

# Neue Wege in der bioinformatischen Phänotypanalyse

## Die Human Phenotype Ontology

### Was ist eine Ontologie?

Wissen heißt zu wissen, wo etwas geschrieben steht. In den Zeiten der weltumspannenden Vernetzung ist diese weitverbreitete Weisheit allgegenwärtig, denn zu wissen, dass sich eine Information beispielsweise im Internet befindet, heißt noch lange nicht, diese dort auch zu finden. Suchmaschinen helfen zwar – aber nur soweit, wie Mensch und Computer einem Begriff dieselbe Bedeutung zuordnen. Dies ist nur eines der vielen Probleme, bei deren Lösung Ontologien einen entscheidenden Beitrag leisten können.

Eine Ontologie beschreibt einen Wissensbereich mit Hilfe eines standardisierten Vokabulars unter Berücksichtigung der Beziehungen zwischen den dort definierten Begriffen. Zum Beispiel können die Begriffe einer Ontologie in Form von Ober- und Unterbegriffen gegliedert und somit durch semantische Beziehungen miteinander verbunden sein. Ein standardisiertes Vokabular und explizite Definitionen der verwendeten Begriffe sind eine essenzielle Voraussetzung für jede computergestützte Datenanalyse. Die Notwendigkeit eines standardisierten Vokabulars lässt sich sehr anschaulich anhand eines Beispiels aus der Gastronomie erklären. Laut Wikipedia sind „Berliner Pfannkuchen“ schwimmend in Fett ausgebackene,

etwa faustgroße Krapfen aus süßem Hefeteig mit einer Füllung aus Konfitüre. Eine einheitliche Bezeichnung für dieses allseits beliebte Gebäck existiert in der Umgangssprache nicht. Vielmehr gibt es eine reiche Anzahl von regional verbreiteten Synonymen: Ob als „Berliner“, „Krapfen“, „Pfannkuchen“ oder „Kreppel“ bezeichnet, alle diese Begriffe beziehen sich auf ein und dieselbe Entität. Zudem wird vielerorts der Begriff „Pfannkuchen“ auch als Synonym für „Eierkuchen“ verwendet, d. h. ein einzelner Begriff kann auch zur Beschreibung von verschiedenen Entitäten eingesetzt werden. Ein Computer kann ohne zusätzliche Informationen nicht erkennen, dass

*Berliner = Krapfen = Pfannkuchen ≠  
 Pfannkuchen = Eierkuchen = dicker Crêpe*

Auch in der medizinischen Terminologie sind synonyme Begriffe sowie Termini mit mehreren möglichen Bedeutungen weit verbreitet. Begriffe wie „generalisierte Amyotrophie“, „generalisierte muskuläre Atrophie“ und „muskuläre Atrophie, generalisiert“ werden beispielsweise bei OMIM (Online Mendelian Inheritance in Man) zur Charakterisierung verschiedener Erkrankungen verwendet. Diese können aber durch computergestützte Analysen nicht ohne Weiteres als

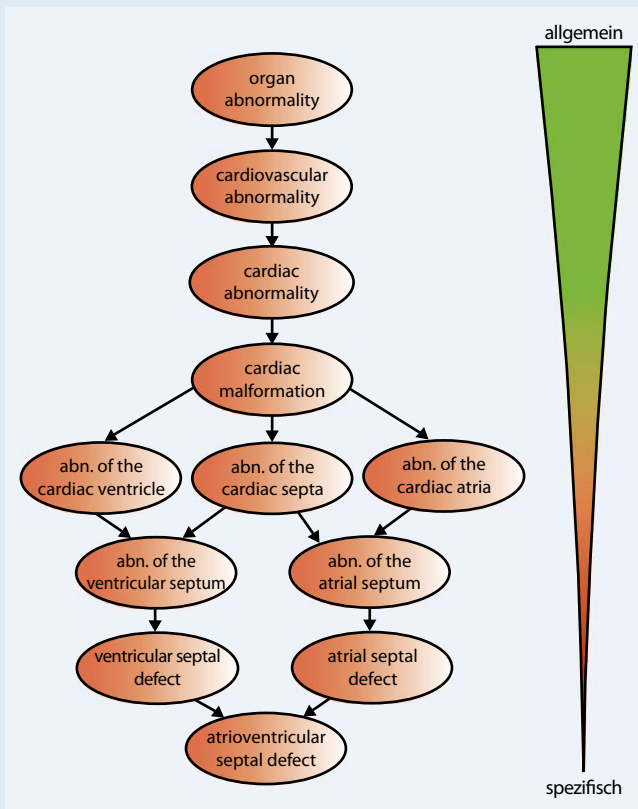
synonyme Begriffe erkannt werden. Andererseits kann zum Beispiel ein medizinischer Begriff wie „Fibrillationen“ für einen Neurologen oder einen Kardiologen durchaus verschiedene Bedeutungen haben. Für einen Computer muss ersichtlich sein, dass

*Zungenfibrillationen = Fibrillationen ≠  
 Fibrillationen = Ventrikelfibrillationen =  
 Kammerflimmern*

Eine Ontologie kann die Grundlage für eine computergestützte Analyse innerhalb eines Wissensgebietes schaffen, indem sie die Terminologie dieses Wissensgebietes für den Computer verständlich und analysierbar festlegt und die semantischen Beziehungen zwischen den Begriffen definiert (■ **Abb. 1**).

### Vokabulare in der Humangenetik

Eine genaue Analyse des klinischen Phänotyps spielt eine fundamentale Rolle in vielen Bereichen der Medizin, insbesondere auch auf dem Gebiet der klinisch-genetischen Diagnostik und Forschung. Angaben zu Phänotypen in Publikationen zu neuen Genmutationen und Genotyp-Phänotyp-Korrelationen bei bereits bekannten Syndromen sind in der Vergangenheit fast ausschließlich in Form von



