Bundesgesundheitsbl 2017 · 60:542–549
DOI 10.1007/s00103-017-2538-5
Online publiziert: 14. März 2017

Sebastian Köhler[1] · **Peter N. Robinson**[2,3]

[1] NeuroCure Cluster of Excellence, Charité-Universitätsmedizin Berlin, Berlin, Deutschland
[2] The Jackson Laboratory for Genomic Medicine, Farmington, USA
[3] Institute for Systems Genomics, University of Connecticut, Farmington, USA

Genetische Diagnostik seltener Erkrankungen

Integration von Phänotyp- und Genomdaten

Einleitung

Menschen mit seltenen Erkrankungen warten 5 bis 30 Jahre auf eine Diagnose und müssen sich während dieser Zeit typischerweise bei drei oder mehr Ärzten vorstellen. Dabei ist die initiale Verdachtsdiagnose in mindestens 40 % der Fälle falsch [1, 2]. Der Einsatz von Next-Generation-Sequencing (NGS) in der Diagnostik und Forschung ist seit der Einführung der NGS vor etwas über einem Jahrzehnt rasant gestiegen. Forschungsprogramme wie das „100.000 Genomes Project" in England, das „Undiagnosed Diseases Network der National Institutes of Health" (NIH) in den USA, das „Kids First Pediatric Research Program" der NIH sowie chinesische, französische amerikanische und britische Programme für genomische und Präzisionsmedizin lassen erwarten, dass über eine Million Genome von Patienten mit Verdacht auf genetisch bedingte seltene Erkrankungen bis zum Ende dieses Jahrzehnts in verschiedenen Forschungsprogrammen sequenziert werden [3–6]. Im diagnostischen Kontext kommen je nach Fragestellung verschiedene NGS-basierte Untersuchungen zum Einsatz. Die Gen-Panel-Diagnostik umfasst die gleichzeitige Sequenzierung von Genen, die jeweils mit einer bestimmten Erkrankungsgruppe assoziiert sind. Zum Beispiel würde ein Gen-Panel für hereditäre Formen der thorakalen Aortendissektion die Gene *ACTA2, COL3A1, FBN1, FLNA, GATA5, MAT2A, MFAP5, MYH11, MYLK, NOTCH1, PRKG1, SMAD3, TGFB2, TGFB3, TGFBR1* und *TGFBR2* beinhalten, weil Mutationen in diesen Genen bei Personen mit hereditären Erkrankungen der Aorta beobachtet werden können [7]. Die Gen-Panel-Diagnostik kann somit die stufenweise Diagnostik durch Sanger-Sequenzierung der einzelnen Gene bei verbesserter Sensitivität und geringeren Kosten ersetzen [8].

Bei der Exomsequenzierung (whole-exome sequencing, WES) werden (nahezu) alle proteinkodierenden Sequenzen angereichert und deren Sequenz ermittelt. Mit der WES konnte eine eindrückliche Anzahl von bislang unbekannten Krankheitsgenen identifiziert werden. Darüber hinaus wird sie in vielen Ländern (z. B. USA, Niederlanden, Schweiz, Kanada) zunehmend bei Kindern und Erwachsenen mit nicht diagnostizierbaren Erkrankungen als diagnostisches Mittel eingesetzt. Bei der Genomsequenzierung (whole-genome sequencing, WGS) werden Sequenzabschnitte (Reads) des gesamten Genoms ermittelt, d. h., auch die der nichtkodierenden Bereiche. Die WGS weist gegenüber der WES einige technische Vorteile auf, darunter eine gleichmäßigere Abdeckung der kodierenden Bereiche als bei der WES.

Im Mittel findet man bei der WES 25.000–100.000 Varianten (je nach Sequenziertiefe und Anreicherungsmethode) und bei der WGS bis über 5 Mio.

Trotz der großen Fortschritte auf dem Gebiet kann die diagnostische Ausbeute der NGS-basierten Diagnostik enttäuschend gering ausfallen. Je unklarer die Diagnose ist, desto schwieriger gestaltet sich die Suche nach den kausalen Mutationen unter all diesen Varianten (nicht ohne Grund wird die bioinformatische Analyse von WES-/WGS-Daten oft mit der Suche nach der Nadel im Heuhaufen verglichen). In der Tat wurde auch festgestellt, dass gerade in Studien mit großen Kohorten von Patienten mit Verdacht auf eine Mendel'sche Erkrankung die diagnostische Aufklärungsrate von WES erst 11 bis 25 % [9–12] betrug. Aus diesem Grund werden zahlreiche bioinformatische Ansätze entwickelt mit dem Ziel, die diagnostische Ausbeute der WES/WGS zu erhöhen.

In dieser Übersicht werden wir nach einer kurzen Zusammenfassung über die bioinformatische Analyse von NGS-Daten für die Diagnostik und im humangenetischen Forschungskontext einen neuen Ansatz vorstellen, der klinische (phänotypische) Daten in die Analyse von WES/WGS Daten mit einbezieht. Mittlerweile wurde der ursprünglich von den Autoren an der Charité Berlin entwickelte phänotypbasierte Ansatz von zahlreichen Gruppen aufgegriffen und in vielen neuen Kontexten eingesetzt.

Die bioinformatische Analyse der bei der WES/WGS-Diagnostik gefundenen Varianten

Die bioinformatische Analyse von NGS-Daten umfasst zahlreiche Schritte, auf die wir hier nicht im Detail eingehen

Die Autoren S. Köhler und P.N. Robinson sind gleichberechtigte Erstautoren.

