Aus dem Charité Centrum 17
Frauen-, Kinder- und Jugendmedizin mit
Perinatalzentrum und Humangenetik
Institut für Medizinische Genetik und Humangenetik
Direktor: Prof. Dr. med. Stefan Mundlos

Aufklärung der genetischen Ursache von Glycosylphosphatidylinositol (GPI)-Ankerstörungen mittels Hochdurchsatz-Sequenzierung

Habilitationsschrift

Zur Erlangung der venia legendi

für das Fach Humangenetik

Vorgelegt dem Fakultätsrat der Medizinischen Fakultät

Charité Universitätsmedizin Berlin

von

Dr. med. Peter Michael Krawitz, Dipl. Phys.

Berlin

Eingereicht: November 2014
Dekanin: Frau Prof. Dr. Annette Grüters-Kieslich
Erstgutachter: Prof. Dr. O. Rieß, Tübingen
Zweitgutachter: Prof. Dr. Bernhard Horsthemke, Essen

„Ελπὶς καὶ κίνδυνος ἐν ἀνθρώποισιν ὁμοῖοι“

Theognis, Vers 637

„Hoffnung und Gefahr sind gleich unter den Menschen“

Theognis, Vers 637

Abkürzungen

1 KGP	1000 Genom Projekt
bp	Basenpaar
CHO	Zelllinie aus Ovarien des chinesischen Hamsters
dbSNP	Datenbank für Sequenzvarianten
DNA	„deoxyribonucleid acid“, Erbsubstanz
ER	Endoplasmatisches Retikulum
FPR	Falsch-Positiv-Rate
GPI	Glykosylphosphatidylinositol
GPI-AP	„GPI-anchored Protein“, GPI-verankertes Protein
HGMD	„Human Gene Mutation Database“, Datenbank pathogener SNVs
HPMRS	„Hyperphosphatasia with Mental Retardation Syndrome“, Hyperphosphatasie mit mentaler Retardierung, Mabry Syndrom
Kb	Kilo Basen, 10^3 bp
MAC	„membrane attack complex“, Membranangriffskomplex
Mb	Mega Basen, 10^6 bp
NGS	„next-generation sequencing“, Hochdurchsatz-Sequenzierung
NMD	„nonsense mediated mRNA decay“, Abbaumechanismus von RNA mit vorzeitigem Stoppcodon
OMIM	„Online Mendelian Inheritance in Man”, Verzeichnis genetischer Erkrankungen
PCR	„polymerase chain reaction“, Verfahren zur DNA Amplifikation
PNH	paroxysmale nächtliche Hämoglobinurie
SNV	„single nucleotide variant“, Einzelbasenaustausch
VUCS	“variants of unknown clinical significance”, Mutationen unbestimmter klinischer Signifikanz

1 Einleitung

1.1 Hochdurchsatz-Sequenzierung

Mit dem in den 90er Jahren des letzten Jahrhunderts gestarteten Humangenomprojekt, HGP, begann ein grundlegender Wandel in der Medizinischen Genetik: Bei der Ursachenklärung seltener Erkrankungen gewinnt die Hochdurchsatz-Sequenzierung, HDS, seitdem zunehmend an Bedeutung. Hierbei werden nicht mehr nur einzelne, bekannte Gene gezielt sequenziert, sondern es kann parallel in vielen Genen oder gar genomweit nach auffälligen Veränderungen gesucht werden.

Ein Höhepunkt des HGP war die Veröffentlichung einer ersten Referenzsequenz des humanen Genoms, die als Blaupause bei der Analyse weiterer Genome diente (International Human Genome Sequencing, 2004). Das im Anschluss gestartete 1000 Genom Projekt, 1KGP, lieferte dann eine wichtige genomweite Übersicht der genetischen Varianten bei gesunden Individuen in unterschiedlichen Bevölkerungsgruppen und trug damit wesentlich zur Beurteilung von Humangenomen bei (Genomes Project, et al., 2010).

Im Rahmen dieser multinationalen Forschungsprojekte kam es auch zu zahlreichen Neu- und Weiterentwicklungen in der DNA-Sequenziertechnologie, über die im Folgenden ein kurzer Überblick gegeben wird. Die Markteinführung moderner Hochdurchsatz-Sequenziergeräte führte auch zu einer drastischen Kostenreduktion. Seit einigen Jahren sinken nun die Preise für die Generierung der Rohsequenzdaten und der aktuelle Preis für ein Humangenom beträgt ca. 1000 € (Abbildung 1). Die Kosten einer Genomsequenzierung liegen damit nun in der Größenordnung von Einzelgenuntersuchungen. Genomweite Suchtests werden daher zunehmend zur Methode der Wahl bei Verdacht auf monogen bedingte Erkrankungen.

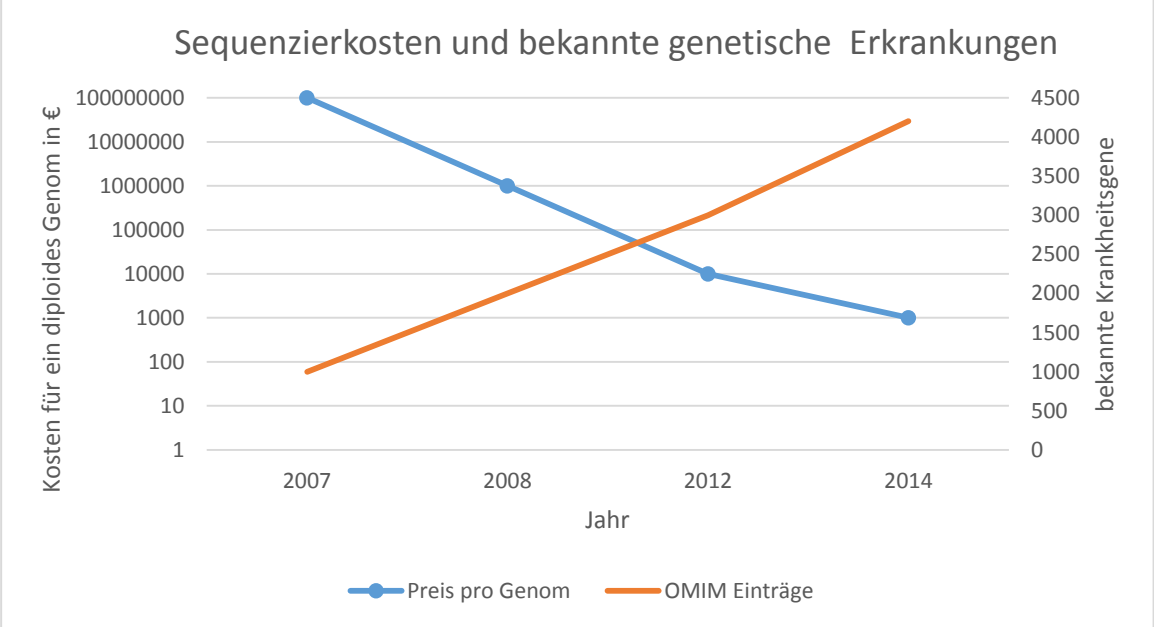


Abbildung 1: Kostenentwicklung in der Humangenomsequenzierung und bekannte, monogene Erkrankungen: Mit der Einführung moderner Hochdurchsatz-Sequenziergeräte kam es zu einer drastischen Preisreduktion bei Humangenomen. Während die Erstellung der ersten diploiden Sequenz eines Individuums 2007 noch ca. 100 Mio. € kostete, belaufen sich die aktuellen Preise auf ca. 1000 €. Durch die immer breitere Verfügbarkeit der kostengünstigen Sequenziermethoden konnten die genetischen Ursachen einer Vielzahl von Erkrankungen, die sich nach den Mendelschen Regeln vererben, in den letzten Jahren geklärt werden. Aktuell sind im online-Verzeichnis genetischer Erkrankungen, OMIM, 4200 monogene Krankheiten beschrieben.

Unter dem Sammelbegriff HDS oder auch „next-generation sequencing“, NGS, versteht man unterschiedliche DNA-Sequenzierverfahren, die es ermöglichen, in einem Ansatz einige Gigabasen Rohsequenzdaten zu generieren. Diese Sequenzmenge liegt um einige Größenordnungen über denen der herkömmlichen Sanger Sequenzierung und eröffnet damit neue Untersuchungsmöglichkeiten wie Genom-, Epigenom- und Transkriptom-Sequenzierungen. Zugleich stellen sich aber auch neue Herausforderungen in der bioinformatischen Datenverarbeitung.

Bei den derzeit verfügbaren Hochdurchsatz-Sequenziergeräten konnte dieser Skaleneffekt durch eine Miniaturisierung der Reaktionsgefäße und den Einsatz von Mikrofluidiksystemen sowie hochauflösender, lichtempfindlicher oder pH-sensitiver elektronischer Bauelemente erreicht werden (Mardis, 2008; Mardis, 2013; Shendure and Ji, 2008). Aktuell am weitesten verbreitet sind Geräte der Hersteller Illumina (MiSeq, HiSeq), Roche (GS Junior, GS FLX), Ion Torrent (Ion PGM, Ion Proton) und Applied Biosystems (SOLiD). Obwohl es plattformspezifische Unterschiede gibt, folgen diese Technologien doch einem gemeinsamen Prinzip, das sich in drei aufeinanderfolgende Schritte untergliedern lässt, die Probenvorbereitung, eine Amplifikation zur Signalverstärkung und letztlich die eigentliche Sequenzierreaktion. Bei der Illumina Technologie sind folgende Bezeichnungen üblich:

- „Library-Preparation“: Hierbei wird die DNA üblicherweise in einem Ultraschallbad zerkleinert und spezielle Oligonukleotide, sogenannte „Adaptors“, werden an die DNA-Fragmente ligiert.
- „Cluster generation“: Die Sequenzfragmente werden mittels der Adaptoren an die Oberfläche einer Flusszelle („flow cell“) gebunden. Es erfolgt eine PCR-basierte „bridge amplification“, bei der an einer umschriebenen Stelle auf der Flusszelle eine Vielzahl sequenzidentischer Oligonukleotide entsteht.
- „Sequencing-by-synthesis“: Die Sequenzierung erfolgt in einem erneuten polymerase-gestützten Syntheseschritt der Cluster, bei dem jedoch nun fluoreszenzmarkierte Nukleotide verwendet werden. Wenn das zugegebene Nukleotid eingebaut werden kann, lässt sich mit einem lichtempfindlichen elektronischen Element (CCD-Sensor) ein Lichtsignal nachweisen. Nach Abschluss der kompletten Synthese können diese Bilddaten dann in eine Sequenz für jeden Cluster auf der Flußzelle umgewandelt werden.

Die einzelnen Sequenzfragmente sind in ihrer Länge meist auf wenige hundert Basenpaare beschränkt. Durch die massive Parallelisierung kann jedoch zeitgleich die Sequenz vieler Millionen Cluster bestimmt und damit eine viele Megabasen umfassende Zielregion untersucht werden. Je nach Fragestellung kann es sich dabei um ein gesamtes Genom handeln, die bekannten, kodierenden Abschnitte des Genoms (Exom-Sequenzierung) oder um ausgewählte Gene (Gen-Panels). Eine Einschränkung der Zielregion erfordert einen zusätzlichen Anreicherungsschritt während oder nach der Library Präparation, der entweder amplikonbasiert sein kann (RainDance, Multiplicon) oder auf Hybridisierung mit Oligonukleotidsonden beruht (Agilent, SureSelect). Dadurch sinkt die erforderliche Datenmenge und es können so Sequenzierkosten und -zeit reduziert werden.

Da die bei allen Verfahren erzeugten Rohsequenzen zufällig über die Zielregion verteilt sind, müssen sie zunächst bioinformatisch möglichst passgenau an einer bereits vorliegenden Referenzsequenz ausgerichtet werden (reference-guided sequence realignment). Die Software Werkzeuge, die hierbei zum Einsatz kommen (short read mapper), wenden ausgefeilte Heuristiken an, um diesen sehr rechenaufwändigen Prozess zu bewältigen.

Bei ausreichender Sequenzqualität kann eine Sequenzvariante üblicherweise gut identifiziert werden, wenn die Abdeckungstiefe über 30 Sequenzfragmente beträgt. Höhere Sequenziertiefen sind erforderlich, wenn größere Insertionen, Deletionen oder strukturelle Veränderungen, die mehrere Exone umfassen können, detektiert werden sollen. Um auch Mosaike und somatische Mutationen aufspüren zu können, ist nochmals eine höhere Abdeckung erforderlich. Die hierauf ausgerichteten Untersuchungsverfahren werden daher auch als deep-sequencing Anwendungen bezeichnet.

Je nach Fragestellung werden bei der Detektion von Sequenzvarianten unterschiedliche Algorithmen („variant caller“) verwendet. Wenn man eine somatische Mutation oder ein Mosaik detektieren möchte, ist ein geringer Frequenzschwellenwert für abweichende Sequenzfragmente bei hoher Abdeckung der sensitivste Ansatz. Bei Trio-Exom-Sequenzierungen, darunter versteht man die Analyse des betroffenen Index-Patienten sowie seiner Eltern, sind hingegen Algorithmen vorteilhaft, die die Wahrscheinlichkeiten bestimmter Genotyp-Kombinationen berücksichtigen.

Die Menge der identifizierten Sequenzvarianten wächst mit der Größe der Zielregion beträchtlich an. Während man bei der herkömmlichen Sequenzierung bei circa einer Kilobase kodierender DNA eine Sequenzvariante erwartet, steigt diese Zahl bei den ca. 30 Megabasen des humanen Exoms bereits auf 15.000-40.000 an. Bei Gesamtgenomsequenzierungen findet man pro Individuum bis zu 3 Millionen Abweichungen von der bekannten Referenz und auch nach Abgleich mit Datenbanken, die alle bislang bekannten Sequenzvarianten enthalten (dbSNP), verbleiben ca. 1% an Mutationen, die gänzlich unbeschrieben sind. Diese Varianten werden häufig als private oder persönliche Mutationen bezeichnet.

Wenn man eine von einer monogenen Erkrankung betroffene Person molekulargenetisch untersucht und bekannte pathogene Mutationen sowie Polymorphismen bereits ausgeschlossen hat, so müssen alle verbleibenden privaten Sequenzvarianten als mögliche Krankheitsursache in Betracht gezogen werden (variants of unknown significance, VUCS). Pro individuellem Exom sind dies einige hundert VUCS, von denen jedoch nur eine der Erkrankung zugrunde liegen dürfte. Die Filterung und Priorisierung dieser Kandidaten stellt eine zentrale Herausforderung der Bioinformatik dar.

1.2 Identifikation pathogener Mutationen in NGS Daten

Pathogene Mutationen, die in Patienten mit genetischen Erkrankungen gefunden wurden, gaben der Grundlagenforschung wiederholt wichtige Impulse und ermöglichten dadurch oftmals erst ein tieferes Verständnis der biologischen Mechanismen. Um bei den von uns mittels HDS untersuchten Patienten die ursächlichen Allele zu identifizierten, verwendeten wir unterschiedliche statistische Verfahren und entwickelten diese zum Teil weiter. Im Folgenden wird ein Überblick über Verfahren gegeben, die bei der Interpretation von potentiell krankheitsverursachenden Sequenzvarianten eingesetzt werden.

Bei der Priorisierung der VUCS können statistische, molekularbiologische sowie phänotypisch motivierte Überlegungen berücksichtigt werden. Unter die Kartierungsverfahren fallen alle Methoden der Kopplungsanalyse („linkage analysis“), mit deren Hilfe Assoziationen zwischen genetischen Markern und phänotypischen Merkmalen aufgespürt werden können. Hierfür ist es zu Beginn erforderlich, in mehreren Individuen derselben Familie, oder in Fall-Kontroll-Gruppen, polymorphe genetische Marker, wie zum Beispiel SNPs, zu bestimmen. Bei großen Familien mit mehreren betroffenen Mitgliedern gibt es häufig klare Hinweise auf ein Vererbungsmodell, z.B. autosomal dominant oder X-chromosomal gebunden rezessiv. In solchen Fällen bieten sich modellbasierte, parametrische Kopplungsverfahren an, die bei vielen Markern zum Teil sehr rechenaufwendig und gegenüber Sequenzierfehlern empfindlich sein können. Wenn kein klarer Vererbungsmodus ersichtlich ist, zum Beispiel aufgrund reduzierter Penetranz, zu geringer

Familiengröße oder in Fall-Kontroll-Gruppen, können modellfreie, nicht-parametrische Assoziationstests angewandt werden.

Zur statistischen Abschätzung einer Kopplung zwischen Marker und Merkmal wird ein Quotenverhältnis der Hypothesen („logarithm of odds“, „LOD Score“) berechnet, wobei üblicherweise ein LOD<-2 eine Kopplung ausschließt und ein LOD>3 als signifikante Assoziation gewertet wird. Oftmals gelingt es durch Kopplungsanalysen den Suchraum im Genom auf einige Megabasen einzuschränken und damit die Anzahl der Kandidaten-Gene zu reduzieren.

Bei den verbleibenden VUCS kann man die Auswirkung auf funktioneller Ebene abschätzen, indem man die evolutionäre Konservierung einer Basen- oder Aminosäuresequenz berücksichtigt und den Effekt einer Änderung auf die Proteinstruktur betrachtet (Cooper and Shendure, 2011). In den letzten Jahren wurden hierfür in der Literatur eine Vielzahl an Vorhersageprogrammen vorgestellt, die bei der Klassifikation der Pathogenität von VCUS eine Sensitivität und Spezifität von bis zu 80% erreichen. Die am weitesten verbreiteten Werkzeuge hierfür sind MutationTaster (Schwarz, et al., 2010) und PolyPhen (Adzhubei, et al., 2010).

Zwei besonders elegante, kürzlich vorgestellte Methoden der Gen-Priorisierung, PHIVE und PhenIX, beziehen in die Einstufung einer VUCS auch die phänotypische Information mit ein, die ein Genetiker bei der Patientenbeschreibung erhoben hat (Robinson, et al., 2014; Zemojtel, et al., 2014). Die Herausforderung besteht darin, die klinischen Auffälligkeiten eines Patienten so zu kodieren, dass Ähnlichkeitsvergleiche mit in der Literatur beschriebenen, phänotypischen Merkmalen möglich werden, auch wenn nicht exakt dieselben Termini in der Beschreibung verwendet wurden. Die Grundlage für diese semantischen Ähnlichkeitsvergleiche bildet die Human Phenotype Ontologie, deren Vokabular über 8000 „phenotypic features“ umfasst und diese Begriffe zueinander in Beziehung setzt (Robinson, et al., 2008). Mit dieser vom Computer interpretierbaren Datenstruktur sind phänotypische Vergleiche mit bekannten Mendelschen Erkrankungen möglich, deren klinische Merkmale in der Enzyklopädie für genetische Erkrankungen, OMIM, erfasst sind (PhenIX). Durch eine gegenseitige Abbildung der Ontologien unterschiedlicher Modellorganismen ist auch eine speziesübergreifende Gegenüberstellung der Phänotypen möglich. Bei den Ähnlichkeitsvergleichen mit PHIVE fließen Informationen über Phänotypen mit ein, die in großen Gen-knockout-Studien in Mäusen und Zebrafischen gewonnen werden konnten.

Die in den Kapiteln 2.1. bis 2.6. zusammengefassten Originalarbeiten beschreiben bioinformatische Verfahren, die im Rahmen der Habilitationsarbeit entwickelt wurden und die bei der Suche nach krankheitsverursachenden Mutationen in vielen Mendelschen Erkrankungen zum Einsatz kamen. Mittels HDS gelang es uns insbesondere bei GPI-Ankerstörungen zahlreiche pathogene Allele in neuen Genen zu identifizieren. Diese Arbeiten sind in den Kapiteln 2.7. bis 2.11. zusammengefasst. Eine Einführung in die Biologie der GPI-Anker und die Pathophysiologie von Synthese- und Reifungsstörungen wird in den Kapiteln 1.2. und 1.3. gegeben.

1.3 Synthese des GPI-Ankers und Pathophysiologie bei GPI-Ankerstörungen

Bei sämtlichen Eukaryoten findet sich in der Plasmamembran ein Proteinkomplex, dessen zentrale Aufgabe es ist, Glycoproteine an der Zelloberfläche zu verankern: Die Glycosylphosphatidylinositol-Anker (GPI-Anker). GPI-verankerte Proteine, GPI-APs, spielen eine bedeutende Rolle bei der Signal-Transduktion, der Zell-Adhäsion und der Antigen-Präsentation (Ikezawa, 2002; Orlean and Menon, 2007).

Ein GPI-AP, dem bei der Charakterisierung einer Subgruppe von GPI-Ankerstörungen eine besondere Bedeutung zukommt, ist die alkalische Phosphatase. Eine vermehrte Sekretion dieses Enzyms führt im Serum von Patienten zu einer erhöhten alkalischen Phosphatase-Aktivität, die sich leicht laborchemisch bestimmen lässt und im klinischen Kontext als Hyperphosphatasie bezeichnet wird.

Die Synthese des GPI-Ankers beginnt im Endoplasmatischen Retikulum, ER, ausgehend von Phosphatidylinositol, PI (Abbildung 2). In mehreren aufeinanderfolgenden, enzymatischen Schritten werden N-Acetylgucosamin, GlcNAc, sowie drei bis vier Mannose- und zwei bis drei Ethanolamin-Moleküle angehängt (Kinoshita, et al., 2008).

Proteine, die GPI-verankert werden, haben ein spezielles Signalpeptid am C-terminalen Ende, das durch die GPI-Transamidase erkannt, abgespalten und durch den zuvor synthetisierten GPI-Anker ersetzt werden kann. Unter physiologischen Bedingungen wird das GPI-AP an das Ethanolamin an der dritten Mannose gebunden.

Die GPI-Transamidase kann jedoch auch durch noch unvollständige Vorstufen des GPI-Ankers aktiviert werden und spaltet dann das Signalpetid eines Proteins ab, ohne dass eine Bindung an den GPI-Anker erfolgen kann. Für eine Aktivierung der GPI-Transamidase scheint mindestens der erste Mannose-Rest am GPI-Anker erforderlich zu sein (Murakami, et al., 2012). Defekte in Genen, die den Transfer der weiteren zwei Mannose-Moleküle und des Ethanolamins am dritten Mannoserest beeinträchtigen, wie *PIGV*, *PIGO*, *PIGB* und *PIGF*, führen dann nach Abspaltung des Signalpeptids zu einer Sekretion des Proteins. Bei Patienten, bei denen pathogene Mutationen in den Genen *PIGV* und *PIGO* identifiziert werden konnten, erklärt dies die beobachtete Hyperphosphatasie.

Wenn die Synthese des GPI-Ankers gar nicht erst bis zu dieser Stufe fortschreiten kann, wird das Signalpeptid nicht abgespalten, die Proteine verbleiben im ER und werden letztlich wieder abgebaut. Demzufolge führen funktionseinschränkende Mutationen in Genen der frühen Ankersynthese, wie *PIGA, PIGQ* und *PIGL* bei Patienten auch nicht zu einer erhöhten Serumaktivität der alkalischen Phosphatase.

Auf ähnliche Weise wirken sich vermutlich auch pathogene Mutationen aus, die Gene betreffen, die am Aufbau der GPI-Transamidase selbst beteiligt sind. Bislang gibt es nur Fallberichte von Patienten, bei denen hypomorphe Mutationen in *PIGT* identifiziert werden konnten. Diese Patienten wiesen keine veränderten Alkalische Phosphatase-Werte auf.

Wenn ein Protein GPI-verankert ist, so erfolgen weitere Reifungsschritte im ER und Golgi-Apparat an den Fettsäureresten des Inositols (Fujita and Kinoshita, 2012). Bislang kennt man drei Gene, *PGAP1*, *PGAP2* und *PGAP3*, die daran beteiligt sind, eine Acylkette zu entfernen und einen ungesättigten Fettsäurerest durch einen gesättigten zu ersetzen. Diese Modifikationen ermöglichen eine Assoziation der GPI-APs mit Lipidflößen (lipid rafts). Wenn eine Organisation der GPI-APs in lipid rafts unterbleibt, wie dies zum Beispiel bei PGAP2 und PGAP3 Defekten beobachtet werden kann, so sind die GPI-APs auf der Zellmembran für Phospholipasen zugänglicher und es kommt zu vermehrter Sekretion. In Übereinstimmung mit den zellbasierten experimentellen Daten zeigen Patienten mit

pathogenen Mutationen in *PGAP2* und *PGAP3* eine sekundäre GPI-AP Defizienz auf ihren Zelloberflächen und eine Hyperphosphatasie.

Ein Defekt in PGAP1 hingegen und ein damit verbundener zusätzlicher Fettsäurerest am Inositol stellt einen Schutz vor Phospholipase C dar. Die Pathophysiologie dieser GPI-Ankerstörung könnte darin bestehen, dass die biologisch bei manchen Prozessen erwünschte Abspaltung eines GPI-APs unterbleibt.

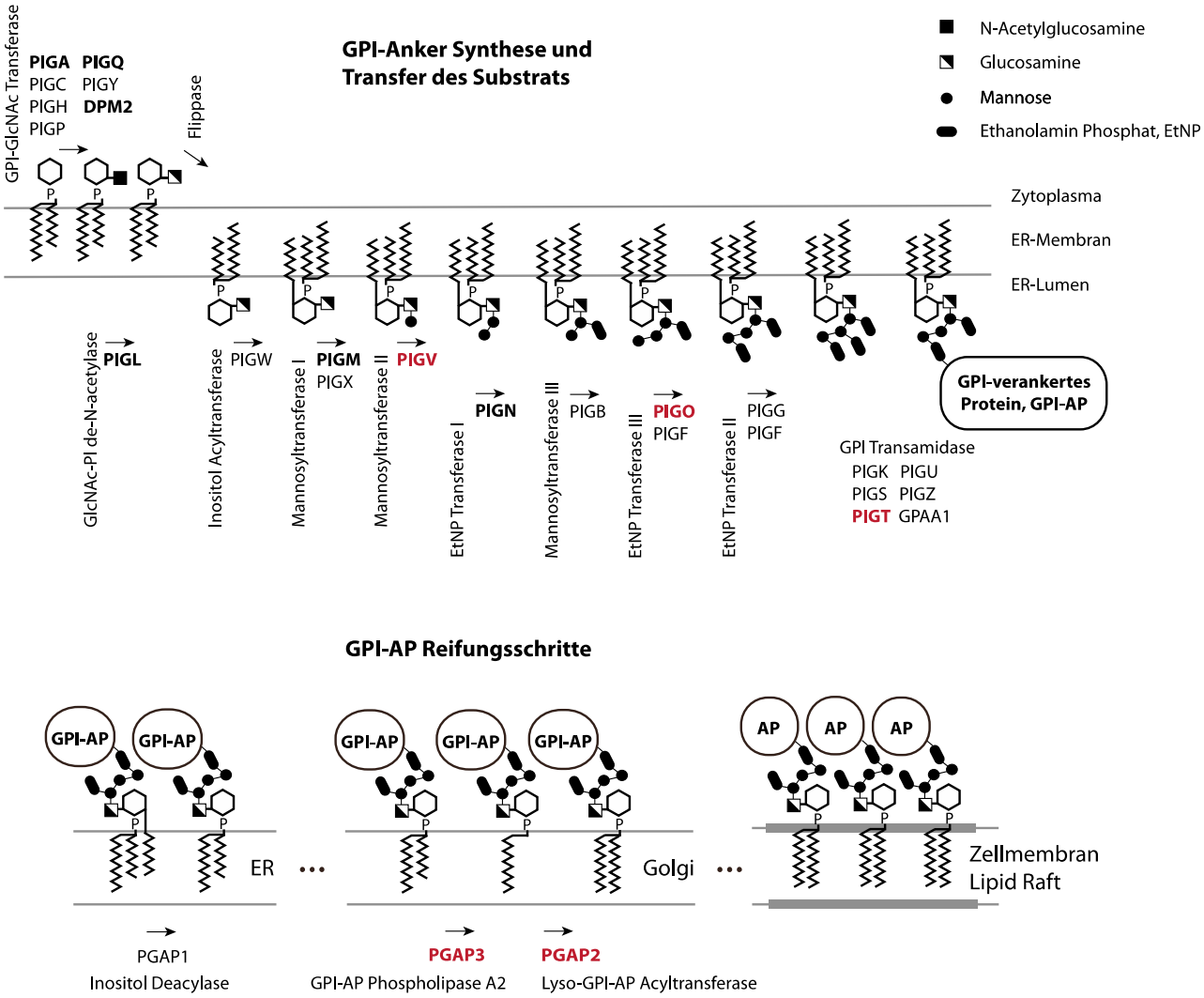


Abbildung 2: Synthese und Reifung von GPI-verankerten Proteinen. Allen GPI-APs ist die Kernstruktur Protein-EtNP-Man-3-Man2-Man-1-GlcN-PI gemeinsam, an deren Synthese 27 Gene beteiligt sind. Als Varianten können ebenfalls GPI-APs auftreten, die eine weitere Ethanolamin-Seitenkette am zweiten Mannoserest aufweisen oder einen zusätzlichen vierten Mannoserest. Bislang wurden pathogene Mutationen in insgesamt 11 dieser Gene in Patienten mit erworbenen sowie angeborenen syndromologischen Krankheitsbildern nachgewiesen (in Fettdruck dargestellt; die im Rahmen meiner Arbeit erstmals beschriebenen sind rot hervorgehoben). Der GPI-Ankerstoffwechsel kann funktionell untergliedert werden in frühe GPI-Ankersynthese (*PIGA*, *PIGC*, *PIGH*, *PIGP*, *PIGQ*, *PIGY*, *DPM2, PIGL, PIGW, PIGM, PIGX*), späte GPI-Ankersynthese (*PIGV, PIGN, PIGB, PIGO, PIGF, PIGG*), GPI-Transamidase (*PIGK, PIGS, PIGT, PIGU, PIGZ, GPAA1*) und GPI-Ankerreifung (*PGAP1, PGAP2, PGAP3*). Defekte in der frühen Ankersynthese und Transamidase führen vermutlich zu einem Abbau der GPI-AP Proproteine

während Defekte in der späten GPI-Ankersynthese und GPI–Ankerreifung zu einer vermehrten Sekretion der Substrate führen.

1.4 Klassifikation der GPI-Ankerstörungen

GPI-Ankerstörungen können hinsichtlich ihres Entstehungszeitpunktes in angeborene und erworbene Störungen unterteilt werden. Des Weiteren ist eine funktionelle Unterteilung möglich, die das betroffene Gen im molekularen pathway berücksichtigt und damit Aussagen über die zu erwartenden Auswirkungen auf GPI-verankerte Proteine, GPI-APs, erlaubt.

PNH1 (MIM #300818) und PNH2 (MIM #615399): Die einzige derzeit bekannte erworbene GPI-Ankerstörung ist die nächtliche paroxysmale Hämoglobinurie, PNH, deren erste Fallbeschreibung auf Strübing im Jahre 1882 zurückgeht (Strübing, 1882). Die PNH ist klinisch gekennzeichnet durch eine chronische Hämolyse, eine verstärkte Thromboseneigung und häufig abdominelle Schmerzen sowie Abgeschlagenheit.

Mahoney und Kollegen wiesen im Jahr 1992 nach, dass der Erkrankung ein Defekt in der GPI-Ankersynthese in Zellen des blutbildenden Systems zugrunde liegt (Mahoney, et al., 1992). Dies führt zu einem Fehlen der GPI-verankerten Proteine CD55 und CD59, die eine wichtige Rolle in der Regulation des Komplementsystems spielen. Erythrozyten, die kein CD55 und CD59 auf der Zelloberfläche exprimieren, werden vermehrt durch den aktivierten Membranangriffskomplex, MAC, zerstört. In der Therapie der komplementvermittelten Symptome wie Hämolyse wird ein humanisierter monoklonaler IgG Antikörper (Eculizumab) eingesetzt, der an das Protein C5 des Komplementsystems bindet und somit die terminale Aktivierung inhibiert.

Die häufigste genetische Grundlage der PNH stellen somatische Mutationen des X-chromosomalen Gens *PIGA* dar, dessen Genprodukt eine Untereinheit der N-Acetylglucosamin-Transferase bildet, die am ersten Schritt der GPI-Ankersynthese beteiligt ist (Takeda, et al., 1993). Eine X-chromosomale somatische Mutation manifestiert sich im männlichen Geschlecht. Aufgrund der zufälligen Inaktivierung eines X-Chromosoms in weiblichen Zellen genügt ebenfalls ein erworbener Defekt in *PIGA*, der das aktive X-Chromosom betrifft, um die GPI-Ankersynthese auch in weiblichen Individuen zu stören. Alle anderen Gene der GPI-Ankersynthese sind autosomal und ein Funktionsverlust setzt Mutationen in beiden Genkopien voraus. Wir identifizierten erstmals zwei Fälle mit PNH, in denen es infolge einer Keimbahn-Mutation und einer somatischen Deletion zu einem Funktionsverlust des autosomalen Gens *PIGT* kam, das eine Untereinheit der Transamidase darstellt.

Alle bislang identifizierten Mutationen des GPI-pathway bei PNH Fällen stellen Nullmutationen dar.

PIGM (MIM #610293): In drei Patienten aus zwei nicht verwandten Familien türkischen und arabischen Ursprungs, die Portalvenenthrombosen und Epilepsien aufwiesen, konnte eine homozygote, hypomorphe Mutation im Promoter des Gens *PIGM* identifiziert werden (Almeida, et al., 2006). Diese Mutation interferiert mit einer Bindestelle für den Transkriptionsfaktor SP1. Die Bedeutung dieser Bindestelle für die transkriptionelle Regulation von *PIGM* scheint zelltypspezifisch zu sein, so dass aktuell noch nicht klar ist, welche phänotypischen Auffälligkeiten bei funktionellen Störungen von PIGM zu erwarten sind, die alle Zellen gleichermaßen betreffen.

HPMRS1 (MIM #239300), HPMRS2 (MIM #614749), HPMRS3 (MIM #614207), HPMRS4 (MIM #615716): Das Mabry-Syndrom oder auch Hyperphosphatasie Syndrom mit mentaler Retardierung ist neben der namensgebenden geistigen Entwicklungsverzögerung und der erhöhten Serumaktivität der alkalischen Phosphatase noch durch folgende phänotypische Auffälligkeiten gekennzeichnet: Hypertelorismus, lange Lidspalten, breite Nasenwurzel und Nasenspitze, schmale Oberlippe, verkürzte Endphalangen und Nagelhypoplasie. Die meisten Patienten weisen zudem eine muskuläre

Hypotonie sowie Epilepsien auf. Variabler ausgeprägt sind weitere Organfehlbildungen wie Hirschsprungkrankheit, vesicouretrale, renale und anorektale Malformationen (Horn, et al., 2011; Horn, et al., 2010; Horn, et al., 2014; Mabry, et al., 1970).

Die molekulare Genese des Mabry Syndroms ist heterogen. Bislang konnten pathogene Mutationen in zwei Genen der späten Ankersynthese, *PIGV* und *PIGO*, sowie pathogene Mutationen in zwei Genen der Ankerreifung, *PGAP2* und *PGAP3*, nachgewiesen werden. Die Mutationen wirken sich funktionseinschränkend auf GPI-APs aus. Sowohl die Oberflächenexpression als auch die Stabilität des GPI-Ankers ist reduziert, so dass GPI-APs, wie zum Beispiel auch die alkalische Phosphatase, vermehrt abgespalten und in den Extrazellularraum abgegeben werden.

(in der Diskussion auf Genotyp-Phänotyp Korrelationen eingehen: In den wenigen bislang beschriebenen compound heterozygoten Fällen, in denen eine der Sequenzvarianten eine Nullmutation darstellt, ist der Verlauf gravierender.

MCAHS1 (MIM #614080), MCAHS2 (MIM #300868), MCHAS3 (MIM #615398), EIEE (MIM #300382): Bei einer Reihe von Patienten mit schweren psychomotorischen Entwicklungsverzögerungen und multiplen angeborenen Fehlbildungen, muskulärer Hypotonie und Epilepsien (multiple congenital anomalies hypotonia, and seizures, MCAHS) konnten pathogene Mutationen in weiteren Genen der GPI-Ankersynthese identifiziert werden (Chiyonobu, et al., 2014; Johnston, et al., 2012; Kato, et al., 2014; Kvarnung, et al., 2013; Martin, et al., 2014; Maydan, et al., 2011; Nakashima, et al., 2014; Ohba, et al., 2014). Die Mutationen betreffen Gene deren Produkte an der frühen und späten Ankersynthese beteiligt sind (*PIGA, PIGQ, PIGW*, *PIGN*) oder eine Untereinheit der Transamidase bilden (*PIGT*). Einige der Patienten mit pathogenen Mutationen in *PIGA, PIGQ* und *PIGW* entstammen Fallgruppen mit frühkindlicher epileptischer Enzephalophatie („early infantile epileptic encephalopahty“, EIEE, West Syndrom), so dass zu dieser Krankheitsgruppe phänotypische Überschneidungen bestehen. Aufgrund der zu geringen Fallzahlen ist aktuell noch nicht klar, ob GPI-Ankerstörungen, die durch Defekte in frühen Syntheseschritten hervorgerufen werden, grundsätzlich schwerer verlaufen als diejenigen, bei denen die späten Prozesse betroffen sind.

CHIME (MIM #280000): Das Zunich neuroektodermale Syndrom ist gekennzeichnet durch Colobome, Herzfehlbildungen, Ichthyosis, mentale Retardierung, Hörstörungen und Epilepsien (CHIME) (Zunich and Kaye, 1983). Alle bislang beschriebenen Fälle mit Zunich Syndrom weisen einen europäischen Bevölkerungshintergrund auf und zeigen homozygote oder compound heterozygote pathogene Mutationen im *PIGL*-Gen. Die missense Mutation, p.Leu167Pro, ließ sich bei allen Patienten nachweisen, so dass es sich offensichtlich um eine Gründermutation handelt. Im Vergleich zu anderen GPI-Ankerstörungen ist die Reduktion von GPI-APs auf der Zelloberfläche gering.

MRT42 (MIM #615802): Murakami und Kollegen beschrieben erstmals zwei Fälle mit einer nicht-syndromalen geistigen Entwicklungsverzögerung, die auf eine Störung der GPI-Ankerreifung zurückzuführen ist (Murakami, et al., 2014). Bei zwei Betroffenen einer syrischen, konsanguinen Familie konnten Nullmutationen im Gen *PGAP1* identifiziert werden. Durch diesen Gendefekt unterbleibt die Abspaltung einer Acylkette vom GPI-AP durch die Phospholipase C. Dies verdeutlicht, dass nicht nur die Präsentation GPI-verankerter Proteine auf der Zelloberfläche sondern auch die regelrechte Reifung des Ankers für eine normale neurologische Entwicklung von Bedeutung ist.

1.5 Überlegungen zur Inzidenz bei GPI-Ankerstörungen

Unter den angeborenen GPI-Ankerstörungen stellt das Mabry-Syndrom mit der Kombination von geistiger Entwicklungsverzögerung, charakteristischem fazialen Aspekt und insbesondere der Hyperphosphatasie eine diagnostisch klar abgrenzbare Entität dar. Bislang konnten wir und Kollegen in über 20 Patienten mit Verdacht auf Mabry Syndrom 27 unterschiedliche, pathogene Sequenzvarianten in den GPI-Ankersynthese-Genen, *PIGV*, *PIGO*, *PGAP2* und *PGAP3* identifizieren (Tabelle 1). Damit stehen für die Subgruppen HPMRS1-4 der GPI-Ankerstörungen die bislang umfangreichsten Mutationsdaten zur Verfügung.

Zur Abschätzung der Inzidenz dieser Erkrankung können die populationsgenetischen Daten, die in einer 6500 Individuen umfassenden Exom-Kohorte erhoben wurden, herangezogen werden (Tennessen, et al., 2012). Alle in dieser Studie untersuchten Individuen sind nicht verwandt, nicht von einer geistigen Behinderung betroffen und keines der in Tabelle 1 aufgeführten pathogenen Allele wurde in dieser Kohorte in homozygotem Zustand beobachtet.

Wenn wir die Heterozygoten-Frequenzen der nachgewiesen pathogenen Allele in dieser Kohorte betrachten und ein Hardy-Weinberg-Gleichgewicht annehmen, so ergibt sich als Näherungswert für die Inzidenz von Mabry-Syndrom:

$$Inzidenz = \left(\frac{7}{6500}\right)^2 + \left(\frac{3}{6500}\right)^2 + \left(\frac{1}{6500}\right)^2 + \left(\frac{1}{6500}\right)^2 = 1.4 \text{ auf } 1.000.000$$

Wenn man nun auch die Heterozygoten-Frequenzen von nur funktionseinschränkend vorhergesagten Allelen in anderen Genen der GPI-Ankersynthese heranzieht, ergeben sich ebenso für diese GPI-Ankerstörungen Inzidenzen, die zum Teil deutlich über eins zu einer Million liegen. Auch wenn einige der Einstufungen fehlerhaft sein mögen, so scheint doch die Dunkelziffer der Erkrankungen, die noch nicht als GPI-Ankerstörung erkannt wurden, beträchtlich zu sein.

Aufgrund einer Vielzahl, zum Teil zellspezifisch exprimierter GPI-APs ist es auch denkbar, dass es neben der PNH, die auf somatische Mutationen in myeloiden Stammzellen zurückzuführen ist, noch weitere, erworbene GPI-Ankerstörungen gibt, die andere Organsysteme betreffen.

Die in den folgenden Kapiteln aufgeführten Arbeiten zeigen, wie die Hochdurchsatz-Sequenzierung zur Aufklärung der genetischen Ursachen von GPI-Ankerstörungen beitragen kann.

Tabelle 1: Pathogene Mutationen für Mabry-Syndrom. Bislang wurden krankheitsverursachende Mutationen in zwei Genen der späten GPI-Ankersynthese, *PIGV* und *PIGO*, sowie in zwei Genen der GPI-Ankerreifung, *PGAP2* und *PGAP3*, beschrieben. Anhand populationsgenetischer Daten, lässt sich die Allel-Frequenz dieser pathogenen Mutationen abschätzen (Tennessen, et al., 2012).

PIGV (NM_017837.3)		Referenz	HPMRS Kohorte	Allel Frequenz in 6500 Exomen
c.53G>A	p.Cys18Tyr	(Horn, et al., 2014)	1	0
c.176T>G	p.Leu59Arg	(Horn, et al., 2014)	1	2/13000
c.467G>A	p.Cys156Tyr	(Horn, et al., 2011)	1	0
c.494C>A	p.Ala165Glu	noch nicht publiziert	1	1/13000
c.607C>T	p.Arg203Glu	noch nicht publiziert	1	1/13000
c.766C>A	p.Gln256Lys	(Krawitz, et al., 2010)	2	0
c.905T>C	p.Leu302Pro	(Horn, et al., 2014)	2	0
c.1022C>A	p.Ala341Glu	(Krawitz, et al., 2010)	17	3/13000
c.1022C>T	p.Ala341Val	(Krawitz, et al., 2010)	1	0
c.1154A>C	p.His385Pro	(Krawitz, et al., 2010)	1	0
c.1405C>T	p.Arg469*	(Horn, et al., 2014)	2	0
PIGO (NM_032634.3)				
c.389C>A	p.Thr130Asn	(Nakamura, et al., 2014)	1	0
c.1288C>T	p.Gln430*	(Nakamura, et al., 2014)	1	1/13000
c.1318A>G	p.Ile440Val	noch nicht publiziert	2	0
c.2361dupC	p.Thr788Hisfs*5	(Krawitz, et al., 2012)	1	1/13000
c.2869C>A	p.Leu957Phe	(Krawitz, et al., 2012)	2	0
c.3069+5G>A	p.Val952Aspfs*24	(Krawitz, et al., 2012)	1	1/13000
PGAP2 (NM_001256235.1)				
c.46C>T	p.Arg16Trp	(Krawitz, et al., 2013)	1	0
c.97G>A	p.Ala33Thr	noch nicht publiziert	2	0
c.296A>G	p.Tyr99Cys	(Hansen, et al., 2013)	2	0
c.380T>C	p.Leu127Ser	(Krawitz, et al., 2013)	2	1/13000
c.479C>T	p.Thr160Ile	(Krawitz, et al., 2013)	1	0
c.530G>C	p.Arg177Pro	(Hansen, et al., 2013)	2	0
PGAP3	(NM_033419.3)			
c.275G>A	p.(Gly92Asp)	(Howard, et al., 2014)	2	0
c.314C>G	p.(Pro105Arg)	(Howard, et al., 2014)	2	1/13000
c.439dupC	p.(Leu147Profs*16)	(Howard, et al., 2014)	1	0
c.914A>G	p.(Asp305Gly)	(Howard, et al., 2014)	1	0

2 Originalarbeiten

2.1 Nachweis kleiner Insertionen und Deletionen in kurzen Sequenzfragmenten

Krawitz, P., Rodelsperger, C., Jager, M., Jostins, L., Bauer, S., and Robinson, P.N. (2010). Microindel detection in short-read sequence data. Bioinformatics 26, 722-729.

Die ersten Geräte für Hochdurchsatzsequenzierung, wie zum Beispiel der Genome Analyzer der Firma Illumina, erzeugten Sequenzdaten einer Leselänge von 36, 76 oder 100 Basenpaaren (bp). Die bioinformatische Prozessierung kann in drei Abschnitte untergliedert werden:

1) Base Calling. Hierbei wird aus den während des Sequenzierprozesses gewonnenen Fluoreszenzsignalen eine Basensequenz abgeleitet.
2) Reference-guided realignment. Hierbei werden die Sequenzfragmente an einer Referenzsequenz des humanen Genoms so ausgerichtet, dass im Vergleich zu einer andersartigen Ausrichtung an der Referenz eine höhere Abweichung entstünde. Der Mapping Score gibt dann Auskunft darüber, wie wahrscheinlich die gemessene Sequenz tatsächlich der zugewiesenen Position im Genom entspricht.
3) Variant Calling. Hierbei werden die zugeordneten Sequenzen (das Sequenzalignment) positionsweise nach Abweichungen von der Referenzsequenz untersucht. Hierzu wird ein Variant Calling Algorithmus eingesetzt, dem ein Wahrscheinlichkeitsmodell zugrunde liegt. Es kann so abgeschätzt werden kann, ob eine Abweichung von der Referenzsequenz heterozygot oder homozygot vorliegt.

Der Smith-Watherman-Algorithmus, der eine optimale Ausrichtung eines Sequenzfragments an einer Referenz erlaubt, ist so rechenaufwendig, dass er bei der Datenmenge einer Hochdurchsatzsequenzierung und den aktuell zur Verfügung stehenden Rechenleistungen nicht anwendbar ist. Die bioinformatische Herausforderung beim Sequenzrealignment besteht daher in der Entwicklung effizienter Heuristiken, mit denen es gelingt, einer optimalen Ausrichtung eines Sequenzfragmentes an einer Referenz möglichst nahe kommen, jedoch mit weniger Rechenschritten.

Die ersten „short-read-mapper“ für NGS Daten waren nur in der Lage, ein Sequenzfragment lückenlos in einem Stück an der Referenz zu positionieren (ungapped mapper). Dies ermöglichte es, Einzelbasensubstitutionen, SNVs, zu detektieren. Erst mit den weiterentwickelten Algorithmen, die ein „gapped alignment“ erlauben, ist es möglich, auch kleine Insertionen und Deletionen, die bis zu 10 Basenpaare umfassen, in einem Sequenzalignment nachzuweisen.

Als diese neuartigen „gapped-short-read-mappers“, wie BWA, Novoalign und RazerS zur Verfügung standen, testeten wir anhand simulierter Sequenzfragmente die Genauigkeit der Detektion von kleinen Insertionen und Deletionen, sog. Microindels. Wir schlugen zudem eine Nomenklatur vor, die eine eineindeutige Beschreibung eines indels gestattete und beschrieben einen Algorithmus, der eine Überführung unterschiedlicher Notationsweisen erlaubte. Die Anwendung dieses Algorithmus ermöglichte es, die Detektionsrate der getesteten Algorithmen zu erhöhen und diese miteinander zu vergleichen.

Mit der von uns entwickelten Detektionspipeline für Microindels reanalysierten wir die im Rahmen der ersten Phase des 1000 Genome Projektes (1 KGP) erhobenen Sequenzrohdaten und

konnten erstmals die Häufigkeit von Microindels in Genomen abschätzen: Das Verhältnis von Indels zu SNVs betrug in unserer Datenanalyse 7:1. Drei Jahre später, am Ende der zweiten Phase des 1 KGP, standen noch umfangreichere Datensätze zur Verfügung. Die mittlere Anzahl autosomaler SNVs eines Individuums ergab 3,6 Millionen im Vergleich zu 344 Tausend Indels, was einem Verhältnis von ca. 10:1 entspricht (Genomes Project, et al., 2012).

BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 6 2010, pages 722–729
doi:10.1093/bioinformatics/btq027

Sequence analysis

Advance Access publication February 9, 2010

Microindel detection in short-read sequence data

Peter Krawitz[1,2,3,†], Christian Rödelsperger[1,2,3,†], Marten Jäger[1,3], Luke Jostins[4], Sebastian Bauer[1,3] and Peter N. Robinson[1,2,3,*]

[1]Institute for Medical Genetics, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, [2]Berlin-Brandenburg Center for Regenerative Therapies, Augustenburger Platz 1, 13353 Berlin, [3]Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin and [4]Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Several recent studies have demonstrated the effectiveness of resequencing and single nucleotide variant (SNV) detection by deep short-read sequencing platforms. While several reliable algorithms are available for automated SNV detection, the automated detection of microindels in deep short-read data presents a new bioinformatics challenge.

Results: We systematically analyzed how the short-read mapping tools MAQ, Bowtie, Burrows-Wheeler alignment tool (BWA), Novoalign and RazerS perform on simulated datasets that contain indels and evaluated how indels affect error rates in SNV detection. We implemented a simple algorithm to compute the equivalent indel region *eir*, which can be used to process the alignments produced by the mapping tools in order to perform indel calling. Using simulated data that contains indels, we demonstrate that indel detection works well on short-read data: the detection rate for microindels (<4 bp) is >90%. Our study provides insights into systematic errors in SNV detection that is based on ungapped short sequence read alignments. Gapped alignments of short sequence reads can be used to reduce this error and to detect microindels in simulated short-read data. A comparison with microindels automatically identified on the ABI Sanger and Roche 454 platform indicates that microindel detection from short sequence reads identifies both overlapping and distinct indels.

Contact: peter.krawitz@googlemail.com; peter.robinson@charite.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 22, 2009; revised on January 14, 2010; accepted on January 16, 2010

1 INTRODUCTION

Microinsertions and microdeletions ('indels') constitute a class of genetic mutations that play an important role in human genetic disease (Ball *et al.*, 2005). The reliable detection of microinsertions and microdeletions is thus a prerequisite for current efforts to assess the medical relevance of genetic variation including small indels across the human genome. Structural variations on the order of kilobases, whose prevalence has long been underestimated because of the lack of appropriate methods of detection, have been recently shown to be responsible for more polymorphism than single nucleotide variants (SNVs) as measured by nucleotide content per genome (Korbel *et al.*, 2007; Redon *et al.*, 2006). Therefore, we hypothesized that also on the scale of only a few nucleotides, the frequency of microindels might have been underestimated. Harismendy *et al.* (2009) performed an analysis on sequences amplified by long-range PCR to compare three next-generation sequencing (NGS) platforms (Mardis, 2008), Illumina GA, Roche 454 FLX and ABI SOLiD, to the *de facto* gold standard of ABI Sanger sequencing. All three NGS platforms showed high sensitivity (>95%) in variant calling for sequence sites covered to saturation. However, for microindel detection, they only compared the results of the automated microindel detection pipeline of the Roche 454 platform to ABI Sanger sequencing. Currently only few software solutions for microindel detection in short read sequence data are available and they do not yet meet the need for unambiguous microindel positioning. In addition, the evaluation of automated microindel detection on NGS data remains difficult, as a gold standard is lacking—on the ABI Sanger platform the automated detection of heterozygous microindels remains highly error prone (Bhangale *et al.*, 2005). In the targeted sequence analyzed in Harismendy *et al.* (2009), 11 indels were identified by ABI Sanger, whereas 43 additional indels that were not found by ABI Sanger were called by Roche 454. On the other hand, there were five single-base indels in homopolymers that were called by ABI Sanger but not by Roche 454. This illustrates that the existing approaches to automated detection of microindels remain a technological challenge with presumably high false positive and negative rates.

Because of the high error rates of automated indel detection with the ABI Sanger platform and the Roche 454 technology, we were motivated to study the potential to identify microindels in short-read sequence data produced on NGS platforms. We introduce a simple indel calling algorithm that is based on the efficient mapping of short reads and which takes into account the fact that short sequence reads containing indels may often not be unequivocally aligned to the reference genome due to the surrounding sequence. Our microindel calling algorithm makes use of gapped alignments produced by efficient short-read mapping tools, such as Burrows-Wheeler alignment tool (BWA), Novoalign or RazerS, in order to call SNVs and indels. As the true distributions and frequencies of microindels in genomic sequences remain unknown due to technological shortcomings, we use a simulation approach to perform an analysis for varying microindel sizes and frequencies and study the effect on SNV

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

as well as microindel detection. Finally, we apply our microindel calling approach for short-read data to the aforementioned cross-platform-validated datasets (Harismendy *et al.*, 2009) and show that microindel detection on Illumina GA short-read sequences is feasible. Although our approach is applicable to any sort of short-read data, we focus in the following analysis on short-read data format of the Illumina GA platform, as the greatest variety of efficient mapping tools is available for this technology.