**Abb. 1** ▲ Die Human Phenotype Ontology (HPO) ist aufgebaut als ein gerichteter, azyklischer Graph, bei dem die Begriffe jeweils einen spezifischeren Aspekt eines allgemeineren Oberbegriffes darstellen. Merkmale können auch mehr als einem Oberbegriff zugeordnet sein, was die komplexen biologischen Beziehungen zwischen den Merkmalen widerspiegelt. Wie hier dargestellt, ist das Merkmal „abnormality of the atrial septum“ ein Unterbegriff von „abnormality of the cardiac septa“ und „abnormality of the cardiac atria“. Dabei sind Begriffe, die sich nahe der Wurzel „organ abnormality“ befinden, weniger spezifisch als die diesem Begriff zugeordneten Unterbegriffe. Die klinische Spezifität eines Merkmals schlägt sich mathematisch in dessen Informationsgehalt nieder. Dieser ist definiert als negativer Logarithmus der Häufigkeit des Auftretens des Begriffes in den Annotationen aller in der Datenbank erfassten Krankheiten. Als Annotationen werden hier die phänotypischen Auffälligkeiten bezeichnet, die einer spezifischen Erkrankung zugeordnet sind. Intuitiv verständlich ist beispielsweise, dass ein häufig beobachtetes Merkmal wie „mental retardation“ einen geringeren Informationsgehalt hat als beispielsweise das Merkmal „calcific stippling“, das derzeit in den Annotationsdateien von nur vier hereditären Erkrankungen aufgeführt ist

freien Textbeschreibungen erfolgt. In den letzten Jahren wurde die wissenschaftliche Analyse des phänotypischen Spektrums von Erkrankungen des Menschen, das sogenannte Phänom, von einer Reihe von Arbeitsgruppen für verschiedenste Fragestellungen aufgegriffen. Unter anderem wurden Konzepte zur Suche von Kandidatengenomen entwickelt und phänotypische Netzwerke verschiedenster genetisch bedingter Erkrankungen genauer untersucht [8, 16].

Ein standardisiertes, kontrolliertes Vokabular erlaubt es, in klarer, eindeutiger Form phänotypische Informationen in medizinischen Gutachten, Publikationen und Datenbanken zu kommunizieren. Zudem eröffnet es unter anderem neue Möglichkeiten, phänotypische Daten computergestützt zu analysieren und auch Vergleiche zwischen den bei verschiedenen Spezies beobachteten Phänotypen zu ziehen, um neue Einblicke in Ätiologie und

Pathogenese hereditärer Erkrankungen zu gewinnen.

Zur Zeit ist die OMIM-Datenbank eine der wichtigsten Quellen für Informationen über genetisch bedingte Erkrankungen des Menschen, in der die meisten beim Menschen bekannten monogenen sowie einige weitere hereditäre Erkrankungen detailliert beschrieben sind [7]. Das Fehlen eines standardisierten medizinischen Vokabulars sowie konsistenter Annotationen mit untereinander gut definierten Beziehungen hat bisher eine computergestützte Analyse dieser Daten erheblich erschwert. Als Annotationen werden hier die phänotypischen Auffälligkeiten bezeichnet, die einer spezifischen Erkrankung zugeordnet sind. Dies entspricht bei OMIM der „clinical synopsis“.

Eine einfache Suche nach dem Begriff „generalized muscle atrophy“ auf der OMIM-Website vom 02.10.2009 ergab beispielsweise 174 Ergebnisse, während eine Suche nach „muscular atrophy, generalized“ lediglich 96 Ergebnisse liefert. Das heißt, dass eine Suche nach ein und demselben Begriff (Semantik) je nach Wortwahl (Syntaktik) eine deutlich abweichende Antwort liefern kann. Ein weiteres Hindernis sind unterschiedliche Schreibweisen, Synonyme, aber auch einfache Rechtschreibfehler, die zum Teil dazu führen, dass trotz Verwendung des korrekten Suchbegriffes das gewünschte Resultat nicht als solches erkannt wird. Ein uneinheitlicher Gebrauch von medizinischen Termini erschwert bzw. verhindert eine effiziente Nutzung der vorhandenen Daten.

Die London Dysmorphology Database (LDDDB) [4], POSSUM [11] und Orphanet [1] verwenden definierte Vokabulare, die speziell als Standard für die jeweilige Datenbank, jedoch nicht als Werkzeug für eine bioinformatische Datenanalyse entwickelt wurden. Es existieren bei Weitem zu wenig Begriffe zur Beschreibung der verschiedenen, beim Menschen beobachteten phänotypischen Auffälligkeiten. Auf der anderen Seite sind einige der in diesen Datenbanken verwendeten Begriffe für eine präzise Analyse des Phänotyps nicht geeignet. Zum Beispiel beinhalten viele der in der LDDDB verwendeten Begriffe eigentlich zwei separate Entitäten. Der Begriff „asternia or bifid sternum“ bein-

haltet eigentlich die zwei separaten Merkmale „asternia“ und „bifid sternum“.

## Die Human Phenotype Ontology (HPO)

Die Analyse phänotypischer Auswirkungen genetischer Mutationen ist seit langem ein essenzieller Bestandteil der Erforschung biologischer Funktionen von Genen. Bei einigen hereditären Erkrankungen des Menschen fällt aufgrund von überlappenden klinischen Merkmalen eine gewisse phänotypische Ähnlichkeit auf. Man kann beispielsweise sagen, dass das Marfan-Syndrom (MIM 154700) und die kongenitale kontraktuelle Arachnodaktylie (MIM 121050) verwandte Erkrankungen zu sein scheinen, da einerseits beide Erkrankungen eine Reihe gleicher skelettaler Symptome aufweisen und andererseits die betreffenden Genmutationen „verwandte“ Gene betreffen (*FBN1* und *FBN2*; [13]). Die Tatsache, dass viele genetisch bedingte Erkrankungen überlappende phänotypische Merkmale aufweisen, hat zu der Entwicklung des Konzeptes der „Krankheitsfamilien“ geführt [2, 10]. Phänotypische Ähnlichkeiten innerhalb einer Gruppe verwandter Erkrankungen können in vielen Fällen auf eine Dysfunktion innerhalb eines regulatorischen Netzwerkes zurückgeführt werden. Ein Beispiel hierfür ist die Dysregulation eines Signaltransduktionsweges, wie sie für das Noonan-Syndrom [2, 10] und verwandte Erkrankungen gezeigt werden konnte [5]. Dies führt uns vor Augen, wie wichtig eine genaue phänotypische Analyse für unser Verständnis der Physiologie und Pathophysiologie zellulärer Netzwerke ist. Eine präzise phänotypische Beurteilung ermöglicht unter Umständen Rückschlüsse auf Gruppen verwandter Gene, die Signalwege und biochemische Module bilden, deren Fehlfunktion wiederum verschiedene, überlappende phänotypische Auswirkungen zur Folge haben kann. In den letzten Jahren wurden von verschiedenen Arbeitsgruppen genomweite Untersuchungen durchgeführt, die das enorme Potenzial einer Korrelation von Phänotyp zu genetischen und zellulären Netzwerken erahnen lassen [6, 8].