Tab. 1 Anatomie eines HPO-Terms. Die Tabelle zeigt die wichtigsten Attribute des HPO-Terms *Atrial septal defect*

Element	Beispiel	Erklärung
id	HP:0001631	Eintragungsnummer für diesen Term
Name	Atrial septal defect	Bevorzugte Bezeichnung
Synonym	ASD; Atrial septum defect	Synonyme
Definition	Atrial septal defect (ASD) is a congenital abnormality of the interatrial septum that enables blood flow between the left and right atria via the interatrial septum	Begriffsbestimmung
Xref	ICD-10:Q21.1	Querverweis auf synonymen Begriff in einem anderen Datenbank
is_a	Abnormality of cardiac atrium (HP:0005120); Abnormality of the atrial septum (HP:0011994)	Einer oder mehrere „Elternterme", d. h., allgemeinere Terme, die direkt oberhalb dieses Terms in der Hierarchie stehen
Logische Definition	„has part" some („closure incomplete" and [„inheres in" some „interatrial septum"] and [„has modifier" some abnormal])	Computer-lesbare Begriffsbestimmung des Terms mittels eines OWL-Klassenausdrucks mit Verweisen auf andere Ontologien

können. Neben der Qualitätskontrolle umfassen die wichtigsten ersten Schritte das Alignieren der Sequenzreads gegen das Referenzgenom und die Bestimmung von Varianten anhand der Basenzusammenstellung der Spalten des Alignments. Dieser Prozess ist mit vielen Fallstricken verbunden, die zu falsch-positiven oder falsch-negativen Ergebnissen führen können. Die Gründe dafür sind vielfältig. Das so genannte „Calling" (also die Bestimmung einer Variante durch Analyse der alignierten Sequenzreads der WES bzw. WGS) von Indels (wenige Basen betreffende Insertionen oder Deletionen) ist technisch schwierig, genauso wie das Einschätzen von in schlecht abgedeckten Regionen gelegenen Varianten (wenige Reads in Region aligniert). Außerdem ist es weiterhin ein Problem, dass verschiedene Zusammenstellungen verschiedener informatischer Analyseschritte zu unterschiedlichen Ergebnissen der gefunden Varianten führen [13].

Ein typisches Exom enthält 30.000 bis 100.000 Variationen [14]. Die Analyse der zehntausenden „gecallten" Varianten eines Exoms beginnt normalerweise mit einem Filtrationsprozess, wobei Varianten von der weiteren Analyse ausgeschlossen werden, die nicht in den angereicherten Regionen liegen. Im Allgemeinen werden in der Bevölkerung häufige Varianten herausgefiltert unter der Annahme, dass eine häufige Variante für eine seltene Erkrankung nicht ursächlich sein könne. Zum Beispiel können Varianten von der weiteren Analyse ausgeschlossen werden, welche eine Häufigkeit (*minor allele frequency*, MAF) von 1 % oder mehr in den Daten des 1000-Genomes-Projekts aufweisen.

Ein weiterer Schritt besteht dann darin, die verbleibenden Varianten nach ihrer vorhergesagten Pathogenität zu sortieren. Dafür ist es unabdinglich, die Varianten den entsprechenden Genen bzw. Transkripten zuzuordnen. Zum Beispiel muss die Information „chr10:g.123256215 T > G" (chromosomal) in eine Gen-basierte Form übertragen werden, also c.518A>C; *p*.Glu173Ala im Gen *FGFR2*. Diese Übertragung ist wichtig, da die klinische Auswertung von genomischen Varianten sich zumeist auf die translatierten Bereiche der Gene beschränkt [15].

Nach diesem Filterschritt bleiben typischerweise bis zu etwa 100 Missense-Varianten übrig, deren Krankheitsrelevanz (Pathogenizität) mit bioinformatischen Verfahren wie *Polymorphism Phenotyping* (PolyPhen) [16] oder *MutationTaster* [17] bewertet werden kann. Stehen Exomproben einer Kernfamilie zur Verfügung, ist es möglich, mit Applikationen wie Jannovar die Varianten je nach Vererbungsmodus zu filtern [15]. Auch nach diesen Filterschritten kann die Liste der verbleibenden, vermutet pathogenen Varianten immer noch hunderte (bei kodierenden Varianten) bis tausende (bei nicht kodierenden Varianten) Einträge enthalten [18]. Falls eine Patientenkohorte zur Verfügung steht, können mehrere betroffene Individuen sequenziert und diejenigen Gene identifiziert werden, die bei allen (oder den meisten) Individuen Varianten aufweisen. Für Familienstudien kann auch eine Kopplungsanalyse angewendet werden [19]. Bei diesen Methoden werden Gene oder genomische Intervalle als Kandidaten herausgefiltert.

Mit Ausnahme von umfassend untersuchten Varianten, die mit gut verstandenen Krankheiten assoziiert sind, bleibt es ein extrem schwieriges Unterfangen einzuschätzen, ob eine gefundene Variante eine Rolle bei der Entstehung einer gegebenen Erkrankung spielt oder nicht [20]. Zum Beispiel wurde kürzlich *TRIM63* als neues Krankheitsgen für die hypertrophe Kardiomyopathie identifiziert. Es wurden zwei Missense-Varianten und eine Deletion in 302 Patienten gefunden, die bei keiner der 229 Kontrollpersonen vorkamen [21]. Es gab allerdings auch Studien, die gezeigt haben, dass Personen mit Nonsense-Variationen in *TRIM63* keinerlei Anzeichen von Kardiomyopathie hatten [22]. Es bleibt daher bis heute unbekannt, ob Mutationen in *TRIM63* eine kausale Rolle bei der Entstehung der hypertrophen Kardiomyopathie besitzen. Existierende Datenbanken zu speziellen Genen und Krankheiten unterscheiden sich sehr stark in Parametern wie Genauigkeit klinischer Information und Abdeckung veröffentlichter und nicht veröffentlichter pathogener Varianten. Des Weiteren können sie auch einen substanziellen Anteil fehlerhafter Information enthalten [23]. Eine 2011 publizierte Untersuchung legte zum Beispiel nahe, dass bis zu 27 % der als pathogen publizierten Varianten in Wirklichkeit keine Mutationen sind, sondern stattdessen häufig auftretende Polymorphismen sind, aus Sequenzierfehlern resultieren oder bei denen einfach nicht genügend Belege für ein klare Pathogenität der Varianten vorliegen [24]. Um schnellere und genauere Diagnosen zu ermög-

lichen, muss daher erreicht werden, dass Daten (und Annotationen) in solchen Genotyp-Phänotyp-Datenbanken maximale Verlässlichkeit und Zugänglichkeit haben [20].