2 METHODS

2.1 Simulating short-read data

The 296 kb reference sequence used for generating sequence reads was constructed by extracting the following sequences from the human genome using the UCSC Genome Bioinformatics site (genome.ucsc.edu) build hg18 (NCBI build 36): chr11:73836950-73862566, chr21:34656259-34672486, chr21:34734911-34819450, chr2:223615214-223628218, chr3:38553978-38665289 and chr7:150268610-150312098. Serial repeats in the sequence fragments were detected with mreps 2.5 (Kolpakov *et al.*, 2003). Altogether 296 repeats of length $\geq$ period $+9$ were reported for the targeted sequence; this is comparable with the repeat frequency of randomly chosen 300 kb sequence fragments of human DNA (e.g. there are $\sim$260 000 repeats in 330 Mb of chr. 1). Deletions and insertions were simulated with frequencies ranging from 0.1 microindels/kb to 10 microindels/kb. The definition of microindels with respect to size varies in the literature and has often been defined in accordance with the detection limits of the technologies used in a study. We studied microindels up to a length of 5 bp for 36 bp reads and up to a length of 10 bp for 76 bp reads. Single nucleotide polymorphisms (SNPs) were simulated with a fixed frequency of 1 SNP/kb in all datasets. Microindels and SNPs were simulated as homozygous or heterozygous variants with a rate of 0.5.

In the simulated sequence, the positions of microindels were randomly chosen with the constraint that neighboring deletions may not overlap. This means that in simulated dataset with microindels of size k bp the positions of two different deletions must at least lie $k+1$ nt apart.

The datasets studied in Harismendy *et al.* (2009) show a distribution of the sequencing depth per chromosomal position that is platform specific. The datasets produced on the Genome Analyzer had a mean sequencing depth of 180-fold. We therefore simulated reads such that the read depth at each chromosomal position follows a Poisson distribution with a mean of 180. For the mapping statistics, the original position of every read with respect to the reference sequence was added to the read identifier. The quality scores for 36 bp reads were randomly picked from the corresponding experimental Illumina GA2 data (Harismendy *et al.*, 2009). Analogous quality scores of an unrelated Illumina GA2 run were used for the simulated 76 bp reads. Nucleotides were then switched to variant bases with probabilities according to their quality scores. For every combination of microindel length and frequency 10 runs were simulated. We compared short-read datasets consisting of 150 000 36 bp reads to 450 000 36 bp reads and 70 000 76 bp reads, corresponding to a mean sequencing depth of 18, 54 and 18.

2.2 Aligning short sequence reads

Short sequence read data was downloaded from ftp://ftp.jcvi.org/pub/data/NGS_cross_validation/ or simulated as described above. Short reads were mapped to the reference genome using BWA 0.4.9 (Li and Durbin, 2010), Novoalign Release 2.05.02 (Hercus, 2009) and RazerS 1.0 (Weese *et al.*, 2009) with default settings for mismatch penalty, gap opening penalty and gap extension penalty:

```
bwa aln -e 5 -t 8 <ref.fa> <reads.fastq>
novoalign -o SAM -d <ref.ndx> -f <reads.fastq>
razers -id -i 80 -rr 100 <ref.fa> <reads.fa>
```

For variant detection, Bowtie 0.11.3 (Langmead *et al.*, 2009) and MAQ 0.7.1 (Li *et al.*, 2008) were also tested with their default settings, which do not allow gapped alignments:

```
maq.pl easyrun <ref.fa> <reads.fastq>
bowtie -S -p 8 <ref.fa> <reads.fastq> <out.sam>
```

MAQ uses a spaced-seed approach to align reads. With default setting only reads that map to the reference genome with less than three mismatched bases in the first 28 bases of the read will be aligned. The ungapped alignment with the best alignment score is reported. Bowtie and BWA are based on backward search schemes with a Burrows–Wheeler transformation to efficiently align short sequencing reads against large reference sequences. Bowtie allows two mismatches or fewer within the high-quality end of each read, and it places an upper limit on the sum of the quality values at mismatched alignment positions. Novoalign finds global optimum alignments using full Needleman–Wunsch algorithm with affine gap penalties. RazerS adapts a q-gram counting technique for read filtering and maps reads using edit or Hamming distance as thresholds.

For all alignments the target sequence was used as reference sequence. When instead the whole genome was used as reference sequence, a certain proportion of reads mapped to locations outside the target region. These reads yielded higher alignment scores at wrong positions due to simulated mutations or sequencing errors.

All alignments were converted to Sequence Alignment/Map (sam) format that codes the position of an indel in the short-read in CIGAR string format. The consensus sequence was called according to the MAQ consensus model (Li *et al.*, 2008) with samtools release 0.1.7:

```
samtools pileup -vcf <ref.fa> <aln.bam>
```

The resulting raw consensus sequence was further filtered with:

```
samtools.pl varFilter -D100
```

This step also filtered out SNVs that are in a 10 bp window around a gap. For SNV detection we only considered reads with a read mapping quality of above 20 and for indel detection with a read mapping quality of above 50.

2.3 SNV and microindel calling

A coverage threshold of at least five reads covering a sequence position was used for variant and indel calling. For SNV calling approach, we used a frequency threshold as filter as described in Harismendy *et al.* (2009): a heterozygous SNV was called when 20–80% of the aligned reads showed the variant nucleotide. The position was called as a homozygous SNV if >80% of aligned reads showed the variant nucleotide.

In contrast to SNVs, the position of an indel with respect to the reference sequence is not necessarily unambiguously defined by a single coordinate, as the example in Figure 1 illustrates. The insertion of an adenosine into the local sequence motive of $C_iA_{i+1}A_{i+2}G_{i+3}$ after position i results in a mutated sequence that is identical to the sequence produced by an inserted adenosine after position $i+1$ or $i+2$. We assume both that each of these insertions has the identical biological meaning and that they are furthermore indistinguishable by mutation detection methods, so that in our example, calls of an insertion at position i, $i+1$ and $i+2$ represent one and the same mutation. An unambiguous annotation for this insertion would therefore have to list all equivalent indel positions, i.e. $+A\{i, i+1, i+2\}$. In a random sequence with all nucleotides occurring with same frequencies, the probability that the position of a single inserted nucleotide is unambiguously defined by a single coordinate is only $9/16$. In genomic sequences, where homopolymers and small tandem repeats are more frequent than randomly expected, it is even less likely that an indel can be defined unambiguously by a single position. Therefore, when a read was aligned with a gap, the equivalent indel region, *eir*, was determined by computing all equivalent positions with respect to the sequence of this specific insertion or deletion. The following

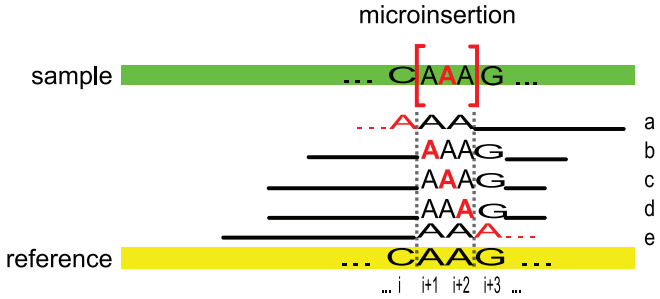


Fig. 1. The sample sequence has an adenosine triplet, compared with the doublet in the reference sequence. A short sequence read can be aligned to the reference with the insertion of an adenosine at any one of three positions (b, c, d). The position of the inserted A is not unambiguously defined by a single coordinate, but only by set of equivalent positions $eir = +A(i, i+1, i+2)$. Depending on the settings of the local alignment algorithm and on the surrounding sequence, a false alignment with mismatched nucleotides may yield a higher alignment score (a, e).

example illustrates how we proceeded with non-homopolymeric indels: if the reference sequence is r = CAGAT, then a called insertion of an AG at position 3 (i.e. following the three nucleotides CAG) leads to the same mutant sequence as a called insertion of GA at position 4: CAGAGAT. Our algorithm therefore identifies all called indel positions that lead to the identical mutated sequence (Fig. 1). To do so, we search for all positions in the reference sequence, where the insertion or deletion of the appropriate sequence pattern will lead to an identical mutated sequence. We refer to the set of all such positions as the *eir*, and consider all reads with called indels in the *eir* as equivalent for the purposes of indel calling. For the above example sequence r = CAGAT, an insertion of AG called at positions 1 and 3, as well as an insertion of GA called at positions 2 and 4, will lead to identical mutated sequences, thus the *eir* is $+AG(1-4)$. The pseudocode for our algorithm is shown in Figure 2. The frequency of an indel was computed by counting all reads that showed the indel sequence in the *eir* and dividing by the total number of reads covering the *eir*. We note that if multiple calls result in two or more distinct overlapping *eir*s, they were treated as separate for the purposes of indel calling. An indel was called if >10% of mapped reads showed the indel sequence in the *eir*.

Algorithm 1: Computation of eir

Input: sequence s, position i_p, pattern p
Output: eir

1 $x \leftarrow p$; // Extend eir to the right
2 $i_r \leftarrow i_p$;
3 $r \leftarrow s_{i_r+1}$;
4 $x' \leftarrow x_2 \ldots x_k.x_1$;
5 **while** $(x.r == r.x')$ **do**
6 $\quad x \leftarrow x'$;
7 $\quad i_r \leftarrow i_r + 1$;
8 $\quad r \leftarrow s_{i_r+1}$;
9 $\quad x' \leftarrow x_2 \ldots x_k.x_1$;
10 $x \leftarrow p$; // Extend eir to the left
11 $i_l \leftarrow i_p$;
12 $l \leftarrow s_{i_l-1}$;
13 $x' \leftarrow x_k.x_1 \ldots x_{k-1}$;
14 **while** $(l.x == x'.l)$ **do**
15 $\quad x \leftarrow x'$;
16 $\quad i_l \leftarrow i_l - 1$;
17 $\quad l \leftarrow s_{i_l-1}$;
18 $\quad x' \leftarrow x_k.x_1 \ldots x_{k-1}$;
19 eir $\leftarrow \{x, i_l, i_r\}$;
20 **return** eir

Fig. 2. An *eir* is computed from the genomic sequence s around an indel of a sequence pattern p after position i_p. i_r denotes the rightmost position of the *eir* and r the nucleotide to the right of i_r. Line 4 computes a cyclic permutation x' of the pattern in x. The '.' operator indicates a string concatenation. Lines 1–9 extend the *eir* to the right. Following the extension to the left (lines 10–18), the left and rightmost positions are returned together with the leftmost pattern.

3 RESULTS

Due to the large amount of short-read data that NGS platforms produce, efficient read mapping tools quickly narrow down the candidate regions where a read possibly maps. In this candidate region, local alignment algorithms are used to minimize the mismatched nucleotides and inserted gaps. An alignment score is finally used to report the best matching alignment. Generally two different terms contribute to the alignment score, a penalty α from mismatched bases and a penalty β in case of gap insertions. The exact values of α and β depend, on the one hand, on global settings of the alignment algorithm that is on the applied similarity matrices and on gap opening and extension penalties. Quality values of the aligned nucleotides as well as their positions in the reads can be taken into account. On the other hand, the alignment score is locally influenced by the surrounding sequence. It is crucial to understand that the optimal alignment score that is reported for a read depends on the algorithmic parameters as well as on the sequence context (Durbin *et al.*, 1999). This explains why one and the same read may be aligned to the very same starting and ending positions by two different mapping tools and yet their alignment shapes and score may differ. A read that covers a certain microindel of a certain size, may also be aligned with a gap in one sequence context, or with mismatches in another, depending on the neighboring nucleotides (Fig. 1). If the position of the inserted or deleted sequence is located at the beginning or end of the short sequence read or if the surrounding nucleotides are similar to the indel sequence an alignment with partially mismatched bases might yield a better alignment score and thus lead to a false alignment with respect to the true sample sequence.

3.1 Mapping efficiency of reads covering indels

The origin of the simulated reads was used to calculate the rates of reads that were mapped to the correct position. A read was counted as correctly mapped, if either its mapped starting or ending position agreed with its original coordinate. Due to the different algorithmic approaches, we expected different mapping efficiencies for the mapping tools we tested. By default, BWA only maps reads that show less than three mismatched bases in the seed to the reference sequence, not counting gaps. This means a read with more than three sequencing errors in the seed sequence cannot be mapped. Novoalign reports the best unique alignment, regardless of the number of mismatched bases. A read that maps to more than one position with the same score is not reported. RazerS maps all reads that can be aligned within a certain editing distance (80% identity).

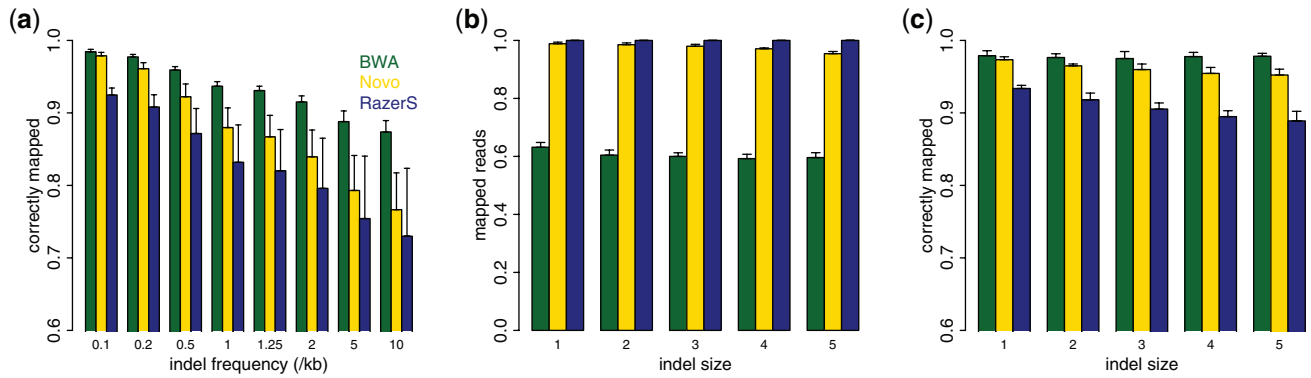


Fig. 3. (**a**) The rate of mapped reads with indels that are correctly mapped decreases with increasing indel frequency. (**b**) In contrast, the rate of all reads with indels that can be mapped shows only a weak dependency on the indel size. (**c**) Also the rate of mapped reads with indels that are *correctly* mapped, shows only a weak dependency on the indel size. As can be seen in (b and c), BWA is able to align a lower proportion of all reads with indels owing to its more stringent criteria, but the accuracy of its alignments is higher for the reads that can be aligned. The height of the barplot indicates the mean proportion of mapped or correctly mapped reads, and error bars indicate 2 SDs. The indel frequency in panels b and c was fixed to 0.2/kb.

Due to the relatively high sequencing error rate in the simulated data, BWA is only able to map ∼80% of reads from regions without indels, compared with >98% by Novoalign and RazerS. Of all mapped reads that did not cover indel sites, >99 % were mapped to the correct coordinates by Novoalign and BWA and >95% by RazerS. In Figure 3a, the rate of mappable reads from regions with at least one indel that were mapped to the correct coordinate is shown. For all mapping tools the rate of correctly mapped reads decreases for an increasing indel frequency. Reads covering more than one indel are poorly mapped by all of the alignment tools tested. For an indel frequency of $f_{\text{indel}} = 0.1$/kb, the probability that two simulated indels have a smaller distance than 36 bp is ∼0.4%. This probability increases to >30% in datasets with an indel frequency of $f_{\text{indel}} = 10$/kb. While many reads that cover more than one indel per site are mapped to incorrect positions by Novoalign and RazerS, these reads are not mapped at all by BWA (Figure 3b and c)

3.2 Microindels affect SNV detection error rates

Many efficient short-read alignment tools that are used for SNV detection map short reads by only allowing a certain number of mismatches. Gapped alignments—that are required for indel detection—are not yet enabled in some widely used alignment tools. The alignment tool MAQ (Li *et al.*, 2008) or Bowtie (Langmead *et al.*, 2009) for instance, will not detect microindels by their default settings and optimize their local alignment only by minimizing the number of mismatched bases. As a consequence ungapped alignments near microindels are prone to false-positive SNV calls. Indeed, Ossiwski *et al.* (2008) found that microindels are a major source of false SNV detection by MAQ. We studied how the frequency of microindels affects SNV detection and analyzed whether SNV calling profits from gapped short-read alignments.

The differences between individuals with respect to SNVs are now adequately known on a genome-wide scale from the comparison of complete diploid genome sequences (Ahn *et al.*, 2009; Bentley *et al.*, 2008; Levy *et al.*, 2007; Wheeler *et al.*, 2008). The haploid genomes of two individuals of Central European descent differ at approximately two million chromosomal positions totalling in about three million homozygous and heterozygous SNVs. In our simulated sequence data, we thus adjusted the SNP frequency between the reference sequence that was used as template for the short-read alignments and the simulated test sequence to a rate of $f_{\text{SNP}} = 1$/kb and assumed that these variants are randomly distributed. For our simulations, we took the frequencies of microindels that were reported in Harismendy *et al.* (2009) as a lower bound to the estimated microindel prevalence. The four analyzed individual samples differed in an average of nine microindels ≤ 5 bp per 88 kb, which translates to a frequency of $f_{\text{indel}} = 0.1$/kb. We arbitrarily chose $f_{\text{indel}} = 10$/kb as an upper bound for the simulations.

Our SNV calling was based on the consensus sequence produced by samtools incorporating a Bayesian model. We further filtered for heterozygous and homozygous variations as described in Harismendy *et al.* (2009): We called a nucleotide a heterozygous variant whenever at least five reads fulfilling the quality criteria covered the sequence position and the variant frequency ranged between 20% and 80%. A homozygous variant was called for a variant frequency >80%.

We compared the effect of different microindel frequencies on SNV detection by mapping tools that do or do not allow gapped alignments. A short sequence read coming from a region in the sample sequence with a microindel compared with the reference sequence may thus be aligned with mismatched bases instead of gaps, resulting in a higher rate of variant bases at indel positions. In Figure 4a and b, the rate of falsely called SNVs at microindel positions versus indel frequency and indel size is shown for 36 bp reads with a mean sequencing depth of 18. The error rate for mapping tools that perform ungapped alignments (Bowtie and MAQ) is higher for SNV detection compared with mapping tools that allow gapped alignments (BWA, Novoalign and RazerS) for all simulated datasets. For increasing indel size the error rate of indel positions that are falsely called as SNVs decreases for ungapped mapping tools, whereas there is an increasing trend for gapped mapping tools (Fig. 4b). For gapped mapping tools, the probability increases with indel size that a read that encompasses a large indel near one of its ends is falsely mismatch aligned. In contrast, ungapped mapping tools tend to not align these reads at all.

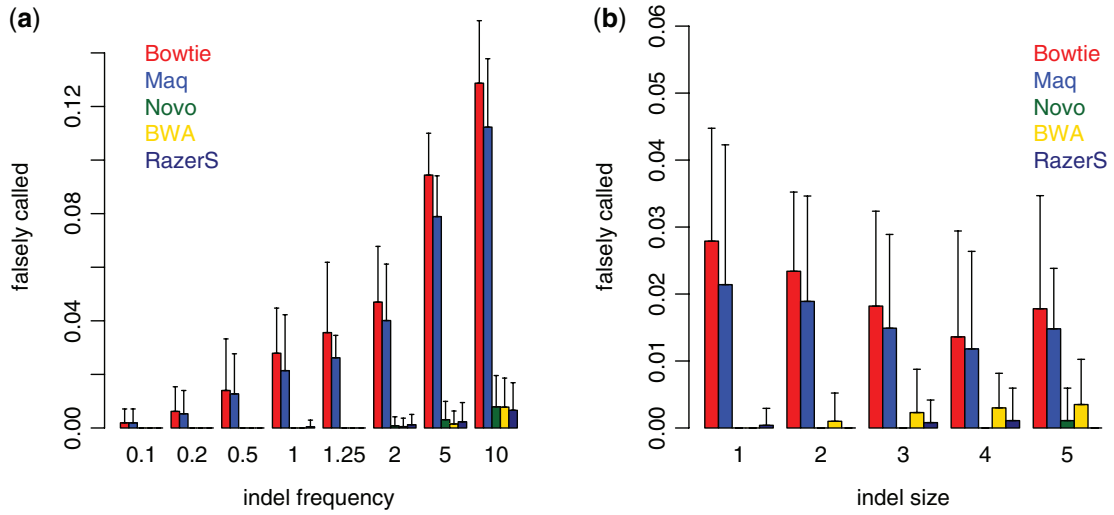


Fig. 4. Thirty-six bp short reads from simulated datasets with a mean sequencing depth of 18 with indel frequencies ranging from 0.1/kb to 10/kb and indel sizes ranging from 1 to 5 were aligned with mapping tools that perform ungapped (Bowtie and MAQ) and gapped (BWA, Novoalign and RazerS) alignments. SNVs were called from the consensus sequence based on a simple frequency threshold. (**a**) The rate of false positive SNVs that are called at simulated indel sites of size one increases with the indel frequency. SNV calling based on ungapped alignments exhibit higher error rates. (**b**) Also for increasing indel sizes the error rates are significantly lower when gapped mapping tools are used. Ungapped mapping tools show a decreasing trend in error rates for increasing indel size, as less of these reads get mapped, whereas gapped mapping tools show an increasing trend, however on a much lower level.

The Bayesian model used by samtools to generate the consensus sequence takes the quality score of the base as well as the mapping score of a read into account. When an indel is falsely aligned as mismatch, the probability increases such that the following bases are also mismatch aligned. A mismatching base in a read with a low-mapping quality will thus be less weighted in calling the consensus base. It has to be noted, that Bowtie and RazerS do not report mapping quality scores in the sam format but report a constant mapping score. Alignments produced by these mapping tools thus benefit less from the Bayesian SNV calling model.

3.3 Detection of microindels in simulated short-read data

In our second experiment, the short-read mapping tools BWA (Li and Durbin, 2010) and Novoalign (Hercus, 2009), which enable gapped alignments, were used to align the same simulated sequence data containing microindels. For each inserted or deleted sequence fragment we computed the equivalent indel region, *eir*. An indel was called if the indel frequency was >10% and at least five reads covered the *eir*. We excluded RazerS from this analysis, as its algorithmic approach that is based on editing distance is not compatible with our indel calling approach based on *eir*. RazerS tends to split up larger indels into smaller subunits of indels. Although these combinations of smaller indels may lead to the same mutated sequence, they will not be recognized as part of the *eir* (see Supplementary Material for further information).

In Figure 5, the sensitivities of indel detection are shown for varying indel frequencies and indel sizes and for datasets of different read length and sequencing depth. An indel was counted as correctly called, when the correct indel sequence was detected in the *eir* at a rate >10%. The sensitivity for detecting indels of size 1 nt depends only weakly on the indel frequency (Fig. 5a–c). In general, the sensitivity of the detection of indels of a certain size is not overly dependent on the indel frequency itself, suggesting that indel detection is quite robust over a wide range of indel frequencies. For increasing indel size, the sensitivities differ for the different mapping tools in datasets of a mean sequencing depth of 18 and 36 bp short reads. While ∼90% of indels of size three are correctly detected in reads aligned by Novoalign, this rate drops to <50% in BWA alignments (Fig. 5d). The sensitivity of detecting larger indels benefits from larger read length and higher sequencing depth. The effect of increased sequencing depth and higher read lengths also outweighs the effects of changed parameter settings of the alignment tools by far (data not shown). For the datasets of 36 bp reads and a mean sequencing depth of 18, we also tested whether a more tolerant mode of indel detection might be used to increase sensitivity rates: in the 10 bp window mode, an indel was counted as correctly called, if an indel was aligned at a rate >10% at any of the 10 bp surrounding a true indel. This also means that the exact indel sequence was not necessary for an indel to be counted as correctly called. In the tolerant window mode, the sensitivity as well as the positive predictive values increase slightly for larger indel sizes (Supplementary Material). However, for indels of size one the sensitivity is higher when indels are called based on the *eir* algorithm. This can be explained as follows: When an indel has an equivalent indel region larger than size one, not all indels are usually placed at the same position by the mapping tool. In some of the cases, the frequency of indels at a single position does not suffice to be called as indel. However, when the indel frequency threshold is based on the equivalent indel region, *eir*, all such indels contribute to the indel frequency, regardless of their position in the *eir*. It should be noted that the tolerant window mode is not able to distinguish between non-equivalent indels occurring within the same window.

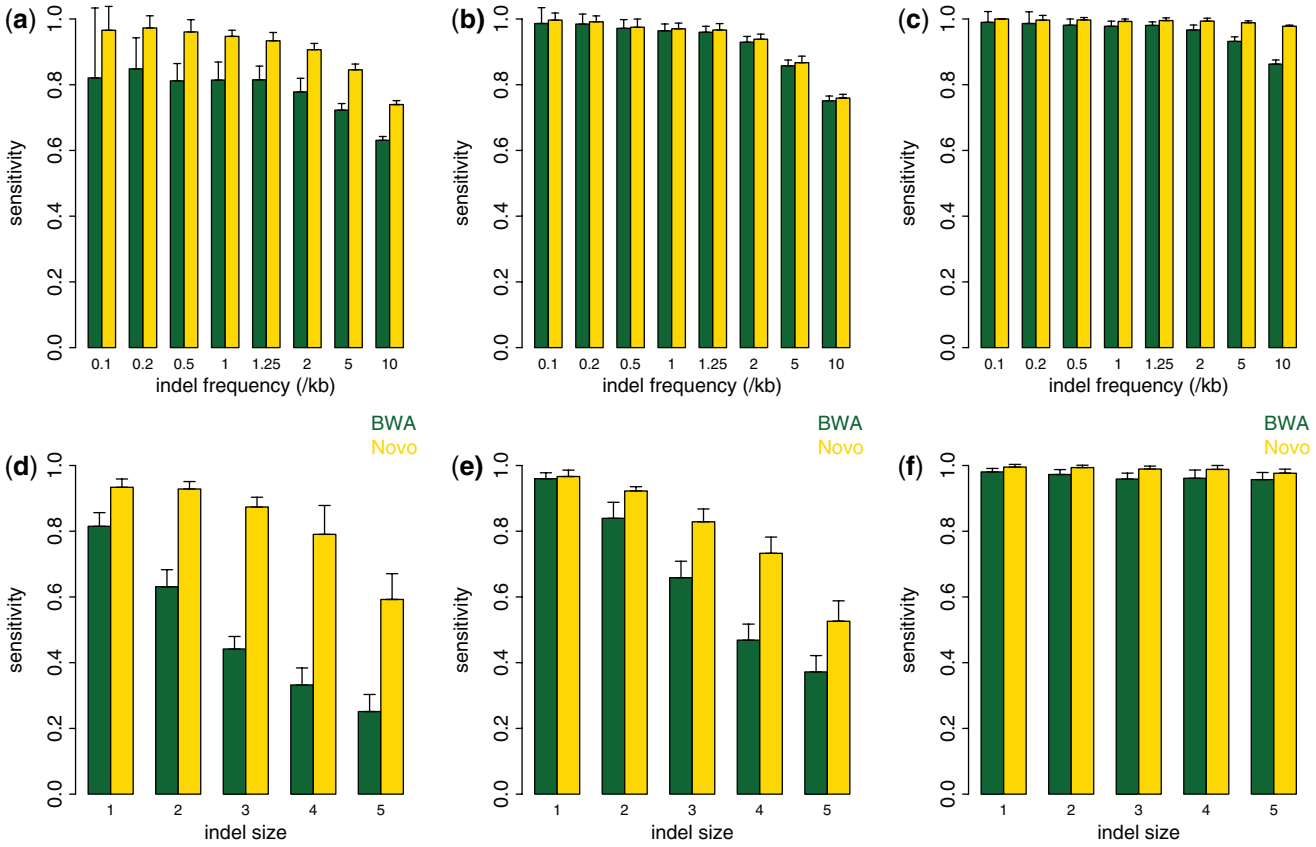


Fig. 5. Short reads containing microindels were mapped with BWA and Novoalign, which enable gapped alignments and indels were called as described in the main text. In (**a, d**) the sensitivity of indel detection was measured for datasets containing 36 bp reads and a mean sequencing depth of 18, in (**b, e**) for 36 bp reads and a mean sequencing depth of 54 and in (**c, f**) for 76 bp reads and a mean sequencing depth of 18. (a–c) The sensitivity for detecting indels of size one decreases for increasing indel frequencies. The sensitivity of indel detection based on Novoalign and especially BWA alignments benefits from higher sequencing depth and read length. (d–f) In datasets of an indel frequency of 1/kb the sensitivity of indel detection of larger indels benefits markedly from longer reads.

3.4 Microindel detection in real data

In Harismendy *et al.* (2009), sequence fragments of a total length of 88 kb were sequenced from four different individuals with ABI 3730xL Sanger, Roche 454, Illumina Genome Analyzer and ABI SOLiD technologies. Indels were automatically detected only on the ABI Sanger and the Roche 454 platform. Altogether 36 microindels of $\leq$5 bp length were detected, 6 by ABI Sanger and an additional 30 by Roche 454. Only 1 out of 6 microindels detected by ABI Sanger were also identified by Roche 454. In Harismendy *et al.* (2009), Illumina and ABI SOLiD short reads were not analyzed for microindels, as no detection algorithms were available at the time of analysis. To evaluate whether microindel detection is also applicable to real short-read data, we mapped the Illumina GA 36 bp short-read data of Harismendy *et al.* (2009) using BWA and Novoalign, and called indels as described. Seven out of the reported 36 microindels could be detected with our approach (Supplementary Material). One of the indels was also detected by ABI Sanger, the other six by Roche 454 (Supplementary Material). In addition a large number of new indels was called for each of the four individuals. For example, in NA17156 11 and 12 additional indels were called from short sequence reads that were aligned by BWA and Novoalign.

When we analyze indels that were called on the total of 296 kb of all four samples covered by short reads, altogether 331 indels were called based on alignments of BWA and Novoalign (Fig. 6). Of these, 138 indels were called in both alignments. We visually inspected all indels that were only called in one of the two alignments: the overwhelming majority of indels that could not be called in both alignments have a low frequency and are not called in one of the two alignments because of the frequency threshold of 10%. Only in three cases different indel sequences were called to the same intervall and were thus counted as different indels (Supplementary Material).

4 DISCUSSION

The genome-wide frequency of small base pair insertions and deletions might have been previously underestimated for a simple reason: traditional sequencing techniques are simply not overly good at automated detection of short indels, especially if they are heterozygous. In Harismendy *et al.* (2009), the ABI Sanger and Roche 454 platforms were used for automated indel detection. Only 1 out of 36 microindels ($\leq$5 bp) was detected by both platforms, suggesting high false negative and false positive rates of both

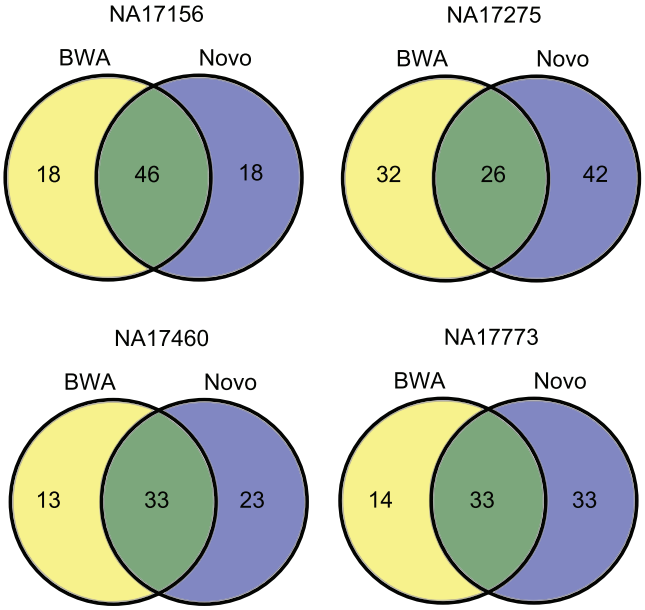


Fig. 6. Venn diagram for microindels called on altogether 296 kb based on the different mapping tools, BWA and Novoalign. In individual NA17156, 46 microindels were detected in alignments of both mapping tools, whereas 18 microindels were only detected in BWA or Novoalign. Most indels that were only detected in one alignment have a low frequency in the aligned reads.

technologies. The pyrosequencing technology of Roche 454 uses the fluorescent signal strength of incorporated nucleotides in a homopolymer to estimate its length. However the signal strength for homopolymer stretches is only linear for up to eight consecutive nucleotides, resulting in a higher error rate for larger homopolymer stretches (Margulies *et al.*, 2005). Five indels detected by ABI Sanger but not by Roche 454 and one indel detected by Roche 454 but not by ABI Sanger are flanked by such homopolymers.

Another issue is whether indels detected in the targeted sequence are actually present in the genomic DNA of the individual, or are artifacts of the long-range PCR amplification process. Especially sites with short tandem repeats exhibit higher mutation rates in PCR reactions due to DNA slippage (Lai *et al.*, 2003; Shinde *et al.*, 2003). Mutations that occurred during the sample amplification will thus be present at frequencies far below 100%. Seven of the 30 indels only detected by Roche 454 are such extensions or contractions of short tandem repeats. Interestingly four of these seven indels could also be seen in alignments of Illumina short-reads, however they were not called as indels by our approach, as their frequency did not pass the frequency threshold of 10%. This might indicate that some indels detected by Roche 454 and other NGS platforms are actually false positives, due to the sample preparation. This error rate might be reduced if DNA enrichment techniques are used that are not based on an in vitro amplification step.

We analyzed the performance of a indel calling algorithm that uses an unambiguous definition of an indel region on simulated datasets containing indels with frequencies ranging from $f_{\text{indel}} = 0.1$/kb to $f_{\text{indel}} = 10$/kb and demonstrated that the sensitivity and positive predictive value are almost constant over a range of two orders of magnitude.

We may now use the positive predictive value that was measured in our simulations for Novoalign to estimate the true microindel frequency in the targeted sequences that were amplified by long-range PCR in the four individuals analyzed in Harismendy *et al.* (2009): $f_{\text{indel}} = (64 + 68 + 56 + 66)/(4 \times 296) \times 0.9 = 0.19$ microindels/kb.

In the first diploid genome that was sequenced using paired-end short reads of the Illumina platform a microindel frequency of 0.033 indels/kb was reported (Bentley *et al.*, 2008). Sequencing of the diploid genome of a famous geneticist using the Roche 454 platform identified an order of magnitude more indels (Wheeler *et al.*, 2008). Therefore, we claim that the range of frequencies of indels used in our simulations are not unrealistic. Additionally, it is plausible that sequencing platform specific as well as algorithmic differences are responsible for at least part of the wide discrepancy of the indel frequencies in these two diploid genome sequences.

Compared with the frequency of SNVs, the microindel frequency seems to be at least an order of magnitude smaller. The effect of microindels on the false positive error rate of SNV detection should thus be relatively small (≤ 0.05) (Fig. 3). However, further studies on real datasets should investigate whether mapping tools that allow gapped alignments reduce false positive SNVs called from short-read data, as our simulations suggest.

We outlined that the unambiguous annotation of an indel may require more than just a single coordinate with respect to the reference depending on the sequence context and suggested the equivalent indel region, *eir*, for this purpose. Databases such as dbSNP have not yet systematically dealt with this annotation problem. For instance, there are two entries in dbSNP that correspond to one of the indels reported in Harismendy *et al.* (2009): rs72552124 and rs41312514 report an inserted guanine at the beginning and alternatively at the end of a 6 base polyguanine tract.

As demonstrated, our method is well suited for automated indel detection in short sequence reads. We also showed that the sensitivity of indel detection benefits considerably from higher sequencing depth and longer reads. This should be considered in the experimental design. For datasets with a low coverage and short reads, the sensitivity may be maximized at the expense of computing time by using more accurate mapping tools. In future work, additional methods to further increase sensitivity and positive predictive values for indel detection will be analyzed. These include the evaluation of paired-end reads and a restriction to indels called only in the center part of a short read.

5 CONCLUSION

Our study has provided insight into systematic errors in SNV detection that is based on short-read sequence alignments. False positive error rates in SNV detection can be markedly reduced by using mapping tools that enable gapped alignments. Microindel detection in short-read alignments using a simple algorithm to calculate the *equivalent indel region* was shown to be technically feasible in simulated datasets. The sensitivity of automated indel detection from short reads is comparable with automated indel detection methods on ABI Sanger or the Roche 454 platform. Continued improvements in our understanding of the technical issues of NGS platforms will allow the development of more sophisticated analysis methodologies.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers that helped in improving our work by their comments. We thank Mathew Wakefield, Nadia Chuzhanova and Marcel Schulz for helpful discussions on the manuscript.

Funding: Berlin-Brandenburg Center for Regenerative Therapies (BCRT) (Bundesministerium für Bildung und Forschung, project number 0313911); Deutsche Forschungsgemeinschaft (DFG SFB 760).

Conflict of Interest: none declared.

REFERENCES

Ahn,S.-M. *et al*. (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.*, **19**, 1622–1629.

Ball,E.V. *et al*. (2005) Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum. Mutat.*, **26**, 205–213.

Bentley,D.R. *et al*. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Bhangale,T.R. *et al*. (2005) Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum. Mol. Genet.*, **14**, 59–69.

Durbin,R. *et al*. (1999) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, The Edinburgh Building, Cambridge CB2 2RU, UK.

Harismendy,O. *et al*. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.*, **10**, R32.

Hercus,C. (2009) www.novocraft.com (last accessed date November, 2009).

Kolpakov,R. *et al*. (2003) mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.*, **31**, 3672–3678.

Korbel,J.O. *et al*. (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.

Lai,Y. *et al*. (2003) The mutation process of microsatellites during the polymerase chain reaction. *J. Comput. Biol.*, **10**, 143–155.

Langmead,B. *et al*. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Levy,S. *et al*. (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.

Li,H. and Durbin,R. (2010) Fast and accurate long read alignment with Burrows-Wheeler transform. *Bioinformatics*. [Epub ahead of print, January 15, 2010]

Li,H. *et al*. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*., **18**, 1851–1858.

Mardis,E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, **9**, 387–402.

Margulies,M. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Ossowski,S. *et al*. (2008) Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res.*, **18**, 2024–2033.

Redon,R. *et al*. (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Shinde,D. *et al*. (2003) Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. *Nucleic Acids Res.*, **31**, 974–980.

Weese,D. *et al*. (2009) RazerS–fast read mapping with sensitivity control. *Genome Res.*, **19**, 1646–1654.

Wheeler, D.A. *et al*. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.

2.2 Die Allel-Verteilung an heterozygoten Positionen in NGS Daten kann durch einen Verzweigungsprozess beschrieben werden

Heinrich, V., Stange, J., Dickhaus, T., Imkeller, P., Kruger, U., Bauer, S., Mundlos, S., Robinson, P.N., Hecht, J., and Krawitz, P.M. (2012). The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. Nucleic Acids Research 40, 2426-2431.

Heterozygotie bei einem diploiden Organismus bedeutet, dass in jeder Zelle zwei unterschiedliche Allele, a_1 und a_2, an einer bestimmten Position im Genom vorliegen. Wenn man nun n Sequenzfragmente, die dieser Position im Genom entstammen und aus einem Pool von Zellen isoliert wurden, zufällig auswählt, so wäre die Anzahl n_1 der Fragmente mit a_1, binomialverteilt, n_1=B(n, 0,5). Die Varianz dieser Binomialverteilung ist 1/4 x n und bei einer Stichprobengröße von n=20 betrüge die Wahrscheinlichkeit >0.95, dass $n_1 \in [6,14]$. Das heißt, die Wahrscheinlichkeit ist größer als 95%, dass man unter den 20 Sequenzfragmenten mindestens 6 aber nicht mehr als 14 mit dem Allel a_1 findet. Hat man hingegen unter den 20 Fragmenten nur 5 a_1 Allele gefunden (Allel Frequenz von a_1/n<0,3 bei einer Abdeckung von n=20), müsste auf dem Signifikanzniveau von 0.05 abgelehnt werden, dass an diesem Lokus tatsächlich Heterozygotie vorliegt.

Bei den meisten NGS Verfahren wird jedoch nicht n_1 bestimmt. Bei der Technologie von Illumina besteht der eigentliche Messvorgang während der Sequenzierung im Nachweis eines Fluoreszenzsignals eines markierten Nukleotids, das während der semikonservativen Synthese des DNA-Stranges eingebaut wird (sequencing-by-synthesis, SBS). Vor dem Sequenzierprozess erfolgen daher unterschiedliche PCR-basierte Amplifikationsschritte der Proben-DNA, um eine hohe Anzahl immer gleicher Nukleotidsequenzen auf eng umschriebenem Raum zu erzeugen, die letztlich eine ausreichende Signalstärke gewährleisten (library preparation and cluster generation). Dies bedeutet aber auch, dass nicht die Sequenz der Erbinformation, die aus dem untersuchten Zellmaterial gewonnen wurde, direkt gemessen wird, sondern erst Kopien davon angefertigt wurden, die schließlich untersucht werden.

Der Amplifikationsprozess, in dem aus den ursprünglich vorliegenden Allelen, n_1 und n_2, die messbaren Kopien, n'_1 und n'_2 entstehen, kann mathematisch als Verzweigungsprozess beschrieben werden, bei der neben n_1 auch die Anzahl der PCR-Zyklen, k, und die Effizienz des Kopiervorgangs, p, mit einfließen. Die Varianz, die in Abhängigkeit dieser Parameter zu erwarten ist, kann aus der momenterzeugenden Funktion Q abgeschätzt werden:

$$Var\big(Q(n,k,p)\big) = \frac{2(1+p)^{-1} - 2(1+p)^{-k-1} + (1+p)^{-k} - 1}{8n}$$

Hierbei haben wir vereinfachend angenommen, dass die initial vorliegende Anzahl der Allele gleich ist, $n=n_1=n_2$, und dass die Amplifikationseffizienzen für die Allele a_1 und a_2 vergleichbar sind, $p=p(a_1)=p(a_1)$. Der Parameter k kann aus den Protokollen bestimmt werden und p und n können aus den während der Probenvorbereitung durchgeführten DNA Konzentrationsbestimmungen abgeschätzt werden. Für eine übliche Exom-Sequenzierung ergibt sich zum Beispiel für n=10, k=20, p=0,4 eine auf n korrigierte Varianz von ca. 0,01. Die Gesamtvarianz setzt sich nun aus der korrigierten Varianz der Binomialverteilung, 0,0125 und der des Amplifikationsprozesses zusammen und 2 Standardabweichungen betragen 2 x sqrt(0,01+0.0125)=0.3. Wenn man nun erneut das 0.95

Konfidenzintervall berechnet, welches bei einer 20-fachen Sequenziertiefe für die Frequenz eines heterozygoten Allels als Resultat eines Verzweigungsprozesses zu erwarten wäre, so fällt es deutlich größer aus: a_1/n=[0.2,0.7]. Damit würde bei einer Sequenziertiefe von 20 eine Heterozygotie bereits angezeigt, wenn 4 Sequenzfragmente ein von der Referenz abweichendes Allel aufweisen würden.

Die durch den Amplifikationsprozesses zusätzlich verursachte Varianz kann also die Detektion heterozygoter Allele deutlich erschweren. Modelle der Genotypisierung hingegen, die nur eine Binomialverteilung der Allele annehmen, können, wenn der Amplifikationsprozess nicht berücksichtigt wird, gegebenenfalls hohe Falsch-Positiv-Raten (FPR) bedingen.

Mit dem von uns hergeleiteten mathematischen Modell, das den Einfluss der unterschiedlichen Parameter *p*, *k* und *n* beschreibt, können auch Empfehlungen für die Probenvorbereitung ausgesprochen werden: Die Varianz reduziert sich, wenn *n* und *p* maximiert und *k* minimiert werden. Nach dem Amplifikationsprozess ist zudem das Verhältnis n_1/n_2 fixiert und auch eine Erhöhung der Sequenziertiefe wird daran nichts ändern. Wenn die Varianz des Amplifikationsprozesses maßgeblich zum Genotypisierungsfehler beiträgt, kann es daher sinnvoll sein, eine komplette Resequenzierung vorzunehmen, bei dem auch der Amplifikationsprozess wiederholt wird, anstatt mehr Rohsequenzen derselben Proben library zu generieren.

Wir konnten zum Beispiel anhand von neun Exomen des gleichen Individuums zeigen, dass es bei den ca. 10.000 heterozygoten Positionen einige Allele gibt, die sich erst nachweisen ließen, wenn technische Replikate berücksichtigt wurden. Im Umkehrschluss kann auch die Verteilung der heterozygoten Allel-Frequenzen in einem Exom darüber Auskunft geben, wie stark der vom Amplifikationsprozess zu erwartende Beitrag zur gesamten Varianz ist.

Da die mittlere Abdeckungstiefe und deren Verteilung zwischen Datensätzen schwanken kann, muss zur Vergleichbarkeit der Gesamtvarianz die Anzahl der Sequenzfragmente, die pro heterozygoter Position berücksichtigt wird, erst angeglichen werden. Wir haben zur Normierung eine Abdeckung von 20 x gewählt. Diese liegt deutlich unter der mittleren Abdeckung, die üblicherweise pro Exom anvisiert wird und ermöglicht so den Einschluss der meisten heterozygoten Positionen der Zielregion. Ist eine Position stärker abgedeckt, so kann durch zufällige Auswahl der Sequenzfragmente künstlich auf 20 x begrenzt werden. Bei dieser Sequenziertiefe steht eine Gesamtvarianz der heterozygoten Allel-Verteilung von unter 0,02 und damit für einen Anteil der Varianz durch den Amplifikationsprozesses von unter 0,0075, für einen Datensatz hoher Güte. Auf der Plattform GeneTalk wird dieser Parameter als Teil eines Qualitätsberichts für NGS-Datensätze berechnet.

Nucleic Acids Research Advance Access published November 29, 2011

Nucleic Acids Research, 2011, 1–6
doi:10.1093/nar/gkr1073

The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process

Verena Heinrich[1], Jens Stange[2], Thorsten Dickhaus[2], Peter Imkeller[2], Ulrike Krüger[1], Sebastian Bauer[1], Stefan Mundlos[1], Peter N. Robinson[1], Jochen Hecht[3] and Peter M. Krawitz[1,*]

[1]Institute for Medical and Human Genetics, Charité Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, [2]Department of Mathematics, Humboldt-University Berlin, Unter den Linden 6, 10099 Berlin and [3]Berlin-Brandenburg Center for Regenerative Therapies (BCRT), Charité Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany

Received July 20, 2011; Revised October 19, 2011; Accepted October 28, 2011

ABSTRACT

With the availability of next-generation sequencing (NGS) technology, it is expected that sequence variants may be called on a genomic scale. Here, we demonstrate that a deeper understanding of the distribution of the variant call frequencies at heterozygous loci in NGS data sets is a prerequisite for sensitive variant detection. We model the crucial steps in an NGS protocol as a stochastic branching process and derive a mathematical framework for the expected distribution of alleles at heterozygous loci before measurement that is sequencing. We confirm our theoretical results by analyzing technical replicates of human exome data and demonstrate that the variance of allele frequencies at heterozygous loci is higher than expected by a simple binomial distribution. Due to this high variance, mutation callers relying on binomial distributed priors are less sensitive for heterozygous variants that deviate strongly from the expected mean frequency. Our results also indicate that error rates can be reduced to a greater degree by technical replicates than by increasing sequencing depth.

INTRODUCTION

Second-generation DNA sequencing has revolutionized many biomedical areas. It especially accelerated the discovery of disease genes in medical genetics (1,2) and is now about to enter diagnostics (3). In order to translate this technology into a reliable tool for molecular diagnostics for human genetics and other fields, it will be necessary to further reduce error rates of sequence variant detection. Understanding the process of how the high-throughput sequencing data arise is crucial for the development of sensitive genotype calling algorithms. It is well known in the field that especially the error rates in detecting heterozygous mutations in diploid genomes are still considerably higher than the comparable error rates of homozygous variants—even at high levels of sequence coverage (4,5).

It is currently widely assumed that the frequencies of calls at heterozygous sites in NGS data are binomially distributed, an assumption that has been incorporated into many variant calling programs for NGS data (6–8). We were motivated to question this assumption by observations of more extreme probability distributions in whole-exome sequencing (WES) data sets, as we will demonstrate in this article. We therefore analyzed the key steps in NGS data generation from a stochastic point of view and identified the amplification of sequence fragments during library preparation before measurement as crucial for the distribution of allele frequencies at heterozygous genomic loci.

We reasoned that the generation of fragments can be described as a Bienaymé–Galton–Watson branching process with discrete time steps, which is a model that has been widely used by physicists and mathematicians in population genetics (9–11). In this work, we provide a detailed description of the fragment amplification process. We then show that our model accurately reflects allele frequencies in real WES data sets. One prediction of our model is that technical replication is more effective in reducing error rates than merely sequencing more reads

*To whom correspondence should be addressed. Tel: +49 30 450 569 122; Fax: +49 30 450 569 915; Email: peter.krawitz@charite.de

from the same library, which we confirmed on a data set with nine technical WES replicates. Our results have important implications for understanding the causes of false-negative errors in NGS diagnostics.

MATERIALS AND METHODS

Exome sequencing and variant detection

Human blood or tissue samples of 17 anonymized donors were used for exome sequencing. For one of these individuals, nine technical replicates were generated. This means nine independent samples of the same individual were collected and further processed independently. For each sample, genomic DNA was enriched for the target region of all human CCDS exons (12) with Agilent's SureSelect Human All Exon Kit and subsequently sequenced on a Illumina Genome Analyzer II with 100 bp single end reads. The enrichment of adapter-modified DNA fragments before sequencing includes an amplification step of 18 cycles of polymerase chain reaction (PCR) in the standard protocol. For one exome, 36 cycles of PCR were run to analyze the effect of the cycle number onto the allele frequency distribution. The cluster generation step follows after the library preparation. Its purpose is to increase the fluorescent signal of a fragment on the sequencing flow cell, so that it becomes detectable. The cluster generation includes another 35 PCR cycles in the standard protocol. The raw data of ~5 GB per exome was mapped to the haploid human reference sequence hg19 with novoalign (13) resulting in a mean coverage of the exome target region of 50x. In this study, heterozygous sequence variant detection was restricted to positions of high human variability as defined by dbSNP132 positions, in order to decrease the probability of false positive calling errors. A genomic position was called as a heterozygous variant if >20 sequence reads covered this position in the reference-based sequence alignment and if the ratio of the non-reference allele to the sum of the non-reference allele and the reference allele was between 0.14 and 0.86. This heterozygous detection algorithm was shown to be highly sensitive for a coverage >20 (14). For the replicates we classified a locus as truly heterozygous, if it was classified as heterozygous by the above described calling criterion and by SAMtools (15) in at least six out of nine replicates.

Heterozygous allele frequencies

The reference allele frequency at a genomic position that was classified as heterozygous as described above is defined as the number of fragments that map to this position, cover the variable base and show the reference allele, divided by all fragments covering this site. There are two well-known biases that shift the detected mean reference allele frequency from the expected value of 0.5 to slightly higher values: (i) SureSelect baits that were used for exon enrichment are designed as 120 bp antisense oligonucleotides to the haploid reference sequence of the latest Human Genome Build. This means DNA hybridization between sample DNA fragments containing common variants, that differ from the reference sequence, may be weaker as compared with hybrids without mismatches. This may lead to a slightly more effective enrichment of sequence fragments containing the reference allele. (ii) After sequencing, all short sequence reads are mapped to the haploid reference sequence. Sequence fragments containing non-reference allele variants have a lower mapping quality. For short read lengths, reads with low base quality and low sequence complexity, this may result in a slightly reduced mapping ratio of non-reference allele fragments (16,17). Due to this *in vitro* enrichment as well as *in silico* read mapping-bias, the allele frequency distribution shifted toward the reference allele (in our analyzed exome data sets from 0.5 to 0.54). However, as these biases are systematic and not stochastic in nature they do not influence the variance of the allele frequency distribution.

Distributions of heterozygous allele frequencies are position- and individual independent

The dependence of the allele frequency distribution on genomic position as well as on the individual was tested on human exome data sets. Position dependence was tested by comparing the distribution of all heterozygous allele frequencies in an individual to a smaller random subset of these positions (Supplementary Figure S3). The comparison between these distributions did not show significant differences by chi-squared testing. The dependence on the individual was tested by comparing the differences of heterozygous allele distributions between different individuals and technical replicates of the same individual. The difference in frequency distributions between different individuals is statistically not significant and fluctuations in these distributions are comparable to those observed in technical replicates of the same individual. Since allele frequencies are position- and individual independent, we computed the heterozygous allele frequency distribution from SNP loci pooled from all sequenced exomes.

RESULTS AND DISCUSSION

Fragment amplification as a stochastic branching process

Suppose that we have a tube that initially contains a set of different alleles such as illustrated in Figure 1A. We now perform K cycles of a PCR on these alleles, which basically means adding a certain number of copies of these alleles to the tube in discrete time steps. This is an essential part of current NGS library preparation protocols that are used to enrich adapter-ligated DNA fragments (18).

For the mathematical description of this process, we will introduce a Markov chain, that corresponds to a Galton–Watson branching process consisting of two populations. Although we will study this process in our work only for biallelic single nucleotide polymorphisms (SNPs), it may be generalized to all sequence variants.