Unter diesen Gesichtspunkten beschlossen wir, eine Ontologie zur Klassi-

## Zusammenfassung · Abstract

medgen 2010 · 22:221–228 DOI 10.1007/s11825-010-0215-2  
© Springer-Verlag 2010

S.C. Doelken · S. Köhler · S. Bauer · C.-E. Ott · P. Krawitz · D. Horn · S. Mundlos · P.N. Robinson  
**Neue Wege in der bioinformatischen Phänotypanalyse.  
Die Human Phenotype Ontology**

### Zusammenfassung

Die präzise Beschreibung phänotypischer Auffälligkeiten ist für die klinische Diagnostik und für unser wissenschaftliches Verständnis von Erkrankungen von grundlegender Bedeutung. Derzeit sind mehrere tausend hereditäre Erkrankungen des Menschen bekannt, die jeweils durch eine mehr oder weniger spezifische Kombination phänotypischer Merkmale charakterisiert sind. Eine besondere Schwierigkeit bei der computergestützten Analyse phänotypischer Daten ergab sich bislang durch das Fehlen eines standardisierten medizinischen Vokabulars und den Mangel an adäquaten Datenstrukturen zur Erfassung phänotypischer Merkmale.

Die Human Phenotype Ontology (HPO) wurde von unserer Arbeitsgruppe mit dem Ziel entwickelt, alle phänotypischen Auffälligkeiten, die bei monogenen Erkrankungen des Menschen auftreten können, zu beschreiben (<http://www.human-phenotype-ontology.org>). Die HPO stellt ein hierarchisch strukturiertes, deskriptives und standardisiertes

Vokabular zur Beschreibung phänotypischer Merkmale bereit und ist somit geeignet, signifikante phänotypische Ähnlichkeiten und Unterschiede verschiedener hereditärer Erkrankungen zu erfassen.

Eine Ontologie wie die HPO eröffnet viele neuartige Möglichkeiten, insbesondere auch auf dem Gebiet der klinisch-genetischen Diagnostik. Ein Beispiel hierfür ist der von uns entwickelte Phenomizer (<http://compbio.charite.de/phenomizer>), ein neuartiges, frei verfügbares ontologisches Suchprogramm. Der Phenomizer nutzt die semantische Struktur der HPO, um klinische Symptome anhand ihrer Spezifität zu wichten, und kann als Werkzeug für eine computergestützte klinische Differenzialdiagnostik verwendet werden.

### Schlüsselwörter

Human Phenotype Ontology · Phenomizer · Biomedizinische Ontologien · Differenzialdiagnostik

## Computational methods for the study of human disease manifestations . The Human Phenotype Ontology

### Abstract

An accurate description of phenotypic abnormalities is the foundation of clinical diagnostics and the basis of our understanding of diseases. There are many thousands of hereditary diseases in humans, each of which displays a more or less specific combination of phenotypic features. Computational analysis of phenotypic data has been hampered by the lack of a controlled medical vocabulary for describing phenotypic abnormalities and adequate computational data structures to capture phenotypic data.

This prompted us to develop the Human Phenotype Ontology (HPO), a hierarchical, descriptive vocabulary for human phenotypes constructed with the goal of covering all phenotypic abnormalities which are commonly encountered in human monogenic diseases (<http://www.human-phenotype-ontology.org>). The HPO provides a standardized

vocabulary for the description of human phenotypes and is able to capture phenotypic similarities between diseases in a highly significant fashion.

Clinical ontologies, such as the HPO allow many novel forms of analyzing the human phenotype for research or clinical purposes. We developed the Phenomizer (<http://compbio.charite.de/phenomizer>), a novel, freely available ontological search routine which uses the semantic structure of the HPO to weight clinical features based on specificity and thereby harnesses the semantic structure of human phenotypic abnormalities to assist with the process of differential diagnostics.

### Keywords

Human Phenotype Ontology (HPO) · Phenomizer · Biomedical ontology · Differential diagnosis

fizierung phänotypischer Auffälligkeiten beim Menschen zu entwerfen. Die Human Phenotype Ontology (HPO) wurde mit dem Ziel entwickelt, alle phänotypischen Auffälligkeiten, die bei genetisch bedingten Erkrankungen des Menschen auftreten können, zu erfassen [14].

Derzeit enthält die HPO über 9500 phänotypische Merkmale sowie Annotationen für 4779 Einträge aus der OMIM-Datenbank, basierend auf dem Vokabular der HPO. Die Datenstruktur der HPO, die phänotypische Auffälligkeiten und deren semantische Beziehungen untereinander erstmals mittels eines einheitlichen Vokabulars beschreibt, ermöglicht einen schnellen, hochspezifischen, computergestützten Vergleich klinischer Merkmale verschiedener hereditärer Erkrankungen.

Jeder Begriff in der HPO beschreibt eine phänotypische Auffälligkeit wie zum Beispiel „atrial septal defect“. Alle Begriffe sind in einem hierarchischen System bestehend aus übergeordneten Begriffen mit verschiedenen zugehörigen Unterbegriffen gegliedert und können durch vielschichtige „Verwandtschaftsbeziehungen“ miteinander verknüpft sein. So ist es auch möglich, dass ein Begriff mehreren übergeordneten Begriffen zugeordnet werden kann. Übergeordnete Begriffe, die nahe der Wurzel stehen, in diesem Fall „organ abnormality“ (■ **Abb. 1**), sind weniger spezifisch als die diesen Begriffen zugeordneten Unterbegriffe.