Genpriorisierung gegen die Datenflut

Es wurden eine Vielzahl von informatischen Ansätzen entwickelt, um die Liste der gefundenen Varianten in einem Genom oder Exom nach ihrer Relevanz für die untersuchte Krankheit zu sortieren [25]. Dieser Vorgang heißt in der Literatur Genpriorisierung (gene prioritization) und stellt ein aktives Forschungsgebiet in der Bioinformatik dar. Die Priorisierungsalgorithmen filtern oder sortieren Kandidatengene nach verschiedenen Kriterien, um sie in die „besten" Plätze der Liste einzuordnen, sodass Forscher oder Kliniker, die in der Regel nicht die Zeit haben, um mehrere hundert Kandidatengene genau zu überprüfen, den „besten" Kandidaten mehr Zeit und Aufmerksamkeit widmen können, was letztlich die Wahrscheinlichkeit erhöht, das richtige Gen zu finden. Bei diversen Priorisierungsansätzen werden die Strukturen und Verbindungen von Protein-Protein-Interaktionsnetzwerken [26, 27] analysiert. Hierbei wird für alle Gene eine Wahrscheinlichkeit berechnet, mit der diese mit dem Krankheitsbild assoziiert sind, wobei bereits bekannte Krankheitsgene als Vorwissen mit einfließen.

Es ist zu beachten, dass die oben beschriebenen informatischen Ansätze die klinischen Auffälligkeiten (Phänotypen) des Patienten nicht mit einbeziehen. Unsere Gruppe hat 2008 mit der Publikation der „Human Phenotype Ontology" (HPO), 2009 mit der Publikation des „Phenomizers" und 2014 mit der Publikation des „Exomisers" Ressourcen und Algorithmen zur Verfügung gestellt, um eine phänotypgetriebene Analyse von genomischen Daten zu ermöglichen, die wir in den folgenden Abschnitten erklären.

Zusammenfassung · Abstract

Bundesgesundheitsbl 2017 · 60:542–549 DOI 10.1007/s00103-017-2538-5

S. Köhler · P. N. Robinson

Genetische Diagnostik seltener Erkrankungen. Integration von Phänotyp- und Genomdaten

Zusammenfassung

Eine Herausforderung für die genomische personalisierte Medizin wird es sein, verlässliche Methoden zur Erfassung und Ähnlichkeitsberechnung von klinischen Phänotypen zu entwickeln, denn eine kombinierte Analyse der phänotypischen Merkmalen sowie der gerade bei Exom- oder Genomsequenzierung sehr zahlreichen genetischen Varianten kann die diagnostische Ausbeute erheblich steigern. Das „Human Phenotype Ontology-Projekt" (HPO-Projekt; www.human-phenotype-ontology.org) stellt eine Ontologie zur Erfassung phänotypischer Auffälligkeiten bereit, womit die klinischen Auffälligkeiten von Patienten und Erkrankungen präzise und umfassend erfasst werden können. Die HPO erlaubt nicht nur eine verlässliche Informationsintegration aus diversen Datenbanken, sondern auch die mathematische Ähnlichkeitsberechnung zwischen Patienten und/oder Krankheiten auf Basis der phänotypischen Profile. Somit bildet die HPO eine robuste Grundlage für differenzialdiagnostische Anwendungen sowie für translationale Forschung und Priorisierungen von bislang unbekannten Krankheitsgenen.

Schlüsselwörter

Ontologie · Phänotyp · Differentialdiagnose · Translationale Forschung · Exomsequenzierung

Diagnostics in human genetics. Integration of phenotypic and genomic data

Abstract

The development of reliable methods for annotation of clinical phenotypes and algorithms to calculate similarity values for clinical phenotype profiles will be a major challenge for genomic personalized medicine, since combined analysis of phenotypic features and genetic variants can increase diagnostic yield, especially with exome or genome sequencing. The Human Phenotype Ontology project (HPO; www.human-phenotype-ontology.org) provides an ontology for capturing phenotypic abnormalities in human disease in a precise and comprehensive fashion. The HPO not only enables reliable integration of disease-relevant information from numerous databases, but it also allows for similarity between patients or between patients and disease descriptions to be calculated algorithmically. The HPO thereby represents a solid foundation for differential diagnostic applications as well as for translational research and prioritization of novel disease genes in exome or genome sequencing projects.

Keywords

Ontology · Phenotype · Differential diagnosis · Translational research · Whole-exome sequencing

Phänotypische Analyse von WES/WGS-Daten

Das Wort Phänotyp kann diverse Bedeutungen haben. In der Biologie wird das Wort meist mit der „Menge aller Merkmale eines Organismus" gleich gesetzt. Diese Merkmale umfassen morphologische und physiologische Eigenschaften auf der Ebene von Zellen, Organen oder des gesamten Organismus. Außerdem schließt die Definition auch Verhaltenseigenschaften ein. Im medizinischen Kontext wird das Wort Phänotyp allerdings meist zur Beschreibung einer Abweichung von normaler Physiologie, Morphologie oder Verhaltensweise verwendet. Dieser Definition wird auch in diesem Artikel gefolgt.

