

# Human Phenotype Ontology: Identifying and Studying Rare Diseases

TMF-Workshop: Registries for patients with undiagnosed rare diseases

Peter N. Robinson

Professor for Medical Genomics  
Institut für medizinische Genetik und Humangenetik  
Charité Universitätsmedizin Berlin

A **subjective list** of goals for improving RD patient care

- 1 Reliably identify pathogenicity of variants in known disease genes
- 2 Quickly identify remaining Mendelian disease genes
- 3 Differential diagnosis and clinical decision support system
- 4 Characterize natural history of RDs and discover clinically actionable complications and risks
- 5 Basis to include clinical aspects in integrative basic science research on disease pathophysiology

# Deep Phenotyping

- the precise and comprehensive analysis of phenotypic abnormalities
- the individual **components** of the phenotype are observed and described
- often for the purposes of scientific examination of human disease.

PN Robinson (2012) Deep phenotyping for precision medicine.  
*Hum Mutat* **33**: 777–780

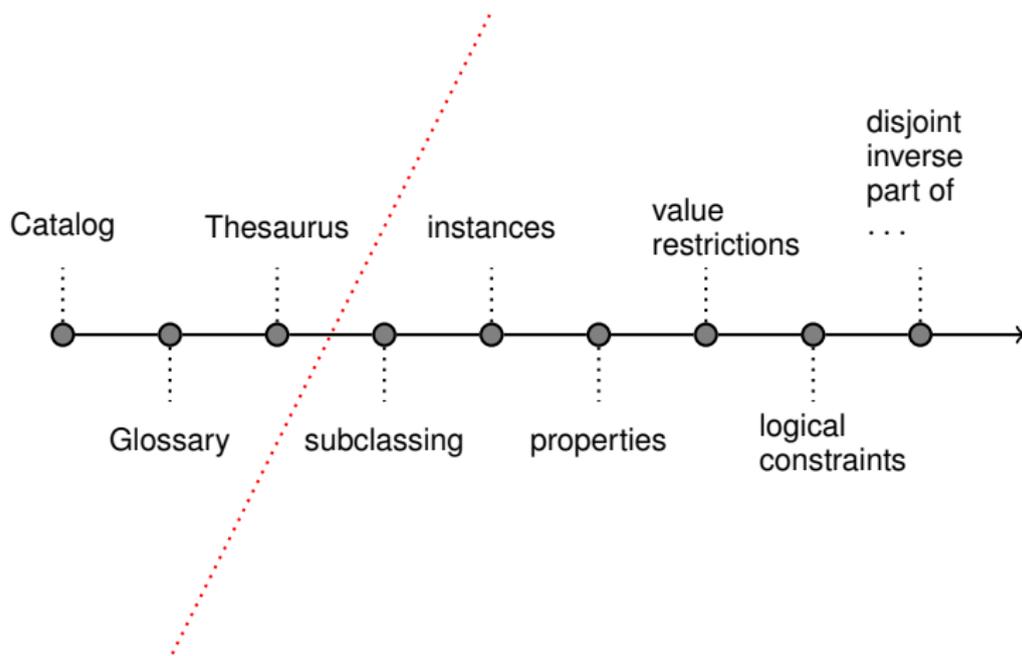
Special Issue of *Human Mutation* on Deep Phenotyping



# What is an Ontology?

“An ontology is a specification of a conceptualization.”