The preparation of a genomic DNA sample starts by shearing the chromosomal DNA into sequence fragments of a few hundred base pairs. We will discuss in the following only fragments that contain a variable base of an SNP, which means we can distinguish between two possible

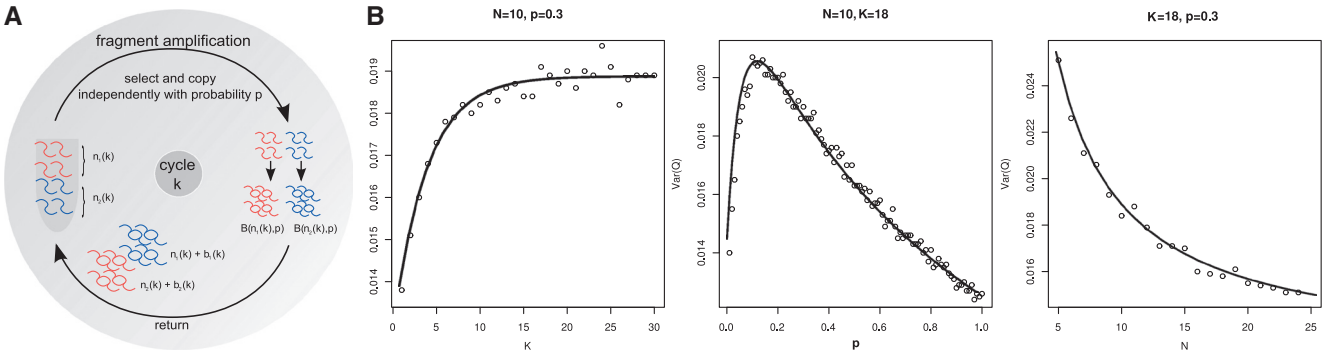


Figure 1. The fragment amplification as a stochastic branching process. (**A**) The distribution of the allele frequencies depends on a parameter P that represents the efficiency of the PCR and the probability that an allele is amplified, the cycle number K, and on the initial number of alleles N. (**B**) The variance of the allele frequency after amplification was sampled from simulations for P ranging from 0 (no amplification) to 1 (perfect duplication in each PCR cycle), for different cycle numbers K and numbers of starting alleles N. The measurement process of sequencing was simulated for a read coverage of 20×. The variance sampled from 10 000 simulated heterozygous SNPs and depicted as black circles (o), is well approximated by the analytical results of Equation (4) (black line). For a cycle number of $K > 20$, the variance does not change significantly. The variance reaches its maximum for an amplification probability around $P = 0.2$. For an increasing number of alleles before amplification, the variance approximates a fixed level, explained solely by the variance introduced by the measurement process of sequencing.

classes of fragments, those containing the base of allele A_1 and those that contain the base of allele A_2. We consider the fragmentation as random and unbiased. This means that the extensions into both directions from the variable position is uniform and only limited by fragment size. We also assume that the numbers n_1 and n_2 of the fragments containing allele A_1 and A_2 are of the same order of magnitude after fragmentation, as the DNA originates from many cells of a single diploid genome (see Supplementary Figure S4 for exceptions from this assumption). Before sequencing (at time step $k = 0$), adaptor oligomers are ligated to the fragments and a PCR is run for K cycles. For successful amplification, adaptors must be attached to both ends of the fragment. The initial number of amplifiable fragments, $n_1 = n_1(0)$ and $n_2 = n_2(0)$, is in the order of dozens. For each such fragment, the attachment of the polymerase to the adaptor is a prerequisite for amplification. We assume that the probability of this event depends only on the total number of polymerase molecules, which remains the same in every PCR cycle k, and the sum of amplifiable fragments, $n_1(k)+n_2(k)$, but is independent of the variant itself. For not too large K, we may assume that polymerase is always in excess of $n_1(k)+n_2(k)$, and thus a constant fraction of fragments will be bound by polymerase. We will use the parameter p in the main manuscript to describe the cycle and allele-independent probability that a fragment is copied (in the Supplementary methods we perform the calculations for allele-specific amplification probabilities, p_1 and p_2). We now describe the probabilities of the three possible transitions of a random allele in PCR cycle k, assuming that the Markov condition holds:

$$\begin{aligned} P((n_1(k), n_2(k)) \to (n_1(k)+1, n_2(k)) &= \frac{n_1(k)}{n_1(k)+n_2(k)}p \\ P((n_1(k), n_2(k)) \to (n_1(k), n_2(k)+1) &= \frac{n_2(k)}{n_1(k)+n_2(k)}p \\ P((n_1(k), n_2(k)) \to (n_1(k), n_2(k)) &= 1-p \end{aligned} \quad (1)$$

The whole system thus transits to:

$$(n_1(k+1), n_2(k+1)) = (n_1(k)+b_1(k), n_2(k)+b_2(k)) \quad (2)$$

where $(b_1(k), b_2(k))$ are realizations of binomially distributed random variables $B(n_1(k), p)$ and $B(n_2(k), p)$ (Figure 1A).

The ratio $n_1(k)/(n_1(k)+n_2(k))$ describes the proportion of allele A_1 after the k-th amplification cycle and this is the allele frequency that we expect to measure by sequencing multiple read fragments of this pool. Note that sequencing itself will contribute to the totally measured variance. Sequencing itself may be understood as a random sample of finite size, which is the sequencing depth, on the allele pool after amplification. We are thus primarily interested in the distribution of the random variable $Q(k)$ describing the ratio of alleles after amplification. The distribution of alleles after step k solely depends on the distribution of alleles in step $k-1$:

$$\begin{aligned} &P((n_1(k), n_2(k))|(n_1(k-1), n_2(k-1)), \\ &\quad (n_1(k-2)n_2(k-2)), ..., (n_1(0), n_2(0))) = \\ &\quad = P((n_1(k), n_2(k)))|(n_1(k-1), n_2(k-1))). \end{aligned} \quad (3)$$

The entire process is determined by the probability generating function of the offspring distribution. Appropriately scaled, the law of $Q(k)$ approaches a normal distribution (10). We derived the first and second moments of the offspring distribution (see Supplementary Methods for a detailed calculus) to compute the asymptotic variance of $Q(k)$:

$$\mathrm{Var}(Q(k)) = \frac{2(1+p)^{-1} - 2(1+p)^{-k-1} + (1+p)^{-k} - 1}{8N} \quad (4)$$

assuming that $n_1(0) = n_2(0) = N$.

According to a standard NGS protocol, we simulated the amplification process of our model depicted in Figure 1A for $K = [1, 30]$, $N = [5, 25]$, for P ranging

from 0 to 1 and a sequencing depth of 20×. We computed the variance of the resulting allele frequency ratio for 10 000 SNPs (Figure 1B) which is the expected order of magnitude for heterozygous variant calls in a human exome. The behavior of the variances sampled from our simulations is well described by function (4) adapted by the additional contribution of variance introduced by sequencing. For fixed P and N, the variance increases with a growing number of PCR cycles K and approaches a constant level for $K > 15$. This also means that increasing the number of cycles in the library preparation above the default value of $K > 18$, as well as amplification of the cluster generation step that succeeds the library preparation will only contribute marginally to the total variance. For fixed K and N, the variance has its maximum around $P = 0.2$ and decreases for P tending to 1. This is clear as with perfect amplification, we expect the initial ratio of $n_1(0)/(n_1(0) + n_2(0)) \approx 0.5$ to remain constant. For fixed K and P, the variance decreases with an increasing number of alleles before amplification. It is easier for one allele to gain predominance in the pool that is sequenced if the initial allele set is small, the amplification efficiency is low and enough PCR cycles are run.

High variance of heterozygous allele frequencies in real human exome data sets

After modeling the amplification step as stochastic process, we analyzed the distribution of allele frequencies at heterozygous genomic loci in real human exome data that were generated following a standard protocol with 18 PCR amplification cycles. In order to compare the empirically measured frequencies with our simulated data, all heterozygous SNP positions that were covered by more than 20 reads were downsampled to 20 reads per position. The allele frequencies were derived from these read sets. The variance of the measured reference allele distribution is 0.017 and thus markedly larger than the variance of 0.012 that is expected for hypothetical sequencing before amplification (this is the variance of a binomial distribution where n represents the sequencing depth and the success parameter is the ratio of the alleles in the starting solution, Figure 2A). Thus, the sequence fragments in a short read alignment, on which the variant call is performed, are not properly represented by a random sample of the initial ratio of $n_1(0)/(n_1(0) + n_2(0))$, but the effect of the amplification process on this distribution has to be taken into account.

Our model assumes a constant amplification efficiency over all PCR cycles, which seems to be a reasonable simplification given the relatively low number of PCR cycles used in NGS library preparation protocols. A value of $P \in [0.3, 0.5]$ yielded a variance for the allele frequencies that is close to the value determined on the real exome data (Figures 1B and 2A). We measured the amount of fragmented DNA used as input in our WES experiments at $k = 0$ (5 ng) and measured about $5 - 10\,\mu g$ after $K = 18$ cycles of amplification. This corresponds to an amplification by a factor of $1 - 2 \times 10^3$, and thus values of $P \in [0.3, 0.5]$ are realistic.

As already discussed, with fixed P and N the variation is approaching a limit for increasing K and for $K > 15$ it hardly changes. To check this experimentally, we sequenced the exome of the same individual that was amplified with 36 PCR cycles instead of 18. As expected by Equation (4), no significant increase in the variance could be detected (Figure 2B). We also studied the effect of the succeeding cluster amplification step by analyzing the variance of the difference of heterozygous allele frequencies of a library preparation that was sequenced after two different cluster generations. In contrast to the library preparation, the effect of the cluster generation on the total variance of the allele frequency is negligible (Supplementary Table S2).

Influence of allele frequency variance on error rates in heterozygous variant detection

Assuming comparable read qualities, the variant call is based on a random sample drawn from the set consisting

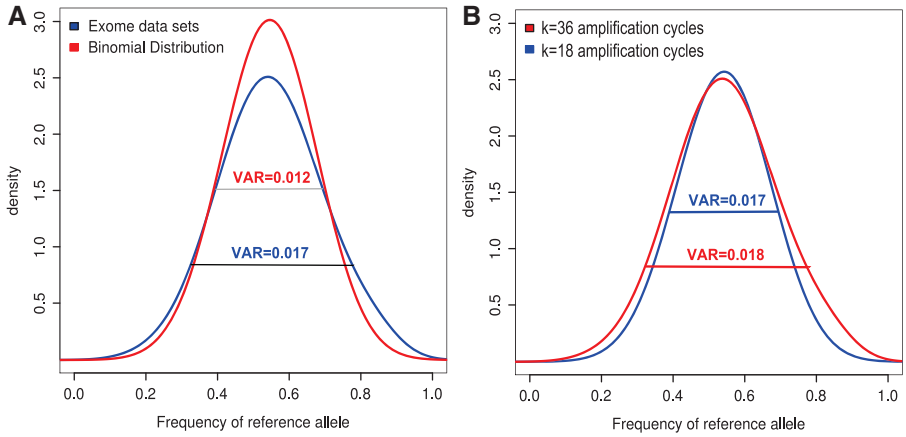


Figure 2. Variance of the measured allele frequency at heterozygous genomic positions in NGS exome data sets. (**A**) The distribution of heterozygous allele frequencies measured in exome data sets at 20× coverage (blue) compared to the theoretical distribution expected before amplification (red). The variance of the real distribution after amplification is significantly larger. (**B**) An exome of the same individual was sequenced following 18 and 36 cycles of amplification. As expected from theory, the variance of the allele frequencies only slightly increases after the additional 18 cycles of amplification.

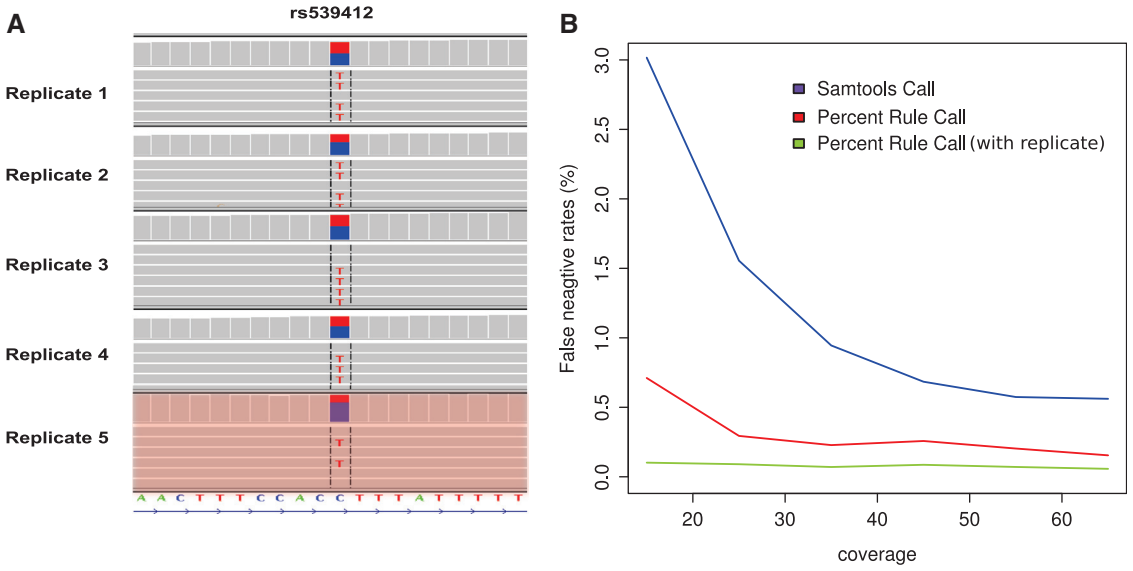


Figure 3. Influence of variance in measured allele frequency on variant calling. (**A**) The genotype at the SNP position rs539412 has been called as heterozygous variant in the first four replicates, but was not detected in the fifth replicate due to low frequency. (**B**) The false negative error rate decreases with increasing sequencing depth. At low total sequencing depth, the error rate is markedly reduced by considering pools of technical replicates. The classification of a genotype as heterozygous based on a simple frequency interval (heterozygous if the non-reference allele frequency is between 14% and 86%) is more sensitive than a calling algorithm that uses a binomial prior distribution as default setting for the allele distribution. The false negative error can be further reduced by considering an additional technical replicate (see also Supplementary Table S1).

of all alleles A_1 and A_2 after amplification which is of size $n_1(k)+n_2(k)$. The coverage or sequencing depth at a variant site is equivalent to the size of the random sample on which the call is based. We hypothesized that a certain rate of true heterozygous alleles will not be called due to the high variance in allele frequencies after amplification (i.e. false-negative calls). To test this, we generated nine exome replicates of the same individual and classified genomic loci as heterozygous if they were called heterozygous in at least six out of nine replicates by two accepted calling algorithms (see 'Material and Methods' section). Figure 3A shows the common polymorphism rs539412, that was called as heterozygous variant in the first four replicates, but failed to be called as heterozygous variant in the fifth replicate due to low frequency. Using this as a gold standard, we then measured the false-negative rate for calls based on each of the single WES data sets. Over the whole exome, we measured a false-negative rate between 1% and 3% depending on the coverage with the default settings of a widely used variant caller (Figure 3B). In a usual exome, one expects between 10 000 and 15 000 heterozygous variants. Our results indicate that one will miss around a hundred heterozygous variants by sequencing an exome only once simply due to the stochastic fluctuation of the allele frequencies after amplification. Surprisingly a variant calling approach that is simply based on a heterozygous allele frequency interval f with $[14\% < f < 86\%]$, as suggested in Ref. (14), has higher sensitivity at a comparable specificity (see ROC analysis in Supplementary Figure S2) than a more sophisticated variant calling algorithm that uses the wrong prior distribution for the allele frequencies independent of the coverage (Figure 3B). Additionally for a sequencing depth above 30× the false negative rate does not decrease further. Thus, once a sufficient sequencing depth has been reached, only technical replication is able to further reduce the total error rates substantially (Figure 3B and Supplementary Table S1).

Final remarks

In this work, we studied the distribution of alleles at heterozygous genomic positions as measured in NGS data sets. A solid knowledge of distribution and variance of allele calls at heterozygous loci is important as it is an essential prior information for many variant calling approaches. Besides, the distribution of the allele frequency also plays a role in algorithms used to detect copy number variations or sample contaminations.

We have demonstrated that amplification steps contribute considerably to the total variance of this distribution. We modeled the fragment generation process as a Bienaymé–Galton–Watson branching process and showed that the variance is accurately described by Equation (4). For typical values of the efficiency P of the amplification process and sequencing depth, this is substantially higher than the variance of the corresponding binomial distribution (Figure 2A). Clearly, the higher the variance of allele calls at heterozygous loci, the higher the false negative error will be. Ultimately, calling errors arising from random events during library preparation and fragment amplification could be avoided in single molecule sequencing techniques of the future (19) and we are eager to see these data.

From our analytical results, one may draw some conclusions about how to reduce the stochastic fluctuactions coming from the amplification step: increasing the efficiency of the adaptor ligation (which is increasing N), increasing p and reducing the number of PCR cycles K

in a second-generation protocol will help to reduce the variance of heterozygous alleles.

NGS technologies such as whole-exome and genome sequencing are beginning to be used for diagnostic purposes. In this setting, it is critical to provide an estimation of the sensitivity of these approaches. Clearly, it is important to report regions of the exome that are not sufficiently covered for reliable variant calling. In addition, our results suggest that it is also important to evaluate the variance at heterozygous SNP positions as it might serve as an indicator of the quality of an experiment and thus for the overall false-negative error rate. The sensitivity of an exome screen that is based on data of a second-generation sequencing platform is not only bound by the coverage of the target region but is also affected by amplification which is inherent to the method.

SUPPLEMENTARY DATA

Supplementary Data is available at NAR Online: Supplementary Tables 1 and 2, Supplementary Figures 1–4, Supplementary Methods.

ACKNOWLEDGEMENTS

P.M.K. wishes to thank Claudia Bickenbach for her invaluable support.

FUNDING

Funding for open access charge: Deutsche Forschungsgemeinschaft (DFG KR 3985/1-1 to P.M.K.).

Conflict of interest statement. None declared.

REFERENCES

1. Ng,S.B., Buckingham,K.J., Lee,C., Bigham,A.W., Tabor,H.K., Dent,K.M., Huff,C.D., Shannon,P.T., Jabs,E.W., Nickerson,D.A. *et al.* (2010) Exome sequencing identifies the cause of a Mendelian disorder. *Nat. Genet.*, **42**, 30–35.
2. Robinson,P.N., Krawitz,P. and Mundlos,S. (2011) Strategies for exome and genome sequence data analysis in disease gene discovery projects. *Clin. Genet.*, doi: 10.1111/j.1399-0004.2011.01713.
3. Choi,M., Scholl,U.I., Ji,W., Liu,T., Tikhonova,I.R., Zumbo,P., Nayir,A., Bakkalolu,A., zen,S., Sanjad,S. *et al.* (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Pro. Natl Acad. Sci. USA*, **106**, 19096–19101.
4. Nothnagel,M., Herrmann,A., Wolf,A., Schreiber,S., Platzer,M., Siebert,R., Krawczak,M. and Hampe,J. (2010) Technology specific error signatures in the 1000 Genomes Project data. *Hum. Genet.*, doi:10.1007/s00439-011-0971-3.
5. Harismendy,O., Ng,P.C., Strausberg,R.L., Wang,X., Stockwell,T.B., Beeson,K.Y., Schork,N.J., Murray,S.S., Topol,E.J., Levy,S. *et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Bio.*, doi:10.1186/gb-2009-10-3-r32.
6. Li,H., Ruan,J. and Durbin,R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genomes Res.*, **18**, 1851–1858.
7. Li,R., Li,Y., Kristiansen,K. and Wang,J. (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.
8. Goya,R., Sun,M.G.F., Morin,R.D., Leung,G., Ha,G., Wiegand,K.C., Senz,J., Crisan,A., Marra,M.A., Hirst,M. *et al.* (2010) SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, **26**, 730–736.
9. Athreya,K.B. and Ney,P.E. (1972) *Branching Processes*. Springer, Berlin.
10. Yakovlev,A.Y. and Yanev,N.M. (2009) Relative frequencies in multitype branching processes. *Ann. Appl. Probab.*, **19**, 1–14.
11. Polya,G. and Szegö,G. (1970) *Problems and Theorems in Analysis I.*. Springer, Berlin.
12. Pruitt,K.D., Harrow,J., Harte,R.A., Wallin,C., Diekhans,M., Maglott,D.R., Searle,S., Farrell,C.M., Loveland,J.E., Ruef,B.J. *et al.* (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.
13. Hercus, C. (2011) Novoalign V2.07. www.novocraft.com (3 August 2011, date last accessed).
14. Bell,C.J., Dinwiddie,D.L., Miller,N.A., Hateley,S.L., Ganusova,E.E., Mudge,J., Langley,R.J., Zhang,L., Lee,C.C., Schilkey,F.D. *et al.* (2011) Carrier testing for severe childhood recessive disease by next generation sequencing. *Sci. Trans. Med.*, **3**, 64–69.
15. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2010) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
16. Degner,J.F., Marioni,J.C., Pai,A., Pickrell,J.K., Nkadori,E., Gilad,Y. and Pritchard,J.K. (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**, 3207–3212.
17. Krawitz,P., Rdelsperger,C., JŁger,M., Jostins,L., Bauer,S. and Robinson,P.N. (2010) Microindel detection in short-read sequence data. *Bioinformatics*, **26**, 722–729.
18. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
19. Timp,W., Mirsaidov,U.M., Wang,D., Comer,J., Aksimentiev,A. and Timp,G. (2010) Nanopore sequencing: electrical measurements of the code of life. *IEEE Trans. Nanotechnol.*, **9**, 281–294.

2.3 GeneTalk: Ein Expertennetzwerk zur Analyse und Interpretation von seltenen Sequenzvarianten in Genomdaten

Kamphans, T., and Krawitz, P.M. (2012). GeneTalk: an expert exchange platform for assessing rare sequence variants in personal genomes. Bioinformatics 28, 2515-2516.

Die Variabilität des menschlichen Genoms ist so hoch, dass sich die Genome zweier nicht verwandter Personen an ca. jeder tausendsten Stelle unterscheiden. Obwohl über die letzten Jahre viele Millionen Sequenzvarianten in großen populationsgenetischen Studien identifiziert wurden, sind die meisten dieser Varianten so selten, dass sich viele davon nicht in den bekannten Datenbanken, wie zum Beispiel dbSNP, finden lassen, wenn ein neues Genom sequenziert wird (Coordinators, 2014; Genomes Project, et al., 2010; Sherry, et al., 2001). Hinzu kommt, dass in jedem Individuum durch die Neumutationsrate von ca. 1,2 $x10^{-8}$ pro Basenpaar und Zellteilung ca. 50-60 neue SNVs entstehen. Dies entspricht pro Exom üblicherweise bis zu drei *de novo* Mutationen (de Ligt, et al., 2012; Rauch, et al., 2012).

Wenn eine Person, bei der eine genetische Ursache für eine Erkrankung vermutet wird, mittels HDS untersucht wird, so muss jede detektierte Sequenzvariante hinsichtlich ihrer medizinischen Relevanz im Kontext der Fragestellung beurteilt werden. Da bei der referenzbasierten Sequenzierung von Exomen (WES) oder Genomen (WGS) ca. 30.000 bzw. 3 Millionen Varianten detektiert werden, stellt deren effiziente Interpretation die zentrale Herausforderung für den forschenden Kliniker dar.

Um auch Personen, die nicht selbst über Programmierkenntnisse verfügen, eine solche Analyse zu ermöglichen, haben wir die online-Plattform GeneTalk entwickelt. Mit Hilfe dieser Software können Genetiker, die mit der Bedienung eines Browsers vertraut sind, WES und WGS Daten auswerten. Die Plattform bietet Filterwerkzeuge mit denen die Varianten auf die klinisch vielversprechendsten Kandidaten reduziert werden können und ist angebunden an bioinformatische Programme, die eine weitere Priorisierung ermöglichen. Mit die wichtigste Funktion von GeneTalk besteht aber darin, dass es seinen Nutzern den mutationsspezifischen Wissensaustausch ermöglicht.

Ein wichtiger Filter beruht auf den Allel- und Genotyp-Frequenzen, die in den populationsgenetischen Untersuchungen erhoben wurden. Wenn der Vererbungsmodus auf eine monogene Erkrankung hinweist und der Phänotyp zudem selten ist, so können mittels der bekannten Genotyp-Frequenzen bereits viele Varianten herausgefiltert werden. Bei vollständiger Penetranz zum Beispiel dürfen die krankheitsverursachenden Allele rezessiver Erkrankungen in gesunden Kontrollen nicht homozygot auftreten. Bei dominant vererbten Erkrankungen reicht bereits ein heterozygotes Vorkommen dieser Allele in gesunden Personen zum Ausschluss.

Anschließend werden die verbleibenden Kandidaten meist auf Varianten beschränkt, die eine funktionseinschränkende Auswirkung auf Proteinebene erwarten lassen. Bioinformatisch muss hierfür zuerst eine Annotation der genomischen Variante auf Transkript-Ebene erfolgen (Jager, et al., 2014). Auch wenn HDS-Daten zu weiteren Familienmitgliedern erhoben wurden oder bereits Intervalle aus Kopplungsanalysen vorliegen, kann die Menge der Varianten weiter eingeschränkt werden.

Nach diesen Filterschritten bleiben üblicherweise noch Kandidaten in einigen Dutzend Genen übrig, für die auch überprüft werden muss, ob es für diese bereits Beschreibungen in der Fachliteratur gibt. Die drei derzeit umfangreichsten Datenbanken, die hierfür zur Verfügung stehen, sind die Human Gene Mutation Database (HGMD), Leiden Open Variant Database (LOVD) und ClinVar (Cooper, et al., 1998; Fokkema, et al., 2011; Landrum, et al., 2014). Insgesamt umfassen diese Datenbanken ca. 100.000 Einträge von SNVs und Indels, die als krankheitsverursachend eingestuft wurden.

Auch wenn eine Übereinstimmung mit einem Datenbankeintrag gefunden wird, ist der Fall damit jedoch noch nicht zwingend gelöst. In der HGMD beispielsweise beruht die Einstufung eines Eintrags als „disease causing mutation“ auf den Angaben der Erstbeschreibung in der wissenschaftlichen Literatur. Bevor die HDS verfügbar war, wurde eine Sequenzvariante oftmals als pathogen einzustuft, wenn sie in einem Patienten nachgewiesen wurde, nicht aber in 200 gesunden Kontrollen. Gerade bei seltenen Erkrankungen ist jedoch bei diesem Stichprobenumfang nicht einmal eine signifikante Aussage über eine Assoziation mit dem Phänotyp gegeben. Die nun vorliegenden populationsgenetischen Daten, die mehrere tausend Kontrollen umfassen, zeigen, dass viele der als krankheitsverursachend eingestuften Mutationen auch bei gesunden Personen anzutreffen sind und die Klassifikation vieler Einträge daher revidiert werden muss (Xue, et al., 2012). In GeneTalk können Nutzer für solche Mutationen eine Annotation hinterlassen, um auf möglicherweise falsche Datenbankeinträge hinzuweisen und somit einen Beitrag für eine schnelle Berichtigung liefern.

Die statistische Evidenz für ein Krankheitsgen nimmt mit jedem weiteren Erkrankungsfall zu, bei dem pathogene Allele am gleichen Lokus identifiziert werden können. Gerade bei den ultra-seltenen Erkrankungen, die eine Inzidenz von weniger als 5 auf 100.000 aufweisen, ist es oftmals schwierig, mehrere Patienten zu rekrutieren. Auch hier ist die globale Zusammenarbeit der Genetiker gefordert und GeneTalk versucht dafür eine Plattform zum Informationsaustausch zu bieten. Mittlerweile ist die Nutzergemeinde auf über 1000 Wissenschaftler und Kliniker angewachsen, die in gemeinschaftlicher Anstrengung versuchen, die Bedeutung unbekannter Sequenzvarianten („variants of unkown significance“, VUCS) zu ergründen.

Bioinformatics Advance Access published July 23, 2012

GeneTalk: an expert exchange platform for assessing rare sequence variants in personal genomes

Tom Kamphans [1], Peter M Krawitz [2,*]

[1]GeneTalk, Finckensteinallee 84, 12205 Berlin, Germany
[2]Department of Medical and Human Genetics, Charité, Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany.

Associate Editor: Prof. Martin Bishop, Prof. Alfonso Valencia

ABSTRACT

Summary Next-generation sequencing (NGS) has become a powerful tool in personalized medicine. Exomes or even whole genomes of patients suffering from rare diseases are screened for sequence variants. After filtering out common polymorphisms, the assessment and interpretation of detected personal variants in the clinical context is an often time consuming effort. We have developed GeneTalk, a web-based platform that serves as an expert exchange network for the assessment of personal and potentially disease relevant sequence variants. GeneTalk assists a clinical geneticist who is searching for information about specific sequence variants and connects this user to other users with expertise for the same sequence variant.

Availability: GeneTalk is available at www.gene-talk.de. Users can login without registering in a demo account.

Contact: peter.krawitz@gene-talk.de

INTRODUCTION

Exome sequencing has become an invaluably powerful tool in the identification of disease causing variants (Bamshad *et al*., 2011; Robinson *et al*., 2011) and the first patients are now already treated based on sequence variant information of their exomes (Worthey *et al*., 2011). To date, the primary bottleneck in such clinical personal genome cases is not anymore data generation but data analysis. Clinical geneticists therefore require efficient tools for filtering and interpreting the clinically meaningful sequence variants.

Currently the most potent filters for reducing the set of novel and potentially causal mutations in rare diseases are based on variation data from population scale sequencing efforts such as the 1000 genomes project (www.1000genomes.org). Only very few variants will be medically relevant, perhaps—in case of a monogenic disorder—just one. Bioinformatics tools such as ANNOVAR (Wang *et al*., 2010) and MutationTaster (Schwarz *et al*., 2010) may then be used for comprehensive annotations and predictions about the expected pathogenicity of a variant. However, such classifiers have high false positive and negative error rates. They may therefore serve only for prioritization but cannot replace the assessment of human experts.

*to whom correspondence should be addressed

APPROACH

In a genetic disease of unknown cause, the association with a new disease gene may be shown either by a functional assessment of the detected variants or by statistical evidence. For a functional assessment, expert knowledge about a suitable test essay or genetically modified model organisms is required. For statistical evidence, a sufficient number of patients of the same disease group with mutations in the same gene are required. However, usually a clinical geneticist analyzing a patient's exome does not have access to many similar cases because the disease is so rare. Further, she may not be skilled to do the functional assessment immediately and happened to identify a new gene of interest after filtering the patients variants. Hence, such a geneticist is interested in finding other individuals with mutations in the same gene or scientists that are performing basic research on this gene. Web-based expert networks proved to be efficient tools for knowledge management in various scientific fields. There are knowledge bases for disease, gene, and protein centered information (www.ncbi.nlm.nih.gov/omim, www.geneontology.org, www.wikiproteins.org). However, there is no platform that allows the scientific exchange of experts about specific variants detected in NGS experiments. GeneTalk aims at providing such a web-based platform that enables to improve expert annotations on human genetic variants in a community approach.

APPLICATION

GeneTalk is an exchange platform that allows users to look for variant specific information and makes human expertise searchable (Figure 1). Any sequence variant with respect to the human reference genome, based on the GRCh37 assembly, is annotatable. The user decides to whom an annotation is visible. One user may link to scientific articles that are relevant in context with a certain variant or that even provide evidence that a mutation is disease causing, A second user might comment on this annotation to express her concern because she views the detected variant as an technical artefact. A third user might state that she has seen patients with this genotype and is not sure about the statistical significance of the association with the phenotype. All annotations and comments of GeneTalk users about a certain genomic position can be read like a locus specific conversation thread. The trustworthiness of an annotations can be rated by users as well as the likelihood of a mutation to be disease causing (Figure 1). If there is consensus

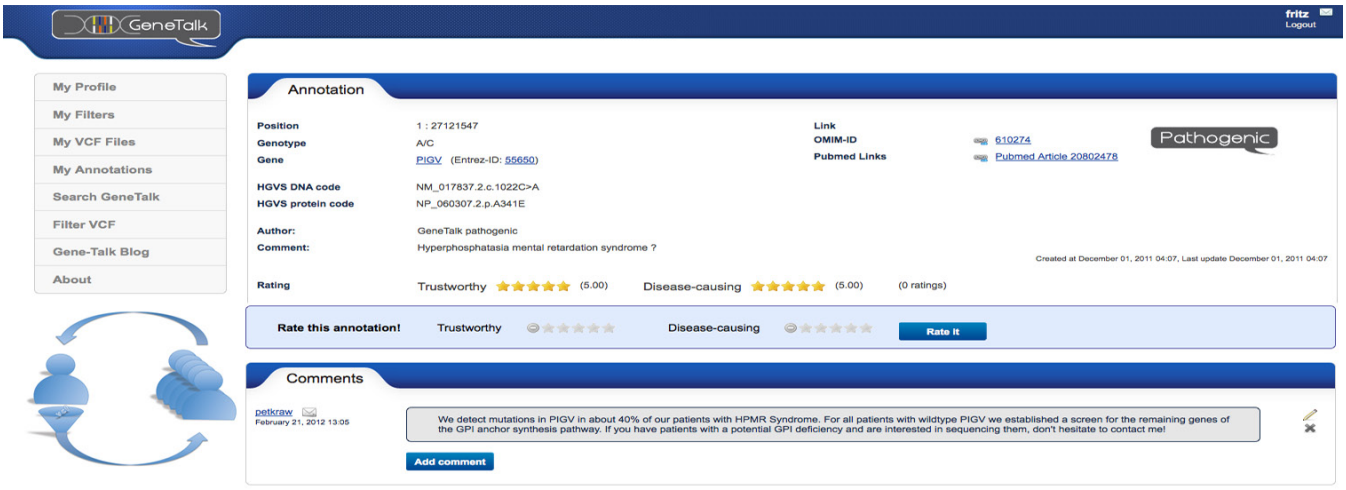


Fig. 1. GeneTalk a communication platform for sequence variants: A user filters sequence variants down to a small set of potentially disease relevant mutations. She then searches for detailed information annotated by the GeneTalk community for these variants. In GeneTalk users may annotate and comment genetic variants. Annotations and comments may link to relevant literature or discuss experimental and clinical findings. Based on this locus specific information GeneTalk users may rate the trustworthiness of an annotation and the potential of a mutation to be disease-causing. This screenshot is taken from fritz' account who is looking at the annotation of a mutation in the gene PIGV. The GeneTalk community finds this annotation trustworthy and rates the described mutation as highly likely to cause a syndrome called hyperphosphatasia with mental retardation. The user petkraw left a comment for this variant. He seems to have some expertise in this disease and might be an interesting person to contact for fritz.

in the GeneTalk community that a certain mutation is pathogenic and its annotation is trustworthy, this mutation is added to the annotation track *pathogenic*. This annotation track is thus curated in a collaborative effort of all GeneTalk users.

GeneTalk also assists users in filtering genetic variants from NGS projects. A user that has a patient's informed consent to analyze the clinical data may upload sequence variants to GeneTalk in variant call format, VCF (Danecek *et al*., 2011), version 4.0 and above. In order to reduce the initial VCF to a set of potentially disease relevant mutations, the user can apply certain filter settings first: The list of variants could be restricted to e.g. only nonsynonymous, homozygous variants with the functional and inheritance filter. Common variants can be filtered out by a genotype frequency filter that is based on high quality NGS data sets from HapMap, the 1000 genomes project and the 5000 exomes project. If a linkage analysis has been performed, a genomic interval may be set to limit the search space or gene panels may be applied as *in silico* filters to restrict the analysis to certain molecular pathways. In a rare recessive monogenic disease, the mode of inheritance and the genotype frequency filter that is set to one per thousand usually reduce the number of candidate mutations down to a few hundreds in a patient's exome. These variants may then be further analyzed for *disease causing* annotations. If the pathogenic mutation of this case has not yet been described in the literature and no *pathogenic* annotation exists, the user can look for annotations that discuss patients with similar phenotypes or basic research scientists that talk about unpublished experimental data for this gene. Such an annotation can serve as a conversation starter and the users can simply contact the author by clicking on the envelope symbol (Figure 1). Currently, the annotation database contains over 32.000 clinically relevant entries from dBSNP. A video tutorial on www.gene-talk.de illustrates how an exome data set may be analyzed in GeneTalk: In a few easy steps the variant data of a simulated patient with Hyperphosphatasia with Mental Retardation sydnrome is filtered down to the disease causing mutation.

CONCLUSIONS

GeneTalk provides an intuitive web-based interface for geneticists that analyze human sequence variants. GeneTalk is a platform for efficient knowledge management of genetic variants and simplifies the scientific discussion and interpretation especially of rare mutations.

ACKNOWLEDGEMENT

P.M.K. thanks Li Chun Su for insightful notes. *Funding*: Deutsche Forschungsgemeinschaft (DFG KR 3985/1-1) grant to P.M.K.

REFERENCES

Bamshad M.J., et al. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Genet Rev*, **1**2: 745-55.

Robinson, P.N., et al. (2011) Strategies for exome and genome sequence data analysis in disease-gene discovery projects. *Clin Genet*, **80:** 127-32.

Worthey, E.A., et al. (2011) Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease, *Genetics in Medicine*, **13**, 255-62.

Wang, K., et al. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data., *Nucleic Acids Res*, **38**, 164.

Schwarz, J.M., et al. (2010) MutationTaster evaluates disease-causing potential of sequence alterations,*Nat Methods*, **7**, 575-6.

Fokkema, I.F.A.C. (2011) LOVD v.2.0:The Nect Generation in Gene Variant Databases,*Human Mutation*, **32**, 557-563.

Danecek, P., et al. (2011) The variant call format and VCFtools,*Bioinformatics*, **27**, 2156-8.

2.4 Reduktion auf Gene mit zusammengesetzt-heterozygoten Varianten in nicht-verwandten Familien

Kamphans, T., Sabri, P., Zhu, N., Heinrich, V., Mundlos, S., Robinson, P.N., Parkhomchuk, D., and Krawitz, P.M. (2013). Filtering for compound heterozygous sequence variants in non-consanguineous pedigrees. PLoS One 8, e70151.

Wenn ein rezessives Allel eines Gens als Folge einer Neumutation entsteht, so ist das Individuum nicht erkrankt, da das zweite Allel mit der Wildtyp-Sequenz ausreicht, um die biologische Funktion zu erfüllen. Eine Person mit einer solchen pathogenen Sequenzvariante wird als Anlageträger bezeichnet. Da diese heterozygoten Neumutationen keinem Selektionsdruck unterliegen und in der Regel an vielen Positionen eines Gens auftreten können, weisen Personen, die von einer rezessiven Erkrankung betroffen sind und deren Eltern nicht verwandt sind oder aus dem gleichen Isolat stammen, üblicherweise zwei verschiedene pathogene Allele auf (Compound Heterozygtie).

Wir haben einen Algorithmus entwickelt, der es ermöglicht, auf NGS-Datensätzen von mehreren Familienmitgliedern, von denen eines von einer rezessiven Erkrankung betroffen ist, gezielt auf Gene zu filtern, in denen compound-heterozygote Sequenzvarianten vorliegen. Bereits bei einer Trio-Exom-Sequenzierung, das heißt, bei einer Bestimmung der Exom-Genotypen von Eltern und Kind, ist es möglich, unter Anwendung eines Frequenzfilters, d. h. eines Filters auf compound-heterozygote Varianten mit negativer Auswirkung auf Proteinebene, die Anzahl der Kandidatengene von ca. 20.000 üblicherweise auf unter zehn zu reduzieren. Mit diesem Priorisierungsansatz ist es uns gelungen, einige Fälle angeborener GPI-Ankerstörungen sowie weiterer seltener monogener rezessiver Erkrankungen zu diagnostizieren.

Mit jedem weiteren betroffenen oder auch nicht betroffenen Familienmitglied, welches für diese Filterung zur Verfügung steht, reduziert sich die Anzahl der Kandidaten-Gene weiter. Im Fall einer Quadro-Sequenzierung, Mutter, Vater und zwei betroffene Kinder mit Mabrys Syndrom, konnte damit die Anzahl der Kandidatengene auf drei, *MUC16*, *NBPF10* und *PIGO*, reduziert werden.

Eine weitere Strategie bei der Priorisierung der verbleibenden Varianten besteht darin, die Länge der kodierenden Sequenz des Gens sowie die Häufigkeit seltener, heterozygoter Varianten, die aus großen populationsgenetischen Studien gesunder Probanden ermittelt wurden, einzubeziehen. In dem genannten Beispiel weisen sowohl *MUC16* als auch *NBPF10* hohe Werte für cDNA Länge sowie eine mittlere Heterozygotierate auf, im Gegensatz zu PIGO, das daher als das wahrscheinlichste Krankheitsgen infrage kam. Das Genprodukt von *PIGO* ist eine Transferase, die Ethanolamin an die letzte der drei Mannose-Reste des GPI-Ankers bindet und stellt damit einen der letzten Schritte in der Ankersynthese dar. Somit ist *PIGO* also auch aus Sicht des molekularen pathways der vielversprechendste Kandidat. Dies konnte in funktionellen Experimenten und bei weiteren Patienten mit Mabry Syndrom verifiziert werden. Es stellt damit ein weiteres Krankheitsgen für Hyperphosphatasie mit mentaler Retardierung dar (HPMRS2, OMIM ID; 614749).

OPEN ACCESS Freely available online

PLOS ONE

Filtering for Compound Heterozygous Sequence Variants in Non-Consanguineous Pedigrees

Tom Kamphans[1], Peggy Sabri[2], Na Zhu[2], Verena Heinrich[2], Stefan Mundlos[2,3], Peter N. Robinson[2], Dmitri Parkhomchuk[2], Peter M. Krawitz[2,3]*

1 Smart Algos, Berlin, Germany, **2** Institute for Medical Genetics and Human Genetics, Charité Universtätsmedizin, Berlin, Germany, **3** Max Planck Institute for Molecular Genetics, Berlin, Germany

Abstract

The identification of disease-causing mutations in next-generation sequencing (NGS) data requires efficient filtering techniques. In patients with rare recessive diseases, compound heterozygosity of pathogenic mutations is the most likely inheritance model if the parents are non-consanguineous. We developed a web-based compound heterozygous filter that is suited for data from NGS projects and that is easy to use for non-bioinformaticians. We analyzed the power of compound heterozygous mutation filtering by deriving background distributions for healthy individuals from different ethnicities and studied the effectiveness in trios as well as more complex pedigree structures. While usually more then 30 genes harbor potential compound heterozygotes in single exomes, this number can be markedly reduced with every additional member of the pedigree that is included in the analysis. In a real data set with exomes of four family members, two sisters affected by Mabry syndrome and their healthy parents, the disease-causing gene *PIGO*, which harbors the pathogenic compound heterozygous variants, could be readily identified. Compound heterozygous filtering is an efficient means to reduce the number of candidate mutations in studies aiming at identifying recessive disease genes in non-consanguineous families. A web-server is provided to make this filtering strategy available at www.gene-talk.de.

Citation: Kamphans T, Sabri P, Zhu N, Heinrich V, Mundlos S, et al. (2013) Filtering for Compound Heterozygous Sequence Variants in Non-Consanguineous Pedigrees. PLoS ONE 8(8): e70151. doi:10.1371/journal.pone.0070151

Editor: Kai Wang, University of Southern California, United States of America

Received April 20, 2013; **Accepted** June 20, 2013; **Published** August 5, 2013

Funding: This work was supported by a grant from the Deutsche Forschungsgemeinschaft grant to P.M.K. (DFG KR 3985/1-1). GeneTalk is supported by a grant from the Bundesministerium für Wirtschaft und Technologie (BMWT; 03EGSBB082). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Tom Kamphans, who is affiliated to SmartAlgos, is a self-employed software developer and consultant (SmartAlgos). This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: peter.krawitz@gmail.com

Background

In recessive genetic disorders, both copies of a certain gene are defective. For autosomal recessive genes, this means that the maternally as well as the paternally transmitted copy of a gene harbors a pathogenic mutation. The occurrence of a pathogenic mutation can be viewed as a random process, and many different pathogenic mutations have arisen for recessive genes in the human population over time. This also means the lower the kinship of the parents the higher is the chance that two different mutant alleles of the disease gene are present in a patient affected by a recessive disease, whereas in a closely related, consanguineous partnership it is more likely that an affected child will inherit the same pathogenic allele from both parents and thus be homozygous for the disease causing mutation. This translates to a simple rule of thumb: If the parents are non-consanguineous, the most likely explanation for a recessive disease is compound heterozygosity for two different pathogenic mutations. Exceptions from this rule of thumb may be founder mutations in certain populations and specific gain of function mutations in certain genes such as e.g. *FGFR2*.

A challenge in filtering sequence variants for compound heterozygotes is that one has to figure out whether the two heterozygous variants affect different copies of a gene or the same copy. Usually, that cannot be determined from a single DNA sequence, if the read length is less than the distance between the variants or if it is not possible to phase the haplotypes by any other means. However, when sequence variants of more than one family member are available, one can exclude certain variants based on rules of Mendelian inheritance. We will describe a set of rules that we used for compound heterozygous filtering and analyze how effectively the sequence variants can be reduced in certain case scenarios.

Methods

We implemented a compound heterozygous filter in Ruby inside the GeneTalk framework [1]. We assume that the phenotype is fully penetrant and that all sequenced individuals are either affected or not affected. The first two rules work on a single variant level:

1) A variant has to be in a heterozygous state in all affected individuals.
2) A variant must not occur in a homozygous state in any of the unaffected individuals.

If a variant were homozygous in an unaffected individual it could not be disease causing, otherwise the individual would have

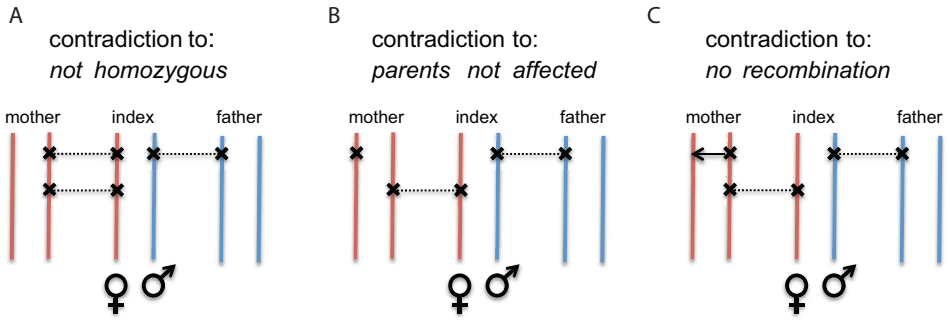


Figure 1. Compound Heterozygote Filtering Rules. If both parents of the index patient are unaffected it is not possible that one of the heterozygous disease causing mutations is present in a heterozygous state in both parents unless a recombination occurred between this variant and the second compound heterozygous mutation.
doi:10.1371/journal.pone.0070151.g001

to be affected, as both copies of the gene harbor the same mutation.

If the genotypes of both parents of an affected child are available and they are both unaffected there is a third rule that is very powerful in reducing the variants:

3) A variant that is heterozygous in an affected child must be heterozygous in exactly one of the parents.

This rule is a compact version of:

3a) The variant must not be heterozygous in both parents.

3b) The variant must be present in at least one of the parents.

Rule 3a is applicable only if no recombination occurred between the tested loci in any of the parents. However, most genes have an extension considerably less than one megabase in the genome and thus the probability of a recombination is usually far below one per cent and the assumption of no recombination is well justified. In Figure 1, we illustrate why a variant that does not fulfill 3a may be removed as not pathogenic. If we keep such a variant it will result in a violation of one of prerequisites for a compound heterozygous mode of inheritance. Without loss of generality we may consider two heterozygous variants in an affected individual. One of them is in a heterozygous state in both unaffected parents. If the variant, for which mother and father are heterozygous, is transmitted by both of them to the index patient, then the variant would be homozygous in the child and would be therefore removed based on the first rule (Figure 1 A). The index patient could be heterozygous for both variants if both variants occurred on different copies of the gene in one of the parents. However, in this case this particular parent has to be affected as well and the third rule does not apply, as its assumption is not fulfilled (Figure 1B). The third case describes a scenario in which one of the parents has the two heterozygous mutations on the same copy of the gene, while the other parent is heterozygous for only one of them. In this scenario the child could be compound heterozygous only if a recombination happened in the germline of its ancestor (Figure 1C). As already mentioned, this case is so unlikely that we exclude it.

Rule 3b is applicable only if we assume that no *de novo* mutations occurred. The number of *de novo* mutations is estimated to be below five per exome per generation [2,3], thus, the likelihood that an individual is compound heterozygous and at least one of these mutations arose *de novo* is low. If more than one family member is affected, *de novo* mutations are even orders of magnitudes less likely as a recessive disease cause. On the other hand, excluding these variants from the further analysis helps to remove many sequencing artifacts. In linkage analysis for example it is common practice of data cleaning to exclude variants as Mendelian errors if they cannot be explained by Mendelian inheritance.

When the filtering rules 1–3 have been applied on a single variant level, the fourth and fifth rule test on a gene level, whether

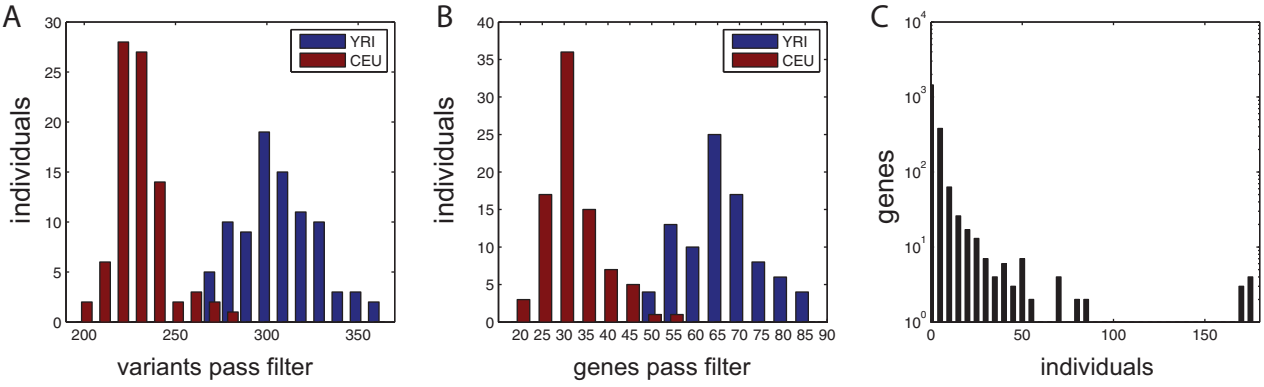
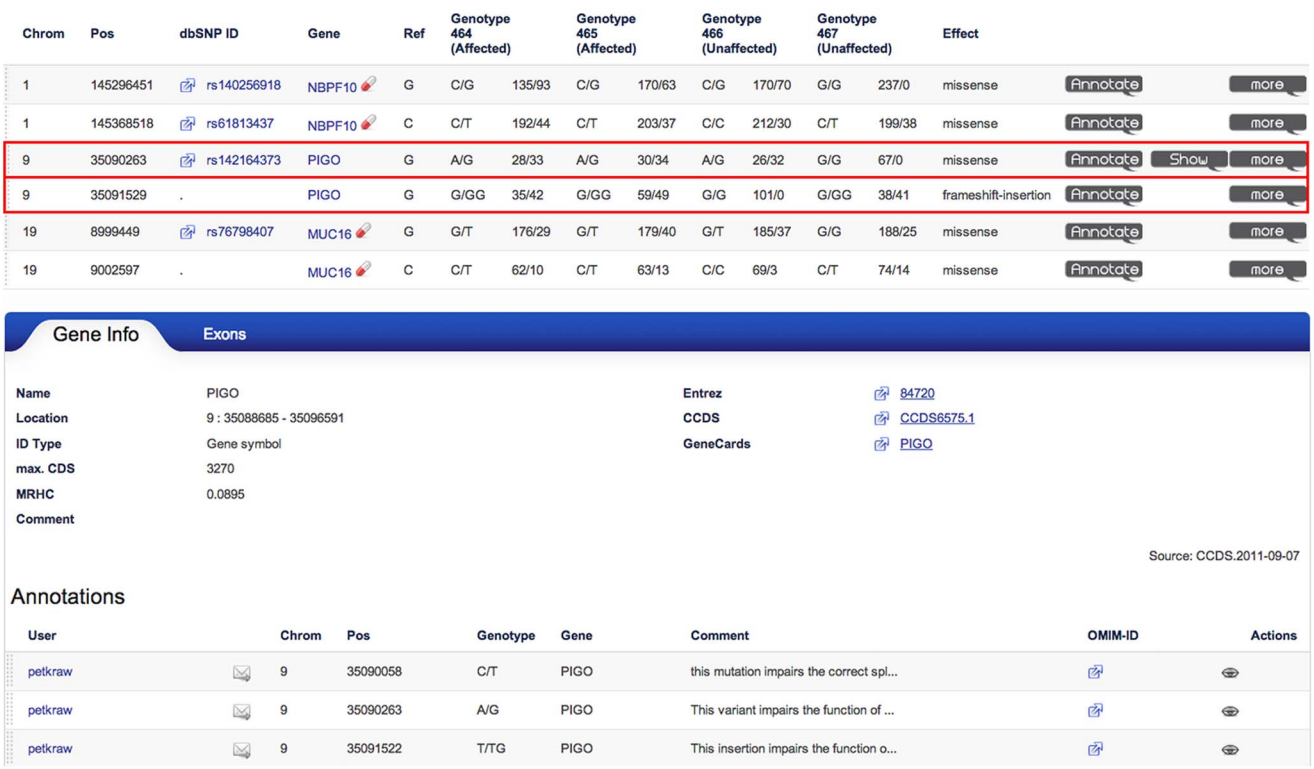


Figure 2. Exomes of 85 European individuals (CEU) as well as 88 African individuals (YRI) were filtered for rare compound heterozygous candidate variants. A) In average around 230 variants pass the filter in CEU exomes and 309 in YRI exomes. B) The potential compound heterozygotes are distributed over 31 genes in CEU individuals and 67 genes in YRI individuals. C) Altogether 1998 genes harbored potential compound heterozygous variants in the tested individuals and compound heterozygotes in 1066 genes occurred only in singular cases.
doi:10.1371/journal.pone.0070151.g002



Chrom	Pos	dbSNP ID	Gene	Ref	Genotype 464 (Affected)		Genotype 465 (Affected)		Genotype 466 (Unaffected)		Genotype 467 (Unaffected)		Effect			
1	145296451	rs140256918	NBPF10	G	C/G	135/93	C/G	170/63	C/G	170/70	G/G	237/0	missense	Annotate		more
1	145368518	rs61813437	NBPF10	C	C/T	192/44	C/T	203/37	C/C	212/30	C/T	199/38	missense	Annotate		more
9	35090263	rs142164373	PIGO	G	A/G	28/33	A/G	30/34	A/G	26/32	G/G	67/0	missense	Annotate	Show	more
9	35091529	.	PIGO	G	G/GG	35/42	G/GG	59/49	G/G	101/0	G/GG	38/41	frameshift-insertion	Annotate		more
19	8999449	rs76798407	MUC16	G	G/T	176/29	G/T	179/40	G/T	185/37	G/G	188/25	missense	Annotate		more
19	9002597	.	MUC16	C	C/T	62/10	C/T	63/13	C/C	69/3	C/T	74/14	missense	Annotate		more

Gene Info | Exons

Name	PIGO	Entrez	84720
Location	9 : 35088685 - 35096591	CCDS	CCDS6575.1
ID Type	Gene symbol	GeneCards	PIGO
max. CDS	3270		
MRHC	0.0895		
Comment			

Source: CCDS.2011-09-07

Annotations

User	Chrom	Pos	Genotype	Gene	Comment	OMIM-ID	Actions
petkraw	9	35090058	C/T	PIGO	this mutation impairs the correct spl...		
petkraw	9	35090263	A/G	PIGO	This variant impairs the function of ...		
petkraw	9	35091522	T/TG	PIGO	This insertion impairs the function o...		