Die HPO selbst beschreibt keine individuellen Erkrankungen, sondern die phänotypischen Merkmale, die mit den Erkrankungen assoziiert sind. Die Mehrzahl der in der HPO enthaltenen Begriffe beschreibt Organauffälligkeiten. Zudem werden jedoch auch verschiedene separate Ontologien zur Beschreibung des Vererbungsmodus und zur Beschreibung des Zeitpunktes des Auftretens und des klinischen Verlaufes von Erkrankungen bereitgestellt. Klinische Entitäten wurden jeweils mit dem spezifischsten Begriff aus der HPO annotiert. Die so genannte „true path rule“ [3] besagt, dass der Weg von einem Nachkommen bis hin zu dem zugehörigen parentalen Begriff der höchsten Ebene immer biologisch korrekt sein muss. Diese Regel gilt für alle Begriffe in der HPO, und für das Beispiel in ■ **Abb. 1** gilt somit, dass für eine Er-

krankung, die auf den Begriff „atrial septal defect“ annotiert ist, auch alle diesem Begriff übergeordneten Termini wie z. B. „abnormality of the cardiac septa“ zutreffen. Dies bedeutet auch, dass bei einer Suche mit dem Suchbegriff „abnormality of the cardiac septa“ alle Syndrome aufgelistet werden, die auf diesen oder einen der diesem Begriff zugeordneten Unterbegriffe annotiert sind. Die Struktur der HPO erlaubt auf diesem Wege erstmals eine flexible Suche nach möglichen Differenzialdiagnosen, je nach Wunsch entweder mit sehr spezifischem oder auch mit breiterem Fokus.

### Anwendung der HPO für die klinische Diagnostik

Im weiteren Verlauf wandten wir uns der Erkundung möglicher klinischer Anwendungsverfahren der HPO zu. Kombinationen verschiedener Merkmale werden auf dem Gebiet der klinischen Genetik regelmäßig bei der Suche nach einer klinischen Diagnose verwendet. Die Erstellung einer differenzierten klinischen Diagnose ist eine der wichtigsten Aufgaben des Arztes. Die klinische Diagnosestellung ist oftmals eine Herausforderung, insbesondere auf dem Gebiet der klinischen Genetik, in dem die Differenzialdiagnostik häufig schon allein durch die Anzahl und Vielfalt hereditärer Erkrankungen, aber auch durch Pleiotropie und variable Expressivität erschwert wird.

Eine zeitnahe und korrekte genetische Diagnosestellung ist entscheidend, um unnötige diagnostische Prozeduren zu vermeiden, wichtige therapeutische Maßnahmen zu ergreifen, klinische Behandlungsstrategien zu entwerfen und um eine adäquate genetische Beratung zu gewährleisten. Trotz aller Bemühungen lässt sich bislang nur bei etwa der Hälfte der Kinder, die mit Dysmorphiezeichen und/oder mentaler Retardierung auffallen, eine ätiologische Diagnose stellen [12, 15].

Ein individueller Patient mit einer hereditären Erkrankung muss nicht zwangsläufig alle Symptome zeigen, die potenziell bei der vorliegenden Erkrankung vorkommen können, oder er oder sie hat möglicherweise weitere klinische Auffälligkeiten, die nicht mit der Erkrankung assoziiert sind. Je nach Situation mag

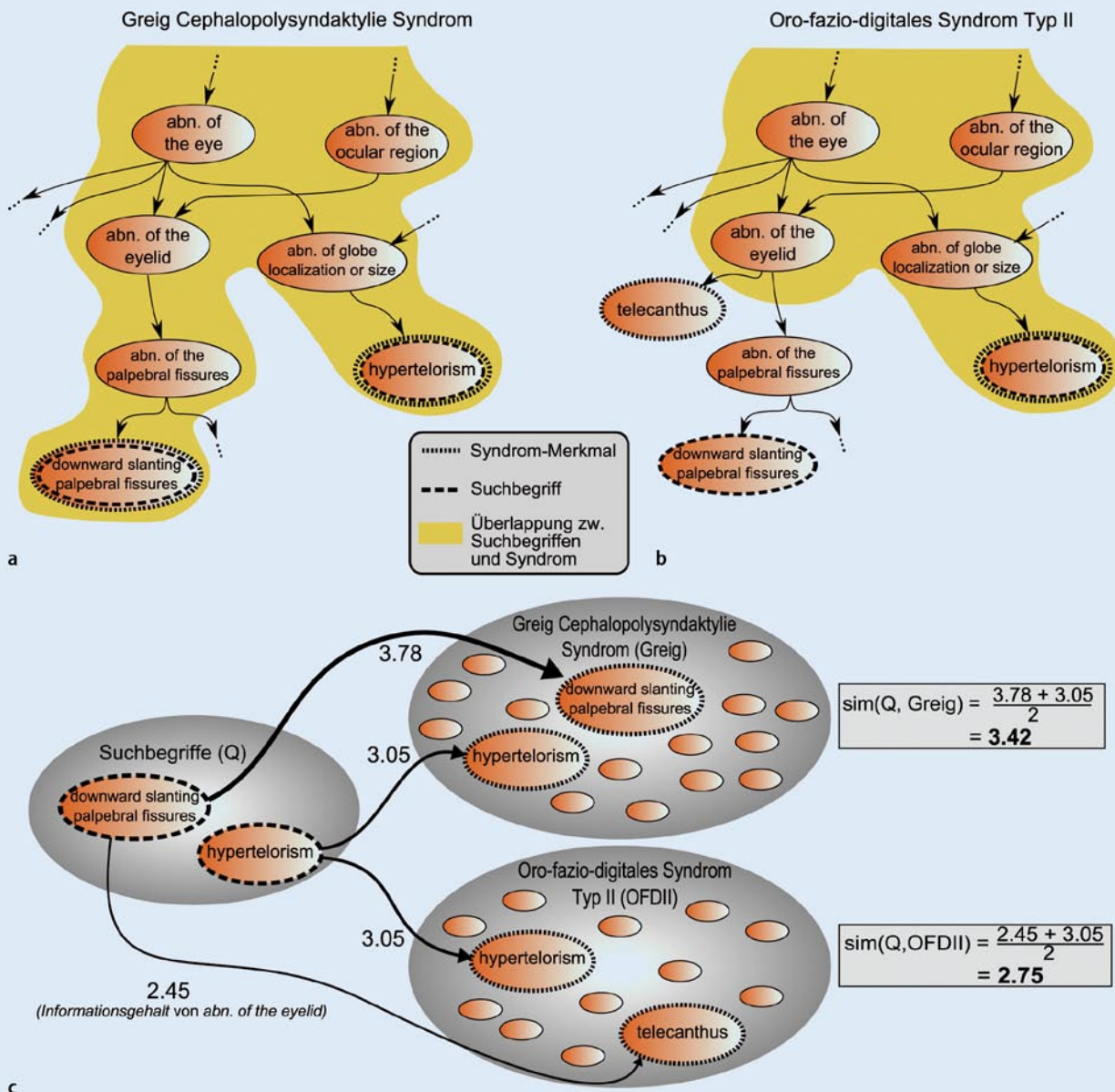
ein behandelnder Arzt in der Lage sein, die klinischen Auffälligkeiten eines Patienten mit einem unterschiedlichen Grad an Spezifität anzugeben. Abhängig vom Grad der Erfahrung des behandelnden Arztes und der Verfügbarkeit von Laboruntersuchungen und technischen Untersuchungsverfahren können klinische Auffälligkeiten sowohl sehr detailliert als auch auf einem eher allgemeineren Niveau beschrieben werden. Optimalerweise sollten diagnostische Algorithmen eine Suche auf unterschiedlich spezifischer Ebene zulassen, spezifische Merkmale stärker wichten als unspezifische und nicht übermäßig sensitiv auf die Tatsache reagieren, dass bei einem individuellen Patienten möglicherweise einzelne Merkmale nicht vorliegen, hingegen andere, nicht mit der Erkrankung assoziierte Merkmale vorhanden sein können.