– Tom Gruber, 1993

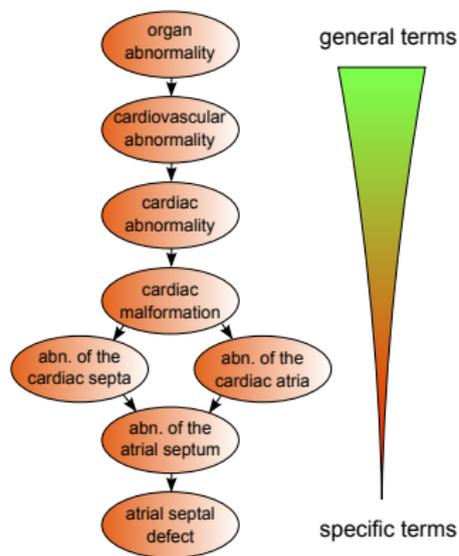


# What is a phenotype ontology?

A phenotype ontology describes not diseases but the individual manifestations of a disease

- symptoms
- signs
- laboratory abnormalities
- anomalies found by imaging studies
- behavioral manifestations

# The Human Phenotype Ontology



- ~ 10, 143 terms, ~ 110, 000 annotations for ~ 7000 mainly monogenic diseases

- <http://www.human-phenotype-ontology.org>

- Robinson PN et al. (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet.* **83**:610–5.

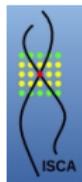
# Uptake in community

## Databases & Bioinformatics Resources Using HPO

DECIPHER (Sanger Institute)  
DDD (Sanger Institute)  
ECARUCA  
FORGE (Genome Canada)  
GWAS Central  
IRDiRC  
ISCA  
NCBI Genetic Testing Registry  
NIH Undiagnosed diseases program  
PhenomeNET  
RIKEN

...

Close integration with other important efforts



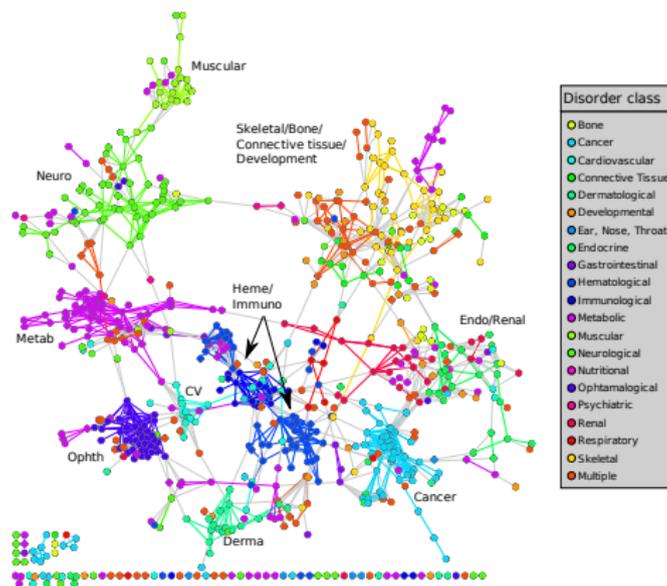
OMIM

orphanet

The screenshot shows a complex web interface for clinical data search. It features a search bar at the top and several filter panels. The main panel lists categories such as 'BEHAVIOR, COGNITION AND DEVELOPMENT', 'NEUROLOGICAL', 'GENETIC DIAGNOSTICS', 'CARDIAC', and 'COMMERICAL'. A 'CURRENT SELECTION' panel on the right highlights 'Delayed gross motor development' and 'Neurological' categories. Below these, there are sections for 'Age of onset', 'Place of progression', 'Clinical course', and 'Medical report (pathology)'. A small image of a brain scan is visible in the 'Medical report (pathology)' section.