Eine der wichtigsten Aufgaben eines Klinikers besteht darin, die Phänotypen eines Patienten aufzuzeichnen und so seine Diagnose zu bestimmen. Das geschieht durch das Stellen der Anamnese, die körperliche Untersuchung, bildgebende Verfahren, Blutanalyse und so weiter. An diesem Punkt wird die Medizin sehr naturwissenschaftlich: es werden

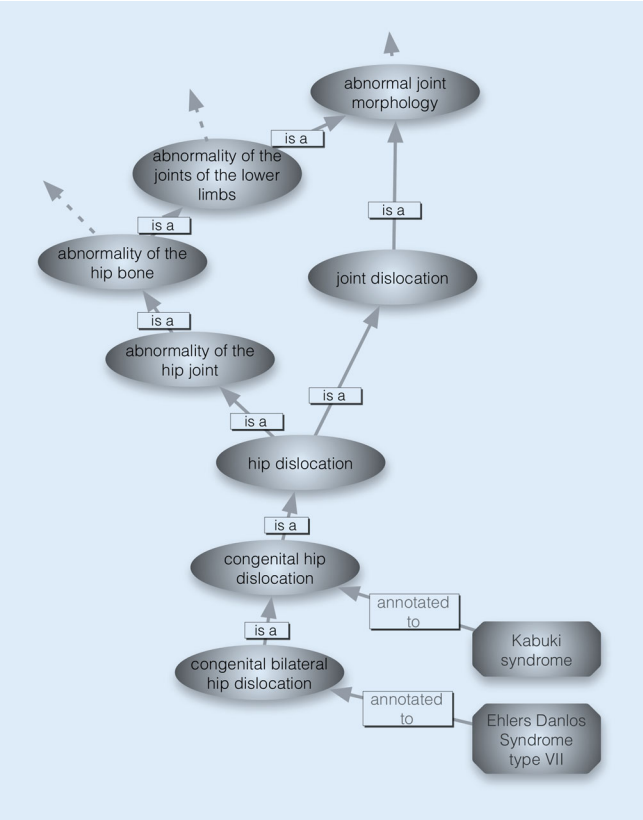

Abb. 1 ◀ Ausschnitt aus der HPO. Kreise stehen für Terme der HPO, wie z. B. „hip dislocation". HPO-Terme können verwendet werden, um Krankheiten oder Patienten phänotypisch zu beschreiben. Hier wird zum Beispiel die Krankheit „Kabuki syndrome" mit dem HPO Term „congenital hip dislocation" annotiert. Die HPO-Terme stehen miteinander in einer Subklassen-Beziehung

Phänotypen beobachtet, daraus wird eine Hypothese generiert („Diagnose"), und schließlich wird die Hypothese getestet, indem ein gewisses Behandlungsschema verschrieben wird.

Die Diagnose zu stellen, kann eine enorm komplexe Aufgabe sein, insbesondere bei seltenen Erkrankungen. Es gibt bereits über 8000 benannte seltene Erkrankungen, und es wird angenommen, dass es tausende weitere zu entdecken und zu klassifizieren gibt. Wie eingangs erwähnt erleben Menschen mit seltenen Erkrankungen oft eine „diagnostische Odyssee", bevor die richtige Diagnose gestellt wird. Auch wenn genaue Daten fehlen, ist anzunehmen, dass die Situation bei Patienten mit sehr seltenen Krankheiten (z. B. „ultrarare diseases" mit Prävalenz von 1:1.000.000 oder weniger [28]) noch schwieriger ist. Klinische Probleme, die mit einer verspäteten oder falschen Diagnose einhergehen, sind verspätete Behandlung, unnötige diagnostische Untersuchungen und psychische Belastungen aufgrund der Unklarheit über Grund und Prognose der Erkrankung.

Ein komplettes und detailliertes Verständnis der mit den jeweiligen Krankheiten assoziierten Phänotypen ist essenziell. um einzuschätzen, ob der gegebene Phänotyp eines Patienten mit der zugrundeliegenden Erkrankung wirklich in Verbindung steht oder nur ein isoliertes Ereignis ist. Das wiederum ist wichtig, um Behandlung und Prognose korrekt einzuschätzen. Das Wissen um alle möglichen mit einer Krankheit assoziierten Phänotypen ist wichtig, um Komplikationen zu verhindern, oder sie zumindest so früh wie möglich zu erkennen.

Die Bedeutung des *Deep Phenotyping* für die Diagnostik

Oftmals werden Phänotypen in der medizinischen Literatur unpräzise oder gar „schlampig" beschrieben. Beispielsweise sind Beschreibungen wie „im EMG myopathisches Bild" für die Differenzialdiagnostik nicht sehr aussagekräftig. Die Angabe der genauen Gründe für die Diagnose wäre unter Umständen nützlicher (z. B. verkürzte Dauer und Amplitude der Aktionspotenziale, positive scharfe Wellen, Verringerung der Anzahl der motorischen Einheiten des Muskels). Mit unpräzisen Beschreibungen des Phänotyps wird zudem der Vergleich verschiedener Studien erschwert.

In Mutationsdatenbanken ist die Situation ähnlich, da diese oft keine oder nur wenig Phänotypinformationen bereitstellen. Es wird meist nur der Fakt abgespeichert, dass eine gewisse Erkrankung in einer Person mit einer bestimmten Genvariante diagnostiziert wurde. Diese Information ist hilfreich für einen Diagnostiker, der einen Bericht über eine Genvariante schreiben will und in der Datenbank sieht, dass es bereits einen unabhängigen Fall mit der gleichen Genvariation gibt. Diese Information hilft allerdings weniger, wenn versucht wird, die Pathogenese der Erkrankung aufzuklären, das Spektrum der Phänotypen einer Erkrankung vollständig zu beschreiben oder genaue Genotyp-Phänotyp-Korrelationen zu erstellen.