Figure 3. Filtering results for compound heterozyotes in a case study. With the filter settings for genotype frequency <0.01, effect on protein level (functional filter: missense, nonsense, stop loss, splice site, insertions or deletions), and compound heterozygous yields six variants in three genes. *MUC16* and *NBPR10* are both genes from large gene families known for their high variability and detection artifacts due to pseudogenes. The heterozygotes in *PIGO* remain as the likeliest candidates. The *Show* icon at the right end of the line links to an expert curated annotation database that indicates that the mutation in *PIGO* is causing Hyperphosphatasia with mental retardation syndrome and has been published in [9]. The gene view for *PIGO* lists all variant annotations for this gene and links to further knowledge bases. The length of the coding sequence of the longest transcript (max. CDS) and the mean number of rare heterozygous variant calls per exome (MRHC) are important parameters for the assessment of candidate genes.
doi:10.1371/journal.pone.0070151.g003

enough variants remained to fulfill the requirements for a compound heterozygous mode of transmission and whether they were transmitted biparentally:

4) A gene must have two or more heterozygous variants in each of the affected individuals.
5) There must be at least one variant transmitted from the paternal side and one transmitted from the maternal side.

We did not use two as an upper bound in these rules as not necessarily all of the variants that pass these rules have to be pathogenic. However, as we will discuss later, a gene with many variants passing all five rules, is less likely to be a disease gene. The fifth rule makes sure that only genes pass the filter for which there are at least two heterozygous variants in the affected individuals that are transmitted in a biparental mode. Another way of phrasing this rule is: There must not be two identical haplotypes around the disease gene in an unaffected and an affected individual. If all heterozygous mutations of a gene in an affected individual match all the heterozygous mutations of the same gene in an unaffected individual of the family, then we exclude this gene. Imagine a scenario where sequence variants of only one grandfather or grandmother are available as unaffected control but not directly the sequence variants of the parents of the affected child. Let us assume that this index patient has exactly two heterozygous variants in a gene that match the genotypes of this gene in the grandmother. Excluding recombination this means that one of the parents of the child was either carrier of these two heterozygous variants or not. However, this means that one of these variants cannot be disease causing or rule three would be broken. The intervals for which we counted and compared the heterozygous genotypes were defined by the gene start and end points that we derived from Ensembl/BioMart [4]. All filters that we applied prior to the compound heterozygote filter are available through the GeneTalk platform. The allele frequency filter is based on genetic variation data from the 1000 genomes project [5] as well as the 5000 exomes project [6]. The effect of the variants on the protein level which is subject to the functional filter was predicted by ANNOVAR [7].

The length of a gene, as well as the variability of its sequence in a healthy reference population affects the probability to observe rare, heterozygous variants in a test individual. We derived the length of the coding sequence of the longest transcript per gene ($\max.CDS$) and determined the mean number of rare, heterozygous variant calls in healthy individuals ($MRHC$) based on the 5000 exomes data [8] to assist in the interpretation of candidate genes after filtering. The $MRHC$ was computed by adding the frequencies of heterozygous variant calls for all positions in a gene that were below 0.01:

$$MRHC = \sum_{i \in gene} \frac{r_i}{c_i}, with \frac{r_i}{c} \leq 0.01$$

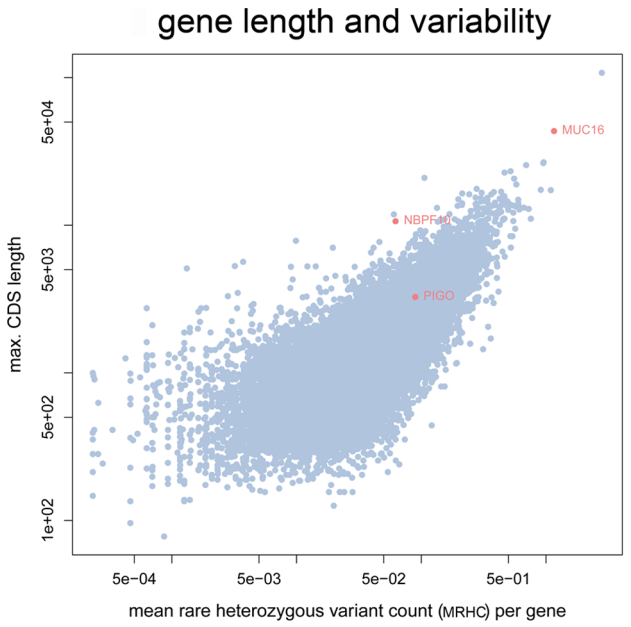


Figure 4. The length of the coding sequence and the mean number of rare alleles per gene. In an average healthy individual from the 5000 exomes project there is more than one rare heterozygous variant in *MUC16* that has an allele frequency below 0.01 in the reference population. In contrast, the coding sequence of *PIGO* is much shorter and rare heterozygous variants occur in less than 8 out of 1000 exomes.
doi:10.1371/journal.pone.0070151.g004

where c_i is the total genotype count at position i and r_i is the heterozygous genotype count at position i.

Results

We tested the effectiveness of the compound heterozygotes filter in GeneTalk for single samples as well as in more complex pedigree structures. If single individuals are analyzed only rules 1 and 4 can be used. Rules 2 and 5 require data of at least one additional unaffected family member and rule 3 is applicable only if sequence data of both parents is available. Any additional filter settings should be used prior to rule 4 as this rule is more effective in reducing the number of candidate genes the lower the expected number of heterozygous variants is per gene. We used an allele frequency cutoff of 1% for heterozygous variants, removed all synonymous variants and known sequencing artifacts before applying the compound heterozygous filter. We choose the frequency cutoff of 1% as an upper bound, as this is above the allele frequency of the most common pathogenic alleles in cystic fibrosis (MIM 219700), one of the more common autosomal recessive disorders.

With these parameter settings, we filtered 85 European exomes (CEU) as well as 88 African exomes (YRI) available from the 1000 genomes project [5]. All these individuals are healthy and the numbers of variants as well as genes passing our filter settings serve as a background distribution that one has to expect when filtering single exome data. In the CEU exomes in average 230 variants distributed over 31 genes passed the filter, whereas in average 309 variants distributed over 67 genes passed the filter in the Yorubian exomes (Figure 2A and 2B). In the 173 tested individuals we identified variants as possible compound heterozygotes in altogether 1998 genes, and 1066 of these genes were unique to only one of the tested samples (Figure 2C).

As all these exomes were of healthy individuals it would be difficult to identify a single additional gene that passes the filters in a patient with a rare recessive disorder due to the true disease causing compound heterozygotes.

We then analyzed the effectiveness of the compound heterozygous filter for cases in which more than only a single exome of a family is available. For this purpose we used two parent-child trios of European (NA12878, NA12891, NA12892) as well as Yoruban (NA19238, NA19239, NA19240) descent from the 1000 genomes project. We assigned the status of the affected index to the offspring (NA12878, NA19238), so that for both trios all five rules were applicable. As before, in advance to the compound heterozygous filter, we reduced the exome variants of the trios to rare variants with a presumable effect on the protein level and removed known calling artifacts. This prior filtering step yielded 1668 variants in the European trio and 2653 variants in the Yoruban trio. The compound heterozygous filter reduced this number down to 48 variants in 17 genes in the European individual NA12878, while 68 variants in 29 genes passed in the Yoruban individual NA19238. If we remove genes for which compound heterozygous variants also occurred in other unrelated individuals (Figure 1C and Table S1), only twelve candidate genes remained in NA12878 *(CACHD1, DPP4, CACNA1D, PLXNA1,*

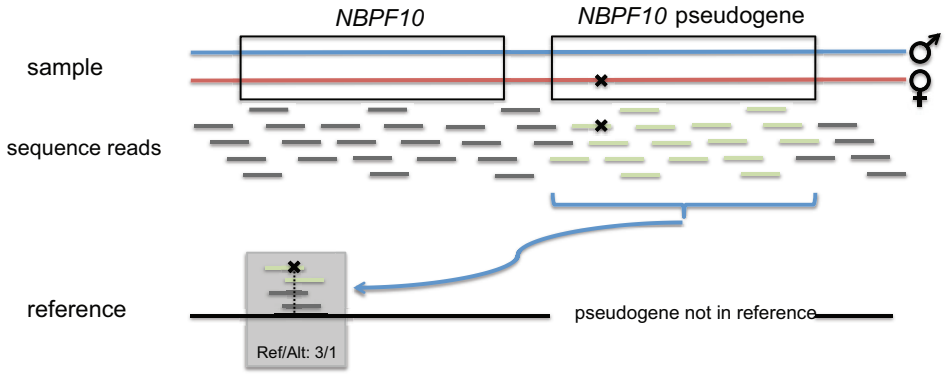


Figure 5. Illustration of mapping artifacts resulting in false positive variant detection. The illustrated sample carries a mutation in the maternal copy of a pseudogene of *NBPF10*. If the pseudogene is not included in the reference sequence, the reads originating from this pseudogene are mismapped. This may result in a false variant call. Indicative for false genotype calls are proportions of reads supporting the alternate allele that strongly deviate from 0.5 or 1.
doi:10.1371/journal.pone.0070151.g005

MECOM, PSMD2, TBC1D4, DCAF5, ZNF614, MYH9, PKDREJ, MAGEC1) and in NA19238 *(CD180, ERAP2, EPB41L2, COL1A2, MUC16, AASS, PZP, ZMYM2, ZNF423, LRRC48, DMC1, GPR64)*. Thus, in comparison to filtering a single individual the inclusion of the parental genotypes reduces the number of variants by a factor of around five. The removal of highly variable genes as observed by the analysis of single individuals cut the remaining variants in half.

We continued our analysis with exome data from a family with two daughters affected by Mabry syndrome also known as hyperphosphatasia with mental retardation syndrome (MIM 614749) [9]. The parents were of European descent and unrelated. Altogether 27584 distinct variants were detected in the coding regions of all four analyzed family members. Applying the genotype frequency filter and removing known sequencing artifacts reduced this number to 2208 variants. 1446 of these variants had a predicted effect on the protein level. The compound heterozygous filter reduces this set to six variants in three distinct genes, *NBPF10*, *MUC16*, and *PIGO* (see Figure 3).

MUC16 is a member of a large family of mucin coding genes and *NBPF10* is from the neuroblastoma breakpoint family, which consists of 22 genes and pseudogenes that arose by gene duplication [10,11]. Both genes are highly variable and harbor many low frequency variants [8]. Genes with a long coding sequence (CDS) and many rare heterozygous variants are more likely to appear as candidate genes after compound heterozygous filtering. In Figure 4 a scatterplot is shown for the mean number of rare heterozygous variants and the CDS length of the longest transcript for each gene. *MUC16* is not only an extraordinarily large gene with over 40kb coding sequence, in average there were also 1.16 rare heterozygous variant calls per healthy individual with an allele frequency below 0.01 in the reference population of the 5000 exomes project. In e.g. NA19238, there were also two such rare heterozygous calls that passed the filter. Genes with pseudogenes, such as *NBPF10* are also prone to genotyping artifacts in reference guided resequencing due to mismapped reads: A variant that is classified as a heterozygous genotype is likely to be a false positive call, if the coverage of this position is high, however the proportion of reads supporting the alternate allele deviates strongly from the value of 0.5 that is expected for a heterozygous genotype. Figure 5 illustrates the mismapping of reads originating from a pseudogene of *NBPF10* that result in such genotyping artifacts. For *NBPF10* many such calling artifacts have been reported in GeneTalk and were automatically excluded. Also the variants detected in *NBPF10* and *MUC16* are highly suggestive for such false calls, for instance in the heterozygously called variant with the dbSNP ID rs61813437 the alternate allele is only supported by 44 out of altogether 236 sequence reads in the first individual and likewise in the others (alternate allele/reference allele: 203/37, 212/30, 199/38) and in dbSNP this variant is also listed as "not validated". After this quick assessment of the trustworthiness of the variant calls, *PIGO* remains as the most promising disease candidate gene in this case and the mutations were indeed confirmed as the causative mutations of the disorder [9].

With the additional second affected individual in this case scenario the number of candidate genes was markedly lower than in the two trios. In autosomal recessive disorders all affected family members share both haplotypes around the disease locus and are therefore identical by descent for both copies of the gene (IBD = 2) [12]. Two siblings are in average IBD = 2 in only one quarter of their genome which reduces the number of candidate genes likewise with every additional affected sibling. If we analyzed in this family only one of the affected sisters at a time in a trio approach, we would have seen the additional candidate genes *HRNR*, *PLCD1*, and *MUC5B*. KGGseq [13], a statistical framework for analyzing exome data of multiple individuals has only implemented rule 2 for compound heterozygous filtering and reduced to 223 rare candidate variants.

Conclusions

In this work we developed a filter for identifying compound heterozygotes in exome data of one or more individuals of a family. The rule set on which our filter is based, is comprehensive for analyzing multiple samples and advances the prioritization of compound heterozygous candidate variants beyond existing tools for analyzing data of exome sequencing studies [14,15]. We showed that filtering for compound heterozygous mutations is an effective means in identifying disease candidate genes especially if multiple family members are available for the analysis. In a trio analysis with exome data of the parents and one affected child, typically, mutations in only about a dozen candidate genes remain. This manageable number of remaining genes can then be assessed based on the expertise of the investigator or further prioritized by suitable tools[16–18]. We implemented the compound heterozygous filter as an intuitively usable web service that allows a quick reduction of the exome variants to such a candidate set. The filter as well as the European and Yoruban trios are accessible via the demo account at www.gene-talk.de [1].

Web Resources

The URL for data presented herein are as follows:

1000 genomes project website. Available: http://www.1000genomes.org. Accessed 2013 May 2.

NHLBI Exome Sequencing Project (ESP) website. Available: http://evs.gs.washington.edu/EVS/. Accessed 2013 May 2.

GeneTalk website. Available: www.gene-talk.de. Accessed 2013 May 2.

Supporting Information

Table S1.
(PDF)

Acknowledgments

The authors would like to thank the NHLBI GO Exome Sequencing Project and its ongoing studies which produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926) and the Heart GO Sequencing Project (HL-103010).

Author Contributions

Conceived and designed the experiments: TK PMK. Performed the experiments: PS NZ VH DP. Analyzed the data: PMK VH DP PNR SM . Wrote the paper: PMK PNR.

References

1. Kamphans T, Krawitz PM (2012) GeneTalk: an expert exchange platform for assessing rare sequence variants in personal genomes. Bioinformatics 28: 2515–2516.

2. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485: 242–245.

3. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328: 636–639.
4. Haider S, Ballester B, Smedley D, Zhang J, Rice P, et al. (2009) BioMart Central Portal–unified access to biological data. Nucleic Acids Res 37: W23–27.
5. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
6. (ESP) NGESP (2013) Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP),. Seattle, WA.
7. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38: e164.
8. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, et al. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337: 64–69.
9. Krawitz PM, Murakami Y, Hecht J, Kruger U, Holder SE, et al. (2012) Mutations in PIGO, a member of the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation. Am J Hum Genet 91: 146–151.
10. Karro JE, Yan Y, Zheng D, Zhang Z, Carriero N, et al. (2007) Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res 35: D55–60.
11. Pei B, Sisu C, Frankish A, Howald C, Habegger L, et al. (2012) The GENCODE pseudogene resource. Genome Biol 13: R51.
12. Rodelsperger C, Krawitz P, Bauer S, Hecht J, Bigham AW, et al. (2011) Identity-by-descent filtering of exome sequence data for disease-gene identification in autosomal recessive disorders. Bioinformatics 27: 829–836.
13. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–2993.
14. Li MX, Gui HS, Kwan JS, Bao SY, Sham PC (2012) A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Res 40: e53.
15. Sifrim A, Van Houdt JK, Tranchevent LC, Nowakowska B, Sakai R, et al. (2012) Annotate-it: a Swiss-knife approach to annotation, analysis and interpretation of single nucleotide variation in human disease. Genome Med 4: 73.
16. Bornigen D, Tranchevent LC, Bonachela-Capdevila F, Devriendt K, De Moor B, et al. (2012) An unbiased evaluation of gene prioritization tools. Bioinformatics 28: 3081–3088.
17. Tiffin N, Andrade-Navarro MA, Perez-Iratxeta C (2009) Linking genes to diseases: it's all in the data. Genome Med 1: 77.
18. Moreau Y, Tranchevent LC (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet 13: 523–536.

2.5 Beurteilung der Datenqualität von Exomen durch Vergleich mit Datensätzen großer populationsgenetischer Studien

Heinrich, V., Kamphans, T., Stange, J., Parkhomchuk, D., Hecht, J., Dickhaus, T., Robinson, P.N., and Krawitz, P.M. (2013). Estimating exome genotyping accuracy by comparing to data from large scale sequencing projects. Genome Medicine 5, 69.

Um NGS-Untersuchungsmethoden sinnvoll in den klinischen Alltag integrieren zu können, ist es wichtig, deren diagnostische Erfolgsrate bei einer spezifischen medizinischen Fragestellung zu bestimmen. Bei vielen allgemeinen Indikationen, zum Beispiel auch einer nicht-syndromalen geistigen Entwicklungsverzögerung, unter die auch GPI-Ankerstörungen fallen können, erhält die Exom-Sequenzierung zunehmend Einzug in die Routinediagnostik. Mit dieser Entwicklung steigt auch die Bedeutung von Qualitätskriterien, die benötigt werden, um Mindestanforderungen an Datensätze zu definieren.

Bei der bioinformatischen Prozessierung der Rohsequenzdaten zu den Sequenzvarianten kommen unterschiedliche Software-Werkzeuge zum Einsatz. Die gesamte Analysestrecke wird häufig als Pipeline bezeichnet. Aufgrund der großen Auswahlmöglichkeit und den häufigen Verbesserungen einzelner Software-Komponenten kann es zum Teil erhebliche Unterschiede zwischen den Ergebnissen eines Rohdatensatzes geben, der mit unterschiedlichen Pipelines verarbeitet wurde.

Eine vergleichbare diagnostische Erfolgsrate ist jedoch nur dann zu erwarten, wenn sich die Genauigkeit der letztlich aufgeführten Sequenzvarianten ähnelt. Wir haben einen Algorithmus entwickelt, der es uns erlaubt, die Güte eines Exom-Datensatzes einzuschätzen, indem das Profil der detektierten Sequenzvarianten einer Person mit den qualitativ hochwertigen Daten der Individuen des 1000 Genom-Projektes verglichen wird.

Der Methode liegt eine Ähnlichkeitsmetrik zugrunde, mit der sich anhand der Positionen, an denen sich die Genotypen in zwei Exomen unterscheiden, ein Wert berechnen lässt, der als Abstand aufgefasst werden kann. Ein geringer Abstand steht dabei für eine hohe Ähnlichkeit der Datensätze und ein großer Abstand für Ungleichheit. Sowohl die tatsächliche genetische Variabilität zwischen zwei Individuen als auch Genotypisierungsfehler beeinflussen damit die Distanz.

Entstammen zwei Individuen nun der gleichen Bevölkerung, so ist es wahrscheinlicher, dass tatsächliche genetische Unterschiede an Positionen auftreten, die in dieser Population polymorph sind. Abweichungen an Positionen geringer Variabilität hingegen sind mit einer höheren Wahrscheinlichkeit auf Sequenzierfehler zurückzuführen.

Wenn der Anteil seltener Genotypen in einem Testdatensatz, der von einem Individuum einer Population stammt, die auch im 1KGP untersucht wurde, überproportional hoch ist, so weist dies auf eine mindere Datenqualität hin. Durch einen Gewichtungsfaktor, der eine negative Korrelation mit der Variabilität einer Position aufweist, kann die Sensitivität der Ähnlichkeitsmetrik für falsche Messergebnisse zusätzlich gesteigert werden. Anhand simulierter Datensätze mit definierter Fehlerrate haben wir für unsere Metrik Schwellenwerte ermittelt, die von Testdaten bestimmter Güte nicht überschritten werden sollten.

Genome Medicine

METHOD Open Access

Estimating exome genotyping accuracy by comparing to data from large scale sequencing projects

Verena Heinrich[1], Tom Kamphans[2], Jens Stange[3], Dmitri Parkhomchuk[1], Jochen Hecht[4,5], Thorsten Dickhaus[3], Peter N Robinson[1] and Peter M Krawitz[1,5*]

Abstract

With exome sequencing becoming a tool for mutation detection in routine diagnostics there is an increasing need for platform-independent methods of quality control. We present a genotype-weighted metric that allows comparison of all the variant calls of an exome to a high-quality reference dataset of an ethnically matched population. The exome-wide genotyping accuracy is estimated from the distance to this reference set, and does not require any further knowledge about data generation or the bioinformatics involved. The distances of our metric are visualized by non-metric multidimensional scaling and serve as an intuitive, standardizable score for the quality assessment of exome data.

Background

In recent years, next-generation sequencing (NGS)-based exome screens have become an invaluable tool in Mendelian disease gene discovery and are now being introduced as clinical diagnostic tools for genetic disorders of high phenotypic and genetic heterogeneity [1,2]. Various solutions for exome enrichment and sequencing exist and numerous algorithms for sequence read mapping and variant detection are in use [3-12]. There are recommendations for sequencing depth and benchmarks for the distribution of sequence read coverage over the target region. The common core of the diverse approaches to sequence exomes represents the consensus coding sequences as defined by the consensus coding sequence (CCDS) project [13]. The majority of publications [12] related to this field seem to confirm that a mean sequencing depth of this target region with high quality short sequence reads should be above 50-fold and more than 90% of the CCDS exons should be covered by at least 10 sequence reads for diagnostic purposes. Another recently introduced parameter for the quality assessment of multiple read alignments is the variance of the ratio of reads that support the alternate allele at heterozygous positions [14]. The lower this variance, the lower the error rate to be expected from amplification artifacts. The ratio of transitions versus transversions (ti/tv) and the proportion of variants that are already listed in databases of genetic variation such as the Single Nucleotide Polymorphism database (dbSNP) [15] are measures of quality that may be applied to the entire variant call set of an exome. The ti/tv ratio should be close to 3:1 for the CCDS exons, and the proportion of singletons should be below 10% [16]. However, the ti/tv ratio is influenced by the target region, whereas the number of novel variants may depend on the background population. The higher the amount of non-coding variants, the lower the ti/tv ratio, and higher proportions of novel variants may be observed if the sequenced individual is from a population that is poorly represented in the variant databases.

Although these parameters may serve as valuable indicators for quality they do not directly indicate the accuracy of a sequenced exome and to our knowledge there is no criteria for assessing whether the variants identified by whole exome sequencing represent a comprehensive list. Specifically, it is not possible to estimate an exome-wide false positive or negative rate for variant detection that is purely based on the quality scores of genotype

* Correspondence: peter.krawitz@charite.de
[1]Institute for Medical Genetics and Human Genetics, Charité Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany
Full list of author information is available at the end of the article

calls. Sequencing technology-specific error signatures [17] can yield artificial variant calls of erroneous high quality and result in an underestimated proportion of false calls, while poorly adjusted bioinformatics pipelines for data processing may lead to missed calls. In most NGS studies, a Phred-like quality score is provided for each called genotype. This score describes the confidence in a genotype call. Based on a certain likelihood model for genotypes, the Phred score represents the probability that the genotype call is wrong, given the reads in an alignment ($\text{Phred score} = -10log_{10}Pr(wrong\ \ genotype)$). In a model that assumes, for example, a Bernoulli distribution of the sequence reads at a heterozygous position, the Phred score of a heterozygous genotype would decrease the more the ratio of reads supporting the alternate allele deviates from the expected value of 0.5. However, this quality score not only depends on the raw data but also on the mapping algorithms and probability models that were used for variant calling. That means that processing the same raw data by different bioinformatics pipelines may result in different distributions of quality scores, suggesting different genotyping error profiles for the same exome. Even variant calling approaches that are based on similar Bayesian methods do not yield the same genotype probabilities due to different priors [18], and methods of quality score recalibration cannot completely adjust for that effect (Additional file 1, Table S1).

In order to enable interoperability and platform independence, we have developed a method to measure the accuracy of a set of variants by assessing its composition. In our metric, the distance between sets of variants from two exome samples is computed without considering genotyping quality scores. The basic idea is that the distance between variant sets of comparable quality is closer than the distance between variant sets of very different quality. In our comparison, the variant data of individuals of the 1000 genomes project [19] serve as a gold standard and we refer to them as the reference set. If the genotype concordance between the reference variant call set and the test variant call set is high, this suggests a comparable sequencing quality. We will show in the following that the distance of a test sample to this high quality reference data is an indicator for the genotyping accuracy of the exome.

Methods

Generation of exome data, accession of reference data and data processing

Genomic DNA of European individuals was enriched for the target region of all human CCDS exons with SureSelect Human All Exon Kit (Agilent, Santa Clara, USA) according to the manufacturer's protocol and sequenced on a HiSeq 2000 (Illumina, San Diego, USA), yielding more than 5 gigabases raw sequence data per exome. The Charité University Medicine ethics board approved this study, which conforms to the Helsinki Declaration, and we obtained informed consent of all participants.

Publicly available NGS raw data and variant calls of 1,063 individuals of different populations were downloaded from the ftp server of the 1000 genomes project [19]. Exome variants of these individuals served as reference variant sets in our work. Exome variants of the 5000 exomes project and of de Ligt *et al.* were used for testing the accuracy predictions [20,21].

Exomes of test samples were enriched with Human All Exon SureSelect baits from Agilent and sequenced on Illumina Genome Analyzer IIx and Illumina HiSeq 2000 as 100 bp single-end reads or paired-end reads according to the manufacturers' protocols. Short sequence reads were mapped by Novoalign (Novocraft, version 2.08) or BWA[22] to the reference sequence GRCh37. Variants were detected with default settings with SAMtools [23] or GATK [10] on bam-formatted alignments [22]. Variant calls in variant call format (vcf) [24] were restricted to single nucleotide changes and to the consensus exome target region of the 1000 genomes project. Additionally, sites that were classified as technical artifacts by the 1000 genomes project were ignored.

Distance function

The distance d_{ij} between any two samples x_i and x_j for all positions k in the target region (exome), where the called genotypes differ from the reference sequence in at least one sample, can be calculated by a weighted indicator function $I\left(x_i(k), x_j(k)\right) * W_{ij}(k)$, with:

$$I_{ij}(k) = I\left(x_i(k), x_j(k)\right) = \begin{cases} 1, if\ x_i(k) == x_j(k) \\ 0, if\ x_i(k) \neq x_j(k) \end{cases}$$

and $W_{ij}(k) = \dfrac{2}{f\left(x_i(k)\right) + f\left(x_j(k)\right)}$.

This means that for the same genotypes the indicator I is weighted by the reciprocal of the genotype frequency $f\left(x_i(k)\right)$, which is based on the reference set with an appropriate background population. To give an example, a genotype for individual i at a given position $x_i(k = chr6:79595096) = C/C$, $x_j(k = chr6:79595096) = C/C$, would refer to a genotype frequency $f\left(x_i(k)\right) = 0.999$, if 1 out of 1,000 individuals in the reference set differs from this genotype.

For genotypes that were present only in the test sample but not observed at all in the reference set, we set their frequency to $1/(n+1)$, where n is the total number of individuals in the reference set.

Based on that the distance d_{ij} is defined as:

$$d_{ij} = 1 - \frac{1}{C_{ij}} \sum_{k} I_{ij}(k) * W_{ij}(k)$$

where $C_{ij} = \sum_{k} W_{ij}(k)$ is used as a normalizing constant.

Therefore, a disagreement at a position of low variability in the reference set contributes more to the total distance than one at a highly variable position.

In the resulting distance matrix, D, pairs of individuals who are 'closely related' can be distinguished from those who are distinctly apart by lower distance values. Thus, a distance $d_{ij} = 0$ means total agreement of all genotypes and a distance value of $d_{ij} = 1$ means total disagreement of all genotypes.

Visualization of distance matrices by non-metric multidimensional scaling

The output of the above-described pairwise comparison of variant sets is a high-dimensional distance matrix with given distances or dissimilarities between pairs of individuals that satisfy all conditions of a metric. To represent the dissimilarities as distances between points in a low-dimensional space, we used a statistical technique named non-metric multidimensional scaling (MDS), that is, a visualization method such as principal component analysis or metric MDS. However, in contrast to principal component analysis (PCA) and metric MDS, non-metric MDS does not make any assumptions about the distribution of the underlying high-dimensional data. With a pre-specified number of dimensions for the embedding ϕ and an appropriate initial configuration, the *isoMDS* function of the *MASS* R-package was used to minimize the goodness of fit, called stress *S*, of Kruskal and Shepard (see [25]). To promote readability and an easy interpretation of the data, we chose a standard two-dimensional embedding with:

$$S(x_1, \ldots, x_n, \varphi) = \sqrt{\frac{\sum_{i=1, i \neq j}^{n} \left(d_{ij} - ||\varphi(x_i) - \varphi(x_j)||\right)^2}{\sum_{i=1, i \neq j}^{n} ||\varphi(x_i) - \varphi(x_j)||^2}}$$

where $||\cdot||$ defines the Euclidean norm.

Down sampling of raw data and simulation of genotyping accuracy

For coverage-adjusted comparisons, we randomly removed sequence reads from the original alignments. Variants were recalled on these down-sampled exomes as described above. As genotyping accuracy we define the percentage of the entire exome that was correctly genotyped, that is the sum of true positive genotype calls (alternate and reference genotypes) divided by the entire size of the exomic target region. For our simulations, we assumed that the reference set had a genotyping accuracy of 100% and introduced genotyping errors at random positions. As most of the exomic positions had low variability in the reference set, the contribution of genotyping errors to the distance function could be approximated by adding twice a binomial distributed random variable, $X \sim B(N, p) * 2$, to the normalizing constant C_{ij}, with probability p equaling the specified genotyping error and the number of trials $N = 2.8 * 10^7$ bp is the total size of the exome,.

Computation of the standardized dissimilarity score and reference curve

Distances between all individuals of the reference set were measured and the averaged values of the median and interquartile range of all columns of the distance matrix were computed to standardize the median of a test sample. The median of the distances from a test sample to all individuals of the reference set was computed and normalized by subtracting the pre-calculated median of the reference set and dividing the interquartile range (IQR) of the reference set. The reference curve and both 5% and 95% quartiles for the standardized dissimilarity score (SDS) were computed for the reference set and simulated data sets of decreasing error groups.

Results

An error sensitive genotype-weighted metric

Like any metric, the distance measure that we used to compare different sequences of a set of test samples induces a topology. Variant calls, which describe the measurable differences between samples, represent true genetic variation, as well as genotyping errors. The subject of our work is the quality assessment of a set of exome genotypes, thus our metric needs to induce a topology that is sensitive to sequencing errors while being robust to true genetic differences. By using a weighting method for genotypes that uses their frequency of occurrence, we achieved higher precision in accuracy prediction compared to an unweighted hamming distance (Figure S1 in Additional file 1). If two samples are not the same at an exonic position, which is highly constrained in the population, this contributes more to the total distance, because such an event has a higher probability of being a genotyping error than divergent genotypes at a highly variable site of the exome. When we compare two exomes, our metric works on all genotypes that have been called in these samples. The genotypes are weighted by the degree to which the genotype is constrained in the population. Though several definitions for measuring the degree of genotype conservation have been

suggested [26,27], most variable positions in the human genome are biallelic and for simplicity we approximate the conservation of a genotype by the inverse of its frequency. Thus the differences in two variant sets are weighted by their respective genotype frequencies, and the detection of many rare variants in a test sample therefore points to a higher proportion of false positive genotype calls. By contrast, if many variants that are common in the population are not detected in a test sample, this points to a high false negative error rate. By this means, we compute a matrix that contains the distances of the test sample to all the samples of a reference set as well as their mutual distances. These distances are a result of a function that works on the entire exomic target region, as defined by the 1000 genomes project, and may therefore be viewed without distortion only in the multivariate, exomic space. We have implemented and tested our method using whole exome data, but it could be applied to other types of high-throughput sequencing data. However, the precision of predicting the accuracy of genotyping decreases for smaller target regions (Figure S2 in Additional file 1).

Non-metric multidimensional scaling is best suited for distance visualization

Because distance is based on multiple variables of weighted categorical data, visualization in a plane requires a transformation. We tested several standard techniques of data visualization and found that non-metric MDS [28] showed the best characteristics in conveying the differences in genotyping accuracy in two dimensions (Figure S3 in Additional file 1). We therefore project the exomic distance matrix into two artificial dimensions of Φ1 and Φ2 that have the smallest loss of information [29-31].

The reference samples of the 1000 genomes project form clusters according to their ethnicity (Figure S4 in Additional file 1) and for any test sample we chose the closest cluster as the best matching reference set. Samples from the same population background form homogeneous clusters in non-metric MDS scaling, indicating a comparable genotyping quality (Figure 1A).

We then analyzed two test samples of European descent but of unknown genotyping accuracy and included them into the MDS projection of all central European (CEU) individuals from the 1000 genomes project, which is shown in Figure 1B. Except for one representative recalled sample, NA06986, all individuals of the CEU reference set are displayed as black circles, whereas the two exome test samples are represented by the colored triangles. Although the mean sequence coverage of the exome target region is above 60× for both test samples, they clearly differ in their mean distance to samples from the reference set: the second sample is close to the cluster formed by the individuals of the reference set, whereas the first sample is an outlier, indicating inferior quality. This considerable difference in the mean distance is remarkable given the high sequencing depth and a comparable ti/tv ratio of 3:2 (Additional file 1, Table S1). Also the proportion of variants found in dbSNP is around 97%, which is comparable to NA06986. Only the variance of the heterozygous allele frequencies, which increases with a growing number of artifacts from the amplification steps during the library preparation, suggests a lower quality for sample 1 with var(het AF) = 0.017 compared to 0.012 in test sample 2 and 0.013 in NA06986 at the same mean coverage [14].

Visualization of exomes of different genotyping accuracy

We measured the mean distance to the reference set for 85 exomes of European descent that were all analyzed in the context of NGS research projects. The two samples displayed in Figure 1B illustrate the extreme spectrum of the mean distance to the reference set that we encountered. We hypothesized that variants often detected in exomes of the 1000 genomes project, but not in the outliers of our test samples, might point to a subset that requires high data quality to be properly detected and that might explain partly the separation from the reference cluster. Figure 2A displays the value pairs of variant allele frequencies that are based on 85 CEU individuals from the 1000 genomes project and 85 exome test samples. For technical replicates one would expect all allele frequency value pairs to lie close to the diagonal in this kind of visualization. For two sample sets of equal size that are drawn from the same population, one would expect a certain degree of variance in the measured allele frequency that is simply based on the finite sample size: given the allele frequencies one would expect for a sample size of 85 that about 95% of the frequency value pairs fall inside the boundaries of the displayed ellipse based on a Bernoulli distribution. However, in our case there are considerably more outliers than expected by chance.

Characteristics of sequence variants with high error rates

We looked for similarities of these outliers and computed the GC content of 100 bases flanking the variant alleles that were present in more than half of the individuals analyzed in the 1000 genomes project but in only one or less of our analyzed samples. The distribution of the GC content of these variants clearly deviates from the distribution that one would expect for randomly located variants in the exome (Figure 2B).

To investigate the reasons for the higher false negative error rate for variants in a GC-rich sequence context, we computed the mean read coverage of the target region depending on the GC-content. Figure 2C shows that the distribution of the coverage is smaller for test

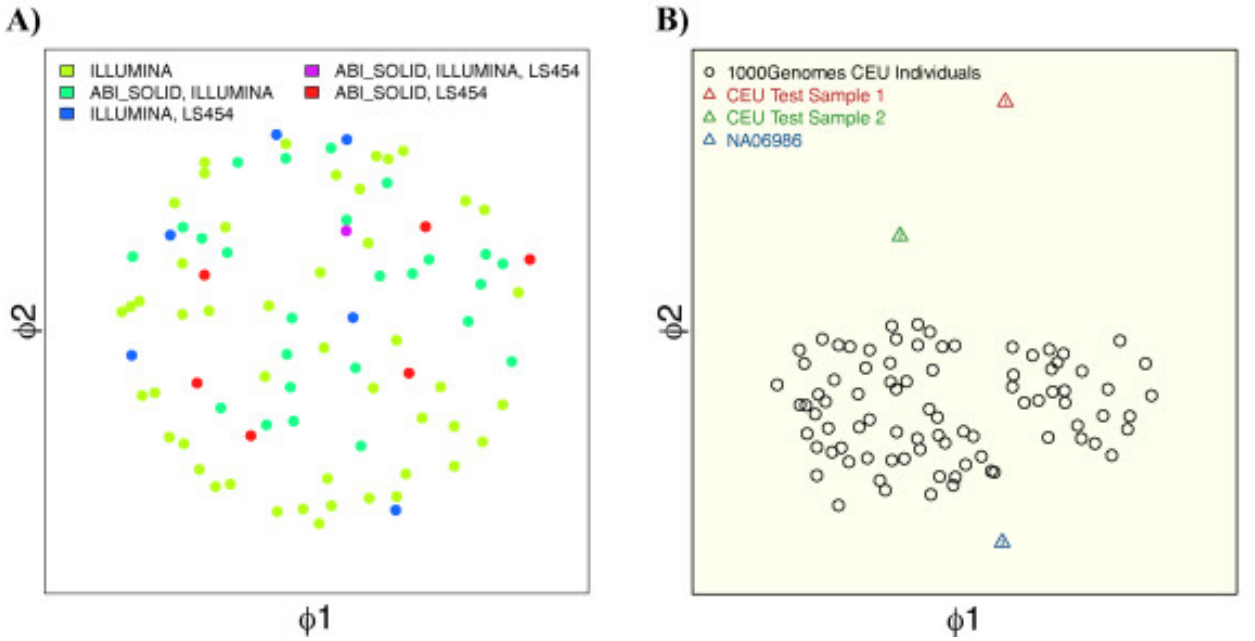


Figure 1 Similarities of exome samples. **(A)** The similarity between exome samples is measured with a genotype frequency weighted metric. The similarity matrix is visualized by non-metric multidimensional scaling in the two-dimensional plane. Variant sets of individuals of European descent that were analyzed by the 1000 genomes project form a homogeneous cluster, irrespective of the applied sequencing technology. **(B)** Two test samples of the same ethnicity that we genotyped according to a standard exome protocol are compared to samples of the reference set. The larger distance of sample 1 to the reference cluster indicates a lower genotyping accuracy, while sample 2 points to a high quality.

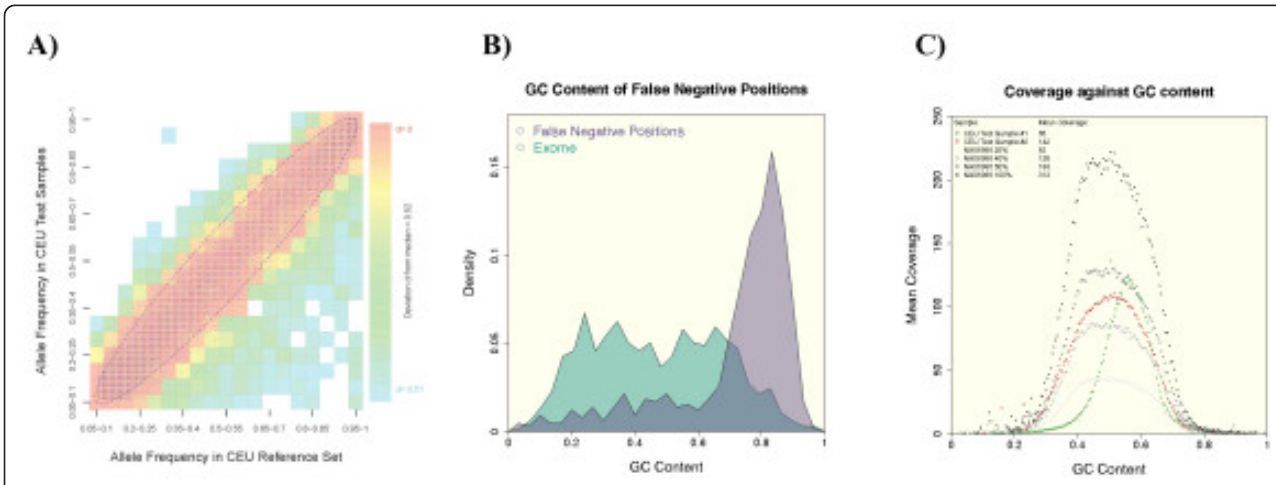


Figure 2 Analysis of low quality samples. **(A)** Comparison of variant allele frequencies in different sample sets. Value pairs of genotype frequencies were computed for exomes of the reference set (CEUs from the 1000 genomes project) and compared to test samples of the same ethnicity that are low quality. The ellipse indicates twice the standard deviation assuming a binomial model for the allele frequency *p*. Variants in the right lower quadrant were called with a lower probability in our test samples and are characterized by a GC content that deviates from the expected mean. **(B)** GC content at false negative positions. Variants that are underrepresented in exomes with a large distance to the high quality reference set are overrepresented in exome regions with high GC content (violet curve). The green curve indicates the distribution of the GC-content that is expected for an equal number of variants that are randomly drawn from the exome. **(C)** Coverage against GC content. The mean sequence coverage of the consensus exome varies with the GC content of the target region. The overall coverage for an exemplary sample from the 1000 genomes project (NA06986) was higher compared to test samples 1 and 2. Test sample 1 has a particularly low coverage in regions with extreme GC content, suggesting a higher error rate.

sample 1 compared to test sample 2 and NA06986 that was sampled down to a comparable mean coverage. Thus, the higher distance of test sample 1 to the reference set is partly due to a critically low sequence read coverage of regions with an extreme GC-content. Benjamini *et al.* studied the bias caused by GC content depending on read coverage in detail for Illumina sequencing data and showed that it also varies between technical replicates [32]. This also means that two samples may have different genotyping accuracy for the whole exome although they have been processed by the same protocol.

The distance to the reference set versus coverage and error rates

We hypothesized that the distance to the high quality reference set should grow when the amount of raw sequence data decreases. We therefore successively reduced the sequence coverage in the raw exome data of NA06986, called variants anew, and observed an increasing distance to the reference set (green to blue circle in Figure 3). A decreasing sequencing coverage will not only yield an increasing rate of false negative genotypes but also increase the rate of false positive calls. It is more likely that a sequencing error will be called as a heterozygous variant, and a heterozygous variant as homozygous, if the position is only covered by a few reads. We then analyzed how an increasing false positive error affects the distance to the reference set by simulating detection artifacts that were randomly distributed over the target region and added to a sample from the reference set. The triangles in Figure 3 show that the data points follow a trajectory that departs from the reference cluster with growing error rate. It has to be noted that the

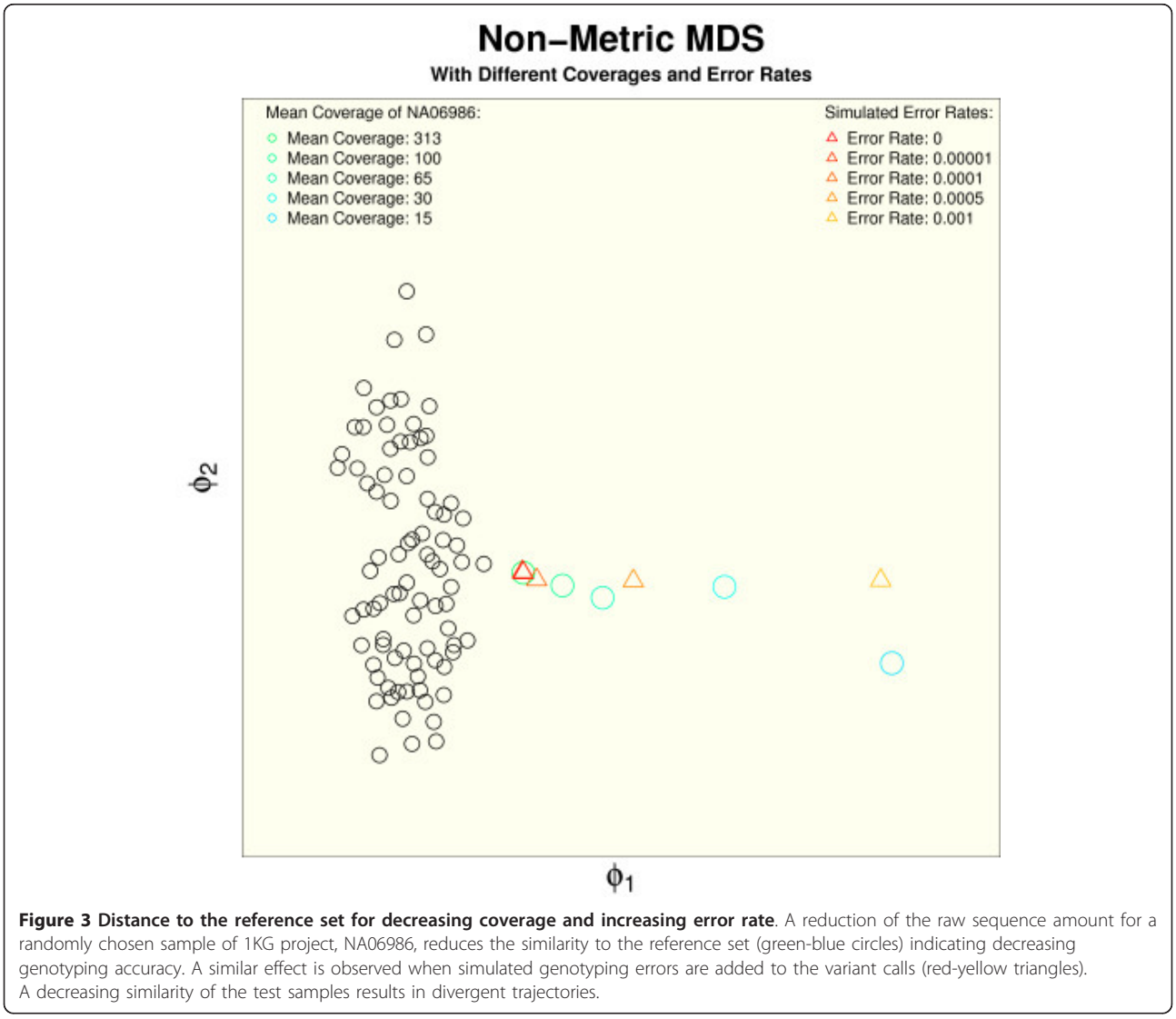


Figure 3 Distance to the reference set for decreasing coverage and increasing error rate. A reduction of the raw sequence amount for a randomly chosen sample of 1KG project, NA06986, reduces the similarity to the reference set (green-blue circles) indicating decreasing genotyping accuracy. A similar effect is observed when simulated genotyping errors are added to the variant calls (red-yellow triangles). A decreasing similarity of the test samples results in divergent trajectories.

visualization of multiple versions of the same test sample that differ only in coverage or error rates in a single MDS plot contorts the relative distances. All the depicted simulated data points in Figure 3 originate from the same data set and are therefore self-similar. The self-similarity is high for a low error rate and high coverage and decreases with increasing error rate and decreasing coverage, as the divergent trajectories of circles and triangles indicate. We also obtained similar results in the analysis of simulated data sets of other ethnic backgrounds (Figure S5 in Additional file 1).

A genotyping error of 0.00001 corresponds to an expectation of one genotyping error in approximately 100 kb of the target region. Two randomly chosen samples from the reference set would differ in about 100 positions in such a window of 100 kb [21]. The samples with the simulated error rate begin to separate from the high quality cluster for error rates above 0.00001, which corresponds to a positive predictive value of 0.99 (number of true positive divided by number of positive calls). Interestingly, the positive predictive value that was reported by Tennessen *et al*. for the variant calls of the 5000 exomes project is between 0.97 and 0.98 [21]. The resolution of our visualization techniques is therefore sufficient to display these qualitative differences.

Comparison of exome data from different next-generation sequencing studies

In contrast to the 1000 genomes project, the genotype calls from the 5000 exomes project were publicly available only in a collapsed form as genotype frequencies for European Americans and African Americans and not as separate variant sets for each sample. In addition to our in-house exome data, we also analyzed the distances to the reference cluster for exomes that we simulated based on the genotype frequencies from these European Americans and 100 exomes that were already studied by de Ligt *et al.* [20]. Figure 4A shows the distribution of the SDS, which represents a normalized distance of a test sample to the reference cluster. The mean SDS for the simulated exomes of the 5000 exomes project is comparable to our exome data and lower than the SDS of the de Ligt *et al*. exomes. The smaller variance of the SDS distribution in the 5000 exomes samples, which points to a higher self-similarity, is due to a simulation process that did not properly represent the haplotype substructure of the data. The higher mean SDS of the de Ligt *et al*. data is mainly explained by outdated genotyping algorithms with higher error rates and a lower mean coverage of the target region. Figure 4B depicts the coverage distribution over the exome and additional quality parameters for an exemplary sample from de Ligt *et al*. and our two test samples.

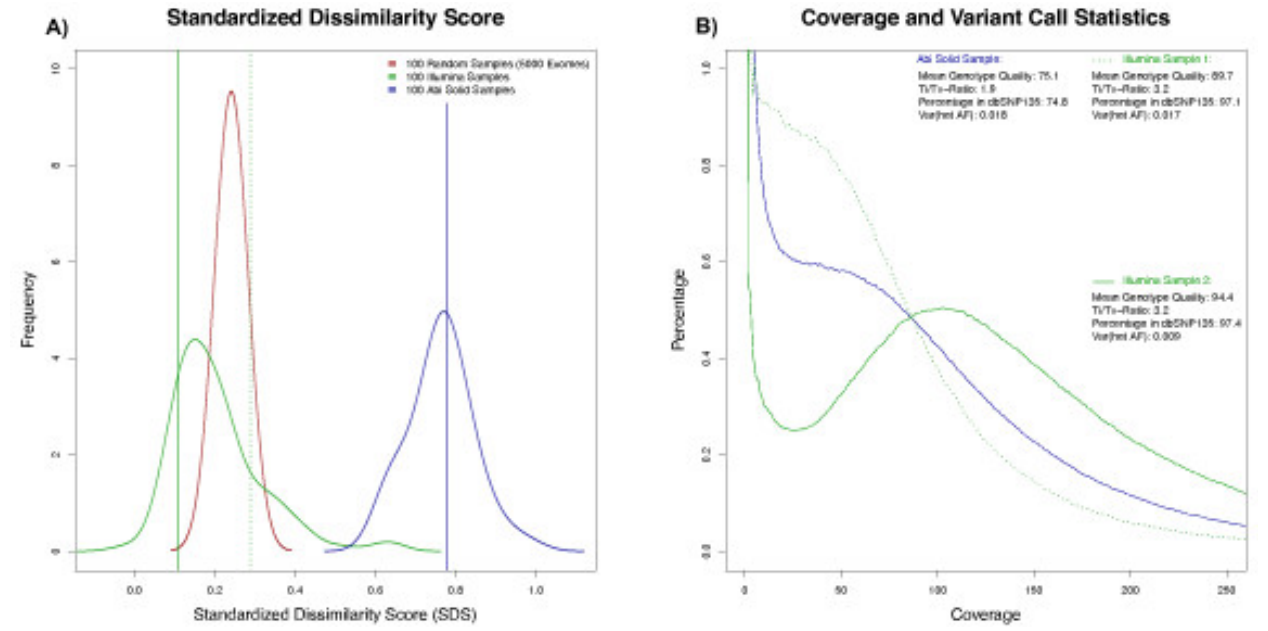


Figure 4 Distributions of standardized dissimilarity scores for different exomes. **(A)** 100 samples, including test samples 1 and 2, were sequenced on the Illumina platform with mean coverage above 60×. The mean SDS for the Illumina samples is comparable to the mean SDS of 100 exomes that were simulated based on genotype frequencies from the 5000 exomes project. The variation of these SDS is smaller due to missing haplotype information. (B) The mean SDS of 100 exomes that were sequenced on a Solid platform is considerably higher due to a lower and less uniform sequence distribution over the target region and due to less accurate variant calls.

By simulating increasing sequencing noise for all exome data sets of the 1000 genomes, we derived distributions for the mean distances of the original data. Based on these distributions, we computed a reference curve for the SDS of an unknown sample that correlates with its exome-wide accuracy (Figure 5). The SDS that is measured for our test samples may be used for estimating their genotyping accuracy by intersecting with the reference curve. For test sample 1, the mean distance to the reference set was 0.29, which corresponds to an estimated genotyping error of 0.0001. By contrast, the SDS of 0.11 for test sample 2 indicates an error rate that is much closer to that of the 1000 genomes project. Interestingly, the distance to the reference set shows characteristics of a phase transition, when the contribution of genotyping errors exceeds the genetic variability between individuals. We also checked the validity of our approach by deriving the genotyping accuracy via a complementary method. In Heinrich *et al.*, we analyzed the effect of amplification steps during sample preparation and derived rates of genotyping errors from technical replication [14]. The SDSs for these replicates indicate genotyping accuracy for the exomes between 99.99% and 99.999%, which is in good agreement with the previously computed accuracy.