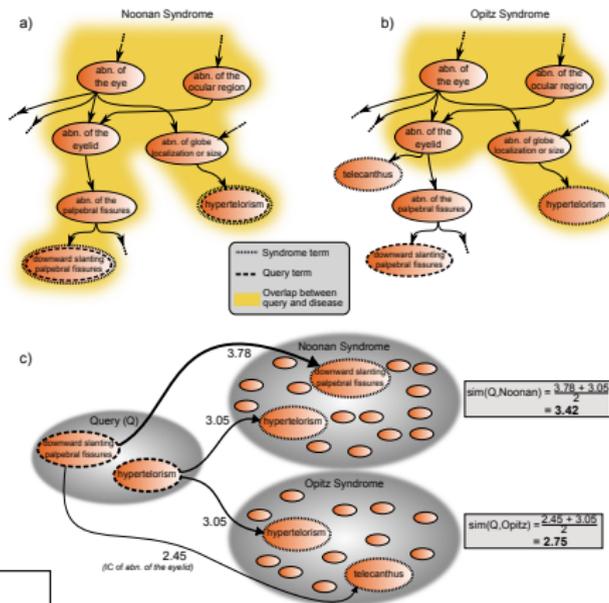
Phenotips (Brudno group, U Toronto)

# The Human Phenome: Network of Human Diseases and Disease Genes



$$\text{sim}(d_1, d_2) = 0.5 \cdot \text{avg} \left[ \sum_{s \in d_1} \max_{t \in d_2} \text{sim}(s, t) \right] + 0.5 \cdot \text{avg} \left[ \sum_{s \in d_2} \max_{t \in d_1} \text{sim}(s, t) \right]$$

# Ontological diagnostics



Q: Query terms

d: Disease terms

$$\text{sim}(Q \rightarrow d) = \text{avg} \left[ \sum_{s \in Q} \max_{t \in d} \text{sim}(s, t) \right]$$

# The Phenomizer

The Phenomizer interface is divided into two main panels. The left panel, titled 'Features', shows search results for 'SCOLIOSIS'. The right panel, titled 'Diagnosis', shows a list of differential diagnoses for the patient's features.

**Features Panel:**

Search term: SCOLIOSIS

HPO id.	Feature.
HP:0008453	CONGENITAL KYPHOSCOLIOSIS
HP:0008458	CONGENITAL SCOLIOSIS, PROGRESSIVE
HP:0002751	KYPHOSCOLIOSIS
HP:0003412	KYPHOSCOLIOSIS MAY OCCUR
HP:0004619	LUMBAR KYPHOSCOLIOSIS
HP:0004626	LUMBAR SCOLIOSIS
HP:0003303	MILD SCOLIOSIS
HP:0004615	MILD THORACIC SCOLIOSIS
HP:0004585	MILD THORACOLUMBAR SCOLIOSIS
HP:0003424	PROGRESSIVE KYPHOSCOLIOSIS
HP:0003317	PROGRESSIVE SCOLIOSIS
HP:0002650	SCOLIOSIS
HP:0004567	SCOLIOSIS, THORACOLUMBAR, SEVERE, PROGRESSIVE
HP:0002770	SEVERE SCOLIOSIS
HP:0004593	SEVERE, PROGRESSIVE KYPHOSCOLIOSIS

Page 1 of 2

Features 1 - 15 of 20

**Diagnosis Panel:**

Algorithm: resnik (Symmetric). 6 Features.

p-value	OMIM name	Genes
<input checked="" type="checkbox"/> 0.0095	LOEYS-DIETZ SYNDROME, TYPE 1A	TGFBR1
<input checked="" type="checkbox"/> 0.2386	MARFANOID HYPERMOBILITY SYNDROME	
<input checked="" type="checkbox"/> 0.3244	MENTAL RETARDATION, X-LINKED, SNYDER-ROBINSON TYPE	SMS
<input checked="" type="checkbox"/> 0.3356	HOMOCYSTINURIA	CBS
<input checked="" type="checkbox"/> 0.3356	MENTAL RETARDATION, X-LINKED, SYNDROMIC 14	UPF3B
<input checked="" type="checkbox"/> 0.3356	BRACHIOSKELETOGENITAL SYNDROME	
<input type="checkbox"/> 0.4783	PECTUS EXCAVATUM	
<input type="checkbox"/> 0.4783	CAMPTODACTYLY WITH FIBROUS TISSUE HYPERPLASIA AND SI	
<input type="checkbox"/> 0.5216	PECTUS EXCAVATUM, MACROCEPHALY, SHORT STATURE, DYS	
<input type="checkbox"/> 0.5277	MARFANOID HABITUS WITH SITUS INVERSUS	
<input type="checkbox"/> 0.5786	SHPRINTZEN-GOLDBERG CRANIOSYNOSTOSIS SYNDROME	FBN1
<input type="checkbox"/> 0.8492	ARTERIAL TORTUOSITY SYNDROME	SLC2A10
<input type="checkbox"/> 0.9119	UVULA, BIFID	
<input type="checkbox"/> 0.9119	CEREBRAL AMYLOID ANGIOPATHY, APP-RELATED	APP
<input type="checkbox"/> 0.9154	CONTRACTURAL ARACHNOACTYLY, CONGENITAL	FBN2