Dagegen bedeutet *deep phenotyping* die präzise und umfassende Analyse phänotypischer Auffälligkeiten, bei der die individuellen Komponenten eines Phänotyps erfasst und beschrieben werden.

Die Human Phenotype Ontology

Die Human Phenotype Ontology (HPO, http://www.human-phenotype-ontology.org) ist eine Ontologie, die es ermöglicht, phänotypische Auffälligkeiten, die bei Erkrankungen eines Menschen auftreten können, zu erfassen. Die HPO stellt umfangreiche bioinformatische Ressourcen für die Analyse von menschlichen Erkrankungen und Phänotypen zur Verfügung und bildet somit eine informatische „Brücke" zwischen der Genombiologie und der klinischen Medizin. Die HPO wurde erstmalig 2008 veröffentlicht [29] mit dem Ziel, die Integration von phänotypischen Informationen zwischen wissenschaftlichen und medizinischen Disziplinen sowie zwischen Datenbanken zu verbessern. Seitdem sind der Umfang und die Verbreitung der Projekts erheblich gewachsen [30].

Tab. 2 Eine Auswahl öffentlicher klinischer Datenbanken, die HPO nutzen, um Patientendaten für Projekte zur Identifizierung von Krankheitsgenen zu annotieren

Name	URL	Ref
PhenomeCentral	www.Phenomecentral.org	[6]
DDD (Deciphering Developmental Disorders)	www.ddduk.org	[38, 39]
DECIPHER (DatabasE of genomiC varIation and Phenotype in Humans using Ensemble Resources)	www.Decipher.sanger.ac.uk	[40]
ECARUCA (European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations)	http://umcecaruca01.extern.umcn.nl:8080/ecaruca/ecaruca.jsp	[41]
The 100,000 Genomes Project	https://www.genomicsengland.co.uk/	[3]
Geno2MP (Exome sequencing data linked to phenotypic information from a wide variety of Mendelian gene discovery projects)	http://geno2mp.gs.washington.edu	[42]
NIH UDP (Undiagnosed Diseases Program)	Available via www.phenomecentral.org	[37]
NIH UDN (Undiagnosed Diseases Network)	Available via www.phenomecentral.org	[6]
HDG (Human Disease Gene Website series)	www.humandiseasegenes.com	–
Phenopolis (aggregation and harmonization of sequencing and phenotype data using HPO)	https://github.com/pontikos/phenopolis.github.io	–
GenomeConnect	www.genomeconnect.org	[43]
FORGE Canada & Care4Rare Consortium	www.care4rare.ca	[44]
RD-Connect	www.Platform.rd-connect.eu	[45]
Genesis	www.thegenesisprojectfoundation.org	–

Ontologie

Die HPO ist in vier unabhängige Hierarchien (*subontologies*) unterteilt, welche die Kategorien *Phänotypische Abnormalität, Vererbungsmodus, Sterblichkeit/Altern* und *Modifikator* abdecken. Sie enthält derzeit 11.813 Terme, deren semantische Beziehungen (15.595) zueinander und 14.328 Synonyme. 8627 Freitextdefinitionen beschreiben genau, was der entsprechende Term meint. Zusätzlich gibt es 5717 auf logischen Ausdrücken basierende „maschinenverständliche" Definitionen, die automatisches Schlussfolgern ermöglichen. Zum Beispiel kann ein Algorithmus auf Grundlage der logischen Definition eines Terms bestimmen, welche Vorfahren (d. h., allgemeinere Terme) dieser Term haben sollte. Jeder HPO-Term stellt also eine für menschliche Leser verständliche, aber auch eine vom Computer verwendbare Definition einer phänotypischen Abnormalität dar, mit Attributen wie Name, Definition, Lage in der ontologischen Hierarchie, synonyme Bezeichnungen, und Querverweise auf andere Termininologien (**Tab. 1**).

Annotationen

Eine Annotation meint in diesem Artikel die Verknüpfung eines HPO-Terms mit einer Erkrankung. Das heißt, dass Annotationen versuchen, das phänotypische Spektrum von Patienten mit einer bestimmten Erkrankung als eine Menge von HPO-Termen zu beschreiben (**Abb. 1**). Dementsprechend stellt die HPO derzeit 123.724 Annotationen von HPO-Termen mit seltenen Erkrankung aus Orphanet, DECIPHER und OMIM bereit [30].

Der initiale Fokus der HPO lag auf genetisch bedingten Erkrankungen, allerdings wird die Ontologie seit einigen Jahren massiv für komplexe und Volkskrankheiten erweitert. Deshalb wurden mittels Textanalyse von publizierten Artikeln in PubMed auch Annotationen für häufige Krankheiten („*common disease*", Volkskrankheiten) erzeugt. Es werden 132.620 HPO-Annotationen für 3145 Volkskrankheiten bereitgestellt [31].

Präzise Annotationen von „deep phenotyping"-Daten

Computerbasierte Suchalgorithmen, die auf Basis von Phänotypdaten (HPO) arbeiten, können deutlich besser funktionieren, wenn ein umfassendes phänotypisches Profil (deep phenotyping) bereitgestellt wird. Die Person, die ein solches HPO-Profil erstellt, sollte sich daher vor Augen halten, dass das Profil gegen alle bekannten HPO-Profile (z. B. Marfan-Profil, Williams-Beuren-Profil, etc.) verglichen wird. Daher ist es essenziell, die auffälligsten und wichtigsten Phänotypen mit größtmöglicher Spezifität anzugeben.