Es existiert derzeit eine Reihe von genetischen Datenbanken, zu denen unter anderem LDDB [4], POSSUM [11], OMIM [7] und Orphanet [1] gehören. Bei diesen Datenbanken gibt der Nutzer eines oder mehrere klinische Merkmale ein und erhält eine Liste von möglichen Differenzialdiagnosen, die durch eines oder mehrere der eingegebenen klinischen Merkmale charakterisiert werden. Die bisher verfügbaren Suchprogramme können keine explizite Rangfolge der auf diesem Wege aufgelisteten Differenzialdiagnosen erstellen. Die potenziell sehr lange Liste der Suchergebnisse ist entweder ungeordnet oder sogar einfach alphabetisch sortiert, das heißt, die Plausibilität der verschiedenen Kandidaten wird nicht weiter bewertet. Aus diesem Grund entwickelten wir den Phenomizer, ein Anwendungsprogramm zur Unterstützung bei der differenzialdiagnostischen Beurteilung von Patienten mit hereditären Erkrankungen, das auch eine Aussage über die Plausibilität einer bestimmten Differenzialdiagnose ermöglicht.

### Der Phenomizer: computergestützte Differenzialdiagnostik hereditärer Erkrankungen

Der Phenomizer bedient sich der semantischen Struktur der HPO [14], um bei der Differenzialdiagnostik die Begriffe anhand ihrer klinischen Spezifität zu wichten. In-





**Abb. 2** ▲ Analyse der phänotypischen Ähnlichkeit der HPO-Suchbegriffe „downward slanting palpebral fissures“ und „hypertelorism“ anhand der Annotierungen des Greig-Cephalopolysyndaktylie-Syndroms (a, MIM 175700) und des Oro-fazio-digitalen Syndroms Typ II (Mohr-Syndrom, b, MIM 252100). Zu beachten ist, dass für oben genannte Syndrome nicht alle annotierten Merkmale aufgeführt sind. Da das Merkmal „hypertelorism“ nicht auf das Oro-fazio-digitale Syndrom Typ II annotiert ist, ist der Deckungsbereich (hier gelb dargestellt) für dieses Syndrom geringer als für das Greig-Cephalopolysyndaktylie-Syndrom. Die daraus resultierenden Folgen für die Kalkulation des Ähnlichkeitswertes sind in c dargestellt. Für das Greig-Cephalopolysyndaktylie-Syndrom stimmen die Suchbegriffe perfekt mit den annotierten Merkmalen überein. Im Gegensatz dazu ist die bestmögliche Übereinstimmung für „hypertelorism“ bei den für das Oro-fazio-digitale Syndrom Typ II annotierten Merkmalen der Begriff „telecanthus“. Der gemeinsame Oberbegriff mit dem höchsten Informationsgehalt für diese beiden Merkmale ist „abnormality of the eyelid“. Der Informationsgehalt der Ähnlichkeit zwischen Suchbegriff und Erkrankung wird als durchschnittlicher maximaler Ähnlichkeitswert für jeden Suchbegriff definiert. Folglich ergibt eine Suche nach „downward slanting palpebral fissures“ und „hypertelorism“ eine bessere Übereinstimmung zum Greig-Cephalopolysyndaktylie-Syndrom als zum Oro-fazio-digitalen Syndrom Typ II. Die mathematischen Einzelheiten werden in [9] näher erklärt

tuitiv lässt sich erkennen, dass der Grad der Ähnlichkeit zweier Erkrankungen, die beide auf ein Merkmal mit hohem Informationsgehalt annotiert sind, wie beispielsweise „downward slanting palpebral fissures“, höher sein sollte, als der Grad

der Ähnlichkeit zweier Erkrankungen, die beide auf ein weniger spezifisches Merkmal annotiert sind, wie zum Beispiel „abnormality of the eyelid“, ein Merkmal mit einem deutlich geringeren Informationsgehalt (■ Abb. 2).

Wir haben ein Suchverfahren entwickelt, das ein über den Informationsgehalt der HPO-Begriffe definiertes semantisches Ähnlichkeitsmaß verwendet, um die phänotypische Ähnlichkeit zwischen den Eingaben des Arztes und den

HPO-Begriffen, auf die einzelne hereditäre Krankheiten annotiert werden, zu erfassen. Zudem haben wir ein statistisches Modell entwickelt, das den Ergebnissen einer Suche p-Werte zuweist, die herangezogen werden, um die Differenzialdiagnosen in eine sinnvolle Reihenfolge zu bringen. Der p-Wert gibt an, mit welcher Wahrscheinlichkeit ein äquivalenter oder besserer Ähnlichkeits-Score erreicht werden kann, wenn vergleichsweise eine zufällige Anfrage mit ebenso vielen willkürlich ausgewählten Symptomen durchgeführt wird. Derzeit liegt dem jeweiligen p-Wert einer Abfrage der Vergleich mit einer Simulation von 100.000 zufälligen Abfragen zugrunde. In der Statistik geht man im allgemeinen von einer Signifikanzschwelle von 0,05 (5%) aus, d. h. ab einem p-Wert von 0,05 gilt das Ergebnis als signifikant.

