP, Yengo L et al (2019) Recovery of trait heritability from whole genome sequence data. *bioRxiv*. <https://doi.org/10.1101/588020>
 58. Walsh T, McClellan JM, McCarthy SE et al (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 320:539–543
 59. Wessel J, Chu AY, Willems SM et al (2015) Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nat Commun* 6:5897
 60. Wu L, Rosa-Neto P, Hsiung GY et al (2012) Early-onset familial Alzheimer’s disease (EOFAD). *Can J Neurol Sci* 39:436–445
 61. Yang J, Lee SH, Goddard ME et al (2011) GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 88:76–82
 62. Yap CX, Sidorenko J, Wu Y et al (2018) Dissection of genetic variation and evidence for pleiotropy in male pattern baldness. *Nat Commun* 9:5407
 63. Zuk O, Hechter E, Sunyaev SR et al (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A* 109:1193–1198