Thus, the SDS is a parameter derived from the composition of a variant set and is even more powerful in predicting the data quality than other quality control parameters such as coverage distributions, which require access to the read alignments (Additional file 1, Table S1).

We tested the influence of the sequencing platform on the error prediction by our approach by restricting the reference set to samples that were sequenced with the same technology (Illumina). Although the visualization in metric MDS clearly shows that the test samples are closest to the Illumina samples from the reference set (Figure S3 in Additional file 1), the effect of the sequencing platform on the accuracy prediction is marginal (Figure S6 in Additional file 1). The SDS is therefore

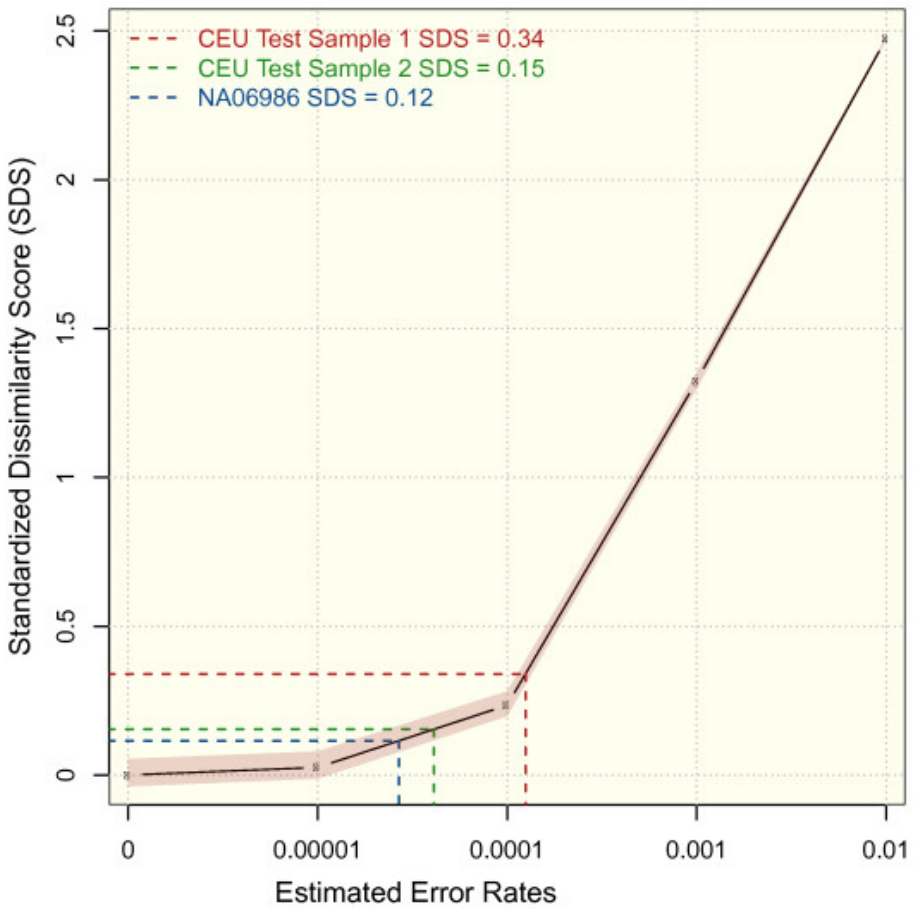


Figure 5 Estimation of genotyping errors from standardized dissimilarity scores. The reference curve with its 5% and 95% quantiles is based on the distances of samples with simulated error rates to the reference set. The SDS of a test sample indicates its error rate by its intersection with the reference curve. The estimated error rate of test sample 1 is considerably higher than of test sample 2 and of NA06986 from the 1000 genomes project. SDS, standardized dissimilarity score.

robust and can be applied to predict the quality of genotyping data from different sequencing technologies. We used the platform independence of our approach to analyze the quality of a sample from the 1000 genomes project, NA12878, that was recently re-sequenced by a new technology (Proton, Ion Torrent, aligned with TMAP and genotyped with variantCaller). In Figure 6, the distance of this Proton variant set of NA12878 to the reference set is compared to a variant set of the same individual based on Illumina raw data that was downsampled to a mean coverage of 30 fold as described above. This figure visualizes how the quality of two exomes of the same individual that were generated on two different sequencing platforms and processed by different bioinformatics pipelines may be interpreted at a glance. We used the SDS to estimate the genotyping accuracy and predicted 99.9% for the low coverage Illumina exome and above 99.99% for the Proton exome, which is in good agreement with the values based on Sanger validation for this sample.

Discussion

We have described a new approach to assess the accuracy of variant calls from NGS studies. The genotyping accuracy for variant calls, that is genotypes that differ from the true sample sequence, has been estimated in large-scale NGS-based projects such as the 1000 genomes project [19] or the exome sequencing project [21] and comparisons of NGS platforms [33]. In these projects, samples were sequenced to a very high mean coverage on different sequencing platforms and the reported variants represent an intersection of technical replicates and independent analysis pipelines. Even in these high quality data sets, up to about 2% of the variants cannot be validated if re-sequenced by a complementary approach such as ABI Sanger. In such a high quality exome, one detects around 15,000 single nucleotide variants per 30 Mb coding sequence and approximately 300 of them are likely false positive calls, which corresponds to an error rate of $300/30*10^6 = 0.00001$.

Based on simulated accuracy groups for variant calls, we were able to assess the quality of an exome test sample without detailed knowledge of the applied enrichment and sequencing technology or of the bioinformatics pipeline that was used to align the reads and call the genotypes. The SDS is therefore suitable for a comprehensive quality control in all exome-based mutation screens and might turn out especially useful as a criterion for data inclusion in studies that combine exome data of different sources, due to its platform independence. The estimated genotyping error in particular might serve as quality criterion before variants detected in an exome are further analyzed: only if the estimated error is comparable to that of a high quality reference set such as the 1000 genomes project would one proceed with variant analysis. We envision that our approach to estimate the genotyping accuracy of exomes will facilitate the quality assessment of NGS data.

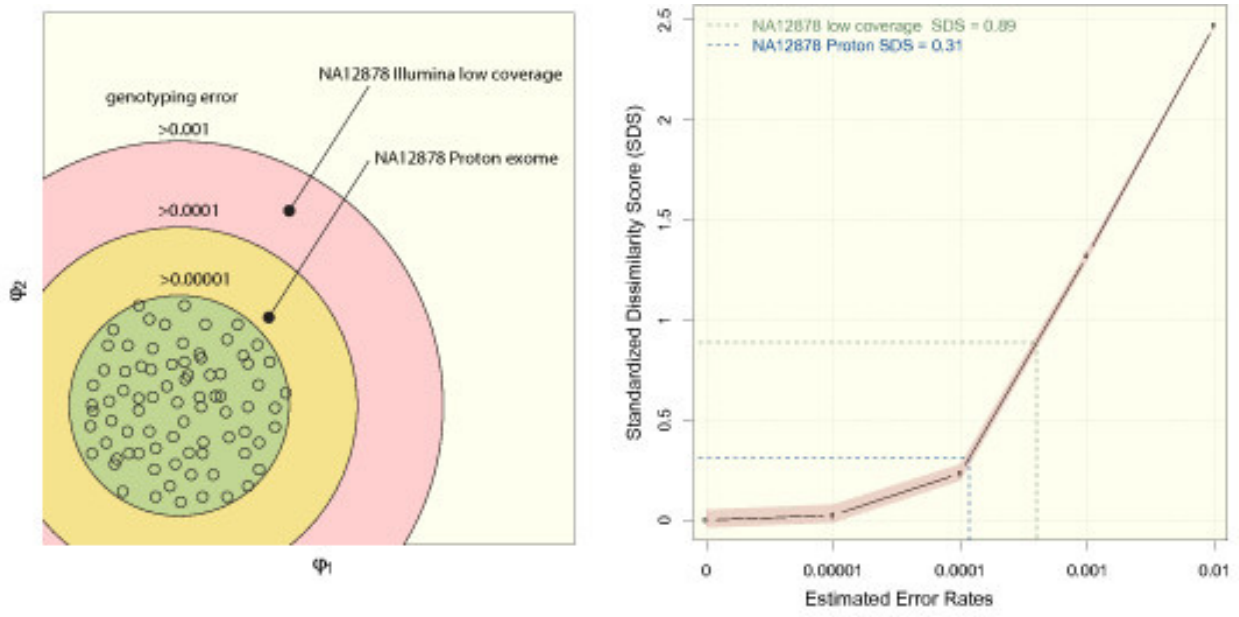


Figure 6 Comparison of exome data of different NGS platforms. A) Exome genotypes of individual NA12878 were called based on an Illumina read data set of low coverage and a Proton dataset. The Proton exome is closer to the CEU reference data than the low coverage Illumina exome. B) The accuracies of the exome genotype sets were estimated based on the mean distance to the reference set.

Software that computes the SDS, visualizes the distances to data of the 1000 genomes project, and predicts the genotyping accuracy is available for download and as a web service at GeneTalk [19].

Web resources

The URLs for data and methods presented herein are as follows:

NHLBI Exome Sequencing Project (ESP) Exome Variant Server, http://evs.gs.washington.edu/EVS/

GeneTalk, https://gene-talk.de/qc

ftp server of the 1000 genomes project, ftp://ftp.1000genomes.ebi.ac.uk./vol1/ftp/

Additional material

Additional file 1: Supplementary material. Figure S1: Distribution of dissimilarities of simulated error groups for different distance metrics. The distance of simulated test samples of different error groups were to the reference set was measured with an unweighted hamming metric and a genotype frequency weighted metric. The variance of the dissimilarities of the test samples is smaller in the genotype frequency weighted metric and thus allows a more precise prediction of the error group. Figure S2: Visualization of distances and estimation of error rates for different target regions. The distances of two test samples (sample 1 of high and sample 2 of medium quality) to the reference data were computed for five different target regions that differ in size. The CCDS exome comprises 29 Mb, the human phenotype ontology (HPO) [34] panel contains all exons of genes associated with phenotypic features (5.8 Mb), the Kingsmore panel comprising 548 genes of known inherited diseases (1.2 Mb) [35], all coding exons of chromosome 22 (600 kb), and the GPI panel that contains all genes involved in the GPI-anchor synthesis (45 kb). The larger the target region, the higher the number of sequence variants for comparison. This increases the precision of the estimation of the error rates. With decreasing size of the target region, the confidence intervals of the reference curve for the standardized dissimilarity score widen. While the different error rates of sample 1 and 2 can be clearly estimated and visualized for the larger target regions, gene panels below 1 Mb do not allow this assessment any more due to the larger confidence intervals. Figure S3: Data visualization techniques. Comparison of ordination methods for the visualization of the distances of exome genotypes of two test samples and high quality reference samples of a matched background population. The mean distance of test sample 2 with the low genotyping accuracy to the reference samples is larger compared to sample 1 with the high genotyping accuracy for all visualization methods. For principal component analysis and metric MDS, a substructure in the reference samples is visible that is specific to the sequencing platform. Figure S4: Exomes of different ethnicities (European (CEU), Yorubian (YRI) and Japanese (JPT)) form distinct clusters based on their similarity. For a test sample the closest cluster from the 1000 genomes project data is chosen as reference set. Figure S5: The distance of a test sample of the Yorubian reference set increases for a growing simulated error rate. Figure S6: Influence of sequencing platform on error prediction. In contrast to non-metric MDS visualization, principal component analysis of the similarities of European samples of the 1000 genomes project reveals some information about the sequencing platform that was used (A). However, the effect of the sequencing platform for predicting the genotyping accuracy is small. The predicted error rates of the test samples are comparable if the reference set is restricted to specific sequencing platforms (B, C). Table S1: Comparison of different parameters for quality assessment. Short sequence reads of test sample 1 and 2 and a sample from the 1000 genomes reference set, NA06986, were sampled to comparable mean coverage over the target region. The total number of variant calls decreases with a decreasing coverage indicating an increasing false negative error rate. Also the mean genotype quality scores decrease with a decreasing coverage indicating an increasing false positive error rate. The ti/tv ratio and the ratio of variants that are present in dbSNP vary only very little with changing coverage. Different priors in the genotyping models of Samtools and GATK result in different mean genotype quality scores for the same alignments. Quality score recalibration with GATK VariantRecalibrator performed on Samtools- and GATK-called variants ads an adjusted quality score, VQSLOD (log odds ratio of being a true variant versus being false under a trained gaussian mixture model). This score is used in GATK ApplyRecalibration to generate a tranche file of highly confidential calls. The percentage of these calls which passed as high-confidential shows an irregular behavior with respect to the mean coverage. The SDS, which is computed for the entire variant call set of a sample, correlates with its accuracy and allows a sample-to-sample comparison.

Abbreviations

bp: base pair; CCDS: consensus coding sequence; CEU: central European; dbSNP: National Center for Biotechnology Information Single Nucleotide Polymorphism Database; Kb: kilobase; Mb: megabase; MDS: multidimensional scaling; NGS: next-generation sequencing; SDS: standardized dissimilarity score; ti/tv: ratio of transitions versus transversions.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

VH, PNR and PMK conceived and designed the study. VH developed the original code, TK and DP implemented modifications for GeneTalk. JH contributed exome data. JS and TD supervised and contributed to the statistical analysis. VH, PNR, TD and PMK wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by a grant of the Deutsche Forschungsgemeinschaft (DFG KR 3985/1-1) to PMK. We thank Christian Gilissen and Joris Veltman for providing solid exome data and Michal R Schweiger and Martin Kerick for cancer exome data for analysis. The authors would like to thank the NHLBI GO Exome Sequencing Project and its ongoing studies, which produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926) and the Heart GO Sequencing Project (HL-103010).

Authors' details

[1]Institute for Medical Genetics and Human Genetics, Charité Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany. [2]Smart Algos, Retzbacher Weg 83, 13189 Berlin, Germany. [3]Department of Mathematics, Humboldt-University Berlin, Unter den Linden 6, 10099 Berlin, Germany. [4]Berlin-Brandenburg Center for Regenerative Therapies (BCRT), Charité Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany. [5]Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, 14195 Berlin, Germany.

Received: 6 February 2013 Revised: 19 July 2013
Accepted: 31 July 2013 Published: 31 July 2013

References

1. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J: **Exome sequencing as a tool for Mendelian disease gene discovery.** *Nat Rev Genet* 2011, **12(11)**:745-755.
2. Ku CS, Cooper DN, Polychronakos C, Naidoo N, Wu M, Soong R: **Exome sequencing: dual role as a discovery and diagnostic tool.** *Ann Neurol* 2012, **71(1)**:5-14.
3. Sulonen AM, Ellonen P, Almusa H, Lepisto M, Eldfors S, Hannula S, Miettinen T, Tyynismaa H, Salo P, Heckman C, *et al*: **Comparison of solution-based exome capture methods for next generation sequencing.** *Genome biology* 2011, **12(9)**:R94.

4. Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R, Euskirchen G, Butte AJ, Snyder M: **Performance comparison of exome DNA sequencing technologies.** *Nature biotechnology* 2011, **29(10)**:908-914.
5. Holtgrewe M, Emde AK, Weese D, Reinert K: **A novel and well-defined benchmarking method for second generation read mapping.** *BMC Bioinformatics* 2011, **12**:210.
6. Ruffalo M, LaFramboise T, Koyuturk M: **Comparative analysis of algorithms for next-generation sequencing read alignment.** *Bioinformatics* 2011, **27(20)**:2790-2796.
7. Goya R, Sun MG, Morin RD, Leung G, Ha G, Wiegand KC, Senz J, Crisan A, Marra MA, Hirst M, *et al*: **SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors.** *Bioinformatics* 2010, **26(6)**:730-736.
8. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: **VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.** *Genome Res* 2012, **22(3)**:568-576.
9. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18(11)**:1851-1858.
10. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20(9)**:1297-1303.
11. Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H: **SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data.** *Nucleic acids research* 2011, **39(19)**:e132.
12. Mardis ER: **Next-generation DNA sequencing methods.** *Annu Rev Genomics Hum Genet* 2008, **9**:387-402.
13. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, *et al*: **The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes.** *Genome Res* 2009, **19(7)**:1316-1323.
14. Heinrich V, Stange J, Dickhaus T, Imkeller P, Kruger U, Bauer S, Mundlos S, Robinson PN, Hecht J, Krawitz PM: **The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process.** *Nucleic acids research* 2012, **40(6)**:2426-2431.
15. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic acids research* 2001, **29(1)**:308-311.
16. Bainbridge MN, Wang M, Wu Y, Newsham I, Muzny DM, Jefferies JL, Albert TJ, Burgess DL, Gibbs RA: **Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities.** *Genome biology* 2011, **12(7)**:R68.
17. Nothnagel M, Herrmann A, Wolf A, Schreiber S, Platzer M, Siebert R, Krawczak M, Hampe J: **Technology-specific error signatures in the 1000 Genomes Project data.** *Hum Genet* 2011, **130(4)**:505-516.
18. O'Rawe J, Guangqing S, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson E, Wei Z, Jiang T, *et al*: **Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing.** *Genome medicine* 2013, **5(3)**:28.
19. **A map of human genome variation from population-scale sequencing..** *Nature* 2010, **467(7319)**:1061-1073.
20. de Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, de Vries P, Gilissen C, *et al*: **Diagnostic exome sequencing in persons with severe intellectual disability.** *The New England journal of medicine* 2012, **367(20)**:1921-1929.
21. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, *et al*: **Evolution and functional impact of rare coding variation from deep sequencing of human exomes.** *Science* 2012, **337(6090)**:64-69.
22. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25(14)**:1754-1760.
23. Li H: **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.** *Bioinformatics* 2011, **27(21)**:2987-2993.
24. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, *et al*: **The variant call format and VCFtools.** *Bioinformatics* 2011, **27(15)**:2156-2158.
25. Venables WNRBD: *Modern Applied Statistics with S* Springer; 2002.
26. Schneider TD: **Information content of individual genetic sequences.** *Journal of theoretical biology* 1997, **189(4)**:427-441.
27. Shannon CE: **A Mathematical Theory of Communication.** *At&T Tech J* 1948, **27(4)**:623-656.
28. Kruskal JB: **Nonmetric Multidimensional-Scaling - a Numerical-Method.** *Psychometrika* 1964, **29(2)**:115-129.
29. Jombart T, Pontier D, Dufour AB: **Genetic markers in the playground of multivariate analysis.** *Heredity* 2009, **102(4)**:330-341.
30. Lessa EP: **Multidimensional-Analysis of Geographic Genetic-Structure.** *Syst Zool* 1990, **39(3)**:242-252.
31. Wang CL, Szpiech ZA, Degnan JH, Jakobsson M, Pemberton TJ, Hardy JA, Singleton AB, Rosenberg NA: **Comparing Spatial Maps of Human Population-Genetic Variation Using Procrustes Analysis.** *Stat Appl Genet Mol* 2010, **9(1)**.
32. Benjamini Y, Speed TP: **Summarizing and correcting the GC content bias in high-throughput sequencing.** *Nucleic acids research* 2012, **40(10)**:e72.
33. Lam HY, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, Gerstein MB, *et al*: **Performance comparison of whole-genome sequencing platforms.** *Nature biotechnology* 2012, **30(1)**:78-82.
34. Robinson PN, Kohler S, Bauer S, Seelow D, Horn D, Mundlos S: **The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease.** *American journal of human genetics* 2008, **83(5)**:610-615.
35. Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, Langley RJ, Zhang L, Lee CC, Schilkey FD, *et al*: **Carrier testing for severe childhood recessive diseases by next-generation sequencing.** *Science translational medicine* 2011, **3(65)**:65ra64.

doi:10.1186/gm473
Cite this article as: Heinrich *et al.*: Estimating exome genotyping accuracy by comparing to data from large scale sequencing projects. *Genome Medicine* 2013 **5**:69.

2.6 Identifikation pathogener Mutationen in *PIGV* in Exom-Daten von Patienten mit Mabry Syndrom

Krawitz, P.M., Schweiger, M.R., Rodelsperger, C., Marcelis, C., Kölsch, U., Meisel, C., Stephani, F., Kinoshita, T., Murakami, Y., Bauer, S., et al. (2010). Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nature Genetics 42, 827-829.

Bereiche, die in Genomen mehrerer Individuen einer Familie von demselben Vorfahren (Gründer) abstammen, bezeichnet man als erbgleich (identity-by-descent, IBD). Wenn sich in unterschiedlichen Familienmitgliedern sowohl die paternale Kopie eines Gens als auch die maternale Kopie auf die gleichen Gründer zurückführen lässt, so spricht man von IBD2. Wenn beide Kopien auf denselben Gründer zurückgehen, liegt zudem in diesem Bereich eine Homozygotie vor, andernfalls eine zusammengesetzte Heterozygotie (Compound Heterozygotie). Die Wahrscheinlichkeit, dass ein beliebiger Lokus im Genom bei *n* Individuen derselben Familie den Zustand IBD2 hat, ist aufgrund der zufälligen Aufteilung der Allele in der ersten Reifeteilung $1/2^{n-1}$.

Bei autosomal rezessiven Erkrankungen weisen beide Kopien des Krankheitsgens eine pathogene Veränderung auf und bei mehreren betroffenen Familienmitgliedern darf angenommen werden, dass die pathogenen Allele erbgleich sind. Wenn der Krankheitslokus noch unbekannt ist, so stellt die Identifikation erbgleicher Bereiche also eine Möglichkeit dar, den Suchraum im Genom einzuschränken.

Wenn an einer Position gleiche Genotypen bestimmt wurden (identity-by-state, IBS), so kann dies durch zwei unterschiedliche Entstehungsmodelle erklärt werden: Erbgleichheit oder Zufall. Dem „Messereignis“ IBS ist nicht eindeutig die Ursache anzusehen und man kann auch sagen, dass das zugrundeliegende Ereignis, also IBD oder IBS aus Zufall, verborgen ist („hidden state“).

Betrachtet man nun zwei benachbarte Genotypen, die beide IBS sind, so unterscheiden sich die Wahrscheinlichkeiten je nachdem, welches Entstehungsmodell angenommen wird. Liegen die Allele jeweils auf gleichen Haplotypen, so ist IBS auf IBD zurückzuführen. Die Wahrscheinlichkeit, dass gleiche Genotypen per Zufall vorliegen, lässt sich aus den Genotyp-Frequenzen in der zugrunde liegenden Population bestimmen.

Der Übergang zwischen zwei verborgenen Zuständen kann nun als dynamischer stochastischer Prozess beschrieben werden (Hidden Markov Model, HMM). Für die Übergangswahrscheinlichkeit von IBD zu nicht IBD, also die Unterbrechung eines gemeinsamen Haplotyps, werden in diesem Modell zusätzlich die Rekombinationsraten berücksichtigt.

Wenn nun die Abfolge übereinstimmender oder abweichender Genotypen in *n* Individuen betrachtet wird, so kann mit Hilfe des Viterbi Algorithmus auf die wahrscheinlichste Kombination der verborgenen Zustände im HMM geschlossen werden.

Der von uns entwickelte Programmcode ermöglicht es, aus Exom-Daten die Bereiche abzuschätzen, die erbgleich sind und kann somit eingesetzt werden, um Kandidaten-Mutationen zu priorisieren (Rodelsperger, et al., 2011). Dieses Verfahren haben wir erfolgreich eingesetzt um erbgleiche Bereiche in drei vom Mabry-Syndrom betroffenen Geschwistern zu identifizieren.

Als wir die im Jahr 2009 erstmals Exom Sequenzierungen durchführten, waren sowohl die Fehlerraten in NGS Datensätzen noch beträchtlich und viele der heute als Polymorphismen bekannten Sequenzvarianten waren noch nicht in den Datenbanken vermerkt. Die Filterung eines einzelnen Exoms auf unbekannte SNVs mit möglicher Auswirkung auf Proteinebene bewirkte eine Reduktion auf ca. 3000 Kandidaten-Gene.

Wir setzten ein Hidden Markov Model ein, um Bereiche im Genom der drei Geschwister zu identifizieren, die erbgleich sind (IBD2). Dadurch konnte die Zahl der Kandidaten-Gene auf zwei reduziert werden, *PIGV* und *SLC9A1,* die sich beide innerhalb eines ca. 13 Mb langen homozygoten Intervalls befanden. In weiteren, nicht verwandten Patienten mit Mabry Syndrom wurden diese Gene mittels herkömmlicher Abi Sanger Sequenzierung untersucht und weitere Mutationen in *PIGV* nachgewiesen. Alle in *PIGV* identifizierten Mutationen bewirken einen Austausch evolutionär hochkonservierter Aminosäuren, p.Glu256Lys, p.Ala341Glu, p.Ala341Val und p.His385Pro. Wir vermuteten, dass durch diese Veränderungen die enzymatische Aktivität von PIGV, der zweiten Methyltransferase in der GPI-Ankersynthese, beeinträchtigt sein könnte.

Die funktionellen Auswirkungen der missense Mutationen wurden in einer Zelllinie überprüft, die von Ovarien des chinesischen Hamsters, CHO, abstammt und in der das Gen *PIGV* ausgeschaltet ist. Da in diesen Zellen keine GPI-Anker synthetisiert werden, fehlen auf den Oberflächen GPI-verankerte Proteine, wie CD55, CD59 oder Alkalische Phosphatase. Durch Transfektion mit humanem *PIGV* Wildtyp-Konstrukt konnte die GPI-Ankersynthese wieder hergestellt werden (rescue). Durchflusszytometrisch lässt sich die Effektivität des rescues quantifizieren, indem die Intensität fluoreszenmarkierter GPI-APs gemessen wird.

Die Intensität der markierten GPI-APs war nach Transfektion mit den mutanten Konstrukten deutlich niedriger und wies damit auf eine verminderte enzymatische PIGV Aktivität hin. Dies bewies zugleich, dass die bei den Patienten gemessene Oberflächenreduktion von GPI-APs auf die in *PIGV* identifizierten Mutationen zurückzuführen war. Wir haben damit *PIGV* als neues Krankheitsgen nachgewiesen.

http://dx.doi.org/10.1038/ng.653

http://dx.doi.org/10.1038/ng.653

http://dx.doi.org/10.1038/ng.653

2.7 Mutationen in *PIGO*, einem Gen der GPI-Ankersynthese, als Ursache für HPMRS

Krawitz, P.M., Murakami, Y., Hecht, J., Kruger, U., Holder, S.E., Mortier, G.R., Delle Chiaie, B., De Baere, E., Thompson, M.D., Roscioli, T., et al. (2012). Mutations in PIGO, a member of the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation. American Journal of Human Genetics 91, 146-151.

Nicht bei allen Patienten mit dem klinischen Verdacht auf Hyperphosphatasie mit Mentaler Retardierung lassen sich pathogene Mutationen in *PIGV* nachweisen. In einer Familie mit zwei betroffenen Schwestern führten wir daher eine Exom-Sequenzierung der Eltern sowie der Kinder durch. Da die Eltern nicht blutsverwandt sind, wurden die Exom-Daten auf zusammengesetzt heterozygote Mutationen gefiltert. Der Filteransatz reduzierte die Anzahl der Kandidaten-Gene auf drei, *PIGO, NBPF10* und *MUC16*.

Neben dem Gen mit den ursächlichen Mutationen, *PIGO*, ist insbesondere für Gene, die eine hohe Anzahl seltener Varianten und Sequenzierartefakte aufweisen, die Wahrscheinlichkeit erhöht, dass Kandidaten Mutationen die Filterkriterien für zusammengesetzte Heterozygotie erfüllen. *MUC16* stellt mit über 40kb kodierender Sequenz eines der längsten Gene im Humangenom dar und auch in Exom-Sequenzierungen gesunder Trios weist dieses Gen meist zahlreiche, seltene heterozygote Mutationen auf, die die Kriterien des Filters auf Compound Heterozygtie erfüllen. *NBPF10* zählt zu einer Genfamilie, bei der es in der menschlichen Stammesgeschichte wiederholt zu Duplikationen kam. Aufgrund der hohen Sequenzähnlichkeit dieser Gene, die sich oftmals über viele hundert Basenpaaren erstreckt, ist die Zuordnung eines Sequenzfragments problematisch. Häufig kommt es zu Mapping-Artefakten, NGS-basierten variant calls, die sich nicht mit herkömmlicher Abi Sanger Sequenzierung bestätigen lassen.

Das Genprodukt von *PIGO* stellt, wie auch *PIGV*, ein Enzym der späten GPI-Ankersynthese dar, das ein Ethanolaminphosphat, EtNP, an den dritten Mannose-Rest des Ankers anfügt. Nach aktueller Wissenslage werden GPI-APs, wie die alkalische Phosphatase, über diesen EtNP-Rest mit dem GPI-Anker verknüpft. Eine Beeinträchtigung der enzymatischen Funktion sollte daher ebenfalls zu einer verminderten Oberflächenexpression von GPI-APs und deren vermehrter Sekretion führen.

Die mittels Exom-Sequenzierung in *PIGO* identifizierten Varianten c.2869C>T und c.2361dup führen auf Proteinebene zu einem Aminosäureaustausch, p.Leu957Phe, und einer Leserasterverschiebung, p.Thr788Hisfs*5. In einem Patienten einer weiteren Familie konnten die compound heterozygoten Mutationen c.2869C>T und c.3069+5G>A nachgewiesen werden. Die in beiden Familien heterozygot aufgetretene missense Mutation wurde in PIGO defizienten CHO Zellen untersucht und zeigt im Vergleich zum Wildtyp-Konstrukt nur eine verminderte Wiederherstellung der GPI-AP Oberflächenexpression. Ähnlich wie bei den in *PIGV* beobachteten pathogenen missense Mutationen belegt dies daher ebenfalls bei p.Leu957Phe eine funktionseinschränkende Wirkung. Der durch die Leserasterverschiebung bewirkte vorzeitige Kettenabbruch an Position 793 führt zu einem kompletten Funktionsverlust von PIGO. Die veränderte Spleißstelle führt dazu, dass während des Spleißvorgangs das 215 bp umfassende Exon 9 übersprungen wird und damit ebenfalls ein funktionsloses Transkript resultiert. Damit wurde *PIGO* erstmals als Krankheitsgen nachgewiesen.

Please cite this article in press as: Krawitz et al., Mutations in *PIGO*, a Member of the GPI-Anchor-Synthesis Pathway, Cause Hyperphosphatasia with Mental Retardation, The American Journal of Human Genetics (2012), doi:10.1016/j.ajhg.2012.05.004

REPORT

Mutations in *PIGO*, a Member of the GPI-Anchor-Synthesis Pathway, Cause Hyperphosphatasia with Mental Retardation

Peter M. Krawitz,[1,2,3] Yoshiko Murakami,[4] Jochen Hecht,[2,3] Ulrike Krüger,[1] Susan E. Holder,[5] Geert R. Mortier,[6] Barbara Delle Chiaie,[7] Elfride De Baere,[7] Miles D. Thompson,[8] Tony Roscioli,[9,10] Szymon Kielbasa,[11] Taroh Kinoshita,[4] Stefan Mundlos,[1,2,3] Peter N. Robinson,[1,2,3,12,*] and Denise Horn[1,12,*]

Hyperphosphatasia with mental retardation syndrome (HPMRS), an autosomal-recessive form of intellectual disability characterized by facial dysmorphism, seizures, brachytelephalangy, and persistent elevated serum alkaline phosphatase (hyperphosphatasia), was recently shown to be caused by mutations in *PIGV*, a member of the glycosylphosphatidylinositol (GPI)-anchor-synthesis pathway. However, not all individuals with HPMRS harbor mutations in this gene. By exome sequencing, we detected compound-heterozygous mutations in *PIGO*, a gene coding for a membrane protein of the same molecular pathway, in two siblings with HPMRS, and we then found by Sanger sequencing further mutations in another affected individual; these mutations cosegregated in the investigated families. The mutant transcripts are aberrantly spliced, decrease the membrane stability of the protein, or impair enzyme function such that GPI-anchor synthesis is affected and the level of GPI-anchored substrates localized at the cell surface is reduced. Our data identify *PIGO* as the second gene associated with HPMRS and suggest that a deficiency in GPI-anchor synthesis is the underlying molecular pathomechanism of HPMRS.

More than 100 cell-surface proteins are attached to the plasma membrane by covalent attachment to a glycosylphosphatidylinositol (GPI) anchor that is assembled in the endoplasmic reticulum (ER) and added to the C terminus of the proteins. Biosynthesis of GPI anchors involves more than 30 different genes.[1] Genetic defects in various components of the GPI-anchor-synthesis pathway have been identified in a number of phenotypically diverse diseases that are now also referred to as deficiencies of the GPI-anchor-glycosylation pathway; these diseases belong to a subclass of congenital disorders of glycosylation.[2] Somatic mutations in the X-linked gene phosphatidylinositol glycan class A (*PIGA*, MIM 311770) in hematopoietic stem cells cause paroxysmal nocturnal hemoglobinuria, which manifests as bone-marrow failure, hemolytic anemia, smooth-muscle dystonias, and thrombosis (MIM 300818);[3] germline mutations in this gene result in a severe neurological phenotype (MIM 300868).[4] Germline mutations in *PIGL* (MIM 605947), a gene of the early GPI-anchor glycosylation, cause CHIME syndrome (MIM 280000).[5] Germline promoter mutations in phosphatidylinositol glycan class M (*PIGM* [MIM 610273]; Figure 1) result in a severe deficiency of GPI-anchored proteins (GPI-AP) and were found in individuals with portal- and hepatic-vein thrombosis and intractable absence seizures (MIM 610293).[6] An autosomal-recessive syndrome caused by mutations in phosphatidylinositol glycan class N (*PIGN* [MIM 606097]; Figure 1) and characterized by dysmorphic features and multiple congenital anomalies, severe neurological impairment, chorea, and seizures leading to early death was described (MIM 614080).[7] We have recently identified mutations in phosphatidylinositol glycan class V (*PIGV* [MIM 610274]; Figure 1) in individuals with HPMRS (MIM 239300).[8–10] However, mutations in this gene are only found in approximately half of the individuals with HPMRS. The purpose of the current study was therefore to investigate the molecular etiology of HPMRS in *PIGV*-negative individuals.

This study was approved by the Charité University Medicine ethics board, and informed consent was obtained from responsible persons (parents) on behalf of all study participants.

We performed whole-exome sequencing of all subjects in family A (see Figure 2 for photos, Figure 4A for a pedigree, and Table S1, available online, for clinical details). The affected sisters, who are 12 and 15 years old, are offspring of nonconsanguineous healthy parents of white British origin.

[1]Institute for Medical Genetics and Human Genetics, Charité Universitätsmedizin, 13353 Berlin, Germany; [2]Berlin Brandenburg Center for Regenerative Therapies, Charité Universitätsmedizin, 13353 Berlin, Germany; [3]Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany; [4]Department of Immunoregulation, Research Institute for Microbial Diseases and World Premier International Immunology Frontier Research Center, Osaka University, Osaka 565, Japan; [5]North West Thames Regional Genetics Service, The North West London Hospitals National Health Service Trust, Harrow HA1 3UJ, UK; [6]Department of Medical Genetics, Antwerp University Hospital and University of Antwerp, 2650 Edegem (Antwerp), Belgium; [7]Center for Medical Genetics Ghent, Ghent University Hospital, B-9000 Ghent, Belgium; [8]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON M5G 1L5, Canada; [9]School of Women's and Children's Health, Sydney Children's Hospital, University of New South Wales, Sydney, Randwick NSW 2031, Australia; [10]Department of Human Genetics, University Medical Centre St. Radboud, 6525 Nijmegen, The Netherlands; [11]Center for Human and Clinical Genetics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands
[12]These authors contributed equally to this work
*Correspondence: peter.robinson@charite.de (P.N.R.), denise.horn@charite.de (D.H.)
DOI 10.1016/j.ajhg.2012.05.004.

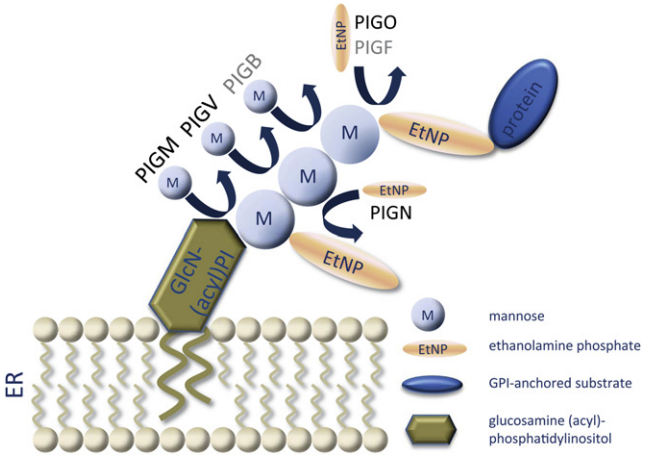


Figure 1. Schematic Illustration of Biochemical Reactions of Late GPI-Anchor Synthesis
The first, second, and third mannose residues are sequentially transferred to GlcN-(acyl)PI by PIGM, PIGV, and PIGB. EtNP is transferred to the first mannose by PIGN and to the third mannose by PIGO and PIGF. Proteins for which the corresponding mutated genes are known to cause congenital disorders of GPI-anchor glycosylation are colored black.

Figure 2. Affected Individuals from Family A
(A) Facial appearance of individual II-1 at the age of 15 years.
(B) Individual II-2 at the age of 12 years.
(C) Nail hypoplasia of the second and fourth digits and absent nail of the fifth digit in individual II-1.
(D) Broad hallux, small nails of the second and third toes, and aplasia of the nails of the fourth and fifth digits in individual II-1.

Genomic DNA of all family members was enriched with the target region of all human consensus coding sequence (CCDS) exons with Agilent's SureSelect Human All Exon Kit according to the manufacturer's protocol. Single-read clusters were then generated on the Cluster Station (Illumina). The captured, purified, and clonally amplified library targeting the exome was then sequenced on an Illumina Genome Analyzer II. Whole-exome sequencing (with two lanes of 120 bp unpaired reads) was performed according to the manufacturer's protocol and resulted in more than 5 Gb of high-quality short-read sequence data.

Novoalign was used for the alignment of the sequence reads to the human genome (GRCh37). The percentage alignment of the reads to the targeted exome was calculated with Perl scripts and Bedtools.[11] In each individual from family A, around 90% of the target region was covered by more than ten unique sequence reads (Figure S1). Samtools and Perl scripts were used for the detection of single-nucleotide variants as well as small indels (<20 bp) on the short-read alignments.[12–14] After common polymorphisms that are also listed in dbSNP (build 132) were filtered out, all detected variants were reduced to family-specific rare variants and annotated with ANNOVAR.[15] Assuming an autosomal-recessive pattern of inheritance with 100% penetrance of the phenotype, we filtered for genes with rare homozygous or compound-heterozygous nonsynonymous variants in the affected siblings and identified phosphatidylinositol glycan class O (*PIGO*) as the single candidate gene (Table S2). The two detected variants in *PIGO* were c.2869C>T (p.Leu957Phe) (NM_032634.3) and c.2361dup (i.e., the mutation inserts an additional cytosine residue into a homopolymer tract consisting of seven cytosine residues), which led to a frameshift (p.Thr788Hisfs*5) (Table S3). These variants were compound heterozygous in the affected sisters and were thus compatible with an autosomal-recessive mode of inheritance. The mother was found to be heterozygous for c.2869C>T, and the father was heterozygous for c.2361dup (Figure 4 A and Figure S2).

After validating these variants by Applied Biosystems Sanger sequencing, we subsequently screened 11 unrelated individuals without *PIGV* mutations for mutations in *PIGO*. We identified the compound-heterozygous candidate mutations c.2869C>T and c.3069+5G>A in individual II-1 from family B (see Figure 3 for photos, Figure 4B for a pedigree, Table S1 for clinical details, and Table S3). Individual II-1 is the second child of nonconsanguineous parents of European descent. The mother is heterozygous for c.3069+5G>A, and two unaffected siblings are also heterozygous for one of the two detected mutations.

We hypothesized that the intronic mutation, c.3069+5G>A, identified in this family would interfere with splicing of the transcript, and we analyzed the effect of this variant on the RNA level. Approximately 3 μg of RNA was isolated from a blood sample of the mother carrying this variant and was used for the first-strand cDNA synthesis. The quality of cDNA was verified by amplification of β-actin cDNA. *PIGO* transcripts were amplified and sequenced from this cDNA pool. The intronic mutation c.3069+5G>A results in an aberrant

Figure 3. Individual II-1 from Family B at Different Ages
(A) When individual II-1 was 6 weeks of age, facial dysmorphism included wide and downward-slanting palpebral fissures, a broad nasal bridge and tip, ptosis of the right eye, a tented upper lip, large ears with fleshy and uplifted ear lobules, and facial asymmetry.
(B) Facial appearance when individual II-1 was 9 months old.
(C) At the age of 18 months.
(D) Nail hypoplasia of the second and fifth digits and clinodactyly V.
(E) Hand radiograph when individual II-1 was 1 week old. Note brachytelephalangy II to V, mostly affecting fingers II and V, and a broad distal phalanx of the thumb.
(F) Nail hypoplasia of all toes.

splicing product with a skipped exon 9 (Figure 4C); this product was not observed in 13 cDNA controls (Figure S3). The deletion of this 215 bp exon causes a frameshift followed by a premature stop codon. According to data from the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project, there is one heterozygous individual for this intronic mutation out of 5,379 tested individuals, which is consistent with the expected incidence of the disease.

In mammals, *PIGO* encodes a 1,089 amino acid protein, GPI ethanolamine phosphate transferase 3 (also known as phosphatidylinositol-glycan biosynthesis class O), that is involved in GPI biosynthesis.[16,17] The substitution p.Leu957Phe affects the second of four leucine residues in a polyleucine stretch within a hydrophobic transmembrane domain of PIGO. The residue is evolutionarily highly conserved in most species, including mammals, frogs, and zebrafish (Figure S4), and the effect of the detected substitution was classified as disease causing by MutationTaster[18] and Polyphen.[19] The heterozygote frequency of all three alleles in the European population is below 0.0005, which is expected for rare recessive disorders.[20]

We first investigated the influence of two *PIGO* mutations on PIGO function. To test the variants p.Leu957Phe and p.Thr788Hisfs*5 for effects on PIGO function, we cloned a human *PIGO* cDNA from a cDNA library derived from Hep3B (a hepatoma cell line) cells, tagged it with

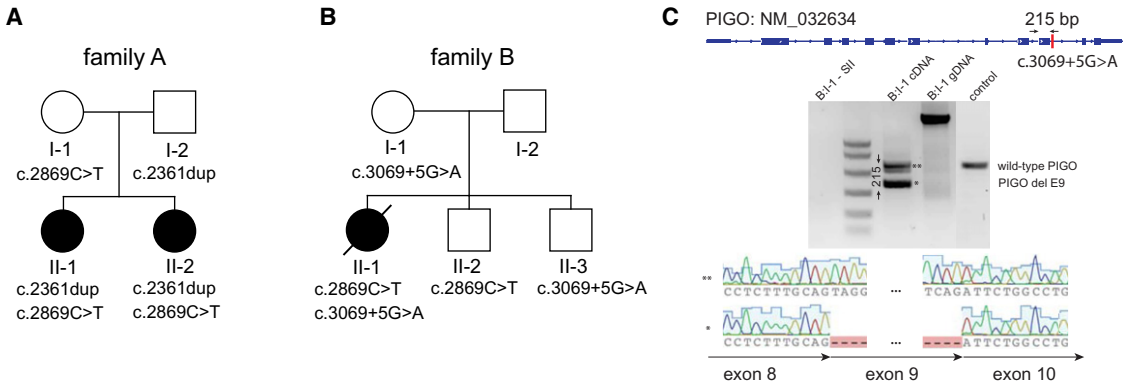


Figure 4. Mutations in *PIGO*
(A) *PIGO* mutations in family A as demonstrated by whole-exome sequencing. Both affected daughters were found to be compound heterozygous for the *PIGO* mutations c.2869C>T and c.2361dup. The father is heterozygous for c.2361dup, and the mother is heterozygous for c.2869C>T.
(B) The affected individual from family B is compound heterozygous for c.2869C>T and c.3069+5G>A. Her mother and healthy brother II-3 are heterozygous only for c.3069+5G>A, and her healthy brother II-2 is heterozygous only for c.2869C>T.
(C)The intronic mutation results in an aberrant splicing product of transcript NM_032634, which is missing 215-bp-long exon 9.

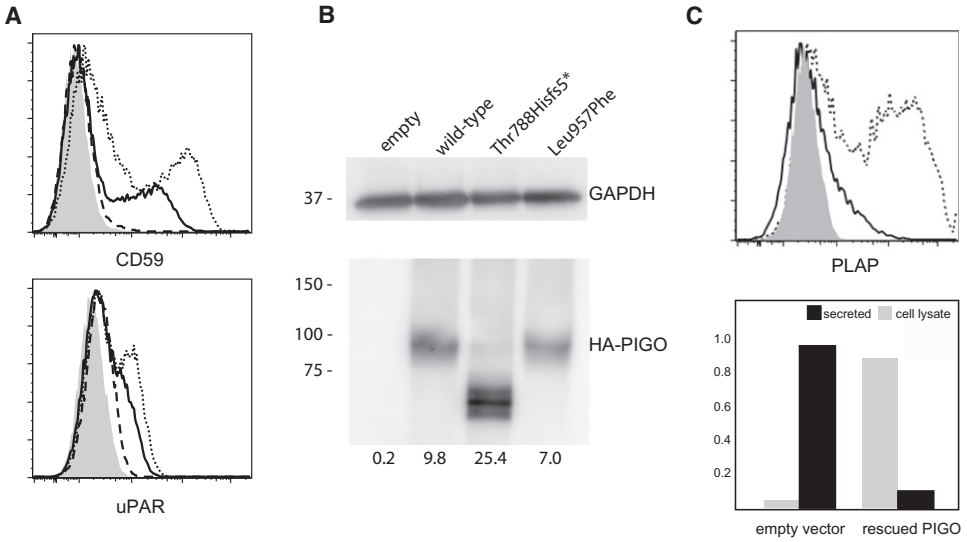


Figure 5. PIGO Activity Is Required for Linking GPI-Anchored Substrates to the Cell Membrane
(A) PIGO-deficient CHO cells were transiently transfected with human wild-type (dotted lines), p.Thr788Hisfs*5 (dashed line), or p.L957F (solid lines) *PIGO* cDNA expression constructs. Restoration of the levels of CD59 at the cell surface and of uPAR was assessed 2 days later. Wild-type PIGO efficiently restored levels of CD59 at the cell surface and of uPAR, whereas Thr788Hisfs*5 PIGO did not restore the level of CD59 at all and the Leu957Phe PIGO induced only very low levels of CD59 and uPAR. The shadowed area indicates an empty-vector transfectant (control).
(B) PIGO levels. The level of the truncated Thr788Hisfs*5 PIGO (lane 3) was about 2.5× higher than that of wild-type PIGO (lane 2), and the level of Leu957Phe PIGO (lane 4) was slightly lower than that of wild-type PIGO (lane 2).
(C) The level of PLAP at the cell surface after cotransfection with *PIGO* into PIGO-deficient CHO cells. PIGO-deficient CHO cells were transiently transfected with pME HA-PLAP together with pME PIGO (dotted line) or an empty vector (solid line). The level of PLAP at the cell surface was analyzed by fluorescence-activated cell sorting. PLAP activity was measured in culture medium and cell lysates after cotransfection of PLAP and PIGO cDNAs into PIGO-deficient CHO cells. Relative ALP activity was measured in culture medium (black bars) and in cell lysates (dark gray bar) against the total ALP activity in PIGO-restored CHO cells. Restoration of PIGO activity reduces ALP activity in the medium and increases activity at the cell membrane.

FLAG at the N-terminus, and subcloned it into pME.[21] PIGO mutants were generated by site-directed mutagenesis. Mutant and wild-type PIGO plasmids were transfected by electroporation into human CD59-expressing PIGO-deficient CHO cells that were derived from aerolysin-resistant clones from chemically mutagenized Chinese hamster ovary (CHO) cells as previously described.[22] We determined the levels of GPI-APs CD59 (at the cell surface) and endogenous urokinase plasminogen activator receptor (uPAR) by staining cells with anti-CD59 (5H8) and anti-hamster uPAR antibodies, and we used a flow cytometer (Cant II; BD Biosciences, Franklin Lakes, NJ) with Flowjo software (Tommy Digital, Tokyo, Japan) to analyze them. Wild-type PIGO efficiently restored the levels of CD59 at the cell surface and of uPAR, whereas p.Thr788Hisfs*5 PIGO did not rescue protein levels at all and the Leu957Phe PIGO induced only very low levels of CD59 and uPAR (Figure 5A). Levels of PIGO in cells were determined by immunoblot analysis. Band intensities of endogenous glyceraldehyde-3-phosphate dehydrogenase (GAPDH) were used as loading controls. Compared with that of wild-type PIGO, the level of Leu957Phe was slightly reduced, whereas the c.2361dup mutation resulted in an increased level of the truncated Thr788Hisfs* protein (Figure 5B).

The three individuals carrying *PIGO* mutations were born with normal measurements, but they had anal stenosis (individuals II-1 and II-2 from family A) or anal atresia with perineal fistula (individual II-1 from family B) (Figures 2 and 3 and Table S1). In individual II-1 from family B, additional malformations included an atrial septal defect, peripheral pulmonary stenosis, left coronal synostosis resulting in plagiocephaly, and an enlarged supratentorial ventricular system. Growth development was delayed in individuals II-1 from family A and II-1 from family B, who also showed marked microcephaly of −5.5 standard deviations (SDs) at the age of 20 months. Psychomotor development was severely retarded in individuals II-1 from family A and II-1 from family B and was moderately delayed in individual II-2 of family A. Individual II-1 from family B developed tonic-clonic seizures at the age of 21 months and died at the age of 22 months as a result of a convulsive crisis. Their common facial signs included wide-set eyes that appeared large because of long palpebral fissures, a short nose with a broad nasal bridge and nasal tip, and a tented mouth. Their fingers showed nail hypoplasia, especially of the second, fourth, and fifth digits, and absent nails of the fifth digits (individuals II-1 and II-2 from family A). Their halluces were broad, but the toes showed small nails or aplasia of nails, especially of the fourth and fifths digits. Serum alkaline phosphatase (ALP) activity was elevated in repeated tests (1,872 U/l in individual II-1 and 1,381 U/l in individual II-2 from family A [the normal range is 200–700 U/l] and 1,436 U/l in individual II-1 from family B [the normal range is 124–341 U/l]).

We next hypothesized that the hyperphosphatasia observed in our patients was due to the *PIGO* defect. We performed measurements of relative placental ALP (PLAP) activity in cell lysates and medium of PIGO-deficient CHO cells, and we compared them with those of wild-type cells. The mutant CHO cells were cotransfected with pME HA-PLAP (a previously described construct[23]), pME Luc, and either a PIGO-expressing plasmid (pME F-PIGO) or an empty vector (pME). Media were changed 6 hr later, and the ALP activity in cell lysates and culture media was measured on the following day with the SEAP assay kit (Clontec). Luciferase activities of cell lysates were measured with the Luciferase assay kit (Promega) and were used for the normalization of gene expression. Relative activity in the culture medium was the ratio of normalized ALP activity in the culture medium to the total normalized ALP activity of PIGO-rescued cells. PLAP failed to express on the cell surface (Figure 5C, upper panel), but 90% of the activity was found to reside in the medium from the PIGO-deficient cells (Figure 5C, lower panel). In contrast, transfection with the PIGO-expressing vector restored the surface expression (Figure 5C, upper panel) and prevented oversecretion of PLAP; the majority of PLAP activity remained in the cell (Figure 5C, lower panel). These results indicate that hyperphosphatasia is a result of the release of ALP into serum, which itself is due to a GPI deficiency caused by the *PIGO* mutation.

A combination of hyperphosphatasia, intellectual disability, and various neurological problems, mainly seizures, has been described by different authors, and the condition has been designated as "hyperphosphatasia with mental retardation."[24,25] In a subset of individuals affected by HPMRS in addition to specific facial features and brachytelephalangy, missense mutations of *PIGV* have been identified.[8–10] Anorectal malformations, Hirschsprung disease, and other organ malformations broaden the clinical spectrum associated with *PIGV* mutations.[8–10] Because *PIGV* mutations are only found in some of the affected individuals with the core manifestations, genetic heterogeneity seems to be likely. Our study identifies compound-heterozygous mutations in *PIGO* as a further genetic cause of HPMRS. The characteristic facial appearance, moderate-to-severe developmental delay, hypoplastic or even absent terminal phalanges (including nails), and hyperphosphatasia were present in all affected individuals studied here. In individual II-1 from family B, a significant deceleration of head growth was seen during infancy (the occipital frontal circumference fell to more than 5 SDs below the mean). Whereas individuals carrying *PIGV* mutations often show growth parameters at or above the mean, individuals with *PIGO* mutations seem to show more pronounced growth delay. Malformations of the urinary system and heart should be considered as part of this condition.

Our results suggest that PIGO is essential for GPI anchoring of GPI-APs alkaline phosphatase, CD59, and uPAR, and we show that the two tested *PIGO* mutations have a deleterious effect on PIGO function in a cell-based assay. We have recently shown that GPI transamidase plays a critical role in the secretion of proproteins in the absence of mature GPI, as observed in PIGV-deficient HPMRS.[23] The presence of mannose residue of immature GPI is critical for cleavage of the proproteins by GPI transamidase, which provides an explanation for the observation that PIGV deficiency, which affects a relatively late step in the GPI-biosynthesis pathway, leads to hyperphosphatasia, whereas PIGM deficiency, which leads to a defect in the first GPI mannosyltransferase, does not.[23] This observation provides a plausible explanation for the fact that mutations in both *PIGV*, which encodes a protein that transfers the second mannose residue to the GPI anchor,[26] and *PIGO*, which encodes a protein that transfers phosphoethanolamine to the third mannose residue of the GPI anchor,[16] are associated with HPMRS. Determining whether there are any clinical differences between these groups will require phenotypic characterization of more individuals with *PIGO* and *PIGV* mutations. Given the fact that neither *PIGV* nor *PIGO* mutations were identified in ten individuals with a similar phenotype, it seems likely that mutations in other GPI-pathway genes might represent further etiologies of HPMRS.