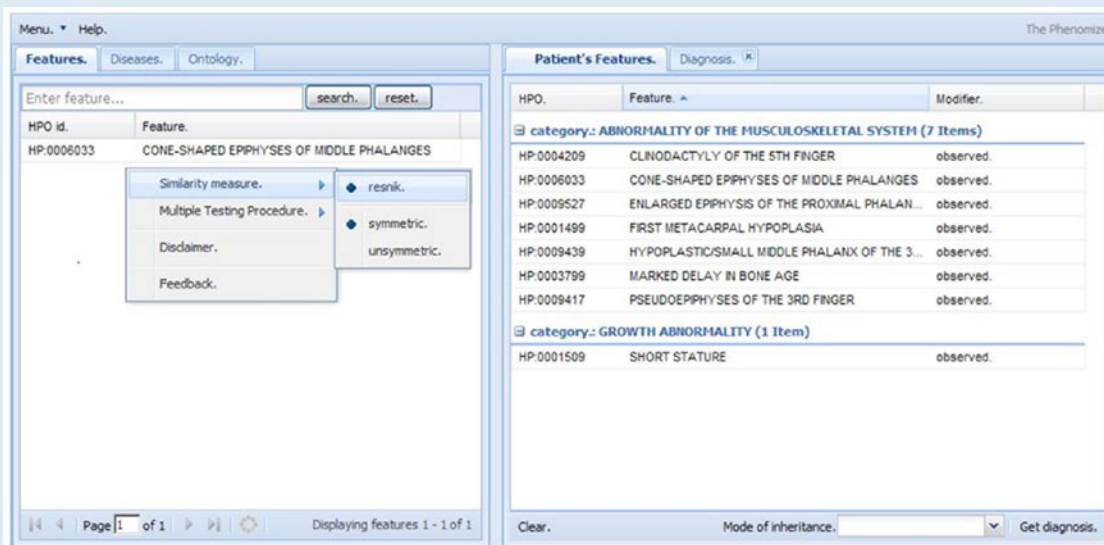
Unsere Methoden zeigen gegenüber anderen Suchverfahren insbesondere dann einen Vorteil, wenn die verwendeten Suchbegriffe unverwandte Symptome enthalten, das heißt, wenn z. B. ein

Patient ein mit der zugrunde liegenden Erkrankung nicht assoziiertes Merkmal aufweist oder wenn ungenaue Suchbegriffe verwendet werden, etwa weil entsprechende Spezialuntersuchungen noch nicht erfolgt sind oder weil dem Arzt die korrekte Bezeichnung eines phänotypischen Merkmals nicht bekannt ist. Wenn ein Arzt z. B. anhand eines Röntgenbildes erkennt, dass eine Veränderung einer Epiphyse eines Fingerknochens vorliegt, er diese Veränderung aber nicht treffsicher als „cone-shaped epiphysis of the middle phalanx of the 3rd finger“ benennen kann, könnte er auch den verwandten Begriff „abnormality of the epiphysis of the middle phalanx of the 3rd finger“ eingeben [9]. Dies unterscheidet den Phenomizer von allen bisher verfügbaren Suchprogrammen. Suchmaschinen wie Pubmed oder Google sind derzeit nur in der Lage, nach den eingegebenen Wörtern oder syntaktischen Varianten derselben zu suchen. Dabei führt die Eingabe eines semantisch identischen Begriffes

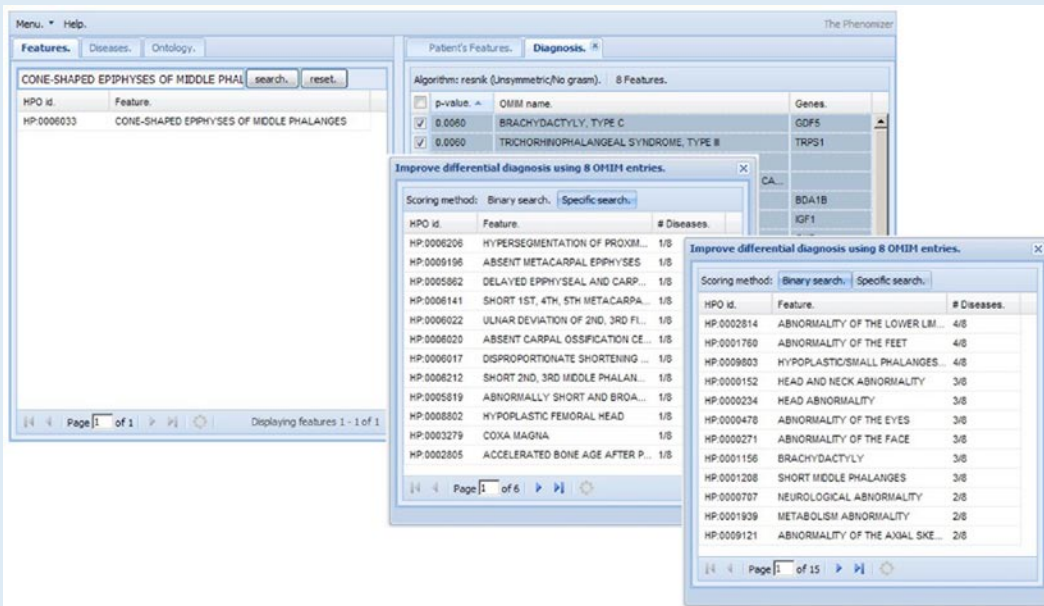
nicht in jedem Fall zu einem identischen Suchergebnis.

Wenn bei einer Suche mit dem Phenomizer die an erster Stelle erscheinende Differenzialdiagnose einen signifikanten p-Wert erreicht, zeigt dies dem Kliniker, dass dieses Syndrom eine mögliche Differenzialdiagnose ist, die gegebenenfalls weiter berücksichtigt werden sollte. Wenn allerdings die an erster Stelle erscheinende Differenzialdiagnose keinen signifikanten p-Wert erreicht, so kann dies einerseits bedeuten, dass die vom Kliniker eingegebene Kombination phänotypischer Merkmale nicht spezifisch genug ist, um eine Diagnosestellung zu ermöglichen. Andererseits kann diese spezielle Kombination phänotypischer Auffälligkeiten auf eine Erkrankung hindeuten, die in der befragten Datenbank nicht vorhanden ist (■ **Abb. 3, 4, 5**).