Die Monarch-Initiative, eine offene Kollaboration zur Zusammenführung von Genotyp- und Phänotypdaten, hat daher ein Werkzeug entwickelt, welches Breite und Tiefe eines gegebenen HPO-Profils mit einem 5-Sterne System bewertet. Das soll dem Benutzer helfen, ein Gefühl dafür zu bekommen, ob das bisher definierte HPO-Profil bereits genau genug ist, um ähnliche (aber nicht die gesuchten) Erkrankungen auszuschließen bzw. phänotypisch passende Modellorganismen zu identifizieren.

Integration der HPO

Die HPO unterscheidet sich von anderen klinischen Terminologien und Ontologien wie SNOMED, deren Zweckbestimmung in erster Linie die Unterstützung von Krankenhaus-IT-Systemen, Abrechnung und Verwaltung umfasst, dadurch, dass die HPO als bioinformatisches Werkzeug für die Differenzialdiagnose und die translationale Forschung entwi-

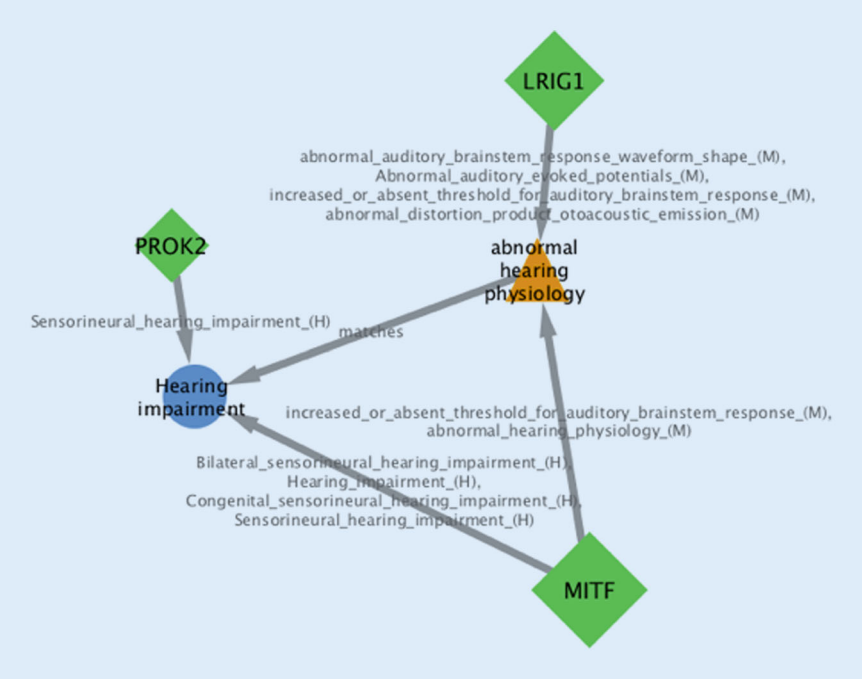

Abb. 2 ▲ Darstellung eines „phenograms" in der Software PhenogramViz. In diesem Beispiel wird ein Patient mit dem phänotypischen Merkmal *Hearing impairment* (*blauer Kreis*) gezeigt, in dessen Genom mehrere Gene (*grüne Vierecke*) verändert sind (z. B. eine Kopienzahlvariante, CNV). Ein Pfeil bedeutet, dass ein verändertes Gen zur Erklärung des Phänotyps herangezogen werden kann. Die Beschriftung des Pfeils zeigt den Grund für die Verbindung an, wobei (*H*) für Human, (*M*) für Maus und (*Z*) für Zebrafisch (Zebrabärbling) steht. Um die Grafik zu erzeugen, werden automatisch bekannte Phänotypassoziationen des Gens aus humanen, Maus- und Zebrafisch-Daten integriert und analysiert. Es ist dadurch möglich, die bekannten Phänotypen der Gene mit den Patientenphänotypen vollautomatisch zu vergleichen und visualisieren. Zum Beispiel ist bekannt, dass ein Ausschalten des Gens *Lrig1* in der Maus den Phänotypen „*abnormal auditory evoked potentials*" verursacht. Dies kann nun verwendet werden, um *LRIG1* mit dem „*Hearing impairment*" des Patienten in Verbindung zu bringen, über den gemeinsamen Vorfahren in der Phänotypontologie (*abnormal hearing physiology*). Auch die Gene *MITF* und *PROK2* können auf analoge Weise mit „*Hearing impairment*" in Verbindung gebracht werden

ckelt wurde. Die HPO ermöglicht die informatische Integration mit zahlreichen anderen bioinformatischen Ressourcen für die Forschung wie Gene Ontology und Phänotypdaten von zehntausenden genetisch modifizierten Maus- und Zebrafischmodellen. Die HPO bietet zudem eine extrem tiefe und breite Abdeckung phänotypischer Begriffe. Das ganze Unified Medical Language System (UMLS), das in sich zahlreiche andere Terminologien wie MeSH und MedDRA vereint, deckte nur 54 % der Begriffe in der HPO an; SNOMED CT nur 20 % [32]. Das UMLS, ein Terminologieintegrationssystem der US-amerikanischen National Library of Medicine [33] hat die HPO seit der 2015AB-Ausgabe komplett integriert. Die HPO wird von zahlreichen Datenbanken in der Humangenetik verwendet, um Patientendaten zu annotieren und phänotypbasierte Suchen zu ermöglichen. Viele dieser Datenbanken folgen dem Zweck, neue Krankheitsgene zu entdecken, und verwenden dabei die HPO-basierte Phänotypanalyse (**Tab. 2**). Es konnten bereits hierdurch mehrere bislang unbekannte Krankheitsgene identifiziert werden [34–37].

