medgen 2010 · 22:209–212
DOI 10.1007/s11825-010-0224-1

H. Runz · Ch. Fischer
Institut für Humangenetik, Universität Heidelberg

Einführung zum Schwerpunktthema Datenbanken

Das Internet bietet Zugang zu einem einzigartigen Wissensfundus, der das gezielte Nachschlagen von Informationen in Lehrbüchern und gedruckten Fachzeitschriften zunehmend ersetzt. „Neue" wissenschaftliche Erkenntnisse sind dank des rasanten Wissenszugewinns in der Genetik zunehmend bereits dann überholt, wenn die zugehörigen wissenschaftlichen Artikel im Druck erscheinen. Und „*Online Supplementary Information*" bietet vielfach die einzige Möglichkeit, genetische Daten in ihrer Komplexität und Vernetzung mit anderen Datensätzen adäquat abzubilden. Darüber hinaus scheitern Printmedien bereits an der Darstellung der schieren Menge der durch neue genetische Technologien erhobenen genetischen Daten: Als im Jahr 2000 die erste Draft-Sequenz des menschlichen Genoms verkündet wurde, waren Sequenzinformation einer Größenordnung von acht Milliarden Basenpaaren (bp) in drei öffentlich zugänglichen und regelmäßig aufeinander abgestimmten Genom-Sequenzdatenbanken hinterlegt: GenBank am NCBI, dem US National Center for Biotechnology Information; RIKEN, der DNA-Datenbank Japans; sowie der Nucleotid-Sequenzdatenbank der Europäischen Molekularbiologischen Laboratorien (EMBL). Aktuell, nur 10 Jahre später, ist die Datenmenge der drei erwähnten Datenbanken auf 270 Milliarden bp angewachsen, und die Datenmenge verdoppelt sich etwa alle 18 Monate. Die Menge der gegenwärtig auf öffentlichen Datenservern hinterlegten Sequenz-Rohdaten wird in der Größenordnung von 25×10^{12} bp angegeben, von denen bereits jetzt der Großteil „new-generation" Sequenzdaten individueller menschlicher Genome ausmachen [Nature Vol. 464, 01.04.2010]. Und die Speicherung von Datenmengen im Petabyte-Bereich (10^{15}), wie sie bei der kompletten Sequenzierung ganzer Patientenkollektive erwartet werden, sind von öffentlich-geförderten Konsortien bereits nicht mehr tragbar, so dass das Datenmanagement kommerziellen Anbietern übertragen wird.

Ist die Humangenetik also einerseits durch riesige, bislang weitgehend nur unzulänglich interpretierbare Sequenz-Datenmengen herausgefordert, hat sie andererseits mit der erheblichen klinischen und biologischen Vielfalt genetischer Krankheitsbilder zu kämpfen, deren komplexe Phänotypen sich einer standardisierten Beschreibung häufig entziehen. Die Anzahl individueller Datenbanken und Algorithmen, die sich beiden Ebenen und ihren Zusammenhängen widmen, ist kaum mehr überschaubar. Durch den Alltag am Computer hat sich die Arbeitsweise in der humangenetischen Beratung und Diagnostik grundlegend verändert: Die Anwendung bioinformatischer Datenbanken reicht von primär krankheitsbezogenen Literaturrecherchen über die Interpretation eigener Laborbefunde bis hin zu komplexen statistischen Analysen in Populationsdatensätzen. Der beschränkte Rahmen einer Ausgabe der Zeitschrift „medizinischegenetik" zum Schwerpunktthema Datenbanken kann daher keinen systematischen und erst recht keinen umfassenden Zugang zu diesem Themengebiet gewährleisten. Stattdessen haben sich die wissenschaftlichen Koordinatoren entschlossen, aktuelle Trends anhand einiger ausgewählter Beiträge von Arbeitsgruppen in Deutschland darzustellen. Darüber hinaus soll eine Toolbox mit einer reichhaltigen Linksammlung das Stöbern im Internet auf eigene Initiative stimulieren.

Bei einer anonymisierten Umfrage im Kollegenkreis des Instituts der Koordinatoren gab jeder der Teilnehmenden zu, humangenetisch-relevante Informationen mit großer Regelmäßigkeit zu „googlen". Auch wenn mehrere Studien zeigen, dass gezielte Google-Anfragen gerade bei seltenen Erkrankungen durchaus zur richtigen Diagnosestellung führen können [z.B. Tang und Ng, 2006 BMJ. 333:1143], basiert das Suchergebnis üblicherweise auf der Strategie, wo und wie man sucht – noch fragen Computer im Allgemeinen nicht nach. Während manchmal ein einzelnes „richtiges" Schlagwort zielsicher zum gewünschten Ergebnis führt, verliert sich der typische User regelmäßig in den Weiten des Netzes.