### Tutorium: Computergestützte differenzialdiagnostische Suche mit dem Phenomizer



**Abb. 3** ▲ Mit Hilfe der Registerkarte „Features“ kann der gewünschte HPO-Begriff durch Eingabe eines beliebigen phänotypischen Merkmals oder auch eines synonymen Begriffes gesucht und durch Doppelklick ausgewählt werden. In diesem Beispiel wurden einige phänotypische Merkmale eines Patienten mit Brachydaktylie Typ C eingegeben. Die so ausgewählten Merkmale werden dann auf der rechten Seite des Fensters aufgelistet. Für jedes Merkmal kann entweder die Eigenschaft „observed“ (beobachtet) oder „mandatory“ (obligat) ausgewählt werden. Wenn ein Merkmal als „mandatory“ eingestuft wird, werden alle Differenzialdiagnosen, bei denen dieses Merkmal nicht vorkommt, herausgefiltert. Falls bekannt, kann weiterhin der Vererbungsmodus angegeben werden. Die phänotypische Ähnlichkeitssuche kann entweder symmetrisch oder asymmetrisch erfolgen. Bei der ontologischen Ähnlichkeitssuche handelt es sich generell um einen asymmetrischen Score, das heißt die Berechnungen zwischen Abfrage und Erkrankung sind nicht zwangsläufig äquivalent zur gegenläufigen Berechnung zwischen Erkrankung und Abfrage. Bei der asymmetrischen Suche wird lediglich die Ähnlichkeit der eingegebenen Suchbegriffe zu der Erkrankung berücksichtigt. Dieser Berechnung liegt die Annahme zugrunde, dass die Abfrage nur einen Teil der Symptome der gesuchten Diagnose darstellt. Bei der symmetrischen Suche wird zusätzlich die Ähnlichkeit der Erkrankung zur Suchabfrage berechnet und der Durchschnitt gebildet. Die symmetrische Suche sollte verwendet werden, wenn man davon ausgeht, dass die eingegebenen Symptome sich vollständig mit den Symptomen der gesuchten Erkrankung decken. Sollten weitere Symptome der Erkrankung existieren, die in der Abfrage nicht enthalten sind, führt dies bei der symmetrischen Suche zu einem schlechteren Score. Generell wird empfohlen, mit einer asymmetrischen Suche zu beginnen, eine symmetrische Suche kann jedoch bei Erkrankungen mit sehr wenigen oder auch bei Erkrankungen mit sehr vielen klinischen Merkmalen möglicherweise besser geeignet sein [9]. Nach Klicken auf die „Get diagnosis“-Schaltfläche erzeugt der Phenomizer eine Liste von möglichen Differenzialdiagnosen, geordnet nach Ihrem jeweiligen p-Wert (■ **Abb. 4**)



**Abb. 4** Der Phenomizer bietet dem Kliniker mittels der Funktion „Improve differential diagnosis“ weiterhin die Möglichkeit, eine Liste derjenigen klinischen Merkmale zu erhalten, die im weiteren diagnostischen Prozess von Bedeutung sein könnten. Zu diesem Zweck können beliebig viele der möglichen Differenzialdiagnosen ausgewählt werden, bevor man die Schaltfläche „Improve differential diagnosis“ anklickt. Der Phenomizer erzeugt daraufhin eine Liste der phänotypischen Merkmale, die für einen Teil („binary search“) oder spezifisch nur für eine der ausgewählten Differenzialdiagnosen („specific search“) annotiert sind. Der Kliniker kann diese Merkmale begutachten und gegebenenfalls zutreffende durch Doppelklick zu den bereits eingegebenen Suchbegriffen hinzufügen. Insbesondere die Liste von spezifischen phänotypischen Merkmalen kann möglicherweise auch Hinweise auf weitere Untersuchungen liefern, die bei einem individuellen Patienten noch veranlasst werden könnten, um zur Diagnosefindung beizutragen. Wenn weitere Suchbegriffe durch Doppelklick hinzugefügt wurden, wird durch erneutes Anklicken von „Get diagnosis“ eine neue Liste möglicher Differenzialdiagnosen in einem separaten Fenster erstellt, und der Kliniker kann diese neue Liste mit der vorherigen vergleichen, um die eventuell resultierenden Veränderungen in der Liste der Differenzialdiagnosen zu überprüfen. Wenn die an erster Stelle erscheinende Differenzialdiagnose einen signifikanten p-Wert erreicht, zeigt dies dem Kliniker, dass dieses Syndrom eine mögliche Differenzialdiagnose ist, die gegebenenfalls weiter berücksichtigt werden sollte. Wenn die an erster Stelle erscheinende Differenzialdiagnose keinen signifikanten p-Wert erreicht, kann dies bedeuten, dass die vom Kliniker eingegebene Kombination phänotypischer Merkmale nicht spezifisch genug ist, um eine Diagnosestellung zu ermöglichen, oder dass diese spezielle Kombination phänotypischer Auffälligkeiten auf eine Erkrankung hindeutet, die in der befragten Datenbank nicht vorhanden ist



**Abb. 5** Für jede Differenzialdiagnose sind über das Kontextmenü weitere Informationen erhältlich, wie z. B. eine Auflistung aller bei dieser Erkrankung bekannten Merkmale („Show annotations“), Links zu den mit der betreffenden Erkrankung assoziierten Genen („Show known genes“), der betreffende OMIM-Eintrag („Show OMIM entry“) und auch eine Visualisierung der Überlappung der eingegebenen Suchbegriffe mit allen für die Erkrankung annotierten Merkmalen („Display overlap“). Das Ergebnis der Phänotypenanalyse kann entweder als CSV- oder als PDF-Datei exportiert werden. Die PDF-Datei enthält Felder zur Eingabe der Patientenkenndaten und weiterer wichtiger Informationen und kann zur Dokumentation des differenzialdiagnostischen Prozesses in der Patientenakte verwendet werden. Zusätzlich zu einer Zusammenfassung der eingegebenen Suchbegriffe und der nach p-Werten sortierten Liste der Differenzialdiagnosen enthält die PDF-Datei für jede Differenzialdiagnose eine Liste derjenigen phänotypischen Merkmale, die spezifisch für dieses Syndrom unter allen gegebenen Differenzialdiagnosen sind