Supplemental Data

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

Acknowledgments

This work was supported by grant 0313911 from the Bundesministerium für Forschung und Technologie and by grants from the Ministry of Education, Culture, Sports, Science, and Technology and the Ministry of Health, Labour, and Welfare of Japan. We wish to thank all patients and their families involved in this study for their generous help.

Received: March 14, 2012
Revised: April 19, 2012
Accepted: May 11, 2012
Published online: June 7, 2012

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes, http://www.1000genomes.org
Agilent eArray, https://earray.chem.agilent.com/earray/
Online Mendelian Ineritance in Man (OMIM), http://www.omim.org
NHLBI Exome Sequencing Project (ESP), http://evs.gs.washington.edu/EVS/

References

1. Kinoshita, T., Fujita, M., and Maeda, Y. (2008). Biosynthesis, remodelling and functions of mammalian GPI-anchored proteins: Recent progress. J. Biochem. *144*, 287–294.

2. Jaeken, J. (2011). Congenital disorders of glycosylation (CDG): It's (nearly) all in it!. J. Inherit. Metab. Dis. *34*, 853–858.
3. Takeda, J., Miyata, T., Kawagoe, K., Iida, Y., Endo, Y., Fujita, T., Takahashi, M., Kitani, T., and Kinoshita, T. (1993). Deficiency of the GPI anchor caused by a somatic mutation of the PIG-A gene in paroxysmal nocturnal hemoglobinuria. Cell *73*, 703–711.
4. Johnston, J.J., Gropman, A.L., Sapp, J.C., Teer, J.K., Martin, J.M., Liu, C.F., Yuan, X., Ye, Z., Cheng, L., Brodsky, R.A., and Biesecker, L.G. (2012). The phenotype of a germline mutation in PIGA: The gene somatically mutated in paroxysmal nocturnal hemoglobinuria. Am. J. Hum. Genet. *90*, 295–300.
5. Ng, B.G., Hackmann, K., Jones, M.A., Eroshkin, A.M., He, P., Wiliams, R., Bhide, S., Cantagrel, V., Gleeson, J.G., Paller, A.S., et al. (2012). Mutations in the Glycosylphosphatidylinositol Gene PIGL Cause CHIME Syndrome. Am. J. Hum. Genet. *90*, 685–688.
6. Almeida, A.M., Murakami, Y., Layton, D.M., Hillmen, P., Sellick, G.S., Maeda, Y., Richards, S., Patterson, S., Kotsianidis, I., Mollica, L., et al. (2006). Hypomorphic promoter mutation in PIGM causes inherited glycosylphosphatidylinositol deficiency. Nat. Med. *12*, 846–851.
7. Maydan, G., Noyman, I., Har-Zahav, A., Neriah, Z.B., Pasmanik-Chor, M., Yeheskel, A., Albin-Kaplanski, A., Maya, I., Magal, N., Birk, E., et al. (2011). Multiple congenital anomalies-hypotonia-seizures syndrome is caused by a mutation in PIGN. J. Med. Genet. *48*, 383–389.
8. Horn, D., Krawitz, P., Mannhardt, A., Korenke, G.C., and Meinecke, P. (2011). Hyperphosphatasia-mental retardation syndrome due to PIGV mutations: Expanded clinical spectrum. Am. J. Med. Genet. A. *155A*, 1917–1922.
9. Krawitz, P.M., Schweiger, M.R., Rödelsperger, C., Marcelis, C., Kölsch, U., Meisel, C., Stephani, F., Kinoshita, T., Murakami, Y., Bauer, S., et al. (2010). Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nat. Genet. *42*, 827–829.
10. Thompson, M.D., Roscioli, T., Marcelis, C., Nezarati, M.M., Stolte-Dijkstra, I., Sharom, F.J., Lu, P., Phillips, J.A., Sweeney, E., Robinson, P.N., et al. (2012). Phenotypic variability in hyperphosphatasia with seizures and neurologic deficit (Mabry syndrome). Am. J. Med. Genet. A. *158A*, 553–558.
11. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.
12. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.
13. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
14. Krawitz, P., Rödelsperger, C., Jäger, M., Jostins, L., Bauer, S., and Robinson, P.N. (2010). Microindel detection in short-read sequence data. Bioinformatics *26*, 722–729.
15. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.
16. Hong, Y., Maeda, Y., Watanabe, R., Inoue, N., Ohishi, K., and Kinoshita, T. (2000). Requirement of PIG-F and PIG-O for transferring phosphoethanolamine to the third mannose in glycosylphosphatidylinositol. J. Biol. Chem. *275*, 20911–20919.
17. Flury, I., Benachour, A., and Conzelmann, A. (2000). YLL031c belongs to a novel family of membrane proteins involved in the transfer of ethanolaminephosphate onto the core structure of glycosylphosphatidylinositol anchors in yeast. J. Biol. Chem. *275*, 24458–24465.
18. Schwarz, J.M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. Nat. Methods *7*, 575–576.
19. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
20. Bell, C.J., Dinwiddie, D.L., Miller, N.A., Hateley, S.L., Ganusova, E.E., Mudge, J., Langley, R.J., Zhang, L., Lee, C.C., Schilkey, F.D., et al. (2011). Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci. Transl. Med. *3*, ra4.
21. Takebe, Y., Seiki, M., Fujisawa, J., Hoy, P., Yokota, K., Arai, K., Yoshida, M., and Arai, N. (1988). SR alpha promoter: An efficient and versatile mammalian cDNA expression system composed of the simian virus 40 early promoter and the R-U5 segment of human T-cell leukemia virus type 1 long terminal repeat. Mol. Cell. Biol. *8*, 466–472.
22. Hong, Y., Ohishi, K., Inoue, N., Kang, J.Y., Shime, H., Horiguchi, Y., van der Goot, F.G., Sugimoto, N., and Kinoshita, T. (2002). Requirement of N-glycan on GPI-anchored proteins for efficient binding of aerolysin but not Clostridium septicum alpha-toxin. EMBO J. *21*, 5047–5056.
23. Murakami, Y., Kanzawa, N., Saito, K., Krawitz, P.M., Mundlos, S., Robinson, P.N., Karadimitris, A., Maeda, Y., and Kinoshita, T. (2012). Mechanism for release of alkaline phosphatase caused by glycosylphosphatidylinositol deficiency in patients with hyperphosphatasia mental retardation syndrome. J. Biol. Chem. *287*, 6318–6325.
24. Mabry, C.C., Bautista, A., Kirk, R.F., Dubilier, L.D., Braunstein, H., and Koepke, J.A. (1970). Familial hyperphosphatase with mental retardation, seizures, and neurologic deficits. J. Pediatr. *77*, 74–85.
25. Kruse, K., Hanefeld, F., Kohlschütter, A., Rosskamp, R., and Gross-Selbeck, G. (1988). Hyperphosphatasia with mental retardation. J. Pediatr. *112*, 436–439.
26. Kang, J.Y., Hong, Y., Ashida, H., Shishioh, N., Murakami, Y., Morita, Y.S., Maeda, Y., and Kinoshita, T. (2005). PIG-V involved in transferring the second mannose in glycosylphosphatidylinositol. J. Biol. Chem. *280*, 9489–9497.

2.8 Mutationen in *PGAP2*, einem Gen der GPI-Anker-Reifung, als Ursache für HPMRS

Krawitz, P.M., Murakami, Y., Riess, A., Hietala, M., Krüger, U., Zhu, N., Kinoshita, T., Mundlos, S., Hecht, J., Robinson, P.N., et al. (2013). PGAP2 mutations, affecting the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation syndrome. American Journal of Human Genetics 92, 584-589.

Die Identifikation pathogener Mutationen in zwei unterschiedlichen Genen des GPI-Ankersynthese-Weges durch uns zeigte bereits, dass es sich bei Hyperphosphatasie mit Mentaler Retardierung um eine heterogene Erkrankung handelt. Um die Suche nach den krankheitsverursachenden Mutationen in Patienten mit Mabry Syndrom, die weder pathogene Veränderungen in *PIGV* noch in *PIGO* aufweisen, effektiver zu gestalten, etablierten wir einen NGS-basierten Screening Ansatz, mit dem die kodierenden Abschnitte aller bekannten Gene, die an der GPI-Ankersynthese und –Reifung beteiligt sind, untersucht werden können. Die gezielte Anreicherung und anschließende Sequenzierung dieser Exons ist deutlich kostengünstiger als eine Exom-Sequenzierung und erlaubte uns daher die gezielte Analyse von 13 Patienten mit Entwicklungsverzögerung und erhöhter Serum-Aktivität der alkalischen Phosphatase.

In dieser Kohorte konnten wir in zwei Patienten Sequenzvarianten in *PGAP2* identifizieren, eine homozygote Mutation, c. 380T>C, sowie die compound-heterozygoten Mutationen c.46C>T und c.479C>T. Zeitgleich fand die Arbeitsgruppe von Rami Abou Jamra in zwei unterschiedlichen konsanguinen Familien, in denen mehrere Mitglieder von nicht-syndromaler Entwicklungsverzögerung betroffen waren, die homozygoten Mutationen c.296A>G und c.539G>C (Hansen, et al., 2013).

Bei allen identifizierten Sequenzvarianten handelt es sich um missense Mutationen (p.Arg16Trp, p.Tyr99Cys, p.Leu127Ser, p.Thr160Ile, p.Arg177Pro), für die im CHO-Test-System eine funktionseinschränkende Wirkung auf PGAP2 nachgewiesen wurde. In Übereinstimmung damit steht auch die bei einem solchen *in vitro* Befund zu erwartende Hyperphosphatasie, die bei allen Patienten bestätigt werden konnte.

Umso bemerkenswerter ist das phänotypische Spektrum der Patienten hinsichtlich der weiteren für das Mabry Syndrom typischen Auffälligkeiten. Der Patient mit der p.Leu127Ser Mutation wies einen Atriumseptumdefekt, eine deutliche muskuläre Hypotonie, sowie eine Nagelhypoplasie des Endgliedes des fünften Fingers auf. Zudem erforderte eine Gaumenspalte und ein Morbus Hirschsprung eine chirurgische Intervention.

Bei der Patientin mit den heterozygoten Mutationen p.Arg16Trp und p.Thr160Ile hingegen liegen keine angeborenen Fehlbildungen vor. Spracherwerb und frühe Schulzeit verliefen normal. Im achten Lebensjahr kam es zu tonisch klonischen Krampfanfällen, die mit Valproatsäure therapiert wurden. Die antiepileptische Medikation konnte jedoch ab dem 22. Lebensjahr abgesetzt werden, ohne dass es zu Rückfällen kam. Ab dem 12. Lebensjahr erfolgte der weitere Bildungsweg auf einer Förderschule und die Patientin befindet sich nun in einem unterstützten Beschäftigungsverhältnis.

Mit dieser Arbeit haben wir daher ein weiteres, neues Krankheitsgen für das Mabry Syndrom beschrieben und zugleich illustriert, wie ausgeprägt die klinische Variabilität sein kann.

PGAP2 Mutations, Affecting the GPI-Anchor-Synthesis Pathway, Cause Hyperphosphatasia with Mental Retardation Syndrome

Peter M. Krawitz,[1,2,3] Yoshiko Murakami,[4,5] Angelika Rieß,[6] Marja Hietala,[7] Ulrike Krüger,[1] Na Zhu,[1] Taroh Kinoshita,[4,5] Stefan Mundlos,[1,2,3] Jochen Hecht,[2,3] Peter N. Robinson,[1,2,3,8,*] and Denise Horn[1,8,*]

Recently, mutations in genes involved in the biosynthesis of the glycosylphosphatidylinositol (GPI) anchor have been identified in a new subclass of congenital disorders of glycosylation (CDGs) with a distinct spectrum of clinical features. To date, mutations have been identified in six genes (*PIGA*, *PIGL*, *PIGM, PIGN*, *PIGO*, and *PIGV*) encoding proteins in the GPI-anchor-synthesis pathway in individuals with severe neurological features, including seizures, muscular hypotonia, and intellectual disability. We developed a diagnostic gene panel for targeting all known genes encoding proteins in the GPI-anchor-synthesis pathway to screen individuals matching these features, and we detected three missense mutations in *PGAP2*, c.46C>T, c.380T>C, and c.479C>T, in two unrelated individuals with hyperphosphatasia with mental retardation syndrome (HPMRS). The mutations cosegregated in the investigated families. *PGAP2* is involved in fatty-acid GPI-anchor remodeling, which occurs in the Golgi apparatus and is required for stable association between GPI-anchored proteins and the cell-surface membrane rafts. Transfection of the altered protein constructs, p.Arg16Trp (NP_001243169.1), p.Leu127Ser, and p.Thr160Ile, into PGAP2-null cells showed only partial restoration of GPI-anchored marker proteins, CD55 and CD59, on the cell surface. In this work, we show that an impairment of GPI-anchor remodeling also causes HPMRS and conclude that targeted sequencing of the genes encoding proteins in the GPI-anchor-synthesis pathway is an effective diagnostic approach for this subclass of CDGs.

In the last 2 years, individuals with characteristic phenotypic features including severe neurological abnormalities were reported to have defects in the GPI-anchor-biosynthesis pathway, representing a new subclass of congenital disorders of glycosylation (CDGs).[1] Mutations in *PIGV* (MIM 610274) and *PIGO* (MIM 614730) were shown to cause hyperphosphatasia with mental retardation syndrome (HPMRS [MIM 239300 and 214749]), which is also referred to as Mabry syndrome.[2–7] Individuals with coloboma, congenital heart disease, ichthyosiform dermatosis, mental retardation, and ear anomalies syndrome (CHIME [MIM 280000]), also known as Zunich neuroectodermal syndrome, were reported to have mutations in *PIGL* (MIM 605947).[8] A hypomorphic promoter mutation in *PIGM* (MIM 610273) causes portal venous thrombosis and absence seizures (MIM 610293).[9] Germline mutations in *PIGN* (MIM 606097) and *PIGA* (MIM 311770) cause severe syndromes with multiple congenital anomalies, hypotonia, and seizures (MCAHS), now referred to as MCAHS1 (MIM 614080) and MCAHS2 (MIM 300868). Similar to other disorders of glycosylation, disorders caused by mutations interfering with the GPI-anchor pathway are characterized by a remarkable phenotypic diversity whereby the clinical impact seems to depend on the severity of the mutation.[10] To date, all identified mutations are hypomorphic and no complete loss of function has been reported in any of these genes. Although distinct phenotypic features seem to be exclusive to single genes or are shared only by a subgroup, the phenotypic features of intellectual disability, seizures, and muscular hypotonia are present in a majority of the individuals described so far.

We therefore included 13 individuals with intellectual disability and elevated serum alkaline phosphatese (ALP) in a mutation screen of all genes encoding proteins in the GPI-anchor-biosynthesis pathway. In these individuals, mutations in *PIGV* had been excluded by Sanger sequencing. The Charité University Medicine ethics board approved this study, and we obtained informed consent from the responsible persons (parents) on behalf of all study participants. In this work, we report the molecular findings in two unrelated individuals (II-1 of family A and II-1 of family B in Figure 1) with the clinical diagnosis of HPMRS but without identifiable mutations in *PIGV* and *PIGO* (Table 1). We performed targeted capture sequencing in the affected individuals of families A and B. Family A is of Finnish origin, and family B is of Turkish origin. For targeted enrichment of exons of all known genes involved in GPI-anchor synthesis, we designed a customized SureSelect library (Agilent) comprising 1,202 different 120 bp oligonucleotide baits in total (see Supplemental Data, available online). Genomic DNA of both individuals was enriched

[1]Institute for Medical Genetics and Human Genetics, Charité Universitätsmedizin, 13353 Berlin, Germany; [2]Berlin Brandenburg Center for Regenerative Therapies, Charité Universitätsmedizin, 13353 Berlin, Germany; [3]Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany; [4]Department of Immunoregulation, Research Institute for Microbial Diseases, Osaka University, Osaka 565, Japan; [5]World Premier International Immunology Frontier Research Center, Osaka University, Osaka 565, Japan; [6]Institute for Human Genetics, Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany; [7]Medical Biochemistry and Genetics, University of Turku, 20520 Turku, Finland
[8]These authors contributed equally to this work
*Correspondence: peter.robinson@charite.de (P.N.R.), denise.horn@charite.de (D.H.)
http://dx.doi.org/10.1016/j.ajhg.2013.03.011.

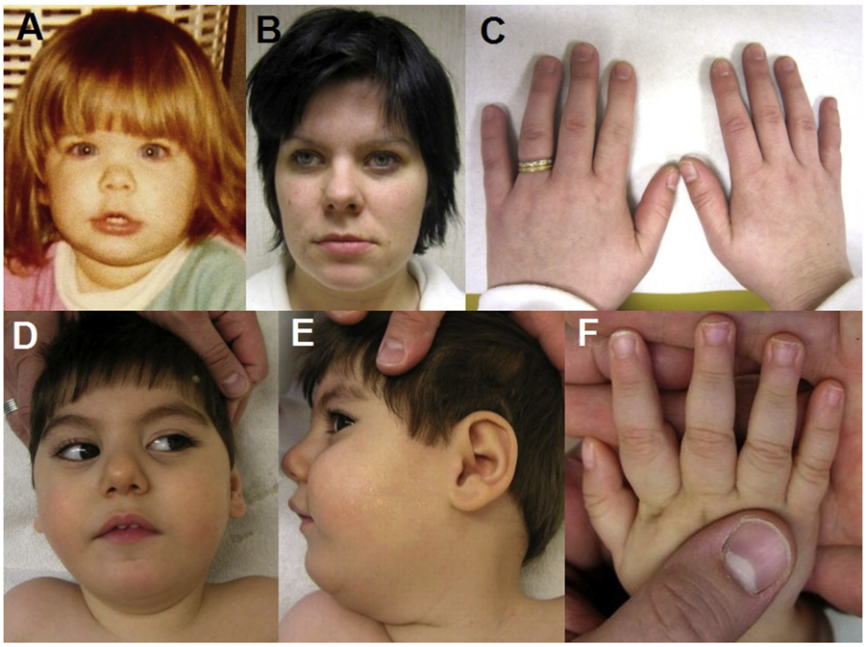


Figure 1. Phenotypic Features of HPMRS Associated with Mutations in *PGAP2*
(A and B) Face of individual A from family A at the ages of 3 (A) and 28 years (B).
(C) Normal-appearing fingernails of the affected individual in family A.
(D and E) Facial dysmorphism of the affected individual in family B at the age of 2 years includes wide palpebral fissures, a short nose with a broad nasal bridge, a tented upper lip, and a small jaw.
(F) Distal tapering of fingers and mild nail hypoplasia of the fifth digit of the affected individual in family B.

for this target region according to the manufacturer's protocol, and this was followed by single-read cluster generation on the Cluster Station (Illumina). The captured, purified, and clonally amplified library was then sequenced on an Illumina Genome Analyzer IIx and mapped to the human reference sequence GRCh37, resulting in a mean coverage of above 300-fold for all exons and more than 10-fold for >95% of the target region (see Figure S1). Variants were detected with SAMtools,[12] annotated with ANNOVAR,[13] and further analyzed in GeneTalk.[14] In individual A, we detected a total of 30 single-nucleotide variants with respect to the reference sequence GRCh37, and these included 14 missense mutations, 9 of which were homozygous (Table S2). Three variants not listed in dbSNP135 coded for heterozygous missense mutations: one in *PIGZ*, c.214G>C (RefSeq accession number NM_025163.2) (p.Asp72His) (RefSeq NP_079439.2), and two in *PGAP2*, c.[46C>T];[479C>T] (RefSeq NM_001256240.1) (p.[Arg16Trp];[p.Thr160Ile]) (RefSeq NP_001243169.1). In individual B, we observed 32 variants, 17 synonymous and 15 homozygous (Table S3). Only one homozygous missense mutation in *PGAP2* (c.380T>C [p.Leu127Ser] [RefSeq NM_001256240.1]) was not listed in dbSNP135.

All missense mutations (c.46C>T, c.380T>C, and c.479C>T) were analyzed for segregation in available family members (Figure 2). In family A, the mother is a carrier for c.46C>T. The healthy brother is a carrier of c.479C>T, allowing us to infer the same genotype for the father, who was not available for analysis. In family B, both parents and one healthy brother are carriers of c.380T>C, whereas one healthy brother has the wild-type sequence.

Individual II-1 of family A is the first child of non-consanguineous Finnish parents. Her younger brother is healthy. There is a family history of febrile seizures and epilepsy, but not of intellectual disability. Her neonatal period was uneventful, and postnatal development was normal. She started to walk at the age of 18 months, and her initial speech development was normal. At that age, her facial dysmorphism was subtle in that she had only a broad nasal bridge and a tented upper lip (Figure 1A and Table 1). From the age of 8 months to the age of 2.5 years, she suffered from febrile seizures. At the age of 8 years, she began to have tonic-clonic seizures, which responded well to valproic acid. At the age of 22 years, her antiepileptic medication was discontinued and she showed no recurrence of seizures. A physical examination at 28 years revealed a height, weight, and head circumference within the normal range. There was no distinctive facial dysmorphism (Figure 1B). Her fingernails appeared to be normal (Figure 1C). A hand radiograph was not available. Individual II-1 of family A started at an ordinary school but has received special education since the age of 12 years, and she currently works in supported employment.

Her serum ALP activity was measured only once during childhood when she was 10 years old. This elevated value (3,470 U/l; the normal range for the corresponding age is 105–400 U/l) was interpreted as a laboratory mistake. When she was 28 years old, ALP was measured again. These values were repeatedly elevated (2,107–2,448 U/l; the normal range is 35–105 U/l).

Individual II-1 of family B is the third child of consanguineous parents of Turkish origin. The family history is unremarkable. Birth length and weight were normal, and the occipitofrontal head circumference (OFC) at birth was 33 cm (−2 SDs). After birth, physical examination revealed a median cleft palate, which was surgically corrected. Chronic constipation and acute illeus led to the diagnosis of Hirschsprung disease, which was histologically confirmed and surgically repaired. Examinations of this tissue or other tissues for intracellular inclusions were not performed. Echocardiography showed an atrial septal defect. Cranial computed tomography revealed hypoplasia of the corpus callosum.

Table 1. Summary of Clinical Findings in HPMRS-Affected Individuals Carrying *PGAP2*, *PIGO* and *PIGV* Mutations

Features	Human Phenotype Ontology ID[11]	Affected Individual in Family A	Affected Individual in Family B	Individuals with *PIGO* Mutations (n = 3)	Individuals with *PIGV* Mutations (n = 14)[a]
Sex	NA	female	male	females	9 females and 5 males
Age at last assessment	NA	28 years	3.5 years	20 months to 15 years	7 months to 17 years
Origin	NA	Finnish	Turkish	European	German, Maroccan, Dutch, Polish, British, and European American
Height (SD)	NA	−0.9	+0.6	−1.4 to −4.2	normal in 13/14
Weight (SD)	NA	normal	−1.0	+0.6 to −3.3	normal in 13/14
OFC (SD)	NA	normal	−4.5	+0.7 to −5.5	normal in 12/14
Hyperphoshatasia[b]	HP:0003155	+	+	3/3	14/14
Intellectual disability[b]	HP:0001263	mild	+	3/3	14/14
Age at walking	NA	18 months	no walking	delayed	delayed
Delayed speech and language development	HP:0000750	−	+	3/3	14/14
Muscular hypotonia	HP:0001252	−	+	3/3	11/12
Seizures	HP:0001250	+	+	1/3	9/12
Apparent hypertelorism	HP:0000316	−	+	3/3	+
Long palpebral fissures	HP:0000637	−	+	3/3	+
Broad nasal bridge	HP: 0000431	+	+	3/3	+
Broad nasal tip	HP:0000455	−	+	3/3	+
Tented upper lip vermilion	HP:0010804	+	+	3/3	+
Brachytelephalangy	HP:0009882	normal appearing fingernails	short fifth fingernail	3/3	14/14
Anorectal abnormalities and/or constipation	HP:0002025 (anal stenosis)	−	+	3/3	6/12
Aganglionic megacolon	HP:0002251	−	+	1/3	2/14
Heart defect	HP:0001631	−	+	+	1/14
Cleft palate	HP:0000175	−	+	0/3	3/14
Hearing impairment	HP:0000365	−	+	0/3	3/14

The following abbreviations are used: NA, not applicable; and OFC, occipitofrontal head circumference.
[a]Not all features were documented in the reported individuals.
[b]Consistent features.

The boy's psychomotor development was severely delayed. At the age of 3.5 years, he was still not able to sit, stand, or walk. At the age of 2 years, he had no speech.

Since he was 7 months old, he has suffered from myoclonic and tonic-clonic seizures, which have responded well to anticonvulsants. Electroencephalography investigations indicated multifocal sharp waves. Brainstem auditory-evoked response demonstrated sensorineural hearing loss. Ophthalmologic examination gave normal results.

Physical examination of this 3.5-year-old male showed a height of 104 cm (+0.6 SD), a weight of 14 kg (−1 SD), marked secondary microcephaly and a head circumference of 45 cm (−4.5 SDs), scoliosis, and severe muscular hypotonia (Table 1). Facial dysmorphism included wide palpebral fissures and a wide mouth (Figures 1D and 1E). His fingers showed broad fingernails and a bilateral hypoplastic fifth fingernail (Figure 1F). ALP activity was elevated in repeated tests (2,022 U/l; the normal range is 120–320 U/l for the corresponding age). Conventional cytogenetic analysis gave normal results. Mutations and deletion of *ZFHX1B* were excluded for ruling out Mowat-Wilson syndrome (MIM 235730).

Thus, both affected individuals presented with intellectual disability, seizures of various degrees, and marked hyperphosphatasia (more than six times the age-adjusted upper limit of the normal range). In addition, mild shortness of fingernails was present in individual II-1 of family B (Table 1).

Whereas individual A, who harbors compound-heterozygous *PGAP2* mutations, shows only mild manifestations regarding neurological involvement and physical features, individual B, who has the homozygous c.380T>C

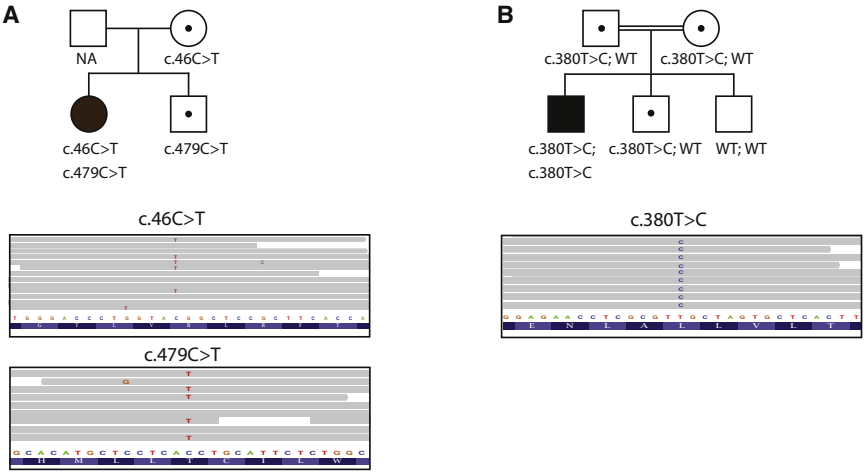


Figure 2. Identification and Segregation of the *PGAP2* Mutations
Pedigrees showing segregation of the HPMRS phenotype with deleterious variants in *PGAP2* in families A (A) and B (B). Circles represent females, squares represent males, filled symbols represent affected individuals, and dots within the symbols represent heterozygotes. Sequence reads show the mutation in short read alignments visualized in integrative genome viewer.

mutation, is severely affected by seizures, muscular hypotonia, and marked intellectual disability, as well as various malformations.

In comparison with the specific phenotypic pattern of all previously reported individuals with *PIGV* and *PIGO* mutations, the phenotype of individual A broadens the clinical range of HPMRS with the absence of syndrome-specific minor anomalies and malformations and only a mild degree of intellectual disability.

PGAP2 is a membrane protein mainly expressed in the Golgi and is required for reacylation of the lysoform intermediate GPI during fatty-acid remodeling.[15] PGAP2 is hypothesized to play a role in the recruitment or recognition of fatty-acid donor substrate.[15]

All three identified PGAP2 alterations, p.Arg16Trp, Leu127Ser, and Thr160Ile, affect evolutionarily highly conserved amino acid residues (Figure 3) and are predicted to be deleterious by MutationTaster.[16] We therefore hypothesized that they might impair the function of PGAP2 (Figure 3). We cloned human *PGAP2* (RefSeq NM_00125640.1) from a cDNA library derived from Hep3B (a hepatoma cell line) cells, tagged with FLAG at the N terminus, and subcloned it into pME.[17] Altered forms of PGAP2 were generated by site-directed mutagenesis. Altered and wild-type PGAP2 plasmids were transfected by electroporation into human-CD59-expressing PGAP2-deficient Chinese hamster ovary (CHO) cells that were derived from aerolysin-resistant clones from chemically mutagenized CHO cells as previously described.[18] The protein levels of CD55 and CD59, both GPI-anchored proteins, at the cell surface were determined by cell staining with anti-FLAG and anti-hamster antibodies and analyzed by flow cytometry (BD FACSCanto II, BD Biosciences) with Flowjo software (Tommy Digital). In PGAP2-deficient cells, fatty-acid remodeling is terminated at the lysoform intermediate GPI as a result of a lack of PGAP2-dependent reacylation. The lysoform GPI-anchored proteins are transported to the cell surface, where they are cleaved by a phospholipase D, resulting in the release of GPI-anchored proteins lacking lipid moiety and a decrease in the cell-surface level of GPI-anchored proteins.[15] After transfection, wild-type PGAP2 restored the levels of CD55 and CD59 at the cell surface more efficiently than did the p.Arg16Trp, p.Leu127Ser, and p.Thr160Ile altered forms (Figure 3). Of all three tested alterations, p.Arg16Trp reduced the levels of CD55 and CD59 to a lesser degree than did p.Leu127Ser or p.Thr160Ile. Although it is uncertain whether this result is relevant for the in vivo situation, it might suggest a less severe impairment of PGAP2 function and might correlate with the milder phenotype in individual II-1 of family A.

Elevated secretion of ALP, which is normally GPI anchored to the cell surface, into the serum leads to hyperphosphatasia. The biochemical mechanisms of hyperphosphatasia in PGAP2-deficient individuals described in this study and in PIGV- or PIGO-deficient individuals reported previously are distinct. In PGAP2-deficient cells, GPI-anchored proteins lacking the lipid moiety and having only the glycan moiety of GPI are released because of a defect in PGAP2-mediated reacylation during fatty-acid exchange in the Golgi and the subsequent cleavage by a phospholipase D after transport to the cell.[15,19] In PIGV- or PIGO-deficient cells, the C-terminal GPI-attachment signal peptide of the GPI-anchored protein precursor tentatively acts as a membrane anchor in the endoplasmic reticulum and is cleaved by GPI transamidase but cannot be replaced by a GPI anchor because of a lack of mature GPI synthesis. This abnormality results in the release of soluble proteins completely lacking GPI moiety.[20]

In summary, we have identified a homozygous missense mutation in *PGAP2* in an affected individual with the specific HPMRS phenotype and compound-heterozygous *PGAP2* mutations causing a nonsyndromic intellectual-disability phenotype in a second individual. These findings suggest that the clinical range associated with *PGAP2* mutations includes severe manifestations of HPMRS and nonsyndromic and mild intellectual disability. Recent data from exome-sequencing studies have shown that mutations of known genes associated with a specific syndrome diagnosis might also be identified in nonsyndromic intellectual disability. This suggests that present syndrome descriptions are strongly biased toward clinically recognizable phenotypes.[21,22]

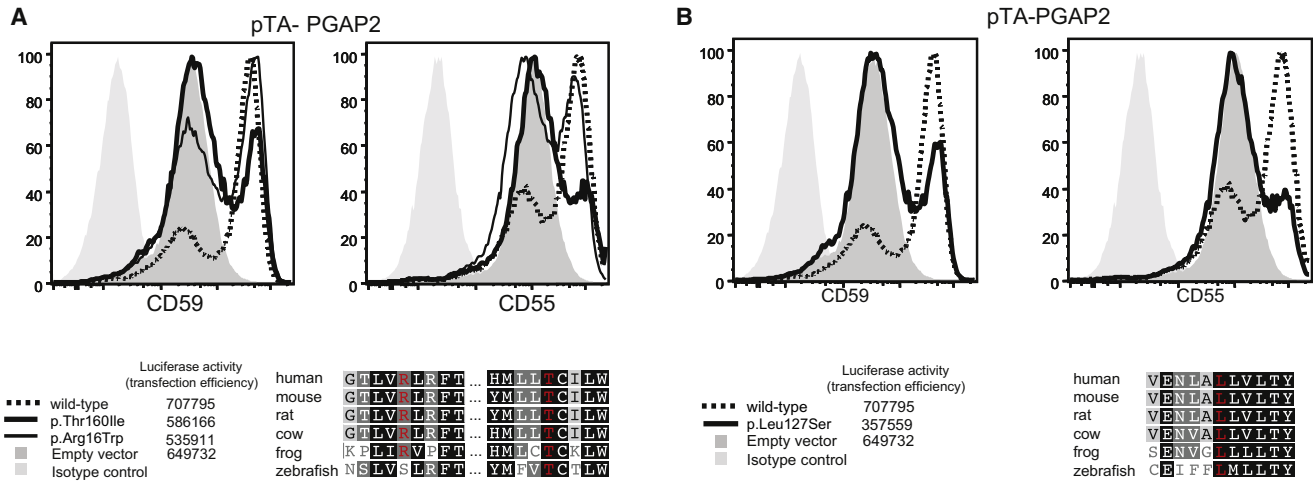


Figure 3. Reduced Activity of Altered Forms of PGAP2 in Restoring Surface Expression of GPI-Anchored Proteins after Transfection into PGAP2-Null Cell Lines

PGAP2-deficient CHO cells were transiently transfected with wild-type or altered forms (p.Arg16Trp, p.Thr160Ile [family A], and p.Leu127Ser [family B]) of pTA Flag-PGAP2 isoform 8 driven by a weak promoter. Restoration of the surface expression was assessed 2 days later by flow cytometry. p.Arg16Trp and p.Thr160Ile detected in family A and p.Leu127Ser detected in family B did not restore the surface expression of CD59 and CD55 as efficiently as the wild-type PGAP2. The reduction of surface protein levels associated with p.Arg16Trp was less severe. This correlates with a lower sequence conservation of this position and a milder phenotype in individual II-1of family A.

Molecular and phenotypic characterization of more individuals with HPMRS will be required for determining whether there are any differences in the phenotypes caused by *PIGV*, *PIGO*, and *PGAP2* mutations. The comprehensive sequence analysis of HPMRS cases, as well as intellectual-disability cases with a suspected GPI-anchor deficiency indicated by, for example, elevated serum ALP activity, will help to elucidate the phenotypic spectrum of mutations affecting this molecular pathway.

Supplemental Data

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

Acknowledgments

This work was supported by a grant from the Bundesministerium für Forschung und Technologie (0313911), by a Deutsche Forschungsgemeinschaft grant to P.M.K. (DFG KR 3985/1-1) and to S.M. (SFB 665), and by grants from the Ministry of Education, Culture, Sports, Science, and Technology and the Ministry of Health, Labour, and Welfare of Japan. We wish to thank all individuals involved in this study for their generous help.

Received: December 13, 2012
Revised: January 28, 2013
Accepted: March 15, 2013
Published: April 4, 2013

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes, http://www.1000genomes.org
Agilent eArray, https://earray.chem.agilent.com/earray/
Human Phenotype Ontology, http://www.human-phenotype-ontology.org
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org
NHLBI Exome Sequencing Project (ESP) Exome Variant Server, http://evs.gs.washington.edu/EVS/
GeneTalk, http://www.gene-talk.de
RefSeq, http://www.ncbi.nlm.nih.gov/RefSeq

References

1. Jaeken, J. (2011). Congenital disorders of glycosylation (CDG): it's (nearly) all in it! J. Inherit. Metab. Dis. *34*, 853–858.
2. Horn, D., Krawitz, P., Mannhardt, A., Korenke, G.C., and Meinecke, P. (2011). Hyperphosphatasia-mental retardation syndrome due to PIGV mutations: expanded clinical spectrum. Am. J. Med. Genet. A. *155A*, 1917–1922.
3. Krawitz, P.M., Murakami, Y., Hecht, J., Krüger, U., Holder, S.E., Mortier, G.R., Delle Chiaie, B., De Baere, E., Thompson, M.D., Roscioli, T., et al. (2012). Mutations in PIGO, a member of the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation. Am. J. Hum. Genet. *91*, 146–151.
4. Krawitz, P.M., Schweiger, M.R., Rödelsperger, C., Marcelis, C., Kölsch, U., Meisel, C., Stephani, F., Kinoshita, T., Murakami, Y., Bauer, S., et al. (2010). Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nat. Genet. *42*, 827–829.
5. Mabry, C.C., Bautista, A., Kirk, R.F., Dubilier, L.D., Braunstein, H., and Koepke, J.A. (1970). Familial hyperphosphatase with mental retardation, seizures, and neurologic deficits. J. Pediatr. *77*, 74–85.
6. Thompson, M.D., Roscioli, T., Marcelis, C., Nezarati, M.M., Stolte-Dijkstra, I., Sharom, F.J., Lu, P., Phillips, J.A., Sweeney, E., Robinson, P.N., et al. (2012). Phenotypic variability in hyperphosphatasia with seizures and neurologic deficit (Mabry syndrome). Am. J. Med. Genet. A. *158A*, 553–558.

7. Thompson, M.D., Nezarati, M.M., Gillessen-Kaesbach, G., Meinecke, P., Mendoza-Londono, R., Mornet, E., Brun-Heath, I., Squarcioni, C.P., Legeai-Mallet, L., Munnich, A., and Cole, D.E. (2010). Hyperphosphatasia with seizures, neurologic deficit, and characteristic facial features: Five new patients with Mabry syndrome. Am. J. Med. Genet. A. *152A*, 1661–1669.
8. Ng, B.G., Hackmann, K., Jones, M.A., Eroshkin, A.M., He, P., Wiliams, R., Bhide, S., Cantagrel, V., Gleeson, J.G., Paller, A.S., et al. (2012). Mutations in the glycosylphosphatidylinositol gene PIGL cause CHIME syndrome. Am. J. Hum. Genet. *90*, 685–688.
9. Almeida, A.M., Murakami, Y., Layton, D.M., Hillmen, P., Sellick, G.S., Maeda, Y., Richards, S., Patterson, S., Kotsianidis, I., Mollica, L., et al. (2006). Hypomorphic promoter mutation in PIGM causes inherited glycosylphosphatidylinositol deficiency. Nat. Med. *12*, 846–851.
10. Freeze, H.H. (2006). Genetic defects in the human glycome. Nat. Rev. Genet. *7*, 537–551.
11. Robinson, P.N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am. J. Hum. Genet. *83*, 610–615.
12. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.
13. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.
14. Kamphans, T., and Krawitz, P.M. (2012). GeneTalk: an expert exchange platform for assessing rare sequence variants in personal genomes. Bioinformatics *28*, 2515–2516.
15. Tashima, Y., Taguchi, R., Murata, C., Ashida, H., Kinoshita, T., and Maeda, Y. (2006). PGAP2 is essential for correct processing and stable expression of GPI-anchored proteins. Mol. Biol. Cell *17*, 1410–1420.
16. Schwarz, J.M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. Nat. Methods *7*, 575–576.
17. Takebe, Y., Seiki, M., Fujisawa, J., Hoy, P., Yokota, K., Arai, K., Yoshida, M., and Arai, N. (1988). SR alpha promoter: an efficient and versatile mammalian cDNA expression system composed of the simian virus 40 early promoter and the R-U5 segment of human T-cell leukemia virus type 1 long terminal repeat. Mol. Cell. Biol. *8*, 466–472.
18. Hong, Y., Ohishi, K., Inoue, N., Kang, J.Y., Shime, H., Horiguchi, Y., van der Goot, F.G., Sugimoto, N., and Kinoshita, T. (2002). Requirement of N-glycan on GPI-anchored proteins for efficient binding of aerolysin but not Clostridium septicum alpha-toxin. EMBO J. *21*, 5047–5056.
19. Maeda, Y., Tashima, Y., Houjou, T., Fujita, M., Yoko-o, T., Jigami, Y., Taguchi, R., and Kinoshita, T. (2007). Fatty acid remodeling of GPI-anchored proteins is required for their raft association. Mol. Biol. Cell *18*, 1497–1506.
20. Murakami, Y., Kanzawa, N., Saito, K., Krawitz, P.M., Mundlos, S., Robinson, P.N., Karadimitris, A., Maeda, Y., and Kinoshita, T. (2012). Mechanism for release of alkaline phosphatase caused by glycosylphosphatidylinositol deficiency in patients with hyperphosphatasia mental retardation syndrome. J. Biol. Chem. *287*, 6318–6325.
21. de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. N. Engl. J. Med. *367*, 1921–1929.
22. Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet *380*, 1674–1682.

2.9 Beeinträchtigung der GPI-Anker-Reifung durch Mutationen in *PGAP3*

Howard, M.F., Murakami, Y., Pagnamenta, A.T., Daumer-Haas, C., Fischer, B., Hecht, J., Keays, D.A., Knight, S.J., Kölsch, U., Krüger, U., Leiz, S., Maeda, Y., Mitchell, D., Mundlos, S., Phillips, JA 3rd, Robinson, P.N., Kini, U., Taylor, J.C., Horn, D., Kinoshita, T., Krawitz, P.M. (2014). Mutations in PGAP3 Impair GPI-Anchor Maturation, Causing a Subtype of Hyperphosphatasia with Mental Retardation. American Journal of Human Genetics 94, 278-87.

Gemeinsam mit einer britischen Arbeitsgruppe konnten wir in drei Patienten mit klassischem Mabry Syndrom pathogene Mutationen in *PGAP3*, einem weiteren Gen der GPI-Anker-Reifung identifizieren. Die ethnischen Hintergründe der Individuen, europäisch-amerikanisch, pakistanisch-britisch und saudi-arabisch, veranschaulichten ebenfalls exemplarisch, dass es sich bei GPI-Ankerstörungen um eine Erkrankung handelt, die weltweit anzutreffen ist. Die bisherigen populationsgenetischen Daten, die eine Abschätzung des Heterozygoten-Risikos erlauben, sowie weitere Fallbeschreibungen aus Japan, legen zudem nahe, dass die Prävalenzen weltweit vergleichbar sein dürften.

Wie für die pathogenen Sequenzvarianten in dem vormals beschriebenen Gen der GPI-Anker-Reifung, *PGAP2*, konnte auch für die in *PGAP3* gefundenen Mutationen, c.275G>A (p.Gly92Asp), c.439dupC (p.Leu147Profs*16), c.914A>G (p.Asp305Gly), c.314C>G (p.Pro105Arg) *in vitro* gezeigt werden, dass sie die Oberflächenexpression von GPI-APs vermindern.

Funktionell wirken PGAP3 und PGAP2 bei der Modifikation der Fettsäurereste des GPI-Ankers zusammen: Nachdem ein Protein an den GPI-Anker angefügt wurde, spaltet zuerst PGAP3 eine ungesättigte Fettsäure am Phosphatidylinositol ab und PGAP2 fügt an gleicher Stelle eine gesättigte Fettsäure an. Dieser Reifungsschritt, der im Golgi Apparat erfolgt, ist notwendig, um die GPI-APs auf der Zelloberfläche mit Lipidflößen zu assoziieren (Maeda, et al., 2007). Unterbleiben diese Schritte, gelangt eine als lyso-GPI-AP bezeichnete Vorstufe auf die Plasmamembran, bei der die Abspaltung des gebundenen Proteins erleichtert ist.

Aktuell sind damit von uns für das Mabry-Syndrom krankheitsverursachende Mutationen in zwei unterschiedlichen Genen der späten GPI-Anker-Synthese (*PIGV*, *PIGO*) und der GPI-Anker-Reifung (*PGAP2* und *PGAP3*) beschrieben worden. Soweit bislang aus den durchflusszytometrischen Untersuchungen ersichtlich, scheinen sich die dadurch bedingten GPI-Anker-Störungen auf alle GPI-APs in ähnlichem Maße auszuwirken.

Eine interessante Fragestellung für die zukünftige wissenschaftliche Arbeit ist es, ob sich, ähnlich wie bei der paroxysmalen nächtlichen Hämoglobinurie, bestimmte klinische Merkmale des Mabry Syndrom durch die Störung spezifischer GPI-APs erklären lassen.

Please cite this article in press as: Howard et al., Mutations in *PGAP3* Impair GPI-Anchor Maturation, Causing a Subtype of Hyperphosphatasia with Mental Retardation, The American Journal of Human Genetics (2014), http://dx.doi.org/10.1016/j.ajhg.2013.12.012

REPORT

Mutations in *PGAP3* Impair GPI-Anchor Maturation, Causing a Subtype of Hyperphosphatasia with Mental Retardation

Malcolm F. Howard,[1,13] Yoshiko Murakami,[2,3,13] Alistair T. Pagnamenta,[1,13] Cornelia Daumer-Haas,[4] Björn Fischer,[5,7] Jochen Hecht,[6,7] David A. Keays,[8] Samantha J.L. Knight,[1] Uwe Kölsch,[9] Ulrike Krüger,[5] Steffen Leiz,[10] Yusuke Maeda,[2,3] Daphne Mitchell,[11] Stefan Mundlos,[5,6,7] John A. Phillips, III,[11] Peter N. Robinson,[5,6,7] Usha Kini,[12,14,*] Jenny C. Taylor,[1,14] Denise Horn,[5,14] Taroh Kinoshita,[2,3,14] and Peter M. Krawitz[5,6,7,14,*]

Glycosylphophatidylinositol (GPI)-anchored proteins play important roles in many biological processes, and mutations affecting proteins involved in the synthesis of the GPI anchor are reported to cause a wide spectrum of intellectual disabilities (IDs) with characteristic additional phenotypic features. Here, we describe a total of five individuals (from three unrelated families) in whom we identified mutations in *PGAP3*, encoding a protein that is involved in GPI-anchor maturation. Three siblings in a consanguineous Pakistani family presented with profound developmental delay, severe ID, no speech, psychomotor delay, and postnatal microcephaly. A combination of autozygosity mapping and exome sequencing identified a 13.8 Mb region harboring a homozygous c.275G>A (p.Gly92Asp) variant in *PGAP3* region 17q11.2–q21.32. Subsequent testing showed elevated serum alkaline phosphatase (ALP), a GPI-anchored enzyme, in all three affected children. In two unrelated individuals in a cohort with developmental delay, ID, and elevated ALP, we identified compound-heterozygous variants c.439dupC (p.Leu147Profs*16) and c.914A>G (p.Asp305Gly) and homozygous variant c.314C>G (p.Pro105Arg). The 1 bp duplication causes a frameshift and nonsense-mediated decay. Further evidence supporting pathogenicity of the missense mutations c.275G>A, c.314C>G, and c.914A>G was provided by the absence of the variants from ethnically matched controls, phylogenetic conservation, and functional studies on Chinese hamster ovary cell lines. Taken together with recent data on *PGAP2*, these results confirm the importance of the later GPI-anchor remodelling steps for normal neuronal development. Impairment of PGAP3 causes a subtype of hyperphosphatasia with ID, a congenital disorder of glycosylation that is also referred to as Mabry syndrome.

Glycosylphophatidylinositol (GPI) anchoring is a post-translational modification that tethers proteins to plasma membranes, and it is thought to play a role in protein sorting and trafficking.[1] The GPI anchor is well conserved among eukaryotes, and there are over 150 mammalian GPI-anchored proteins (GPI-APs), including receptors, adhesion molecules, and enzymes.[1] Many of these proteins are critical for normal neural and embryonic development.[2,3] There are around 30 genes known to be involved in the biosynthesis and remodelling of the GPI anchor, which is formed in the endoplasmic reticulum (ER), where it is attached by the GPI transamidase to a protein showing a specific C-terminal signal before it is transported to the Golgi apparatus for fatty acid remodelling and cellular export.

In the last 4 years, germline mutations in eight genes selectively involved in the GPI-anchor-synthesis pathway have been shown to cause a wide phenotypic spectrum of disorders with intellectual disability (ID) and seizures; these range from syndromic forms with characteristic malformations and minor anomalies to nonsyndromic forms. Congenital disorders that are caused by an impairment of GPI-anchor synthesis and maturation are now classified as congenital disorders of glycosylation (CDGs), a diverse class of metabolic diseases.[4]

PIGV (MIM 610274), *PIGO* (MIM 614730), and *PGAP2* (MIM 615187), three different genes in the GPI-anchor-synthesis pathway, have been implicated in hyperphosphatasia with mental retardation syndrome (HPMRS [MIM 239300]), also known as Mabry syndrome, a recently delineated autosomal-recessive form of ID with a distinct facial dysmorphism, consistently elevated serum alkaline phosphatase (ALP) (hyperphosphatasia), brachytelephalangy, and seizures. All individuals with congenital impairment of GPI-anchor synthesis have residual surface levels of GPI-anchored proteins, and it is hypothesized that complete-loss-of-function mutations in genes involved with GPI-anchor synthesis are embryonically lethal.[5]

[1]National Institute for Health Research Biomedical Research Centre, Wellcome Trust Centre for Human Genetics, University of Oxford, OX3 7BN Oxford, UK; [2]Department of Immunoregulation, Research Institute for Microbial Diseases, Osaka University, Osaka 565-0871, Japan; [3]World Premier International Immunology Frontier Research Center, Osaka University, Osaka 565-0871, Japan; [4]Pränatal-Medizin München, 80637 München, Germany; [5]Institute for Medical Genetics and Human Genetics, Charité Universitätsmedizin, 13353 Berlin, Germany; [6]Berlin Brandenburg Center for Regenerative Therapies, Charité Universitätsmedizin, 13353 Berlin, Germany; [7]Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany; [8]Research Institute of Molecular Pathology, 1030 Vienna, Austria; [9]Institute of Medical Immunology, Charité Universitätsmedizin, 13353 Berlin, Germany; [10]Department of Pediatrics, Klinikum Dritter Orden, 80638 München, Germany; [11]Division of Medical Genetics and Genomic Medicine, Department of Pediatrics, Vanderbilt University School of Medicine, TN 37232-2578, USA; [12]Department of Clinical Genetics, Oxford University Hospitals NHS Trust, OX3 9DU Oxford, UK
[13]These authors contributed equally to this work
[14]These authors contributed equally to this work and are co-senior authors
*Correspondence: usha.kini@ouh.nhs.uk (U.K.), peter.krawitz@charite.de (P.M.K.)
http://dx.doi.org/10.1016/j.ajhg.2013.12.012.

The effect of mutations causing GPI-anchor deficiencies can be subdivided into groups on the basis of the serum activity of the marker enzyme ALP. Mutations in *PIGA* (MIM 311770), *PIGL* (MIM 605947), *PIGM* (MIM 610273), *PIGN* (MIM 606097), and *PIGT* (MIM 610272) affect early GPI-anchor synthesis and the attachment of proteins to the GPI anchor.[6–10] These variants result in primary reduced surface levels of GPI-APs because of their increased intracellular degradation and are not associated with high serum ALP. Mutations in *PIGV* and *PIGO* affect late GPI-anchor synthesis.[11,12] However, GPI transamidase recognizes the incomplete GPI anchor and cleaves the GPI attachment signal, leading to the generation of non-GPI-anchored soluble proteins.[13] Individuals with mutations in *PIGV* and *PIGO* have primary reduced GPI-AP surface levels as a result of increased secretion into the extracellular space, resulting in high serum ALP. Recently, mutations in *PGAP2* have been reported to affect the final GPI-anchor fatty-acid-remodelling step and to thus result in abnormal GPI-APs that are more prone to cleavage.[14–16] Individuals with *PGAP2* mutations have secondary reduced GPI-AP surface levels also as a result of increased secretion into the extracellular space, resulting in high serum ALP. GPI remodelling occurs in the Golgi and involves the removal of an unsaturated fatty acid at the sn-2 position by PGAP3 and its subsequent substitution with a saturated fatty acid by PGAP2. Fatty acid remodelling is critical for proper association between GPI-APs and lipid rafts.[17]

A number of deficient GPI-pathway genes were first identified via traditional autozygosity mapping in combination with next-generation sequencing methodologies.[10,14] This methodology has become a key strategy for the identification of disease-associated genes, particularly those related to genetically heterogeneous conditions such as ID, and was the approach taken for the first family described here.[18]

Family A originates from the Sargodha district of Punjab in Pakistan. The parents are second cousins, and all three children (Figures 1A and 2A) presented with profound developmental delay, severe learning disability, no speech, psychomotor delay, and postnatal microcephaly (−2 to −3 SDs) at the ages of 17, 8, and 4 years. Further clinical details are summarized in Table 1. After a *ASPM* test that did not detect pathogenic variants, they were recruited into the ongoing Structural Brain Abnormalities and Learning Disabilities Study (see Oxford Brain Abnormality Research Group in Web Resources), which received UK ethics approval from the Wales Research Ethics Committee (12/WA/0001) and obtained informed consent from the responsible persons on behalf of all study participants. Blood DNA samples for IV-2, IV-3, V-1, V-2, and V-3 were run on CytoSNP12v2 arrays (Illumina), which confirmed familial relationships and that there were no copy-number changes of likely clinical relevance (Table S1, available online). Autozygosity mapping was then used for identifying a single 13.8 Mb candidate region in 17q11.2–q21.32 (Figure S1). None of the 359 annotated genes appeared to be obvious candidates, and so an exome sequencing approach was adopted.