Im ersten Artikel dieses Hefts stellen Rommel und Kollegen daher eine Option zum „Wo" vor. **Orphanet** ist eine europäische Datenbank für seltene Krankheiten, zu der Partner aus über 30 Ländern Krankheits- und Patienteninformationen beitragen. Ärzte, Forscher und vor allem Patienten selbst können sich in verschiedenen Sprachen nicht nur über Details einer Vielzahl seltener Krankheitsbilder informieren, sondern werden darüber hinaus auf diagnostische Tests, Spe-

zialambulanzen, Selbsthilfegruppen oder verfügbare Therapieoptionen hingewiesen. Hinsichtlich des „Wie" dürfte sich die Google-Trefferquote bei Verwendung eines standardisierten Vokabulars zur präzisen Beschreibung humangenetischer Sachverhalte wesentlich erhöhen. Die systematische Erfassung eines komplexen Krankheits-Phänotyps stellt generell jedoch eine ungleich größere Herausforderung dar, als die Erstellung primärer Sequenzinformation. Über einen ersten Schritt dazu – der Erstellung einer gemeinsamen Sprache, der **Human Phenotype Ontology** – berichten Doelken und Kollegen in ihrem Artikel.

Biomaterial-Datenbanken – strukturierte, qualitativ hochwertige Sammlungen nicht nur virtueller Datensätze, sondern auch biologischer Materialien, möglichst gemeinsam mit Detailinformationen zum Phänotyp der zugehörigen Individuen, sind eine wichtige Quelle für die populationsgenetische und genetisch-epidemiologische Forschung. Krawczak und Kollegen geben in ihrem Beitrag einen Überblick und diskutieren die rechtlichen, ethischen und strukturellen Herausforderungen von Biobanken. Die langfristige Verfügbarkeit und die Bearbeitung auch zukünftiger Forschungsfragen erfordern spezifische datenschutzrechtliche Erwägungen, die im Voraus sicherstellen müssen, dass die Interessen aller Beteiligten gewahrt bleiben. Aktuell sind Biobanken die Grundlage vieler genom-weiter Assoziationsstudien (GWAS), deren unmittelbare Bedeutung für die genetische Beratung und Diagnostik gegenwärtig lebhaft diskutiert wird. So beziehen sich kommerzielle Anbieter bereits jetzt auf GWAS-Resultate, um Krankheitsprädispositionen zu diagnostizieren und Strategien zur Risikoreduktion zu empfehlen. Es ist daher davon auszugehen, dass die Interpretation von GWAS-Ergebnissen für die genetische Beratung zunehmende Relevanz erhalten wird. Eine systematische Auswertung von GWAS-Studien durch **GWAS-Metaanalysen** erlaubt nicht nur eine Objektivierung der tatsächlichen Signifikanz einzelner Loci, sondern könnte zur Verschmälerung der Lücke zwischen Grundlagenwissenschaft und klinischer Anwendung beitragen. Dies stellen Lill und Bertram in ihrem Artikel etwa am Beispiel der Alzheimerkrankheit auf beeindruckende Weise dar.

Es ist davon auszugehen, dass die Diskrepanz zwischen erhobener Datenmenge und ihrer Interpretierbarkeit durch die nächste Stufe der technologischen Entwicklung, der schnellen und kostengünstigen Hochdurchsatz-Sequenzierung individueller Genome, sogar noch zunehmen wird. Stütz und Korbel beschäftigen sich in ihrem Artikel zum **1000-Genome-Project** schwerpunktmäßig mit der informationstechnologischen Infrastruktur zur Archivierung riesiger Datenmengen und deren Bearbeitung, die bei der vollständigen Genomsequenzierung von bis zu 1000 Individuen anfallen – ein Projekt, dass bis 2011 weitgehend abgeschlossen sein soll. Die Autoren prognostizieren, dass sich die Arbeit der Humangenetiker durch zukünftig zu erwartende sehr viel umfangreichere Datenanalysen noch mehr als bisher wandeln wird.

Mögliche Szenarien, mit denen sich der Genetische Berater der (nahen) Zukunft auseinandersetzen zu haben dürfte, sind dem Beitrag von Krawitz zu entnehmen, in dessen Mittelpunkt die Bedeutung von **„Personal Genomics"** für die Prognose und zukünftige Behandlung des individualisierten Patienten steht. Während eine maßgeschneiderte Therapie von Volkskrankheiten gegenwärtig wohl noch als Zukunftsmusik erachtet werden kann, zeigen einige Beispiele aus der Pharmakogenomik, dass bestimmte Genvarianten große Effekte auf Medikamentenwirkungen haben können, so dass individualisierte Genominformation durchaus von Nutzen sein kann.