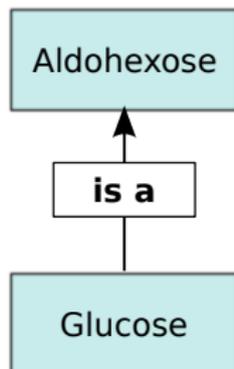
Page 1 of 160

Improve Differential Diagnosis. Download Results

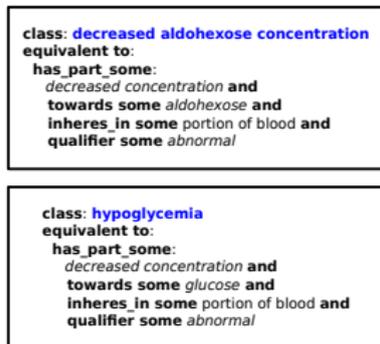
- Sebastian Köhler et al. (2009) Clinical Diagnostics with Semantic Similarity Searches in Ontologies. *Am J Hum Genet*, **85**:457–64.

# Reasoning over Phenotypes

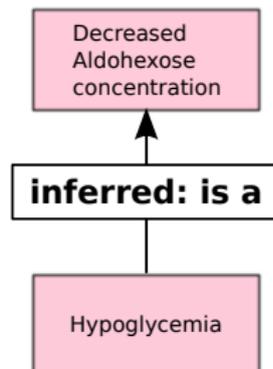
## Building Block Ontology (ChEBI)



## Logical Definitions



## Inference



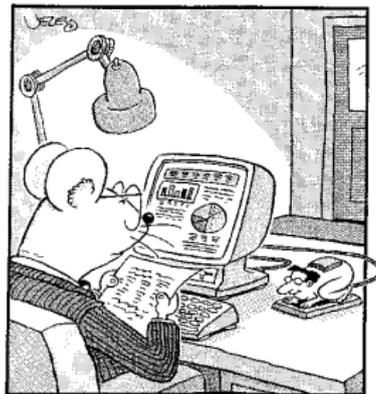
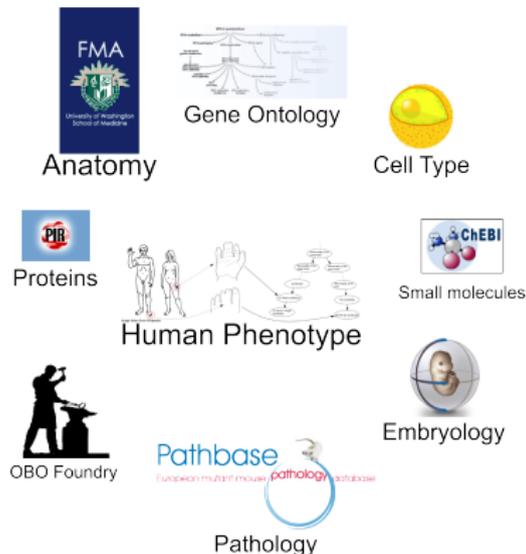
Köhler et al (2013) *F1000Research* 2:30

Gkoutos GV et al. (2009) Entity/Quality-Based Logical Definitions for the Human Skeletal Phenome using PATO. *IEEE Engineering in Medicine and Biology* (EMBC 2009)

Köhler S et al. (2011) Improving ontologies by automatic reasoning and evaluation of logical definitions.