Phänotypgetriebene Analyse genomischer Variationen

Die HPO ist ein Werkzeug mit vielen Funktionalitäten, für diesen Artikel sind zwei Anwendungen besonders wichtig. (1) Die HPO erlaubt es, zwei mit HPO kodierte Phänotypprofile mathematisch zu vergleichen und einen Ähnlichkeitswert (phänotypischer Relevanz-Score) zu berechnen. Dies ist vergleichbar mit der BLAST-Suche (Basic Local Alignment Search Tool) in einer Datenbank mit DNA-Sequenzen. (2) Über die Assoziation von Krankheiten sowohl mit HPO-Termen als auch mit Krankheitsgenen, ist es möglich, HPO-Profile für einzelne Gene zu erstellen. So kann z. B. das Gen *APC* mit dem Phänotyp „*small intestinal carcinoid*" verknüpft werden, da es eine Erkrankung gibt, die durch Mutationen in *APC* verursacht wird und diesen Phänotyp ausprägt.

In der phänotypgetriebenen Analyse von genomischen Varianten werden die Punkte (1) und (2) nun verwendet. Bioinformatische Applikationen wie PhenIX [46] oder Exomiser [47, 48] benutzen einerseits das oben beschriebene Verfahren. um den Varianten einen jeweiligen Variantenscore zuzuweisen, andererseits wird parallel ein Score für die phänotypische Relevanz berechnet. Zunächst werden dafür den Varianten HPO-Profile zugeordnet, unter Verwendung des HPO-Profils des entsprechenden Gens. Dann wird der phänotypische Relevanz-Score zwischen den Varianten und dem Patienten berechnet. Im letzten Schritt werden der Variantenscore und der phänotypische Relevanzscore kombiniert und somit solche Varianten als besonders wichtig erachtet, die bei beiden Scores einen hohen Wert erzielen. Es wurde gezeigt, dass genau diese Kombination besonders hilfreich bei der Identifikation krankheitsrelevanter genomischer Variationen ist.

Ein weiteres Beispiel sind Priorisierung und Interpretation von Kopienanzahl-Variationen (copy-number variation, CNV), bei denen oft multiple Gene betroffen sind. Die vorgestellten Ressourcen erlauben einen automatischen Abgleich der Patientenphänotypen mit den betroffenen Genen und die grafische Darstellung gefundener Hypothesen über Gen-zu Phänotypen-Beziehungen der gefundenen CNV und der Phänotypen (**Abb. 2**).

Ausblick

Der initiale Fokus der HPO betraf in erster Linie genetisch bedingte Erkrankungen, allerdings wird die Ontologie seit einigen Jahren massiv für komplexe und Volkskrankheiten erweitert [49]. Die HPO wird von einer zunehmenden Anzahl von Forschungsgruppen über-

nommen und kommt in zahlreichen klinischen Kontexten zur Anwendung. Die HPO hat sich als leistungsstarkes bioinformatisches Werkzeug erwiesen, um die Differenzialdiagnose und translationale Forschung zu unterstützen. Forschungsgruppen in China, Japan, Frankreich, Spanien, Portugal und Italien haben Übersetzungen der HPO angefertigt. Eine deutsche und niederländische Übersetzung befinden sich in Arbeit. Künftige Arbeit betrifft vor allem die Erweiterung der HPO für die Präzisionsmedizin, Krebs und Fehlbildungen mit nicht-mendelscher Vererbung.

Korrespondenzadresse

Dr. rer. nat. S. Köhler
NeuroCure Cluster of Excellence, Charité-Universitätsmedizin Berlin
Charitéplatz 1, 10117 Berlin, Deutschland
sebastian.koehler@charite.de

Danksagung. Die Arbeit der Autoren wird durch folgende Förderungen unterstützt: E-RARE 2015, Hipbi-RD (Harmonising phenomics information for a better interoperability in the RD field).

Einhaltung ethischer Richtlinien

Interessenkonflikt. S. Köhler und P.N. Robinson geben an, dass kein Interessenkonflikt besteht.

Dieser Beitrag beinhaltet keine von den Autoren durchgeführten Studien an Menschen oder Tieren.