## Ausblick

Derzeit existieren, basierend auf dem Vokabular der HPO, Annotationen für 4779 Einträge aus der OMIM-Datenbank. In den nächsten Jahren wird neben der Komplettierung und Überarbeitung der Begriffe und Annotationen verschiedenster spezifischer klinischer Bereiche eine Ausweitung der bisher existierenden Annotationen auf weitere hereditäre Erkrankungen wie Mikrodeletions-syndrome und chromosomale Aberrationen erfolgen. Zudem sollen für die annotierten Krankheitsbilder weitere Informationen wie Auftretenswahrscheinlichkeit und Schweregrad verschiedener klinischer Merkmale in den Annotationsdateien mit einbezogen werden. Die HPO kann nicht nur für die klinische Diagnostik, sondern auch als eine Basis für die bioinformatische Forschung auf dem Gebiet der menschlichen Phänotypenanalyse verwendet werden. Ein Ziel dieses Forschungszweiges wird es sein, das menschliche Phänom in Beziehung zur molekularen Pathophysiologie der Zelle zu stellen. Die HPO und der Phenomizer sind unter <http://www.human-phenotype-ontology.org> und <http://compbio.charite.de/phenomizer> verfügbar.

## Korrespondenzadresse

**Dr. P.N. Robinson**  
 Institut für Medizinische Genetik  
 Charité – Universitätsmedizin Berlin  
 Augustenburger Platz 1  
 13353 Berlin  
[peter.robinson@charite.de](mailto:peter.robinson@charite.de)

**Interessenkonflikt.** Der korrespondierende Autor gibt an, dass kein Interessenkonflikt besteht.

## Literatur

1. Orphanet: Das Portal für seltene Krankheiten und Orphan Drugs. Inserm, Paris. <http://www.orphanet>
2. Brunner HG, Driell MA van (2004) From syndrome families to functional genomics. *Nat Rev Genet* 5(7):545–551
3. Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res* 11(8):1425–1433
4. Fryns J-P, Ravel TJL de (2002) London Dysmorphology Database, London Neurogenetics Database and Dysmorphology Photo Library on CD-ROM (Version 3) 2001. In: Winter RM, Baraitser M (eds), Oxford University Press. *Hum Genet* 111(1):113

5. Gelb BD, Tartaglia M (2006) Noonan syndrome and related disorders: dysregulated ras-mitogen activated protein kinase signal transduction. *Hum Mol Genet* 15(Spec No 2):R220–R226
6. Goh KI, Cusick ME, Valle D et al (2007) The human disease network. *Proc Natl Acad Sci U S A* 104(21):8685–8690
7. Hamosh A, Scott AF, Amberger JS et al (2005) Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 33(Database issue): D514–D517
8. Köhler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 82(4):949–958
9. Köhler S, Schulz MH, Krawitz P et al (2009) Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet* 85(4):457–464
10. Oti M, Brunner HG (2007) The modular nature of genetic diseases. *Clin Genet* 71(1):1–11
11. POSSUM. <http://www.possum.net.au/>
12. Rauch A, Hoyer J, Guth S et al (2006) Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation. *Am J Med Genet A* 140(19):2063–2074
13. Robinson PN, Arteaga-Solis E, Baldock C et al (2006) The molecular genetics of marfan syndrome and related disorders. *J Med Genet* 43(10):769–787
14. Robinson PN, Köhler S, Bauer S et al (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet* 83(5):610–615
15. Srour M, Mazer B, Shevell MI (2006) Analysis of clinical features predicting etiologic yield in the assessment of global developmental delay. *Pediatrics* 118(1):139–145
16. Wu X, Jiang R, Zhang MQ, Li S (2008) Network-based global inference of human disease genes. *Mol Syst Biol* 4:189

## TMF: Gemeinsame Probleme gemeinsam lösen

Das Rad nicht immer wieder neu erfinden und Doppelarbeit vermeiden, das ist das Leitmotiv der TMF (Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V.). Als Dachorganisation für die vernetzte medizinische Forschung in Deutschland hat sie die Aufgabe, für übergreifende Probleme der medizinischen Forschung jeweils eine professionelle Lösung zu erarbeiten, diese innerhalb der Forschergemeinschaft und mit relevanten weiteren Partnern zu konsentieren und die konsequente Nutzung und Nachhaltigkeit der Lösung zu gewährleisten.

In den interdisziplinären Arbeitsgruppen der TMF identifizieren Forscher übergreifenden Lösungsbedarf, diskutieren Projektideen, stimmen sie ab und schieben die Finanzierung der Vorhaben an. Wissenschaftler finden hier Unterstützung und Erfahrungsaustausch beispielsweise bei der Umsetzung von Datenschutz- und ethischen Richtlinien, bei Fragen zum Aufbau von Forschungsinfrastrukturen wie Daten- und Biobanken, zur strategischen Nutzung von Informationstechnologie, zu Patienteneinwilligungen und Verwertungsfragen. In den regelmäßigen Sitzungen der Arbeitsgruppen entstehen sowohl strategische Impulse für die Forschungsinfrastruktur als auch konkrete Hilfen, Produkte und Services für Forscher.

Zum Thema Biobanken arbeitet eine TMF-Arbeitsgruppe, die eine rechtssichere und langfristige Nutzung von Biobanken durch die Forschung ermöglichen will. Regelmäßig veranstaltet die Arbeitsgruppe zur Fortbildung ihrer Mitglieder Vor-Ort-Besuche von Biobanken. Zuletzt besuchte sie die Biobank an der Medizinischen Universität Graz - eine der größten Biobanken Europas. Den Bereich der Biobanken thematisch ergänzend, arbeitet die TMF-Arbeitsgruppe Molekulare Medizin an der Qualitätskontrolle von molekularen Daten. Dabei geht es unter anderem um die IT-gestützte Verknüpfung klinischer und hochdimensionaler molekularer Daten im Patienten-, Studien- und Zentrums-kontext. Auch an Konzepten zur Qualitätskontrolle der molekularen Daten auf den Ebenen der Erzeugung, Interpretation, Speicherung und Validierung arbeiten Forscher gegenwärtig unter dem Dach der TMF.

Geschäftsstelle TMF e.V.  
 Tel.: 030 31 01 1950  
 E-Mail: [info@tmf-ev.de](mailto:info@tmf-ev.de)  
[www.tmf-ev.de](http://www.tmf-ev.de)