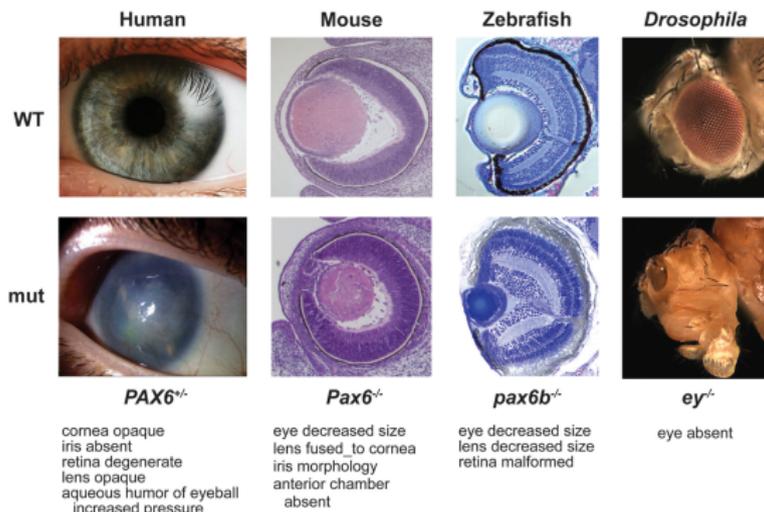
*BMC Bioinformatics* 12:418.

# A semantic web of the human phenotype



- HPO defined using 11 other ontologies  $\Rightarrow$  Semantic network
- Connections to genes, diseases, anatomy, . . .
- Cross phenotype species analysis

# Of mice, fish, flies, and men



- Mutations in orthologous genes are often associated with similar phenotypes
- Phenotype information for >5300 genes currently available **only** in model organisms

# Finding the needle: 3,000,000,000 bases → 1 mutation

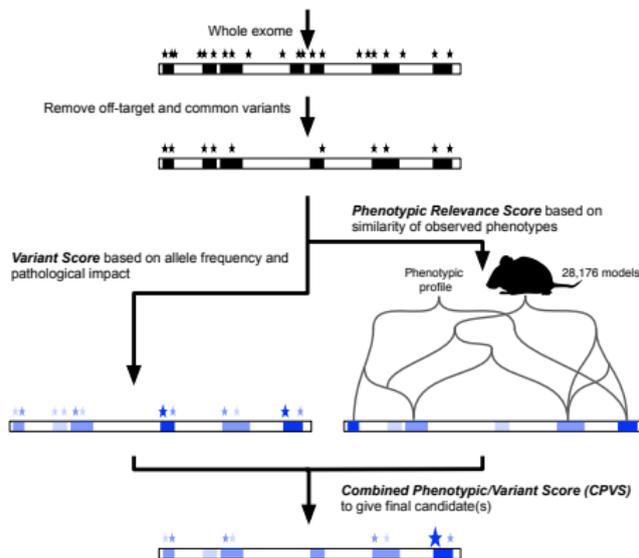
	X-ome	Exome	Genome
SNVs	800–1200	20,000–30,000	2–3 Mio.
↪ dbSNP	100–300	1,000–3,000	100K–300K
Indels (<10bp)	100–200	3,000	600K
↪ dbSNP	50	1,500	150K

- Each genome: Lots of “private” variants with ~ 100 genuine loss of function variants with ~20 genes completely inactivated.
- ∴ sequence-based prioritization alone will struggle to identify the disease-associated mutation