Literatur

1. EURORDIS (2007) What is a rare disease? http://www.eurordis.org/sites/default/files/publications/Fact_Sheet_Eurordiscare2.pdf. Zugegriffen: 17. Sept. 2016
2. Molster C, Urwin D, Di Pietro L, Fookes M, Petrie D, van der Laan S et al (2016) Survey of healthcare experiences of Australian adults living with rare diseases. Orphanet J Rare Dis 11:30
3. Marx V (2015) The DNA of a nation. Nature 524:503–505
4. Cyranoski D (2016) China embraces precision medicine on a massive scale. Nature 529:9–10
5. Tifft CJ, Adams DR (2014) The National Institutes of Health undiagnosed diseases program. Curr Opin Pediatr 26:626–633
6. Buske OJ, Girdea M, Dumitriu S, Gallinger B, Hartley T, Trang H et al (2015) PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. Hum Mutat 36:931–940
7. Arslan-Kirchner M, Arbustini E, Boileau C, Charron P, Child AH, Collod-Beroud G et al (2016) Clinical utility gene card for: Hereditary thoracic aortic aneurysm and dissection including next-generation sequencing-based approaches. Eur J Hum Genet 24:e1–5
8. Proost D, Vandeweyer G, Meester JAN, Salemink S, Kempers M, Ingram C et al (2015) Performant mutation identification using targeted next-generation sequencing of 14 thoracic aortic aneurysm genes. Hum Mutat 36:808–814
9. Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T et al (2012) Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet 380:1674–1682
10. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA et al (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 369:1502–1511
11. Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y et al (2014) Molecular findings among patients referred for clinical whole-exome sequencing. JAMA 312:1870–1879
12. Zhu X, Petrovski S, Xie P, Ruzzo EK, Lu Y-F, McSweeney KM et al (2015) Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. Genet Med 17:774–781
13. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J et al (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 5:28
14. Cooper GM, Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 12:628–640
15. Jäger M, Wang K, Bauer S, Smedley D, Krawitz P, Robinson PN (2014) Jannovar: a java library for exome annotation. Hum Mutat 35:548–555
16. Adzhubei I, Jordan DM, Sunyaev SR (2013) Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet 7(Unit7):20. doi:10.1002/0471142905.hg0720s76
17. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D (2010) MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 7:575–576
18. Smedley D, Schubach M, Jacobsen JOB, Köhler S, Zemojtel T, Spielmann M et al (2016) A whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease. Am J Hum Genet 99:595–606
19. Smith KR, Bromhead CJ, Hildebrand MS, Shearer AE, Lockhart PJ, Najmabadi H et al (2011) Reducing the exome search space for mendelian diseases using genetic linkage analysis of exome genotypes. Genome Biol 12:R85
20. Brookes AJ, Robinson PN (2015) Human genotype-phenotype databases: aims, challenges and opportunities. Nat Rev Genet 16:702–715
21. Chen SN, Czernuszewicz G, Tan Y, Lombardi R, Jin J, Willerson JT et al (2012) Human molecular genetic and functional studies identify TRIM63, encoding Muscle RING Finger Protein 1, as a novel gene for human hypertrophic cardiomyopathy. Circ Res 111:907–919
22. Ploski R, Pollak A, Müller S, Franaszczyk M, Michalak E, Kosinska J et al (2014) Does p.Q247X in TRIM63 cause human hypertrophic cardiomyopathy? Circ Res 114:e2–5
23. Gout AM, ADPKD Gene Variant Consortium, Ravine D, Harris PC, Rossetti S, Peters D et al (2007) Analysis of published PKD1 gene sequence variants. Nat Genet 39:427–428
24. Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J et al (2011) Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med 3:65ra4
25. Moreau Y, Tranchevent L-C (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet 13:523–536
26. Köhler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 82:949–958
27. Smedley D, Köhler S, Czeschik JC, Amberger J, Bocchini C, Hamosh A et al (2014) Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. Bioinformatics 30:3215–3222
28. Richter T, Nestler-Parr S, Babela R, Khan ZM, Tesoro T, Molsen E et al (2015) Rare disease terminology and definitions – a systematic global review: report of the ISPOR rare disease special interest group. Value Health 18:906–914
29. Robinson PN, Köhler 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 83:610–615
30. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I et al (2014) The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res 42:D966–D974
31. Groza T, Köhler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T et al (2015) The Human Phenotype Ontology: semantic unification of common and rare disease. Am J Hum Genet 97:111–124
32. Rainer W, Bodenreider O (2014) Coverage of phenotypes in standard terminologies. In: Proceedings of the Joint BioOntologies and BioLINK ISMB'2014 SIG session "Phenotype Day.", S 41–44
33. Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 32:D267–D270
34. Turro E, Greene D, Wijgaerts A, Thys C, Lentaigne C, Bariana TK et al (2016) A dominant gain-of-function mutation in universal tyrosine kinase SRC causes thrombocytopenia, myelofibrosis, bleeding, and bone pathologies. Sci Transl Med 8:328ra30
35. Stritt S, Nurden P, Turro E, Greene D, Jansen SB, Westbury SK et al (2016) A gain-of-function variant in DIAPH1 causes dominant macrothrombocytopenia and hearing loss. Blood 127:2903–2914
36. Westbury SK, Turro E, Greene D, Lentaigne C, Kelly AM, Bariana TK et al (2015) Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. Genome Med 7:36
37. Bone WP, Washington NL, Buske OJ, Adams DR, Davis J, Draper D et al (2016) Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. Genet Med 18:608–617
38. Firth HV, Wright CF, Study D (2011) The Deciphering Developmental Disorders (DDD) study. Dev Med Child Neurol 53:702–703
39. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M et al (2015) Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet 385:1305–1314
40. Chatzimichali EA, Brent S, Hutton B, Perrett D, Wright CF, Bevan AP et al (2015) Facilitating collaboration in rare genetic disorders through effective matchmaking in DECIPHER. Hum Mutat 36:941–949
41. Silfhout ATV, van Ravenswaaij CMA, Hehir-Kwa JY, Verwiel ETP, Dirks R, van Vooren S et al (2013) An

update on ECARUCA, the European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations. Eur J Med Genet:. doi:10.1016/j.ejmg.2013.06.010
42. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD et al (2015) The genetic basis of mendelian phenotypes: discoveries, challenges, and opportunities. Am J Hum Genet 97:199–215
43. Kirkpatrick BE, Riggs ER, Azzariti DR, Miller VR, Ledbetter DH, Miller DT et al (2015) GenomeConnect: matchmaking between patients, clinical laboratories, and researchers to improve genomic knowledge. Hum Mutat 36:974–978
44. Beaulieu CL, Majewski J, Schwartzentruber J, Samuels ME, Fernandez BA, Bernier FP et al (2014) FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. Am J Hum Genet 94:809–817
45. Thompson R, Johnston L, Taruscio D, Monaco L, Béroud C, Gut IG et al (2014) RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. J Gen Intern Med 29(Suppl 3):S780–S787
46. Zemojtel T, Köhler S, Mackenroth L, Jäger M, Hecht J, Krawitz P et al (2014) Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci Transl Med 6:252ra123
47. Robinson PN, Köhler S, Oellrich A, Sanger Mouse Genetics Project, Wang K, Mungall CJ et al (2014) Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res 24:340–348
48. Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M et al (2015) Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc 10:2004–2015
49. Robinson PN, Köhler 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 83:610–615