The exome of proband V-2 was enriched (TruSeq, Illumina) and sequenced on a HiSeq2500 (Illumina) with the use of standard settings and the 100 bp paired-end read format. Reads were mapped to the human reference sequence (UCSC Genome Browser, hg19) with the use of Stampy,[19] and variants were called with Platypus. The resulting VCF file was uploaded into Ingenuity Variant Analysis (v.2.1.20130711) and filtered for rare variants that were in the homozygous region on chromosome 17 and predicted to be deleterious (Table S2 and Web Resources).

The single remaining variant was a c.275G>A change in *PGAP3* (RefSeq accession number NM_033419.3). Sanger sequencing confirmed that this variant cosegregated with disease, consistent with an autosomal-recessive mode of inheritance (Figure 1B). The variant is predicted to cause a p.Gly92Asp alteration to the protein sequence at a highly conserved position (Figure 1C) and is not present at any frequency in the NHLBI Exome Sequencing Project Exome Variant Server (EVS), 1000 Genomes (version 3), or 274 in-house genomes of mixed ancestry. The mutation was also not detected in 108 Punjabi individuals from Lahore, suggesting that c.275G>A is unlikely to be a variant specific to this ethnicity.

In parallel, we performed targeted sequencing in 19 individuals with ID and elevated ALP and in whom HPMRS was suspected. The Charité University Medicine ethics board approved this study, and we obtained informed consent from the responsible persons (parents) on behalf of all study participants. After excluding pathogenic *PIGV* mutations that are the most common cause of HPMRS by Sanger sequencing, we subjected DNA to enrichment of all 30 known genes associated with GPI-anchor synthesis by using a customized SureSelect library (Agilent), as well as subsequent sequencing on a HiSeq2000, as previously described.[15] Sequence variants were filtered under an autosomal-recessive model of inheritance in GeneTalk,[20] which yielded potentially pathogenic variants in *PGAP3* in two individuals. Compound-heterozygous variants c.439dupC (p.Leu147Profs*16) and c.914A>G (p.Asp305Gly) were detected in individual II-1 from family B, and homozygous variant c.314C>G (p.Pro105Arg) was identified in individual II-1 from family C (Figure 1B).

The detailed clinical findings of individual II-1 from family B were reported previously (affected individual 4 in Thompson et al.)[21] together with a single rare heterozygous variant in *PIGV*, c.1369C>T.[21] However, in *PIGV*-defective Chinese hamster ovary (CHO) cells, the altered protein restored the surface levels of GPI-APs as efficiently as the wild-type protein, suggesting no functional impairment of *PIGV* (Figure S4). In contrast, the duplication of c.439dupC in *PGAP3* causes a frameshift that introduces a premature stop codon, most likely resulting in nonsense-mediated decay of the transcript. To exclude further candidate mutations, we exome sequenced DNAs of the affected

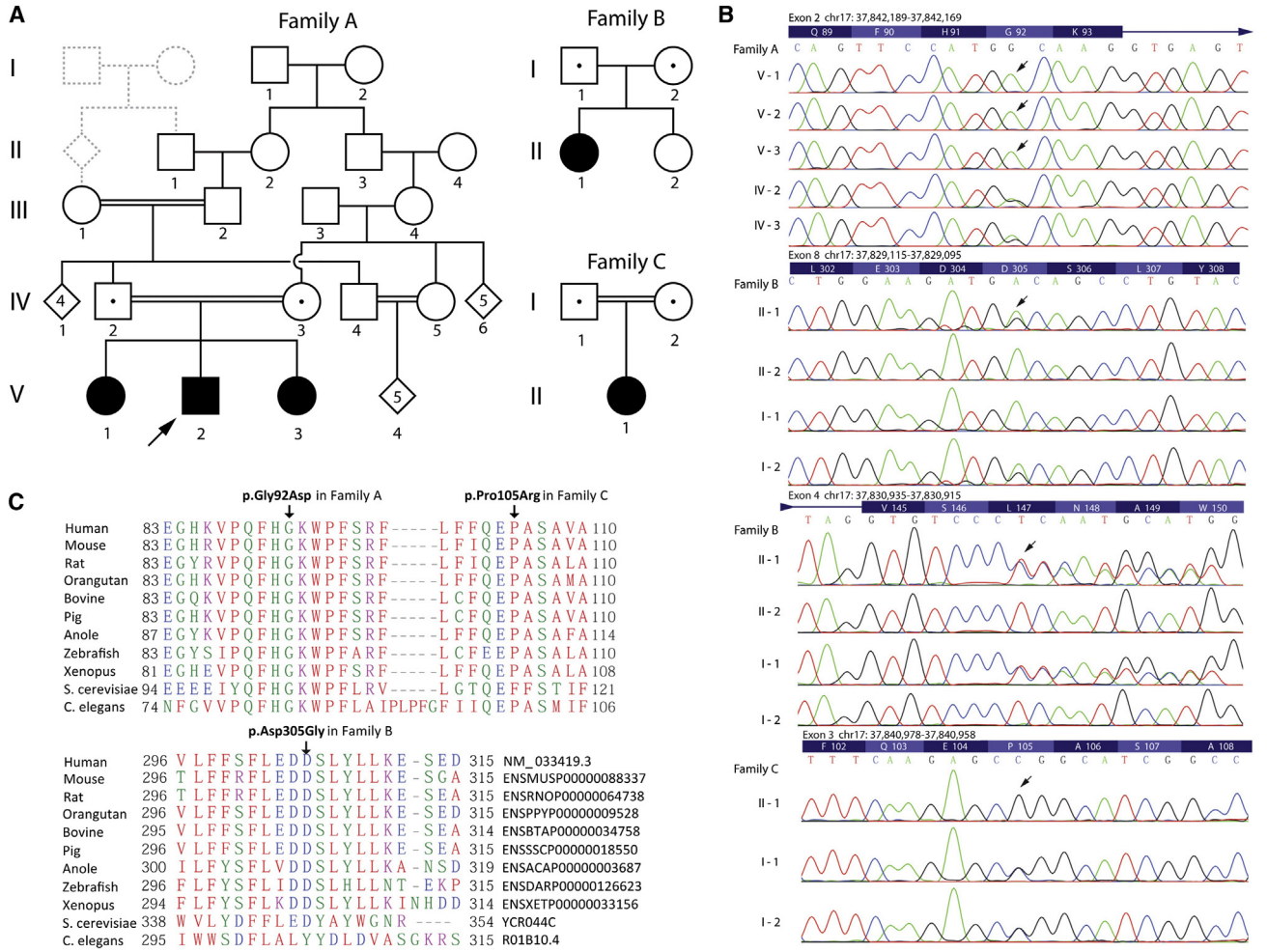


Figure 1. Pedigrees, Sanger Validation, and Phylogenetic Conservation of the Variant Filtering Cascade and Molecular Characterization of *PGAP3* Mutations

(A) Pedigrees are shown for family A (of Pakistani descent), family B (of European descent), and family C (of Saudi Arabian origin). Shading indicates ID and hyperphosphatasia, and proven heterozygote carriers are shown by a dot. In family A, the gray dotted lines indicate an additional consanguineous loop that was not described in the clinical notes but that was inferred by the high inbreeding coefficient detected in IV-2 from the SNP array data. The proband is indicated with an arrow.

(B) Sanger validation and segregation testing of *PGAP3* variants. DNA for the unaffected double first cousins in family A (collectively labeled as V-4) was not available.

(C) ClustalW alignment of amino acid sequence shows high evolutionary conservation.

individual of family B and her parents and confirmed the compound-heterozygous variants in *PGAP3* as the only candidate mutations for an autosomal-recessive model of inheritance.[22,23]

Individual II-1 of family B is the first child of nonconsanguineous European-American parents (Table 1 and Figures 1A and 2B). Her psychomotor development was severely delayed, and she developed tonic-clonic seizures. Physical examination at the age of 10 years showed postnatal short stature, dystrophy, normal occipitofrontal circumference (OFC), and dysmorphic facial features. ALP levels were repeatedly elevated, and a GPI-anchor deficiency on granulocytes and monocytes could be measured by flow cytometry with CD16, CD24, and fluorescence-labeled aerolysin staining (Figure 3A). In family C, individual II-1 is the first child of consanguineous parents of Saudi Arabian origin (Table 1 and Figures 1A and 2C). She had severe psychomotor delay. Myoclonic seizures started in her second year of life. Physical examination of this 2-year-old female showed normal growth parameters and OFC but axial muscular hypotonia, uncoordinated movements, and facial dysmorphism. ALP activity was elevated in repeated tests. Sanger sequencing confirmed that the parents are heterozygous for c.314C>G, and the mutation was not detected in 52 Arabic controls or the NHLBI Exome Sequencing Project EVS, suggesting that it might be a private mutation specific to this family (Figure 1B).

The most common features noted in all affected individuals were normal growth parameters and OFC at birth, severe psychomotor delay with no speech and marked motor delay, and characteristic facial features. Four were unable to walk, whereas the other began to walk with support at the age of 3. Seizures starting between the ages of 18 months and 12 years developed in four of five affected

Figure 2. Photographs of Affected Individuals
Facial features seen in (A) individual V-3 from family A, (B) individual II-1 from family B, and (C) individual II-1 from family C at the ages of 4, 10, and 2 years, respectively. These individuals bear a striking resemblance with a broad nasal bridge, long-appearing palpebral fissures, a broad nasal tip, a short nose, a long philtrum, a thin and wide upper lip, full cheeks, and large fleshy ear lobes.

individuals. Three were described as having tonic-clonic seizures, and the other had myoclonic seizures. Consistent ALP elevation varied between 1.4 and 5 times the age-adjusted upper limit of the normal range. The facial gestalt was similar to that of individuals with HPMRS and included features such as a broad nasal bridge, a short nose with a broad nasal tip, a thin and wide upper lip, and large fleshy ear lobes. In contrast, postnatal development of head circumference and growth parameters differed among the affected individuals studied here. Affected individuals in family A developed microcephaly, whereas OFC was in the mean range in the other individuals. In the affected individual from family B, marked short stature was present, whereas the other affected individuals had a mean height according to age. Two of five individuals had cleft palate, whereas abnormal MRI findings such as thin corpus callosum and dilated lateral ventricles were observed in the affected individual from family C.

The common features of all individuals reported with mutations in *PIGV*, *PIGO*, *PGAP2*, and *PGAP3* are severe psychomotor delay and ID, epilepsy, elevated ALP, and a distinctive facial gestalt. In contrast, brachytelephalangy, which is an important diagnostic sign in other types of HPMRS, is not present in any affected individuals with *PGAP3* mutations. Marked variability regarding the postnatal growth and OFC has also been observed in the groups of affected individuals carrying *PIGV*, *PIGO*, and *PGAP2* mutations. Also, cleft palate has been previously documented in other individuals with HPMRS due to mutations in these genes. In individuals affected by HPMRS and mutations in genes other than *PGAP3*, associated malformations seem to appear more frequently and their spectrum seems to be broader.

To determine the functional consequences of the four *PGAP3* mutations, we used a mutant CHO cell line defective in both *PGAP3* and *PGAP2* (see Figure S3 for principles of the assay).[17] The mutant cells have GPI-APs at mildly reduced levels because of a lack of GPI fatty acid remodelling. When wild-type *PGAP3* cDNA was transfected, the first step in the fatty acid remodelling was restored, whereas the second step remained defective, leading to the release of lyso-GPI intermediates and resulting in a severe reduction in the surface levels of GPI-APs (Figure 3Bi). If the mutant *PGAP3* cDNA has decreased activity, reduction in the surface GPI-AP levels would be partial. Mutant *PGAP3* cDNA bearing the mutation found in family A (c.275G>A [p.Gly92Asp]) either did not reduce or reduced only slightly the surface levels of three GPI-APs: CD59, CD55 (DAF), and urokinase plasminogen activator receptor (uPAR), indicating that the substitution caused a null or nearly null phenotype (Figure 3Bii). One of the substitutions found in family B (p.Leu147Profs*16) also caused a null phenotype, whereas the other (p.Asp305Gly) significantly reduced levels of all three GPI-APs, indicating some residual activity (Figure 3Biii). The substitution found in family C (p.Pro105Arg) caused less efficient reduction of GPI-APs than did the p.Asp305Gly substitution in family B, indicating an even lower residual activity (Figure 3Biv).

To clarify the mechanisms of these functional losses, we expressed mutant *PGAP3* clones that were hemagglutinin (HA) tagged at the N terminus in CHO cells and analyzed them by SDS-PAGE and immunoblotting (Figure 4) and immunofluorescence microscopy (Figure 5) with anti-HA. On SDS-PAGE and immunoblotting, HA-PGAP3 appeared as a smear band at around 37–45 kDa and several clear bands at around 23–35 kDa (Figure 4A, lane 3). After treatment with PNGase F to remove N-glycan, the smear band disappeared and a band at 33 kDa corresponding to de-N-glycosylated full-size protein became a major one (lane 1). The altered (p.Gly92Asp) protein in family A showed similar profiles (lanes 7 and 9), suggesting normal levels of PGAP3 bearing normally matured N-glycan. Consistently, on immunofluorescence microscopy,

Table 1. Summary of Clinical Findings in Individuals with *PGAP3* Mutations

	Individual				
Clinical Findings[a]	**V-1 (Family A)**	**V-2 (Family A)**	**V-3 (Family A)**	**II-1 (Family B)**	**II-1 (Family C)**
Ethnicity	Pakistani (Sargodha district of Punjab)	Pakistani (Sargodha district of Punjab)	Pakistani (Sargodha district of Punjab)	American (European descent)	Saudi-Arabian
Consanguinity	yes	yes	yes	no	yes
Age of last assessment (years)	17	8	4	10	2
OFC at birth	normal	normal	normal	normal	normal
OFC	−2 to −3 SDs	−2 to −3 SDs	−3 SDs	mean	mean
Height	normal	normal	normal	−4 SDs	normal
Weight	normal	normal	normal	−2 SDs	normal
Global developmental delay (HP:0001263)	yes	yes	yes	yes	yes
Motor delay (HP:0001270)	severe (unable to walk)	severe (unable to walk)	severe (unable to walk)	severe (walk with support at age of 3 years)	severe (unable to walk)
Speech and language development	none	none	none	none	none
Muscular hypotonia (HP:0001252)	yes	yes	yes	yes	yes
Seizures (HP:0001250)	yes	yes	no	yes	yes
Age of onset of seizures (years)	12	4	NA	4	1.5
Type of seizures	generalized tonic-clonic	generalized tonic-clonic	NA	tonic-clonic and cluster	myoclonic
Antiepileptic drugs	valproate	valproate	NA	rufinamide, pregabaline	valproate, levetiracetam
Behavioral abnormalities	involuntary midline hand movements, bruxism	involuntary midline hand movements, bruxism	involuntary midline hand movements, bruxism	no	hyperactivity
Apparent hypertelorism (HP:0000316)	yes	yes	yes	yes	yes
Upslanting palpebral fissures (HP:0000582)	yes	yes	yes	no	no
Broad nasal bridge (HP:0000431)	yes	yes	yes	yes	yes
Broad nasal tip	yes	yes	yes	yes	yes
Short nose (HP:0003196)	yes	yes	yes	yes	yes
Tented upper-lip vermilion (HP:0010804)	yes	yes	yes	yes	yes
Large, fleshy ear lobes (HP:0009748)	yes	yes	yes	no	yes
Cleft palate (HP:0000175)	yes	yes	no (high palate)	no	no
Brachytelephalangy (HP:0009882)	no	no	no	no	no
Serum total ALP (U/l) (HP:0003155)	739	1,407	1,452	926–2,000	928–1,370
Upper limit in ALP test (U/l)	525–600	525–600	525–600	400	386

(*Continued on next page*)

Table 1. ***Continued***

	Individual				
Clinical Findings[a]	**V-1 (Family A)**	**V-2 (Family A)**	**V-3 (Family A)**	**II-1 (Family B)**	**II-1 (Family C)**
Further anomalies	no	no	no	no	thin corpus callosum, dilated lateral ventricles
PGAP3 variants (RefSeq NM_033419.3)	homozygous c.275G>A (p.Gly92Asp)	homozygous c.275G>A (p.Gly92Asp)	homozygous c.275G>A p.Gly92Asp)	compound-heterozygous c.914A>G (p.Asp305Gly) and c.439dupC (p.Leu147Profs*16)	homozygous c.314C>G (p.Pro105Arg)

Abbreviations are as follows: ALP, alkaline phosphatase; OFC, occipitofrontal circumference; and NA, not applicable.
[a]Human phenotype ontology IDs are provided if applicable.

wild-type and p.Gly92Asp PGAP3 were found mainly in the Golgi and to a lesser extent in the ER (Figure 5, first and second rows). Only faint bands were seen with p.Leu147Profs*16 from family B, indicating that the level of altered protein was not significant (Figure 4C). The other family B substitution, p.Asp305Gly, showed a main band at 35 kDa (Figure 4A, lane 6) that was shifted to the 33 kDa de-N-glycosylated full-size protein after treatment with either PNGase F or Endo H, the latter of which eliminated only immature N-glycan (lanes 4 and 5). This profile indicated that the p.Asp305Gly altered protein had only immature ER-form N-glycan. Consistently, the altered

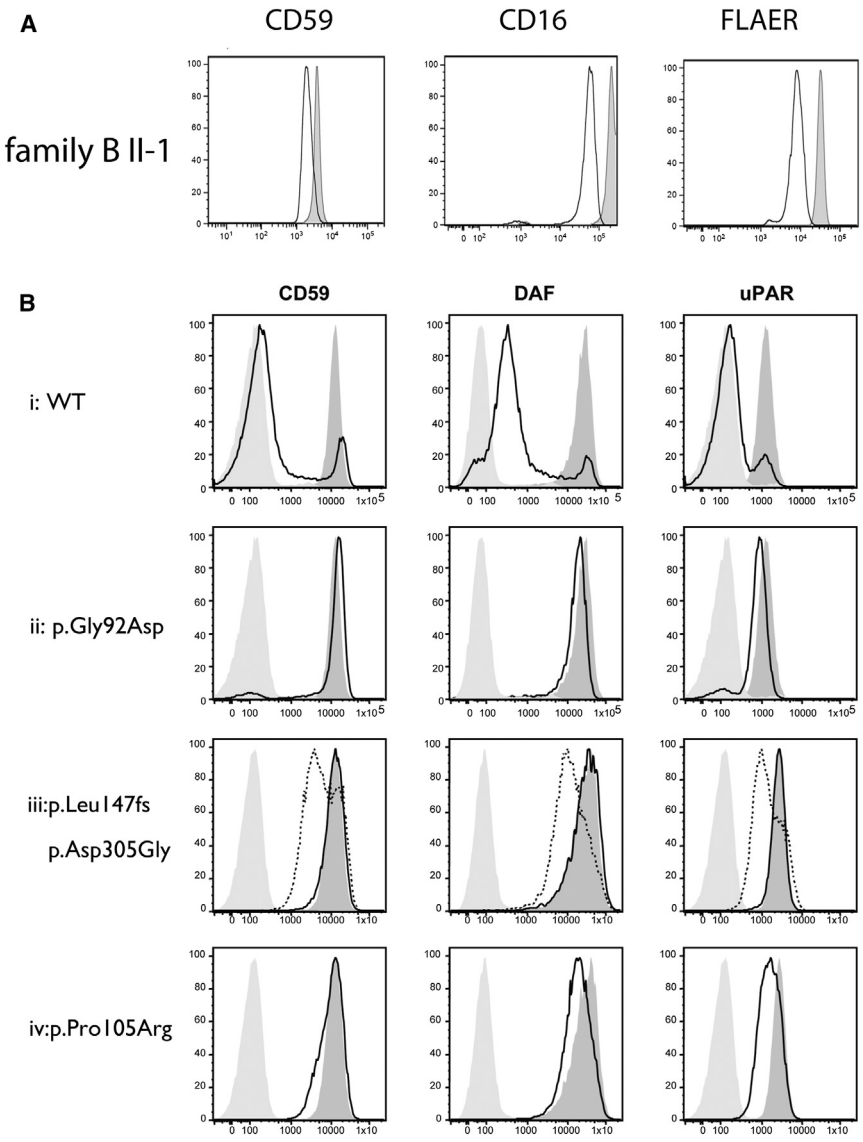


Figure 3. Flow Cytometric Analysis of Granulocyte Surface GPI-APs and Flow Cytometric Assay for Functions of Altered PGAP3

(A) Blood granulocytes from the affected individual in family B (solid line) were stained with anti-CD24 and anti-CD16 and fluorescence-labeled aerolysin (FLAER). The light-gray area represents a healthy control.

(B) *PGAP3* and *PGAP2* double-mutant CHO cells were transfected with *PGAP3* cDNA and 2 days later were stained for CD59, CD55 (DAF), and uPAR. (Bi) After transfection with wild-type *PGAP3* cDNA, the surface levels of three GPI-APs were severely reduced (solid lines). The dark area shows the original surface levels of GPI-APs on *PGAP3* and *PGAP2* double-mutant CHO cells, whereas the light gray area represents the isotype-matched control. (Bii) The *PGAP3* cDNA bearing mutation c.275G>A (p.Gly92Asp) in family A either did not reduce or only slightly reduced GPI-AP levels (solid lines). (Biii) The *PGAP3* cDNA bearing one mutation, c.439dupC (p.Leu147Profs*16), in family B did not reduce GPI-AP levels (solid lines), whereas that bearing the other mutation, c.914A>G (p.Asp305Gly), significantly reduced GPI-AP levels, indicating a hypomorphic mutant phenotype (dotted lines). (Biv) The *PGAP3* cDNA bearing mutation c.314C>G (p.Pro105Arg) in family C slightly reduced the levels of three GPI-APs, indicating a hypomorphic but very severe loss-of-function phenotype (solid lines).

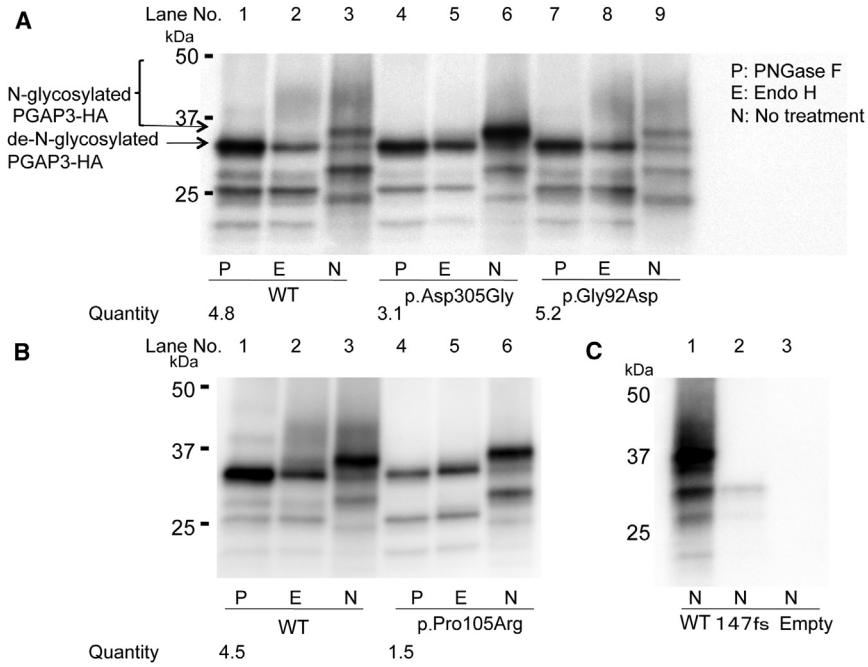


Figure 4. SDS-PAGE and Immunoblotting of HA-Tagged PGAP3

(A) Lanes 1–3, wild-type; lanes 4–6, p.Asp305Gly; lanes 7–9, p.Gly92Asp. Protein levels shown as quantity were normalized with the intensities of GAPDH for the loading control and luciferase activities for the transfection efficiencies. The antibody showed no reactivity in lysates from empty-vector-transfected cells.

(B) Lanes 1–3, wild-type; lanes 4–6, p.Pro105Arg.

(C) Lane 1, wild-type; lane 2, p.Leu147Profs*16; lane 3, empty vector. There are two N-glycosylation sites in PGAP3. The mature PGAP3 appears as a smear at 37–45 KDa (A, lane 3). This smear is sensitive to PNGase F, suggesting heterogeneous N-glycans. Bands seen below the 35 KDa position represent degradation products because they are not seen in the empty-vector transfectant (C, lane 3) or the transfectant with the truncated p.Leu147Profs*16 protein (C, lane 2). Abbreviations are as follows: N, no treatment; P, PNGase F treatment; and E, Endo H treatment.

protein was found in the ER, but not the Golgi (Figure 5, third row). Similarly, the p.Pro105Arg substitution in family C generated only an immature protein with ER-form N-glycan, as shown by a lack of a smear band at around 37–45 kDa (Figure 4B, lane 6) and a shift of the clear major band at 35 kDa to 33 kDa after Endo-H treatment (lanes 5 and 6). Consistent with these results, p.Pro105Arg PGAP3 was localized to the ER, but not to the Golgi (Figure 5, fourth row).

Therefore, the four substitutions that we describe had different effects on PGAP3, the Golgi-resident GPI-specific phospholipase A_2 consisting of an N-terminal luminal domain, seven transmembrane domains with conserved catalytic amino acids, and a C-terminal cytoplasmic tail (Figure S4). (1) The p.Gly92Asp substitution did not affect protein levels or Golgi localization, yet the altered protein had only negligible activity. Because Gly92 resides in a juxta membrane position on the luminal side, a negative charge caused by this substitution might interfere with the association between PGAP3 and GPI anchor substrate. (2) The p.Leu147Profs*16 protein was not detected significantly, consistent with nonsense-mediated mRNA degradation. (3) The p.Asp305Gly and p.Pro105Arg proteins were readily detectable but had immature N-glycan and were mislocalized in the ER. Asp305 resides in the cytoplasmic tail, whereas Pro105 resides in the first transmembrane domain. The cytoplasmic tail and the first transmembrane domain might be important for Golgi localization.

The characteristic biochemical phenotype of individuals with *PGAP3* deficiency is hyperphosphatasia, a sign of GPI-AP release from the cell surface. However, the reduction of GPI-AP levels is expected to depend on the cell type and species, as studies in *Pgap3*-knockout mice suggest.[24,25] In fact, flow cytometric analysis from the affected individual in family B demonstrated significant reduction in the cell-surface levels of GPI-APs in blood granulocytes (Figure 3A), whereas the CD59 surface levels in erythrocytes of an affected individual in family A were found to be normal. This confirms that GPI remodelling by PGAP3 is not essential for CD59 surface levels in erythrocytes, given that the GPI-anchor structure in erythrocytes maintains an unsaturated fatty acid at the sn2 position.[26]

The exact mechanism of release has yet to be characterized, but GPI-APs bearing unremodelled fatty acids are not associated well with lipid rafts.[17] This abnormal membrane distribution might affect the stability of GPI-APs, resulting in release from the cell surface. Although both PGAP3 and PGAP2 are involved in fatty acid remodeling, the above mechanism of hyperphosphatasia or GPI-AP release from *PGAP3*-deficient cells is different from that observed in *PGAP2*-deficient cells. In the latter cells, only the first reaction by PGAP3, elimination of unsaturated fatty acid, occurs and the resulting lyso-GPI intermediate is cleaved and released by a putative lyso-GPI-specific phospholipase D.[15]

In summary, we identified four different *PGAP3* mutations in three unrelated families by using two independent strategies. In family A, recruited on account of postnatal microcephaly, a combination of autozygosity mapping and exome sequencing identified a missense variant in *PGAP3* as the only likely candidate. In families B and C, mutations were uncovered in the same gene via a targeted sequencing approach in a cohort of individuals ascertained specifically for ID and hyperphosphatasia. In spite of the different approaches, the convergent findings, taken together with the segregation, the absence of these

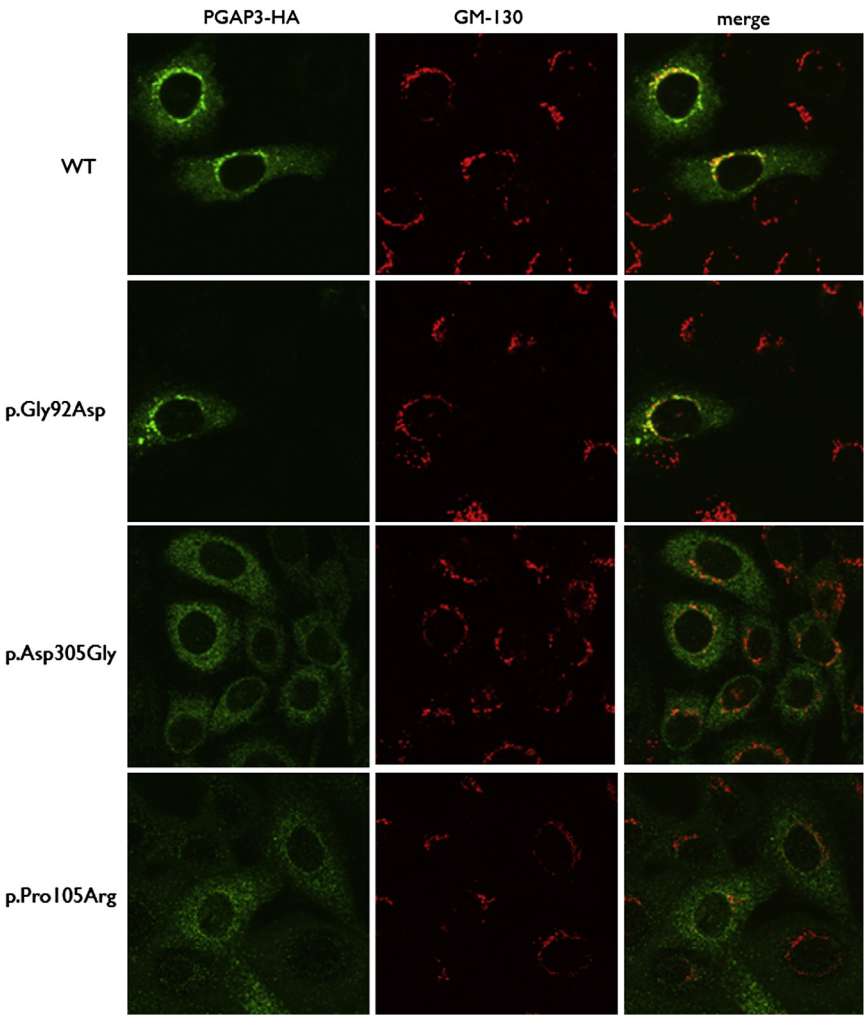


Figure 5. Subcellular Localization of HA-Tagged PGAP3 in CHO Cells
(First row) Wild-type.
(Second row) p.Gly92Asp in family A.
(Third row) p.Asp305Gly in family B.
(Fourth row) Pro105Arg in family C.
GM-130 is the Golgi marker. The relatively high levels of the ER form of HA-tagged wild-type PGAP3 seen in the top panels and also in lane 3 in Figure 4A might have been due to overexpression.

variants in suitable controls, phylogenetic conservation, and functional studies on CHO cell lines, provide strong evidence supporting the etiological role of these mutations. Along with *PGAP1* and *PGAP2*, *PGAP3* is responsible for the modification of the fatty acid residues on the GPI anchor in a maturation process that occurs in the ER and Golgi. Our results and previously reported data on *PGAP2* mutations suggest that impairment of fatty acid remodelling results in GPI-APs that are more prone to cleavage on the plasma membrane and clinical features that are similar to HPMRS. Our findings widen both the phenotypic and the genotypic spectra of hyperphosphatasia and ID syndromes that are caused by mutations in *PIGV*, *PIGO*, and *PGAP2* and suggest a heterogeneous etiology caused by impairment of late GPI-anchor synthesis and the GPI-AP-maturation pathway. These functional mutations define this condition as a CDG, PGAP3-CDG. In contrast to classical CDGs, which affect N-glycosylation and O-glycosylation, or both, defects in the GPI-anchor-biosynthesis pathway cannot be detected with a transferrin or APOCIII glycosylation assay. This underlines the importance of comprehensive genetic testing of individuals with suspected CDGs.

Supplemental Data

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

Acknowledgments

This work was supported by the National Institute for Health Research (NIHR) Biomedical Research Centre Oxford with funding from the Department of Health's NIHR Biomedical Research Centres funding scheme. The views expressed in this publication are those of the authors and not necessarily those of the Department of Health. The work was also supported by a grant from the German Ministry of Research and Education to S.M. (0313911), by Deutsche Forschungsgemeinschaft grants to P.M.K. (KR 3985/1-1) and S.M. (SFB 665), and by grants from the Ministry of Education, Culture, Sports, Science, and Technology and the Ministry of Health, Labour, and Welfare of Japan to Y.M. and T.K. We thank the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics (funded by Wellcome Trust grant reference 090532/Z/09/Z and Medical Research Council Hub grant G0900747 91070) for generating the sequencing data, John Broxholme for assisting with downloading BAM files from the 1000 Genomes Project, Kevin Leyden for performing flow cytometry, Christian Babbs for sharing DNA from Arabic

controls, and Kana Miyanagi for assisting in functional analysis of PGAP3. We would also like to thank all three families for their participation in this study and the 500 Whole-Genome Sequences Consortium and Illumina for use of the in-house database of variant calls for 274 individuals.

Received: September 25, 2013
Accepted: December 11, 2013
Published: January 16, 2014

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes, http://www.1000genomes.org/
ANNOVAR, http://www.openbioinformatics.org/annovar/
Burrows-Wheeler Aligner, http://bio-bwa.sourceforge.net/
ClustalW, http://www.clustal.org/clustal2/
Coriell Cell Repositories, http://ccr.coriell.org/
dbSNP, http://www.ncbi.nlm.nih.gov/snp/
GATK, http://www.broadinstitute.org/gatk/index.php
GeneTalk, http://www.gene-talk.de
Ingenuity, https://variants.ingenuity.com/MutationsinPGAP3causeARID
HomozygosityMapper, http://www.homozygositymapper.org/
KEGG: Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/
LifeScope, http://www.lifetechnologies.com/lifescope.html
MutationTaster, http://www.mutationtaster.org
NHLBI Exome Sequencing Project (ESP) Exome Variant Server, http://evs.gs.washington.edu/EVS/
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org
Oxford Brain Abnormality Research Group, http://www.brainabnormalities.org.uk
Platypus, http://www.well.ox.ac.uk/platypus
PolyPhen-2, http://genetics.bwh.harvard.edu/pph2/
SIFT, http://sift.jcvi.org/
UCSC Genome Browser, http://www.genome.ucsc.edu

References

1. Fujita, M., and Kinoshita, T. (2012). GPI-anchor remodeling: potential functions of GPI-anchors in intracellular trafficking and membrane dynamics. Biochim. Biophys. Acta *1821*, 1050–1058.
2. McKean, D.M., and Niswander, L. (2012). Defects in GPI biosynthesis perturb Cripto signaling during forebrain development in two new mouse models of holoprosencephaly. Biol. Open *1*, 874–883.
3. Park, S., Lee, C., Sabharwal, P., Zhang, M., Meyers, C.L., and Sockanathan, S. (2013). GDE2 promotes neurogenesis by glycosylphosphatidylinositol-anchor cleavage of RECK. Science *339*, 324–328.
4. Jaeken, J. (2011). Congenital disorders of glycosylation (CDG): it's (nearly) all in it!. J. Inherit. Metab. Dis. *34*, 853–858.
5. Nozaki, M., Ohishi, K., Yamada, N., Kinoshita, T., Nagy, A., and Takeda, J. (1999). Developmental abnormalities of glycosylphosphatidylinositol-anchor-deficient embryos revealed by Cre/loxP system. Lab. Invest. *79*, 293–299.
6. Johnston, J.J., Gropman, A.L., Sapp, J.C., Teer, J.K., Martin, J.M., Liu, C.F., Yuan, X., Ye, Z., Cheng, L., Brodsky, R.A., and Biesecker, L.G. (2012). The phenotype of a germline mutation in PIGA: the gene somatically mutated in paroxysmal nocturnal hemoglobinuria. Am. J. Hum. Genet. *90*, 295–300.
7. Ng, B.G., Hackmann, K., Jones, M.A., Eroshkin, A.M., He, P., Wiliams, R., Bhide, S., Cantagrel, V., Gleeson, J.G., Paller, A.S., et al. (2012). Mutations in the glycosylphosphatidylinositol gene PIGL cause CHIME syndrome. Am. J. Hum. Genet. *90*, 685–688.
8. Almeida, A.M., Murakami, Y., Layton, D.M., Hillmen, P., Sellick, G.S., Maeda, Y., Richards, S., Patterson, S., Kotsianidis, I., Mollica, L., et al. (2006). Hypomorphic promoter mutation in PIGM causes inherited glycosylphosphatidylinositol deficiency. Nat. Med. *12*, 846–851.
9. Maydan, G., Noyman, I., Har-Zahav, A., Neriah, Z.B., Pasmanik-Chor, M., Yeheskel, A., Albin-Kaplanski, A., Maya, I., Magal, N., Birk, E., et al. (2011). Multiple congenital anomalies-hypotonia-seizures syndrome is caused by a mutation in PIGN. J. Med. Genet. *48*, 383–389.
10. Kvarnung, M., Nilsson, D., Lindstrand, A., Korenke, G.C., Chiang, S.C., Blennow, E., Bergmann, M., Stödberg, T., Mäkitie, O., Anderlid, B.M., et al. (2013). A novel intellectual disability syndrome caused by GPI anchor deficiency due to homozygous mutations in PIGT. J. Med. Genet. *50*, 521–528.
11. Krawitz, P.M., Murakami, Y., Hecht, J., Krüger, U., Holder, S.E., Mortier, G.R., Delle Chiaie, B., De Baere, E., Thompson, M.D., Roscioli, T., et al. (2012). Mutations in PIGO, a member of the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation. Am. J. Hum. Genet. *91*, 146–151.
12. Krawitz, P.M., Schweiger, M.R., Rödelsperger, C., Marcelis, C., Kölsch, U., Meisel, C., Stephani, F., Kinoshita, T., Murakami, Y., Bauer, S., et al. (2010). Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nat. Genet. *42*, 827–829.
13. Murakami, Y., Kanzawa, N., Saito, K., Krawitz, P.M., Mundlos, S., Robinson, P.N., Karadimitris, A., Maeda, Y., and Kinoshita, T. (2012). Mechanism for release of alkaline phosphatase caused by glycosylphosphatidylinositol deficiency in patients with hyperphosphatasia mental retardation syndrome. J. Biol. Chem. *287*, 6318–6325.
14. Hansen, L., Tawamie, H., Murakami, Y., Mang, Y., ur Rehman, S., Buchert, R., Schaffer, S., Muhammad, S., Bak, M., Nöthen, M.M., et al. (2013). Hypomorphic mutations in PGAP2, encoding a GPI-anchor-remodeling protein, cause autosomal-recessive intellectual disability. Am. J. Hum. Genet. *92*, 575–583.
15. Krawitz, P.M., Murakami, Y., Rieß, A., Hietala, M., Krüger, U., Zhu, N., Kinoshita, T., Mundlos, S., Hecht, J., Robinson, P.N., and Horn, D. (2013). PGAP2 mutations, affecting the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation syndrome. Am. J. Hum. Genet. *92*, 584–589.
16. Tashima, Y., Taguchi, R., Murata, C., Ashida, H., Kinoshita, T., and Maeda, Y. (2006). PGAP2 is essential for correct processing and stable expression of GPI-anchored proteins. Mol. Biol. Cell *17*, 1410–1420.
17. Maeda, Y., Tashima, Y., Houjou, T., Fujita, M., Yoko-o, T., Jigami, Y., Taguchi, R., and Kinoshita, T. (2007). Fatty acid remodeling of GPI-anchored proteins is required for their raft association. Mol. Biol. Cell *18*, 1497–1506.
18. Najmabadi, H., Hu, H., Garshasbi, M., Zemojtel, T., Abedini, S.S., Chen, W., Hosseini, M., Behjati, F., Haas, S., Jamali, P.,

et al. (2011). Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature *478*, 57–63.

19. Lunter, G., and Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. *21*, 936–939.
20. Kamphans, T., and Krawitz, P.M. (2012). GeneTalk: an expert exchange platform for assessing rare sequence variants in personal genomes. Bioinformatics *28*, 2515–2516.
21. Thompson, M.D., Roscioli, T., Marcelis, C., Nezarati, M.M., Stolte-Dijkstra, I., Sharom, F.J., Lu, P., Phillips, J.A., Sweeney, E., Robinson, P.N., et al. (2012). Phenotypic variability in hyperphosphatasia with seizures and neurologic deficit (Mabry syndrome). Am. J. Med. Genet. A. *158A*, 553–558.
22. Heinrich, V., Kamphans, T., Stange, J., Parkhomchuk, D., Hecht, J., Dickhaus, T., Robinson, P.N., and Krawitz, P.M. (2013). Estimating exome genotyping accuracy by comparing to data from large scale sequencing projects. Genome Med. *5*, 69.
23. Kamphans, T., Sabri, P., Zhu, N., Heinrich, V., Mundlos, S., Robinson, P.N., Parkhomchuk, D., and Krawitz, P.M. (2013). Filtering for compound heterozygous sequence variants in non-consanguineous pedigrees. PLoS ONE *8*, e70151.
24. Murakami, H., Wang, Y., Hasuwa, H., Maeda, Y., Kinoshita, T., and Murakami, Y. (2012). Enhanced response of T lymphocytes from Pgap3 knockout mouse: Insight into roles of fatty acid remodeling of GPI anchored proteins. Biochem. Biophys. Res. Commun. *417*, 1235–1241.
25. Wang, Y., Murakami, Y., Yasui, T., Wakana, S., Kikutani, H., Kinoshita, T., and Maeda, Y. (2013). Significance of glycosylphosphatidylinositol-anchored protein enrichment in lipid rafts for the control of autoimmunity. J. Biol. Chem. *288*, 25490–25499.
26. Rudd, P.M., Morgan, B.P., Wormald, M.R., Harvey, D.J., van den Berg, C.W., Davis, S.J., Ferguson, M.A., and Dwek, R.A. (1998). The glycosylation of the complement regulatory protein, human erythrocyte CD59. Adv. Exp. Med. Biol. *435*, 153–162.

2.10 Ein Fall von PNH, mit Keimbahn- und somatischer Mutation in *PIGT*

Krawitz, P.M., Höchsmann, B., Murakami, Y., Teubner, B., Krüger, U., Klopocki, E., Neitzel, H., Höllein, A., Schneider, C., Parkhomchuk, D., et al. (2013). A case of paroxysmal nocturnal hemoglobinuria caused by a germline mutation and a somatic mutation in PIGT, Blood 122, 1312-1315.

Paroxysmale nächtliche Hämoglobinurie, PNH, ist eine erworbene, klonale Erkrankung, bei der eine hämatopoetische Stammzelle ihre Fähigkeit zur GPI-Anker-Synthese verloren hat. Die Routine-Diagnostik beruht auf einer durchflusszytometrischen Messung verschiedener GPI-APs, wie CD55 oder CD59, auf unterschiedlichen myeloischen Zelltypen. In Patienten mit PNH kann man üblicherweise Granulozyten-Subpopulation identifizieren, die einen kompletten Verlust der GPI-verankerten Marker aufweisen. In ca. 90% der Patienten mit PNH lassen sich in diesen Zellklonen somatische loss-of-function Mutationen im X-chromosomalen Gen *PIGA* nachweisen. Auch bei weiblichen Personen geht man aufgrund der zufälligen Inaktivierung eines X-Chromosoms davon aus, dass ein einzelnes Mutationsereignis, welches das aktive X-Chromosom betrifft, ausreicht, um die GPI-Anker-Synthese zu stören.

Wir untersuchten eine Kohorte von Patienten mit PNH, bei denen keine Mutation im Gen *PIGA* mittels Sanger Sequenzierung identifiziert werden konnte. Bei einigen Fällen gelang es uns durch die hohe Abdeckung mit der NGS Methode (ultra-deep-sequencing) somatische Mutationen in *PIGA* zu detektieren, die unter der Nachweisgrenze bei der herkömmlichen Sequenzierung lagen.

In zwei Fällen[1] jedoch fanden wir heterozygote, somatische Deletionen, die Regionen auf dem Chromosom 20 betrafen und das Gen *PIGT* beinhalteten, arr20q11.23q13.12 und arr20q11.22q13.12. Das gesamte Ausmaß der Deletionen wurde mit der array-CGH Methode erfasst. Die Deletionen der Exone von *PIGT* konnten auch anhand einer verminderten Abdeckung in den NGS-Daten bestätigt werden.

Im Gegensatz zum X-chromosomalen Gen *PIGA* kann jedoch eine somatische Mutation in einem der autosomalen Gene der GPI-Anker-Synthese nur zu einer GPI-Anker-Störung führen, wenn bereits eine prädisponierende Mutation auf dem anderen Allel vorliegt.

Tatsächlich konnten wir neben den somatischen Deletionen zudem in beiden Patienten je eine heterozygote Mutation in *PIGT* identifizieren, die alle Zellen betreffen und damit vererbte Mutationen darstellen. In dem ersten Patienten handelte es sich um eine Basen-Substitution, c.1401-2A>G, die eine kanonische Spleiß-Akzeptor-Stelle betrifft und dadurch zum Wegfall eines Exons führt. In dem zweiten Patienten fanden wir eine das Leseraster verschiebende 4 bp Deletion, c.761_764delGAAA. Beide Keimbahn-Mutationen bewirken einen Funktionsverlust des resultierenden Genprodukts.

In DNA, die aus einem Schleimhautabstrich gewonnen wurde, und somit größtenteils nicht myeloiden Zellen entstammt, zeigt sich ein heterozygotes Bild der Mutation. In der Erbsubstanz, die präferentiell aus den somatisch mutierten, myeloiden Zellklonen gewonnen wurde, lässt sich hingegen in der Mehrzahl der Sequenzfragmente die Keimbahn-Mutation nachweisen. Diese relative Depletion des Wildtyp Allels ist bereits ein deutlicher Hinweis darauf, dass die mehrere Megabasen

[1] In der aufgeführten Publikation wird der PNH Fall mit der Keimbahnmutation NM_015937.3:c.1401-2A>G und der somatischen Deletion arr20q11.23q13.12 beschrieben. Der Fall mit der Keimbahnmutation c.761_764delGAAA und der somatischen Deletion arr20q11.22q13.12 ist bislang noch nicht veröffentlicht.

umfassende somatische Deletion auf dem Haplotyp mit der Wildtyp-Sequenz von *PIGT* liegt. Desweiteren konnte die somatische Entstehung der 8 Mb umfassenden Deletion durch eine Fluoreszenz *in situ* Hybridisierung mit einer Sonde, die im dem fraglichen Bereich bindet, an Interphase-Kernen gezeigt werden: Während über 90 % der untersuchten Granulozyoten nur ein Leuchtsignal pro Zellkern zeigten, ließen sich in allen Lymphozyten zwei Bindungsereignisse nachweisen.

Eine PNH, die durch einen Funktionsverlust des autosomalen Gens *PIGT* verursacht wird, setzt damit zwei Mutationen voraus, von denen eine angeboren ist und die zweite erworben. Die geringe Prävalenz von heterozygoten loss-of-function Mutationen in den autosomalen Genen der GPI-Anker-Synthese erklärt, warum somatische Mutationen im X-chromosomalen Gen *PIGA* die bei weitem häufigere molekulargenetische Ursache der PNH darstellen.

Wir haben damit erstmals Mutationen in einem autosomalen Gen als ursächlich für eine PNH nachgewiesen.

RED CELLS, IRON, AND ERYTHROPOIESIS

A case of paroxysmal nocturnal hemoglobinuria caused by a germline mutation and a somatic mutation in *PIGT*

Peter M. Krawitz,[1] Britta Höchsmann,[2] Yoshiko Murakami,[3] Britta Teubner,[1] Ulrike Krüger,[1] Eva Klopocki,[4] Heidemarie Neitzel,[1] Alexander Hoellein,[5] Christina Schneider,[2] Dmitri Parkhomchuk,[1] Jochen Hecht,[6] Peter N. Robinson,[1] Stefan Mundlos,[1] Taroh Kinoshita,[3] and Hubert Schrezenmeier[2]

[1]Institute for Medical Genetics and Human Genetics, Charité Universitätsmedizin, Berlin, Germany; [2]Institute for Clinical Transfusion Medicine and Immunogenetics, German Red Cross Blood Transfusion Service Baden Württemberg – Hessen, University of Ulm, Ulm, Germany; [3]Research Institute for Microbial Diseases and World Premier International Immunology Frontier Research Center, Osaka University, Osaka, Japan; [4]Institute for Medical Genetics, University of Würzburg, Würzburg, Germany; [5]Medizinische Klinik III, Technische Universität München, Munich, Germany; and [6]Next-generation Sequencing Facility, Berlin-Brandenburg Center for Regenerative Therapies, Berlin, Germany

Key Points

- A carrier of a deleterious splice site mutation in *PIGT* acquired a second hit in *PIGT* and developed PNH.

To ascertain the genetic basis of a paroxysmal nocturnal hemoglobinuria (PNH) case without somatic mutations in *PIGA*, we performed deep next-generation sequencing on all exons of known genes of the glycosylphosphatidylinositol (GPI) anchor synthesis pathway. We identified a heterozygous germline splice site mutation in *PIGT* and a somatic 8-MB deletion in granulocytes affecting the other copy of *PIGT*. *PIGA* is essential for GPI anchor synthesis, whereas *PIGT* is essential for attachment of the preassembled GPI anchor to proteins. Although a single mutation event in the X-chromosomal gene *PIGA* is known to cause GPI-anchored protein deficiency, 2 such hits are required in the autosomal gene *PIGT*. Our data indicate that PNH can occur even in the presence of fully assembled GPI if its transfer to proteins is defective in hematopoietic stem cells. (*Blood*. 2013;122(7):1312-1315)

Introduction

Paroxysmal nocturnal hemoglobinuria (PNH) is an acquired hemolytic anemia that results from the expansion of hematopoietic stem cells that are deficient for glycosylphosphatidylinositol (GPI), a glycolipid moiety that anchors >100 different proteins to the cell surface.[1-5] PNH patients were reported to be deficient for an initial step in the GPI anchor synthesis that is catalyzed by the GPI-GlcNAc transferase,[3,6,7] and somatic mutations in the X-chromosomal gene *PIGA* that encodes a subunit of this transferase complex[8] are regarded as the causative event in the predominant number of PNH cases.[2-5,7,9] However, in a small number of PNH cases with a clear GPI anchor deficiency, no mutations in *PIGA* have been found.

In this work, we report about 2 mutation events, a germline splice site mutation and a somatic deletion in *PIGT*, which is another gene of the GPI anchor synthesis pathway, that we identified performing next-generation sequencing in a PNH patient with wild-type *PIGA*.

Study design

Patient sample

This study was conducted in accordance with the Declaration of Helsinki. Genetic analysis was performed after approval by ethical committee and informed consent.

Targeted genomic sequencing, array comparative genomic hybridization, and fluorescence in situ hybridization

For the targeted enrichment of exons of all known GPI anchor synthesis genes (supplemental Table 1 on the *Blood* website), we used a customized SureSelect library (Agilent) as previously described.[10] Genomic DNA of the patient and 9 controls was isolated from whole blood and enriched for GPI pathway exons according to the manufacturer's protocol, followed by single-read cluster generation on a Cluster Station (Illumina). The captured, purified, and clonally amplified library was then sequenced on an Illumina Genome Analyzer IIx and mapped to the human reference sequence GRCh37, resulting in a mean coverage of >300-fold for all exons and >10-fold coverage for >95% of the target region. Variants were detected with SAMtools,[11] annotated with ANNOVAR,[12] and further analyzed in GeneTalk.[13]

For the detection of exon deletions, we first counted the reads per exon and normalized this value for each sample by the total number of reads that were mapped to the target region. This normalized read count per exon was used to compute the mean and variance for the coverage per exon in all analyzed samples. Exons with a normalized coverage that was 2 standard deviations below the mean were classified as partially deleted in a subpopulation of cells and further analyzed.

Array comparative genomic hybridization (arrayCGH) was carried out on genomic DNA isolated from a peripheral blood draw using whole-genome 1 M Oligonucleotide-Array (Agilent) to confirm the deletion of *PIGT* and to characterize its extent. Analysis was performed with Feature Extraction and CGH Analytics software (Agilent) as described previously.[14] The copy number variants (CNV) involving *PIGT* was further analyzed with

Submitted January 28, 2013; accepted May 23, 2013. Prepublished online as *Blood* First Edition paper, June 3, 2013; DOI 10.1182/blood-2013-01-481499.

The online version of this article contains a data supplement.

There is an Inside *Blood* commentary on this article in this issue.