Während das typische humangenetische Institut der Gegenwart von solchen Fragestellungen zumindest wohl noch eine Zeit lang verschont bleiben wird, gehört das informationstechnologische Management erhobener und gespeicherter Daten in jeder genetischen Beratungsstelle und Labor zum Alltag. Die abschließenden Artikel von Schlott/Schröck und Schröder/Müller-Reible zeigen am Beispiel zweier ausgewählter **Patienten- und Labordaten-Managementsysteme** wie Patienten-, Familien- und Labordaten in humangenetischen Praxen und Einrichtungen verwaltet werden können. Beide vorgestellten Systeme wurden im akademischen Umfeld aus der Praxis heraus zum Management der eigenen Daten entwickelt und sind daher direkt auf die Erfordernisse der gemeinsamen Verwaltung von Prozessen wie Genetische Beratung, Probeneingang, Analyseverlauf, Ergebnisdokumentation, Befunderstellung und Abrechnung zugeschnitten. Bei konsequenter Umsetzung der verfügbaren Möglichkeiten lässt sich durch den Einsatz derartiger Programme sehr viel doppelte Datenführung vermeiden, was zu einem weitgehend papierlosen und insgesamt sehr viel effektiveren Arbeiten beitragen sollte.

Korrespondenzadresse

H. Runz, Ch. Fischer
Institut für Humangenetik
INF 366
Universität Heidelberg
69120 Heidelberg
heiko.runz@med.uni-heidelberg.de

Datenbanken-Toolbox

(Stand: 04/2010)

Die nachfolgende Toolbox listet Links zu einer Auswahl an online frei verfügbaren (open-source) Datenbanken, die Informationen und Analyse-Tools für häufige Fragestellungen in der Humangenetik anbieten. Die Auswahl erhebt keinen Anspruch auf Vollständigkeit, und im schnelllebigen World Wide Web kann eine Garantie weder für die Qualität des Seiteninhalts, noch für die Aktualität der aufgelisteten Internet-Adressen übernommen werden. Die Liste richtet sich an den „typischen" humangenetischen User, der als Anwender Informationen aus bestehenden Datensätzen zusammentragen oder eigene kleinere Datensätze (z.B. mögliche funktionelle Konsequenzen einer bestimmten DNA-Sequenzvariante in einem einzelnen Gen) selbst analysieren möchte. Von der Auflistung der zahlreichen und häufig exzellenten Genlokus-spezifischen Datenbanken wurde ebenso abgesehen wie von der Anführung von Tools (z.B. zum Sequenz-Alignment) die wesentliche bioinformatische Vorkenntnisse voraussetzen. Für die Mehrzahl der Seiten sind ausführliche Online-Tutorials verfügbar, die eine schnelle Einführung in deren adäquate Nutzung erlauben. Die aufgeführte Liste ist unter der URL http://www.gfhev.de/de/links/fachinformationen.htm auch online verfügbar. Verbesserungsvorschläge und Tipps zur Aktualisierung und Ergänzung weiterer relevanter Links sind per Email an organisation@gfhev.de herzlich willkommen. Viel Erfolg beim Surfen!

1. Literatur- und Meta-Suchmaschinen:

Google	http://www.google.com
Google Scholar	http://scholar.google.com
Google Books	http://books.google.com
NCBI PubMed	http://www.ncbi.nlm.nih.gov/PubMed
Gopubmed	http://www.gopubmed.com

2. Datenbanken mit Krankheitsinformationen/ Relevanz für die klinische Genetik:

OMIM	http://www.ncbi.nlm.nih.gov/omim
GeneReviews	http://www.ncbi.nlm.nih.gov/sites/GeneTests
Orphanet	http://www.orpha.net/consor/cgi-bin/index.php
NORD	http://www.rarediseases.org
EURORDIS	http://www.eurordis.org
Arzneimittelwirkungen	http://www.arzneimittel-in-der-Schwangerschaft.de

3. Datenbanken mit Relevanz für die klassische Zytogenetik / FISH:

Mendelian Cytogenetics Network	http://www.mcndb.org
ENSEMBL Cytoview	http://www.ensembl.org/Homo_sapiens/cytoview
Zytogenetik-Atlas (Onkologie)	http://atlasgeneticsoncology.org/
Mitelman-db (Onkologie)	http://cgap.nci.nih.gov/Chromosomes/Mitelman
SKY/M-FISH, CGH-db (Onkologie)	http://www.ncbi.nlm.nih.gov/projects/sky
Progenetix (Onkologie)	http://www.progenetix.net/progenetix/index.html
Array-CGH Tumor-db (Onkologie)	http://amba.charite.de/~ksch/cghdatabase/index.htm