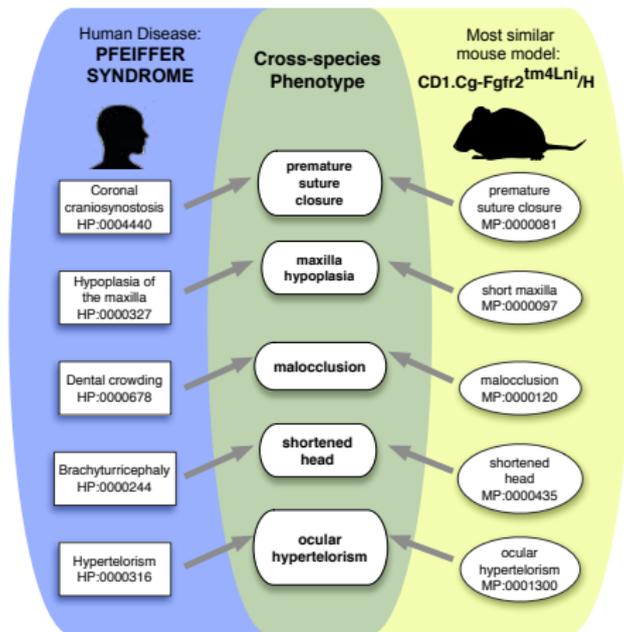


# Exome Sequencing

- at least 30,000 variants per exome
- We all have ~ 100 genuine loss of function variants
- Prioritization based purely on sequence variant pathogenicity will struggle to correctly identify the disease-associated mutation from other variants with a deleterious biochemical effect.