The publication costs of this article were defrayed in part by page charge payment. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

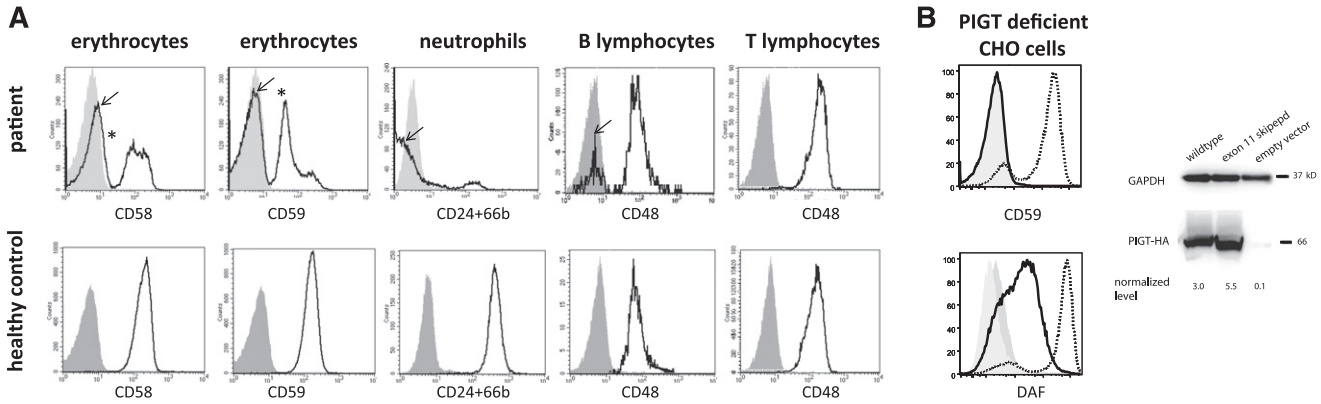


Figure 1. Expression of GPI-anchored proteins on patient's peripheral blood cells; and reduced activity of PIGT mutant in restoring surface expression of GPI-anchored proteins after transfection into *PIGT*-null cell lines. (A) Expression of (left) CD58 and CD59 on red cells, (center) CD24 and CD66b on neutrophil granulocytes, and (right) CD48 on B and T lymphocytes. The first row shows the expression of GPI-anchored protein (AP) at the time of ultradeep sequencing (4.5 years after start of eculizumab); the second row shows expression of GPI-AP in a healthy control. In healthy controls, CD58 and CD59 are expressed on >99.9% and >99.5% of red cells, respectively, and CD24/CD66b is expressed on >99.8% of neutrophil granulocytes (second row). In contrast, the patient shows a mosaic of cells with normal expression of GPI-anchored proteins and cells with reduced or completely missing expression of GPI-AP on (left) erythrocytes or (center) neutrophil granulocytes. The cell populations that completely lack expression of the respective GPI-AP are indicated by arrows; the populations with reduced GPI-AP expression are marked by asterisks. The patient did not receive any blood transfusions over a period of 3 months before this measurement. Expression of the GPI-AP CD48 on T lymphocytes was normal, whereas a subpopulation of B lymphocytes did not express the GPI-AP CD48. The percentages of cells with reduced or absent GPI-AP, ie, PNH cells, and normal range is shown in the supplemental Materials. (B) *PIGT*-deficient Chinese Hamster Ovary cells were transiently transfected with wild-type or a mutant version skipping exon 11 of transcript NM_0015937. (Left) Restoration of the cell surface protein levels of wild-type PIGT and the mutant PIGT lacking 28 amino acids encoded by exon 11 was assessed by flow cytometry. Wild-type PIGT efficiently restored expression levels of CD59 and CD55 at the cell surface (dotted black lines), whereas the mutant PIGT did not rescue CD59 and only partially rescued CD55 expression (solid black lines). Dark shading, empty vector; light shading, isotype-matched control. (Right) Expression levels of transfected wild-type and the mutant HA-tagged PIGT. PIGT proteins were determined by western blotting with anti-HA; GAPDH, loading control. Normalized PIGT levels are shown at the bottom.

fluorescence in situ hybridization (FISH) using BAC clone RP3-337O18 in metaphases of phytohemagglutinin-stimulated T lymphocytes and granulocytes that were enriched by a Ficoll gradient.

Cell culture and fluorescence-activated cell sorter

We cloned a coding region of human *PIGT* (NM_015937) from a cDNA library derived from placenta,[15] tagged with FLAG at the *N* terminus, and subcloned it into plasmid mammalian expression.[16] A *PIGT* mutant with skipped exon 11 was generated by site-directed mutagenesis. Mutant and wild-type *PIGT* plasmids were transfected by electroporation into PIGT-deficient Chinese Hamster Ovary cells expressing human GPI-anchored proteins, CD55 and CD59, as previously described.[17] Two days later, lysates were applied to sodium dodecyl sulfate-polyacrylamide gel electrophoresis and western blotting against anti-FLAG antibody to determine levels of expressed PIGT. The levels of CD55 and CD59 restored at the cell surface were determined by fluorescence-activated cell sorter.

Results and discussion

We performed targeted enrichment of all exons of genes involved in GPI anchor synthesis followed by ultradeep sequencing in a female patient with classical hemolytic PNH that is negative for mutations in *PIGA*. The patient was diagnosed with hemolytic anemia with a negative direct antiglobulin test at the age of 44 years and experienced frequent hemolytic crises, abdominal pain, diarrhea, headache, arthralgia, dyspnea, and fatigue in the following years. At the age of 49 years, a flow cytometric analysis was performed that showed reduced expression of GPI-anchored proteins on blood cells (Figure 1A). DNA was isolated from blood at that time and subjected to ultradeep sequencing. The patient was started on eculizumab due to PNH-related symptoms soon after it became available 6 years ago and responded to this treatment (see supplemental Materials for a detailed clinical description of the patient).

We detected a significant reduction in the coverage of all *PIGT* exons in the DNA extracted from blood compared with other genes of the GPI anchor synthesis pathway, which suggested a deletion of this gene in a subpopulation of cells (Figure 2A). We performed array CGH to measure the full extent of the CNV and detected an 8-MB deletion, arr20q11.23q13.12 (Figure 2A). To clarify which subpopulation was affected by the deletion, we used a FISH probe (RP3-337O18) targeting the CNV interval in T lymphocytes and granulocytes. Although we did not observe any deletion in full metaphases of T lymphocytes, 92% of the evaluated granulocyte interphase nuclei showed only a single signal for RP3-337O18, suggesting a heterozygous deletion including *PIGT* in a myeloid stem cell that occurred as a somatic event (Figure 2B).

The mutation analysis of the deep sequencing data revealed a single nucleotide substitution in *PIGT* affecting the splice acceptor site of intron 10: NM_015937:c.1401-2A>G (Figure 2C). From 1463 sequence reads that cover the splice site, 1239 showed the base substitution, suggesting that the mutation is present on the chromosome without the somatic deletion involving *PIGT*. We also measured the splice site mutation in a heterozygous state in ABI Sanger sequences of DNA that was extracted from epithelial cells of a buccal swab providing further evidence that c.1401-2A is the germline event (Figure 2D). Based on these findings, we hypothesized that the somatic deletion of the wild-type allele of *PIGT* occurred in a myeloid stem cell and resulted in a clone that is hemizygous for *PIGT*. In this clone, the single remaining copy of *PIGT* is functionally impaired due to the splice site mutation that results in skipping 84 bp of exon 11 and deleting 28 highly conserved amino acids in PIGT.

We analyzed the functional effect of the germline splice site mutation in *PIGT*-null Chinese Hamster Ovary cells. Although the transfection of wild-type *PIGT* into these cells restored the levels of wild-type GPI-linked proteins CD55 and CD59 at the cell surface, the transfection of the mutant only leads to a minor increase of CD55 surface expression but almost no CD59 expression at comparable PIGT protein levels (Figure 1B).

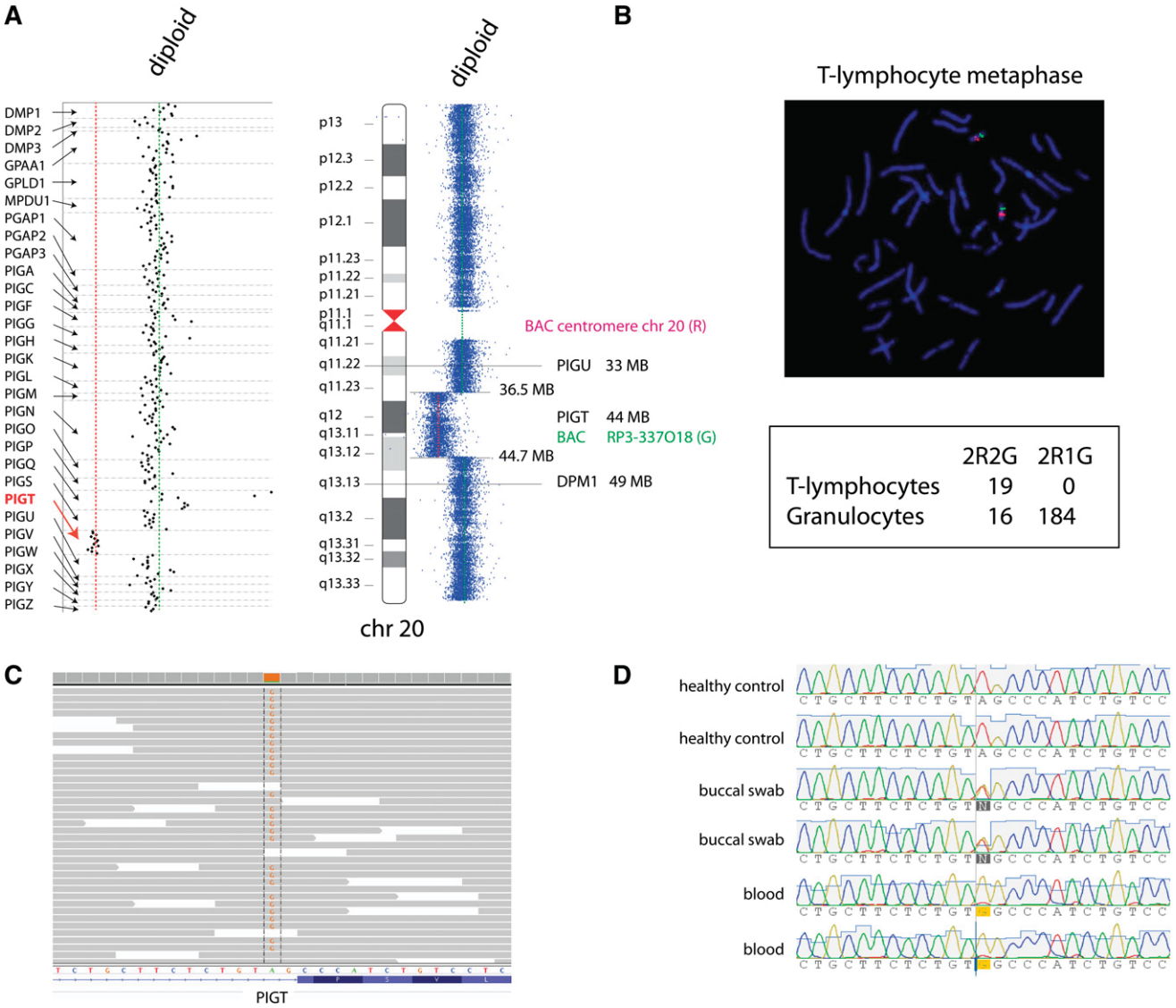


Figure 2. Ultradeep sequencing of all exons of genes involved in GPI anchor synthesis reveals two mutation events in *PIGT*: a germline splice site mutation and a somatic deletion. (A) DNA was isolated from whole blood and enriched for all exons of genes involved in GPI anchor synthesis and subjected to ultradeep sequencing. The coverage of *PIGT* exons was significantly reduced compared with exons of all other GPI anchor synthesis genes, suggesting a deletion involving *PIGT*. The extent of the deletion was further characterized by arrayCGH comprising in total 8 MB, arr20q11.23q13.12. (B) FISH with BAC clone RP3-337O18 (G) and a probe targeting the centromere of chromosome 20 (R) was used to analyze the deletion in T lymphocytes and granulocytes. Although 2 signals of RP3-337O18 were present in all complete metaphases of T lymphocytes, the majority of granulocytes showed only 1 signal for RP3-337O18, indicating a somatic deletion in a myeloid lineage. (C) A single nucleotide substitution in *PIGT* affecting the splice acceptor site of intron 10, NM_015937:c.1401-2A>G, was observed in the ultradeep sequencing data of DNA extracted from whole blood. In total, 1463 sequence reads covered the canonical splice site, and 85% of these reads showed the alternate base, indicating that the mutation is present on the undeleted haplotype of *PIGT*. (D) The splice site mutation was validated by ABI Sanger sequencing and shown to be heterozygous in DNA extracted from epithelial cells of a buccal swap, confirming its presence in different tissues.

In contrast to the X-chromosomal *PIGA,* all other known genes involved in the GPI anchor synthesis pathway, including *PIGT*, are found on autosomes, and inactivating mutations in these genes have to occur on both alleles in the same cell to result in a GPI anchor deficiency. The co-occurrence of 2 mutations in the same gene is a situation that is similar to hereditary cases of retinoblastoma that have been explained by a 2-hit model of 1 inherited mutation and 1 somatic mutation in *RB1*.[18] Therefore, individuals that are heterozygous for mutations in autosomal genes that impair GPI anchor synthesis, such as the reported splice site mutation in *PIGT*, might have an increased risk to develop PNH.

Although PIGA catalyzes the first step of the GPI anchor synthesis,[8] PIGT is a component of the transamidase complex that is required for attachment of preassembled GPI to proteins.[15] Therefore, even in the presence of fully assembled GPI anchors, PNH can occur. This suggests that not only the specific defect in the GPI anchor synthesis that is caused by *PIGA* mutations but also a GPI-anchored protein deficiency that is due to mutations in other genes of the pathway may predispose for PNH. Interestingly, a deletion on 20q is also a recurrent somatic abnormality in myelodysplastic syndrome; however, it is currently not clear whether the loss of heterozygosity of other genes in this region besides PIGT, contributes to the clonal expansion.

Recent findings of congenital GPI deficiencies also shed new light on the clinical feature of hemoglobinuria. Although no hemolysis was reported for patients with germline mutations in *PIGN*,[19] *PIGM*,[20] *PIGO*,[21] *PIGL*,[22] *PIGV*,[23] and even *PIGA*[24] and *PIGT*,[25] chronic hemolysis was described in patients with a congenital CD59

deficiency[26] that responds to eculizumab therapy.[27] Further studies are therefore required to elucidate how mutations in GPI pathway genes contribute to the different phenotypic features and to what extent additional somatic events occur.

Acknowledgments

The authors thank the reviewers for valuable comments and Seval Turkmen for helpful discussions.

This work was supported by a grant from the Bundesministerium für Forschung und Technologie (0313911), Deutsche Forschungsgemeinschaft grants KR 3985/1-1 (to P.M.K.) and SFB 665 (to S.M.), and grants from the Ministry of Education, Culture, Sports, Science and Technology, and the Ministry of Health, Labour and Welfare of Japan.

Authorship

Contribution: P.M.K. and D.P. performed research and analyzed the data; U.K., J.H., and C.S. performed sequencing studies; B.H. and H.S. provided patient samples and characterized the patient; A.H. provided patient samples, performed research, and analyzed data; E.K. performed arrayCGH; B.T. performed the FISH analysis; Y.M. performed cell culture experiments; H.S., B.H., and P.M.K. designed the study; and E.K., H.N., P.N.R., Y.M., J.H., T.K., and S.M. wrote the paper.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

Correspondence: Peter Krawitz, Institute for Medical Genetics and Human Genetics, Augustenburger Platz 1, 13353 Berlin, Germany; e-mail: peter.krawitz@gmail.com.

References

1. Brodsky RA. Advances in the diagnosis and therapy of paroxysmal nocturnal hemoglobinuria. *Blood Rev.* 2008;22(2):65-74.
2. Bessler M, Mason P, Hillmen P, Luzzatto L. Somatic mutations and cellular selection in paroxysmal nocturnal haemoglobinuria. *Lancet.* 1994;343(8903):951-953.
3. Hillmen P, Lewis SM, Bessler M, Luzzatto L, Dacie JV. Natural history of paroxysmal nocturnal hemoglobinuria. *N Engl J Med.* 1995;333(19):1253-1258.
4. Rosse WF. Epidemiology of PNH. *Lancet.* 1996;348(9027):560.
5. Socié G, Mary JY, de Gramont A, et al; French Society of Haematology. Paroxysmal nocturnal haemoglobinuria: long-term follow-up and prognostic factors. *Lancet.* 1996;348(9027):573-577.
6. Takahashi M, Takeda J, Hirose S, et al. Deficient biosynthesis of N-acetylglucosaminyl-phosphatidylinositol, the first intermediate of glycosyl phosphatidylinositol anchor biosynthesis, in cell lines established from patients with paroxysmal nocturnal hemoglobinuria. *J Exp Med.* 1993;177(2):517-521.
7. Takeda J, Miyata T, Kawagoe K, et al. Deficiency of the GPI anchor caused by a somatic mutation of the PIG-A gene in paroxysmal nocturnal hemoglobinuria. *Cell.* 1993;73(4):703-711.
8. Miyata T, Takeda J, Iida Y, et al. The cloning of PIG-A, a component in the early step of GPI-anchor biosynthesis. *Science.* 1993;259(5099):1318-1320.
9. Nafa K, Bessler M, Castro-Malaspina H, Jhanwar S, Luzzatto L. The spectrum of somatic mutations in the PIG-A gene in paroxysmal nocturnal hemoglobinuria includes large deletions and small duplications. *Blood Cells Mol Dis.* 1998;24(3):370-384.
10. Krawitz PM, Murakami Y, Rieß A, et al. PGAP2 mutations, affecting the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation syndrome. *Am J Hum Genet.* 2013;92(4):584-589.
11. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics.* 2011;27(21):2987-2993.
12. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38(16):e164.
13. Kamphans T, Krawitz PM. GeneTalk: an expert exchange platform for assessing rare sequence variants in personal genomes. *Bioinformatics.* 2012;28(19):2515-2516.
14. Klopocki E, Lohan S, Brancati F, et al. Copy-number variations involving the IHH locus are associated with syndactyly and craniosynostosis. *Am J Hum Genet.* 2011;88(1):70-75.
15. Ohishi K, Inoue N, Kinoshita T. PIG-S and PIG-T, essential for GPI anchor attachment to proteins, form a complex with GAA1 and GPI8. *EMBO J.* 2001;20(15):4088-4098.
16. Takebe Y, Seiki M, Fujisawa J, et al. SR alpha promoter: an efficient and versatile mammalian cDNA expression system composed of the simian virus 40 early promoter and the R-U5 segment of human T-cell leukemia virus type 1 long terminal repeat. *Mol Cell Biol.* 1988;8(1):466-472.
17. Hong Y, Ohishi K, Inoue N, et al. Requirement of N-glycan on GPI-anchored proteins for efficient binding of aerolysin but not Clostridium septicum alpha-toxin. *EMBO J.* 2002;21(19):5047-5056.
18. Knudson AG Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci USA.* 1971;68(4):820-823.
19. Maydan G, Noyman I, Har-Zahav A, et al. Multiple congenital anomalies-hypotonia-seizures syndrome is caused by a mutation in PIGN. *J Med Genet.* 2011;48(6):383-389.
20. Almeida AM, Murakami Y, Layton DM, et al. Hypomorphic promoter mutation in PIGM causes inherited glycosylphosphatidylinositol deficiency. *Nat Med.* 2006;12(7):846-851.
21. Krawitz PM, Murakami Y, Hecht J, et al. Mutations in PIGO, a member of the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation. *Am J Hum Genet.* 2012;91(1):146-151.
22. Ng BG, Hackmann K, Jones MA, et al. Mutations in the glycosylphosphatidylinositol gene PIGL cause CHIME syndrome. *Am J Hum Genet.* 2012;90(4):685-688.
23. Krawitz PM, Schweiger MR, Rödelsperger C, et al. Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat Genet.* 2010;42(10):827-829.
24. Johnston JJ, Gropman AL, Sapp JC, et al. The phenotype of a germline mutation in PIGA: the gene somatically mutated in paroxysmal nocturnal hemoglobinuria. *Am J Hum Genet.* 2012;90(2):295-300.
25. Kvarnung M, Nilsson D, Lindstrand A, et al. A novel intellectual disability syndrome caused by GPI anchor deficiency due to homozygous mutations in PIGT [published online ahead of print May 1, 2013]. *J Med Genet.*
26. Yamashina M, Ueda E, Kinoshita T, et al. Inherited complete deficiency of 20-kilodalton homologous restriction factor (CD59) as a cause of paroxysmal nocturnal hemoglobinuria. *N Engl J Med.* 1990;323(17):1184-1189.
27. Höchsmann B, Dohna-Schwake C, Rojewski M, et al. Targeted Therapy with Eculizumab for inherited CD59 Deficiency, submitted

3 Diskussion

Insbesondere auf dem Gebiet der geistigen Behinderungen haben die Hochdurchsatz-Sequenzierverfahren in den letzten Jahren nicht nur dazu beigetragen, eine Vielzahl neuer pathogener Mutationen zu identifizieren, sondern sie haben auch dabei geholfen unser Verständnis der Pathogenese dieser Erkrankungen zu erweitern.

Mit Hilfe der Exom- und Genom-Trio-Sequenzierung konnte in großen Kohorten von Patienten mit mentaler Retardierung gezeigt werden, dass die Behinderung in bis zu zwei Dritteln der Patienten auf einzelne Neumutationen zurückzuführen ist (de Ligt, et al., 2012; Gilissen, et al., 2014; Rauch, et al., 2012): Im Mittel treten pro Kontrollindividuum 0,75 Neumutationen in den 32 MB des Exoms auf. In den untersuchten Patientengruppen mit mentaler Retardierung liegt diese Rate zwischen 1,2 und 1,7 pro Exom.

Es wurde auch deutlich, dass die Anzahl der autosomal rezessiven Gene bei nicht-syndromaler mentaler Retardierung (ARID) beträchtlich ist. In einer großen Studie, in der mental retardierte Kinder aus mehr als 300 konsanguinen Familien mittels HDS untersucht wurden, konnten 50 neue Krankheitsgene identifiziert werden (Najmabadi, et al., 2011). Man nimmt an, dass selbst in exogamen Gesellschaften bis zu einem Fünftel der geistigen Behinderungen auf Mutationen in rezessiven Krankheitsgenen zurückzuführen ist (Musante and Ropers, 2014). Der Anteil multifaktoriell bedingter mentaler Retardierungen scheint also geringer zu sein als noch vor einigen Jahren angenommen (Ellison, et al., 2013). Dies bedeutet, dass ein monogenetischer Defekt in einer Vielzahl unterschiedlicher Gene eine geistige Behinderung zur Folge haben kann.

Da ein experimenteller Nachweis für die Ursächlichkeit einer Sequenzvariante bei nicht-syndromaler mentaler Retardierung meist nicht möglich ist, betrachtet man die Evidenz als ausreichend, wenn mehrere Patienten mit als funktionseinschränkend eingestuften Mutationen im gleichen Gen gefunden werden können.

Eine übliche Definition ist es, von einem „intellectual disability gene“ oder ID Gen zu sprechen, wenn als schädlich eingestufte Neumutationen in mindestens fünf nicht verwandten, betroffenen Individuen gefunden werden konnten. Liegt die Zahl der Betroffenen zwischen 1-5 spricht man von ID Kandidaten-Genen. Gilissen und Kollegen kommen mit dieser Definition auf 538 ID Gene und weitere 628 Kandidaten. Eine Sichtung dieser Listen zeigt jedoch schnell ihre Unvollständigkeit (Supplementary Table 9 und 10 in Gilissen *et al.*). Dies ist nicht der mangelnden Qualität der Arbeit geschuldet, sondern veranschaulicht vielmehr, wie schwierig es ist, bei der raschen Entwicklung auf diesem aktuellen Gebiet die Anzahl von ID Genen zu benennen. Gilissen *et al*. führten Ihre Literaturrecherche Anfang 2014 durch und konnten daher zum Beispiel die Gene *PGAP1* und *PGAP3* noch nicht berücksichtigen. Wenn man den wissenschaftlichen Fortschritt in den kommenden Jahren antizipiert, so erscheint eine Zahl von ca. 2000 ID Genen als durchaus realistisch. Dies würde bedeuten, dass pathogene Mutationen in jedem zehnten Gen als mögliche Ursache einer geistigen Behinderung infrage kommen.

Die außerordentliche genetische Heterogenität bei mentaler Retardierungen erklärt, warum gerade bei dieser Fragestellung die Genomsequenzierung die geeignete Diagnostik darstellt. Für den klinischen Genetiker wird die Herausforderung dann zunehmend darin bestehen z.B. die Neumutationen, die bei einer Trio-Sequenzierung eines Patienten mit mentaler Retardierung identifiziert wurden, zu interpretieren.

Wie bereits angedeutet, wird bei der Klärung des Status von Kandidaten-Genen die internationale Zusammenarbeit von großer Bedeutung sein, da es für einzelne genetische Einrichtungen fast unmöglich ist, eine ausreichende Anzahl von Patienten mit pathogenen Mutationen in den gleichen Genen zu finden.

Aber auch die Interpretation von Varianten in bekannten Krankheitsgenen kann mitunter schwierig sein. Wenn es sich um eine Sequenzvariante handelt, die in ihren vorhergesagten Auswirkungen den in der Literatur beschriebenen pathogenen Mutationen ähnelt und der Patient auch phänotypisch mit den bereits publizierten Fällen weitgehend übereinstimmt, kann ein ursächlicher Zusammenhang der identifizierten, genetischen Veränderung mit der Erkrankung angenommen werden.

Es ist aber auch möglich, dass mittels einer Genom-Sequenzierung pathogene Mutationen in bekannten, syndromalen ID Genen bei Patienten identifiziert werden, bei denen der Phänotyp diese Zuordnung nicht nahegelegt hätte. Auf diese Weise kann die Genom-Sequenzierung eines Patienten auch dabei helfen, das Genotyp- und Phänotyp-Spektrum eines Gens zu erweitern. So könnte, wie bereits von Vogel und Motulsky vor annähernd dreißig Jahren erwähnt, das Studium genetischer Erkrankungen durch phänomenologische Hinweise der Grundlagenforschung neue Fragestellungen aufzeigen, die andernfalls gar nicht erkannt worden wären (Vogel, 1997). Eine besonders interessante Frage ist es, ob sich für die vielen unterschiedlichen Mutationen, die eine geistige Entwicklungsstörung zur Folge haben, charakteristische, kognitive Defizite abgrenzen lassen.

Die Untersuchungsergebnisse der bisher untersuchten Patienten mit GPI-Anker-Störungen deuten darauf hin, dass es sich hierbei um eine Krankheitsgruppe mit äußerst hoher klinischer Variabilität handelt: Wir haben sowohl Patienten mit multiplen Fehlbildungen und schwerer mentaler Retardierung, als auch Patienten mit nur mittelgradiger, nicht-syndromaler, geistiger Entwicklungsstörung beschrieben.

Trotz der hohen Variabilität scheinen sich angeborene funktionseinschränkende Mutationen in Genen, die an frühen Schritten der GPI-Anker-Synthese beteiligt sind, schwerwiegender auf den Phänotyp auszuwirken, als solche, die erst spätere Schritte der Synthese stören. Alle Patienten mit Mutationen in den Gene *PIGA*, *PIGQ*, *PIGN* und *PIGW* wiesen multiple Fehlbildungen und Epilepsien auf, die nur schwer therapierbar waren. Zudem zeigten sich in der radiologischen Bildgebung enzephalopatische Veränderungen.

Patienten mit pathogenen Mutationen in den Genen *PIGV* und *PIGO* hingegen, die für späte Schritte der GPI-Ankersynthese verantwortlich sind, weisen keine Veränderungen der Hirnstruktur auf, soweit dies anhand von MRT und CT Aufnahmen beurteilt werden kann. Von prognostischer Relevanz ist also möglicherweise, ob die Synthese des GPI-Ankers bis zum ersten Mannose-Rest fortschreiten konnte. Dann kann vermutlich die GPI-AP-Transamidase aktiviert werden und es entstehen zumindest instabilere GPI-AP Formen.

Obwohl unser biochemisches Verständnis noch unvollständig ist, erlaubt uns das aktuelle Modell für Mutationen in klinisch noch unbeschriebenen Genen der GPI-Anker-Synthese die Auswirkungen vorherzusagen: Bei einer Funktionseinschränkung von *PIGF, PIGG* oder *PIGB*, die wie auch *PIGV* und *PIGO* zu den Genen der späten Ankersynthese zählen, wäre zum Beispiel eine Hyperphosphatasie zu erwarten (Murakami, et al., 2012).

Pathogene Mutationen in Genen, die an der GPI-Anker-Reifung beteiligt sind, beeinflussen die Stabilität der GPI-APs und deren Organisation auf der Zelloberfläche. Während eine Störung der PGAP2- und PGAP3-Funktion zu einer verminderten Assoziation der GPI-APs mit „lipid rafts“ führt und damit die Abspaltung des Proteins erleichtert, verringert eine gestörte PGAP1-Funktion die

Zugänglichkeit des GPI-APs für Proteine mit Phospolipase-Aktivität. Die Anzahl von GPI-APs auf den Zelloberflächen ist also in Patienten mit einem PGAP1-Defekt nicht reduziert. Für die Pathogenese der mentalen Retardierung könnte dies bedeuten, dass auch der Spaltung bestimmter GPI-APs auf der Zellmembran von Nervenzellen eine biologisch wichtige Funktion zukommt. Dies wäre sowohl für die GPI-verankerten Ephrin A-Liganden, die während der Embryonalentwicklung eine wichtige Rolle spielen, als auch für Enzyme, die am Neurotransmitter Stoffwechsel beteiligt sind, denkbar. In einer noch nicht veröffentlichten Fallstudie wird bei einem Patienten mit angeborener PGAP1-Defizienz eine Störung des visuellen Kortex beschrieben. Es wäre interessant zu untersuchen, ob GPI-verankerte Enzyme, wie zum Beispiel die Acetylcholinesterase, am synaptischen Spalt eine veränderte Aktivität zeigen.

Aufgrund der geringen Fallzahlen ist es bislang noch nicht möglich gewesen, Genotyp-Phänotyp Korrelationen ausführlich zu studieren. Dennoch zeichnen sich erste Hinweise ab: Bei den erworbenen GPI-Anker-Störungen trifft man meist Mutationen an, die mit einem kompletten Funktionsverlust der GPI-Anker-Synthese einhergehen (nonsense, splice site, frameshift Mutationen oder Deletionen, die mehrere Exons bzw. das gesamte Gen umfassen). Bei den angeborenen GPI-Anker-Störungen findet sich immer mindestens eine missense Mutation, so dass noch eine Restfunktion erhalten bleibt und die GPI-Anker-Synthese nicht ganz zum Erliegen kommt. Dies ist in Übereinstimmung mit den Beobachtungen im Tiermodell, die nahe legen, dass ein vollständiges Fehlen der GPI-Anker-Synthese embryonal letal ist (Tarutani, et al., 1997).

Mit den durchflusszytometrischen Messverfahren kann die Reduktion der Oberflächen-Expression von GPI-APs quantifiziert werden. Die im CHO-Modell untersuchten missense Mutationen weisen zum Teil deutliche Unterschiede auf. Die in 2.8 beschriebene missense Mutation p.Arg16Trp in PGAP2 zum Beispiel scheint eine größere Restfunktion zu besitzen als die missense Mutation p.Leu127Ser im gleichen Gen. Da dies auch mit der Schwere der beobachteten Phänotypen korreliert, hat die Durchflusszytometrie möglicherweise auch eine prognostische Aussagekraft.

Das CHO-Testsystem ist jedoch mit erheblichem Arbeitsaufwand verbunden, da für jede identifizierte Sequenzvariante erst ein mutantes Sequenz-Konstrukt erzeugt werden muss, das dann in die CHO-Zellen transferiert wird. Es sollte daher untersucht werden, ob eine Quantifizierung der GPI-Anker-Reduktion alternativ auch in Fibroblasten der Patienten möglich ist, die sich unter kontrollierten Bedingungen einfach in Kultur halten lassen.

Bei der PNH, der einzigen bislang bekannten erworbenen GPI-Ankerstörung, hat der humanisierte IgG Antikörper Eculizumab, der eine Inhibition des aktivierten Komplementsystems bewirkt, die Therapie revolutioniert. Höchsmann und Kollegen beschrieben kürzlich einen Patienten mit einem angeborenen Defekt des GPI-verankerten CD59 Moleküls, der neben der für die PNH typischen Hämolyse ebenfalls Epilepsien aufwies, wie sie auch bei vielen Patienten mit angeborenen GPI-Ankerstörungen beobachtet werden (Hochsmann, et al., 2014). Die Gabe von Eculizumab führte bei diesem Patienten ebenfalls zu einer Besserung der neurologischen Symptomatik. Es erscheint damit möglich, dass der mangelnde Schutz vor Angriffen des Komplementsystems bei der Entstehung der Epilepsien bei angeborenen GPI-Ankerstörungen von Bedeutung ist. Auch im Tiermodell konnte gezeigt werden, dass die Injektion einer hohen Dosis des aktivierten Membranangriffskomplexes Krampfanfälle evozieren kann (Xiong, et al., 2003). Bei Patienten mit einer angeborenen GPI-Ankerstörung mit therapieresistenten Epilepsien und infaustem Verlauf wäre daher auch der Einsatz von Eculizumab im individuellen Heilversuch denkbar.

In dieser Arbeit wurde die Bedeutung der HDS speziell für die Diagnostik der GPI-Ankerstörungen beschrieben. Diese Methode hat darüber hinaus jedoch weitreichende Folgen für die gesamte Medizin. Hierzu wurde bereits in vielfältigen Veröffentlichungen detailliert Stellung bezogen, zum Beispiel auch durch den Deutschen Ethikrat und die Berlin-Brandenburgische Akademie der Wissenschaften („Zukunft der genetischen Diagnostik – von der Forschung in die klinische Anwendung“, 2013, „Neue Sequenzierungstechniken: Konsequenzen für die genetische Krankenversorgung“, 2013).

Die HDS und die dadurch erzeugten Daten stellen hinsichtlich ihres Umfanges eine neue Dimension in der Genetik dar und die durch diese Screening-Verfahren erhobenen Zusatzbefunde werden durchaus kritisch auch als Gefahr betrachtet.

Verfechter des genetischen Exzeptionalismus verweisen in diesem Zusammenhang gern auf die in der griechischen Mythologie beschriebene Büchse der Pandora und vergleichen das Öffnen der Büchse mit der Genomsequenzierung. Nach Hesiods Überlieferung schenkte Zeus der Pandora eine kunstvoll gearbeitete Büchse, die alle erdenklichen Übel enthielt. Ohne Pandora über den Inhalt der Büchse aufzuklären, wies er sie an, diese immer verschlossen zu halten. Durch Neugierde getrieben, öffnete Pandora die Büchse jedoch und so kamen die Übel in die Welt. Die negative Konnotation, die das „Öffnen der Büchse der Pandora“ in unserem Sprachgebrauch hat, geht auf diese Fassung des Mythos zurück. Theognis hingegen überliefert eine ganz andere Version dieser Geschichte, in der die Büchse nur Gutes enthielt. Als die Büchse versehentlich geöffnet wurde entwichen all die segensreichen Dinge, bis auf die Hoffnung, die allein in der Hand des Menschen blieb. In diesem Sinne könnte man also in den durch die Genomsequenzierung gewonnenen Daten nicht nur eine Bürde sehen sondern auch eine Chance, sofern es der Forschung gelingt, diese Vielfalt weiter zu enträtseln und therapeutische Möglichkeiten zu schaffen.

4 Zusammenfassung

Diese Habilitationsschrift fasst Arbeiten zusammen, die zum einen Beiträge zum bioinformatischen Methodenspektrum lieferten, welches bei der Analyse von umfangreichen DNA Sequenzdaten zum Einsatz kommt. Die entwickelten Algorithmen dienen ganz allgemein der Beurteilung der Datenqualität und der Priorisierung möglicherweise krankheitsverursachender Sequenzvarianten bei Patienten mit seltenen genetischen Erkrankungen, die mittels Hochdurchsatz-Sequenzierverfahren untersucht werden.

Zum anderen fasst die vorgelegte Schrift Arbeiten zusammen, in denen auch mit Hilfe der zuvor entwickelten Verfahren pathogene Mutationen in verschiedenen Genen der GPI-Ankersynthese identifiziert werden konnten: Bis 2010 war die paroxysmale nächtliche Hämoglobinurie, PNH, die einzige bekannte GPI-Ankerstörung. Molekulargenetisch konnten die meisten dieser Fälle durch somatische Nullmutationen im X-chromosomalen Gen *PIGA* in Blutstammzellen erklärt werden.

Uns gelang es vor fünf Jahren erstmals auch eine angeborene GPI-Ankerstörung aufzuklären: In Patienten mit mentaler Retardierung und Hyperphosphatasie, HPMRS, konnten wir funktionseinschränkende Mutationen in *PIGV*, einem weiteren Gen der GPI-Ankersynthese, mittels Hochdurchsatz-Sequenzierung identifizieren. Dies hat unser klinisches Bild von GPI-Ankerstörungen grundlegend erweitert.

Die Verfahren der DNA Anreicherung und parallelen Sequenzierung von Gen-Panels haben sich in der Diagnostik heterogener Erkrankungen bewährt und stellen nun oftmals die erste Stufe der molekulargenetischen Untersuchung dar. Wir haben speziell für die GPI-Ankerstörungen ein diagnostisches Gen-Panel etabliert, mit dem krankheitsverursachende Mutationen in allen bekannten Genen der GPI-Ankersynthese gefunden werden können.

Wir wiesen erstmals pathogene Mutationen in den Genen *PIGO*, *PGAP2* und *PGAP3* bei Patienten mit angeborenen GPI-Ankerstörungen nach und trugen damit zu einem besseren molekulargenetischen Verständnis dieser neuen Untergruppe angeborener Glykosylierungstörungen bei.

Mit der Identifikation von somatischen Mutationen im Gen *PIGT* in zwei Patienten mit PNH konnten wir zudem belegen, dass es sich auch bei einem erworbenen GPI-Ankerverlust um eine heterogene Erkrankung handelt. Durch einen Defekt in der GPI-Ankersynthese fehlen auf den Oberflächen der betroffenen Blutzellen die GPI-verankerten Proteine CD55 und CD59, die vor einem Angriff des körpereigenen Komplementsystems schützen.

Im Gegensatz zur PNH sind die Pathomechanismen bei den angeborenen GPI-Ankerstörungen jedoch noch weitestgehend unklar. Es gibt allerdings erste Hinweise darauf, dass - ähnlich wie bei der PNH an den Blutzellen - auch an anderen Zellsystemen das Zusammenspiel zwischen Proteinen der Zellmembran und dem Komplementsystem gestört ist. Dies könnte u.a. auch Epilepsien bedingen, die gehäuft bei Patienten mit angeborenen GPI Ankerstörungen auftreten. Hierbei und bei den weiteren klinischen Auffälligkeiten der Patientengruppe müssen die Auswirkung pathogener Mutationen auf weitere GPI-verankerte Proteine untersucht werden. Da circa jedes zehnte Membranprotein GPI-verankert ist, stellt dies eine immense Herausforderung dar.

In meiner zukünftigen Arbeit möchte ich daher auch Verfahren der Proteomanalyse anwenden, um diese Wechselwirkungen systematisch zu erfassen. Zudem möchte ich die identifizierten pathogenen Mutationen im Tiermodell charakterisieren und damit auch die Voraussetzungen zur Austestung möglicher therapeutischer Ansätze schaffen.

5 Literaturverzeichnis aus freiem Text

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7(4):248-9.

Almeida AM, Murakami Y, Layton DM, Hillmen P, Sellick GS, Maeda Y, Richards S, Patterson S, Kotsianidis I, Mollica L and others. 2006. Hypomorphic promoter mutation in PIGM causes inherited glycosylphosphatidylinositol deficiency. Nature medicine 12(7):846-51.

Chiyonobu T, Inoue N, Morimoto M, Kinoshita T, Murakami Y. 2014. Glycosylphosphatidylinositol (GPI) anchor deficiency caused by mutations in PIGW is associated with West syndrome and hyperphosphatasia with mental retardation syndrome. J Med Genet 51(3):203-7.

Cooper DN, Ball EV, Krawczak M. 1998. The human gene mutation database. Nucleic Acids Res 26(1):285-7.

Cooper GM, Shendure J. 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 12(9):628-40.

Coordinators NR. 2014. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 42(Database issue):D7-17.

de Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, de Vries P, Gilissen C and others. 2012. Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med 367(20):1921-9.

Ellison JW, Rosenfeld JA, Shaffer LG. 2013. Genetic basis of intellectual disability. Annu Rev Med 64:441-50.

Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT. 2011. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat 32(5):557-63.

Fujita M, Kinoshita T. 2012. GPI-anchor remodeling: potential functions of GPI-anchors in intracellular trafficking and membrane dynamics. Biochim Biophys Acta 1821(8):1050-8.

Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. Nature 467(7319):1061-73.

Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56-65.

Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BW, Willemsen MH, Kwint M, Janssen IM, Hoischen A, Schenck A and others. 2014. Genome sequencing identifies major causes of severe intellectual disability. Nature 511(7509):344-7.

Hansen L, Tawamie H, Murakami Y, Mang Y, ur Rehman S, Buchert R, Schaffer S, Muhammad S, Bak M, Nothen MM and others. 2013. Hypomorphic mutations in PGAP2, encoding a GPI-anchor-remodeling protein, cause autosomal-recessive intellectual disability. Am J Hum Genet 92(4):575-83.

Hochsmann B, Dohna-Schwake C, Kyrieleis HA, Pannicke U, Schrezenmeier H. 2014. Targeted therapy with eculizumab for inherited CD59 deficiency. N Engl J Med 370(1):90-2.

Horn D, Krawitz P, Mannhardt A, Korenke GC, Meinecke P. 2011. Hyperphosphatasia-mental retardation syndrome due to PIGV mutations: expanded clinical spectrum. American journal of medical genetics. Part A 155A(8):1917-22.

Horn D, Schottmann G, Meinecke P. 2010. Hyperphosphatasia with mental retardation, brachytelephalangy, and a distinct facial gestalt: Delineation of a recognizable syndrome. European journal of medical genetics 53(2):85-8.

Horn D, Wieczorek D, Metcalfe K, Baric I, Palezac L, Cuk M, Petkovic Ramadza D, Kruger U, Demuth S, Heinritz W and others. 2014. Delineation of PIGV mutation spectrum and associated

phenotypes in hyperphosphatasia with mental retardation syndrome. Eur J Hum Genet 22(6):762-7.

Howard MF, Murakami Y, Pagnamenta AT, Daumer-Haas C, Fischer B, Hecht J, Keays DA, Knight SJ, Kolsch U, Kruger U and others. 2014. Mutations in PGAP3 Impair GPI-Anchor Maturation, Causing a Subtype of Hyperphosphatasia with Mental Retardation. Am J Hum Genet.

Ikezawa H. 2002. Glycosylphosphatidylinositol (GPI)-anchored proteins. Biol Pharm Bull 25(4):409-17.

International Human Genome Sequencing C. 2004. Finishing the euchromatic sequence of the human genome. Nature 431(7011):931-45.

Jager M, Wang K, Bauer S, Smedley D, Krawitz P, Robinson PN. 2014. Jannovar: a java library for exome annotation. Hum Mutat 35(5):548-55.

Johnston JJ, Gropman AL, Sapp JC, Teer JK, Martin JM, Liu CF, Yuan X, Ye Z, Cheng L, Brodsky RA and others. 2012. The phenotype of a germline mutation in PIGA: the gene somatically mutated in paroxysmal nocturnal hemoglobinuria. Am J Hum Genet 90(2):295-300.

Kato M, Saitsu H, Murakami Y, Kikuchi K, Watanabe S, Iai M, Miya K, Matsuura R, Takayama R, Ohba C and others. 2014. PIGA mutations cause early-onset epileptic encephalopathies and distinctive features. Neurology 82(18):1587-96.

Kinoshita T, Fujita M, Maeda Y. 2008. Biosynthesis, remodelling and functions of mammalian GPI-anchored proteins: recent progress. Journal of biochemistry 144(3):287-94.

Krawitz PM, Murakami Y, Hecht J, Kruger U, Holder SE, Mortier GR, Delle Chiaie B, De Baere E, Thompson MD, Roscioli T and others. 2012. Mutations in PIGO, a member of the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation. Am J Hum Genet 91(1):146-51.

Krawitz PM, Murakami Y, Riess A, Hietala M, Kruger U, Zhu N, Kinoshita T, Mundlos S, Hecht J, Robinson PN and others. 2013. PGAP2 mutations, affecting the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation syndrome. Am J Hum Genet 92(4):584-9.

Krawitz PM, Schweiger MR, Rodelsperger C, Marcelis C, Kolsch U, Meisel C, Stephani F, Kinoshita T, Murakami Y, Bauer S and others. 2010. Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nature genetics 42(10):827-9.

Kvarnung M, Nilsson D, Lindstrand A, Korenke GC, Chiang SC, Blennow E, Bergmann M, Stodberg T, Makitie O, Anderlid BM and others. 2013. A novel intellectual disability syndrome caused by GPI anchor deficiency due to homozygous mutations in PIGT. J Med Genet.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42(Database issue):D980-5.

Mabry CC, Bautista A, Kirk RF, Dubilier LD, Braunstein H, Koepke JA. 1970. Familial hyperphosphatase with mental retardation, seizures, and neurologic deficits. The Journal of pediatrics 77(1):74-85.

Maeda Y, Tashima Y, Houjou T, Fujita M, Yoko-o T, Jigami Y, Taguchi R, Kinoshita T. 2007. Fatty acid remodeling of GPI-anchored proteins is required for their raft association. Mol Biol Cell 18(4):1497-506.

Mahoney JF, Urakaze M, Hall S, DeGasperi R, Chang HM, Sugiyama E, Warren CD, Borowitz M, Nicholson-Weller A, Rosse WF and others. 1992. Defective glycosylphosphatidylinositol anchor synthesis in paroxysmal nocturnal hemoglobinuria granulocytes. Blood 79(6):1400-3.

Mardis ER. 2008. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387-402.

Mardis ER. 2013. Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto Calif) 6:287-303.

Martin HC, Kim GE, Pagnamenta AT, Murakami Y, Carvill GL, Meyer E, Copley RR, Rimmer A, Barcia G, Fleming MR and others. 2014. Clinical whole-genome sequencing in severe early-onset

epilepsy reveals new genes and improves molecular diagnosis. Hum Mol Genet 23(12):3200-11.

Maydan G, Noyman I, Har-Zahav A, Neriah ZB, Pasmanik-Chor M, Yeheskel A, Albin-Kaplanski A, Maya I, Magal N, Birk E and others. 2011. Multiple congenital anomalies-hypotonia-seizures syndrome is caused by a mutation in PIGN. Journal of medical genetics 48(6):383-9.

Murakami Y, Kanzawa N, Saito K, Krawitz PM, Mundlos S, Robinson PN, Karadimitris A, Maeda Y, Kinoshita T. 2012. Mechanism for release of alkaline phosphatase caused by glycosylphosphatidylinositol deficiency in patients with hyperphosphatasia-mental retardation syndrome. The Journal of biological chemistry.

Murakami Y, Tawamie H, Maeda Y, Buttner C, Buchert R, Radwan F, Schaffer S, Sticht H, Aigner M, Reis A and others. 2014. Null mutation in PGAP1 impairing Gpi-anchor maturation in patients with intellectual disability and encephalopathy. PLoS Genet 10(5):e1004320.

Musante L, Ropers HH. 2014. Genetics of recessive cognitive disorders. Trends Genet 30(1):32-9.

Najmabadi H, Hu H, Garshasbi M, Zemojtel T, Abedini SS, Chen W, Hosseini M, Behjati F, Haas S, Jamali P and others. 2011. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature 478(7367):57-63.

Nakamura K, Osaka H, Murakami Y, Anzai R, Nishiyama K, Kodera H, Nakashima M, Tsurusaki Y, Miyake N, Kinoshita T and others. 2014. PIGO mutations in intractable epilepsy and severe developmental delay with mild elevation of alkaline phosphatase levels. Epilepsia 55(2):e13-7.

Nakashima M, Kashii H, Murakami Y, Kato M, Tsurusaki Y, Miyake N, Kubota M, Kinoshita T, Saitsu H, Matsumoto N. 2014. Novel compound heterozygous PIGT mutations caused multiple congenital anomalies-hypotonia-seizures syndrome 3. Neurogenetics 15(3):193-200.

Ohba C, Okamoto N, Murakami Y, Suzuki Y, Tsurusaki Y, Nakashima M, Miyake N, Tanaka F, Kinoshita T, Matsumoto N and others. 2014. PIGN mutations cause congenital anomalies, developmental delay, hypotonia, epilepsy, and progressive cerebellar atrophy. Neurogenetics 15(2):85-92.

Orlean P, Menon AK. 2007. Thematic review series: lipid posttranslational modifications. GPI anchoring of protein in yeast and mammalian cells, or: how we learned to stop worrying and love glycophospholipids. J Lipid Res 48(5):993-1011.

Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, Albrecht B, Bartholdi D, Beygo J, Di Donato N and others. 2012. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet 380(9854):1674-82.

Robinson PN, Kohler S, Bauer S, Seelow D, Horn D, Mundlos S. 2008. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 83(5):610-5.

Robinson PN, Kohler S, Oellrich A, Sanger Mouse Genetics P, Wang K, Mungall CJ, Lewis SE, Washington N, Bauer S, Seelow D and others. 2014. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res 24(2):340-8.

Rodelsperger C, Krawitz P, Bauer S, Hecht J, Bigham AW, Bamshad M, de Condor BJ, Schweiger MR, Robinson PN. 2011. Identity-by-descent filtering of exome sequence data for disease-gene identification in autosomal recessive disorders. Bioinformatics 27(6):829-36.

Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 7(8):575-6.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nat Biotechnol 26(10):1135-45.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1):308-11.

Strübing P. 1882. Paroxysmale Hämoglobinurie. Dscht Med Wochenschr 8:17-21.

Takeda J, Miyata T, Kawagoe K, Iida Y, Endo Y, Fujita T, Takahashi M, Kitani T, Kinoshita T. 1993. Deficiency of the GPI anchor caused by a somatic mutation of the PIG-A gene in paroxysmal nocturnal hemoglobinuria. Cell 73(4):703-11.

Tarutani M, Itami S, Okabe M, Ikawa M, Tezuka T, Yoshikawa K, Kinoshita T, Takeda J. 1997. Tissue-specific knockout of the mouse Pig-a gene reveals important roles for GPI-anchored proteins in skin development. Proc Natl Acad Sci U S A 94(14):7400-5.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G and others. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337(6090):64-9.

Vogel FMAG. 1997. Human genetics : problems and approaches. Berlin; New York: Springer.

Xiong ZQ, Qian W, Suzuki K, McNamara JO. 2003. Formation of complement membrane attack complex in mammalian cerebral cortex evokes seizures and neurodegeneration. J Neurosci 23(3):955-60.

Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, Mort M, Phillips AD, Shaw K, Stenson PD, Cooper DN and others. 2012. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. Am J Hum Genet 91(6):1022-32.

Zemojtel T, Kohler S, Mackenroth L, Jager M, Hecht J, Krawitz P, Graul-Neumann L, Doelken S, Ehmke N, Spielmann M and others. 2014. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci Transl Med 6(252):252ra123.

Zunich J, Kaye CI. 1983. New syndrome of congenital ichthyosis with neurologic abnormalities. Am J Med Genet 15(2):331-3, 335.

Danksagung

Ich möchte mich bei Herrn Prof. Dr. Stefan Mundlos bedanken, der mir die Freiheit ließ, neben den täglichen Aufgaben die Arbeitsthemen selbst zu wählen und der mich auf meinem wissenschaftlichen Weg stets kritisch begleitete und doch motivierend unterstützte.

Herzlich danken möchte ich Frau Prof. Dr. Denise Horn, die mich in die klinische Genetik einführte und die durch die Auswahl eines Falles mit Mabry Syndrom den glücklichen Zufall begünstigte, durch den GPI-Ankerstörungen zu meinem Forschungsthema wurden.

Ich möchte mich ausdrücklich bei Herrn Prof. Dr. Peter Robinson bedanken, der mir die Vielfalt bioinformatischer Forschung aufzeigte und durch wissenschaftliche Diskussionen oftmals meine Neugierde weckte.

Danken möchte ich auch Dr. Jochen Hecht. Mit seiner Expertise im Naßlabor und seinem unermüdlichen Arbeitseinsatz ließ er die NGS Datensätze entstehen, in denen ich nach pathogenen Mutationen suchen durfte.

Ein herzlicher Dank gilt Dr. Tom Kamphans, der entscheidend dazu beitrug, die Werkzeuge zu schaffen, die einige der wissenschaftlichen Entdeckungen erst ermöglichten. Zudem war es eine große Freude neben der deduktiven wissenschaftlichen Tätigkeit durch ihn auch konstruktiv tätig geworden zu sein.

Ein besonderer Dank gilt auch den Mitgliedern meiner Gruppe, Verena Heinrich, Na Zhu und Alexej Knaus, die durch ihre Eigeninitiative und Kreativität nicht nur inhaltlich zu den Ergebnissen beitrugen, sondern mich stets durch die gemeinschaftliche Arbeit motivierten.

Ich möchte mich auch bei den technischen Mitarbeitern insbesondere Ulrike Krüger, deren Sachkenntnis und technische Expertise unentbehrlich für meine Arbeit waren, bedanken.

Schließlich möchte ich mich auch bei den vielen Kollegen bedanken, die mit zum Erfolg meiner Arbeit beitrugen, insbesondere Dr. Christian Rödelsperger, Prof. Dr. Michal Schweiger, Dr. Sebastian Köhler, Dr. Sebastian Bauer, Dr. Marten Jäger, Dr. Dmitri Parkhomchuk, Prof. Dr. Eva Klopocki, Dr. Nadja Ehmke und Dr. Malte Spielmann.

Einen herzlichen Dank möchte ich auch an Prof. Dr. Karl Sperling richten, der mir als Mentor zur Seite stand und der mich lehrte durch Gelassenheit nicht an den alltäglichen Widrigkeiten zu verzagen.

Durch die Forschungsförderung der Charité und durch die Deutsche Forschungsgemeinschaft standen mir finanzielle Mittel zur Verfügung, die zur Verwirklichung der Projekte unerlässlich waren. Auch hierfür möchte ich mich bedanken.

Bedanken möchte ich mich auch bei den Patienten und deren Angehörigen, die wohlwissentlich eines nicht unmittelbaren, eigenen Nutzens meine Arbeit unterstützten.

Zuletzt möchte ich mich bei meinen Eltern bedanken, die mich in all meinen Vorhaben immer unterstützten und für mich da waren.

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