4. Humanes Genom Annotations-Browser / Analyse-Tools:

Ensembl	http://www.ensembl.org/index.html
UCSC Genome Browser	http://genome.ucsc.edu/
NCBI Map Viewer	http://www.ncbi.nlm.nih.gov/projects/mapview
Galaxy Genome Software	http://bitbucket.org/galaxy/galaxy-central/wiki/Home
Biomart	http://www.biomart.org
DAS	http://biodas.org

5. Gen-basierte Datenbanken/ Meta-Suchmaschinen:

HGNC	http://www.genenames.org
NCBI Entrez Gene	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
GeneCards	http://www.genecards.org
Bioinformatic Harvester	http://harvester.fzk.de
Genedistiller	http://www.genedistiller.org
iHOP	http://www.ihop-net.org/UniPub/iHOP/

6. Genom-Variations Datenbanken / Genotyp-Phänotyp-Korrelation:

NCBI dbSNP	http://www.ncbi.nlm.nih.gov/SNP
Hapmap	http://hapmap.ncbi.nlm.nih.gov/
Database of Genomic Variants	http://projects.tcag.ca/variation
Decipher	https://decipher.sanger.ac.uk/application
Segmental duplication db	http://humanparalogy.gs.washington.edu
NCBI dbGAP	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap
NCBI Genetic Association db	http://geneticassociationdb.nih.gov
Gen2Phen Knowledge Center	http://www.gen2phen.org/
European Genome-Phenome Archive	http://www.ebi.ac.uk/ega
Welcome Trust Case Control Consortium	http://www.wtcc.org.uk
Human Epigenome Project	http://www.epigenome.org
Human Gene Mutation Database	http://www.hgmd.cf.ac.uk/ac/index.php
HGVS (Nomenklatur)	http://www.hgvs.org/mutnomen
Mutalyzer (Nomenklatur)	http://www.LOVD.nl/mutalyzer
Locus reference Genomic (Nomenklatur)	http://www.lrg-sequence.org

7. Gensequenzanalyse-Tools:

7.1. Sequenzvarianten-Interpretations-Tools:

Polyphen2	http://genetics.bwh.harvard.edu/pph2
SIFT	http://sift.jcvi.org/
Mutation Taster	http://neurocore.charite.de/MutationTaster/
SNPS3D	http://www.snps3d.org
SNAP	http://www.rostlab.org/services/SNAP
Pmut	http://mmb2.pcb.ub.es:8080/PMut
AGVGD	http://agvgd.iarc.fr/index.php
Modbase/LS-SNP	http://modbase.compbio.ucsf.edu/LS-SNP//About.html
SNPeffect	http://snpeffect.vib.be/index.php
TopoSNP	http://gila-fw.bioengr.uic.edu/snp/toposnp
MutDB	http://www.mutdb.org
pupaSNP	http://pupasnp.org
FastSNP	http://fastsnp.ibms.sinica.edu.tw
ssahaSNP	http://www.sanger.ac.uk/resources/software/ssahasnp/

7.2. SpliceSite-Vorhersage-Tools:

HSF	http://www.umd.be/HSF/
ACESCAN	http://genes.mit.edu/acescan/
Splice Site Prediction (D. melanogaster)	http://www.fruitfly.org/seq_tools/splice.html

7.3. DNA-Sequenzmotiv-Vorhersage-Tools:

TFSEARCH	http://www.cbrc.jp/research/db/TFSEARCH.html
JASPAR	http://jaspar.cgb.ki.se
MEME	http://meme.sdsc.edu/meme
TRAWLER	http://ani.embl.de/trawler
WEEDER	http://159.149.109.9/modtools/

8. ausgewählte Gen-/ Proteinfunktions Datenbanken:

GeneOntology	http://www.geneontology.org
SRS 3d	http://srs3d.org
Pfam	http://pfam.sanger.ac.uk/
Expasy	http://www.expasy.ch
PDB	http://www.pdb.org
FOLDX	http://foldx.crg.es/about.jsp
TOPO2	http://www.sacs.ucsf.edu/TOPO2
KEGG pathway analysis	http://www.genome.ad.jp/kegg
Proteinatlas	http://www.proteinatlas.org
4Dxpress	http://4dx.embl.de/4DXpress/welcome.do
Mitocheck	http://www.mitocheck.org/

Sammlung von Programmen zur statistischen Analyse von genetischen Daten

incl. Stammbaumzeichenprogramme:	http://linkage.rockefeller.edu/soft/