*The Exomiser*



## ● PHIVE: *P*henotypic Interpretation of Variants in Exomes

The Exomizer: Annotate and Filter Variants

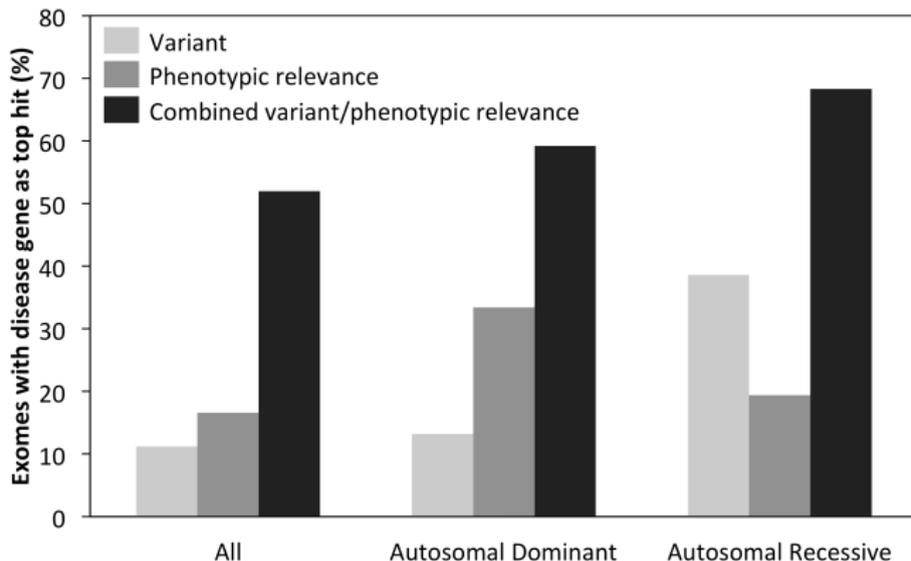
New Query

Filtering summary Distribution of variants Prioritized gene/variant list

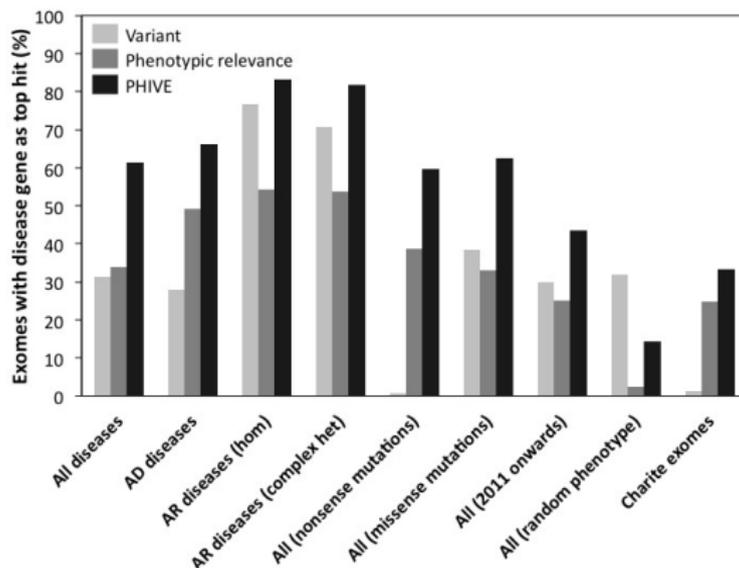
227 entries 1 2 3 ... 23 10 per page TXT CSV XLS

Gene	Score	Variant	Variant analysis	Mouse Phenodigm score	Phenotypic analysis
<a href="#">FGR2</a>	1.81	chr10:g.1232562157>G [Het] <a href="#">View in UCSC Browser</a> uc001bn.4 : c.518A>C (p.E173A; exon6) uc010qtm.2 : c.1343A>C (p.E448A; exon11) uc010qtl.2 : c.1346A>C (p.E449A; exon11) uc021qzw.1 : c.1349A>C (p.E450A; exon10) uc021qzv.1 : c.1358A>C (p.E453A; exon10) uc021qzx.1 : c.1427A>C (p.E476A; exon12) uc021qyb.1 : c.1430A>C (p.E477A; exon12) uc021qzc.1 : c.1481A>C (p.E494A; exon12) uc021qze.1 : c.1694A>C (p.E565A; exon13) uc021qsf.1 : c.1697A>C (p.E566A; exon12) uc021qsv.1 : c.1697A>C (p.E566A; exon13)	Pathogenicity: <b>Missense</b> Mutation Tester: 0.999 (P) Polyphen2: 0.998 (D) SIFT: 0.000 (D) Path score: 1.000 Frequency <a href="#">rs121918506</a> (no frequency data)	0.81	Mouse phenotype data for <a href="#">Fgr2</a> OMIM: Antley-Bixler syndrome without genital anomalies or abnormal steroidogenesis [MIM:207410]; gene: MIM:176943 OMIM: Aeger syndrome [MIM:101200]; gene: MIM:176943 OMIM: Beare-Stevenson cutis gyrata syndrome [MIM:123748]; gene: MIM:176943 OMIM: Bent bone dysplasia syndrome [MIM:614592]; gene: MIM:176943 OMIM: Craniofacial-skeletal-dermatologic dysplasia [MIM-10]; gene: MIM:176943 OMIM: Crouzon syndrome [MIM:123500]; gene: MIM:176943 OMIM: Gastric cancer, somatic [MIM:137215]; gene: MIM:176943 OMIM: Jackson-Watts syndrome [MIM:123150]; gene: MIM:176943 OMIM: LADD syndrome [MIM:149730]; gene: MIM:176943 OMIM: Pfeiffer syndrome [MIM:101600]; gene: MIM:176943 OMIM: Saethre-Chotzen syndrome [MIM:101400]; gene: MIM:176943 OMIM: Scaphocephaly, maxillary retrusion, and mental retardation [MIM:609579]; gene: MIM:176943 Mouse phenotype data for <a href="#">Apar1</a> No OMIM disease entry
<a href="#">APAF1</a>	1.62	chr12:g.99053109G>A [Het] <a href="#">View in UCSC Browser</a>	Pathogenicity: <b>Missense</b> Mutation Tester: 1.000 (P)	0.72	

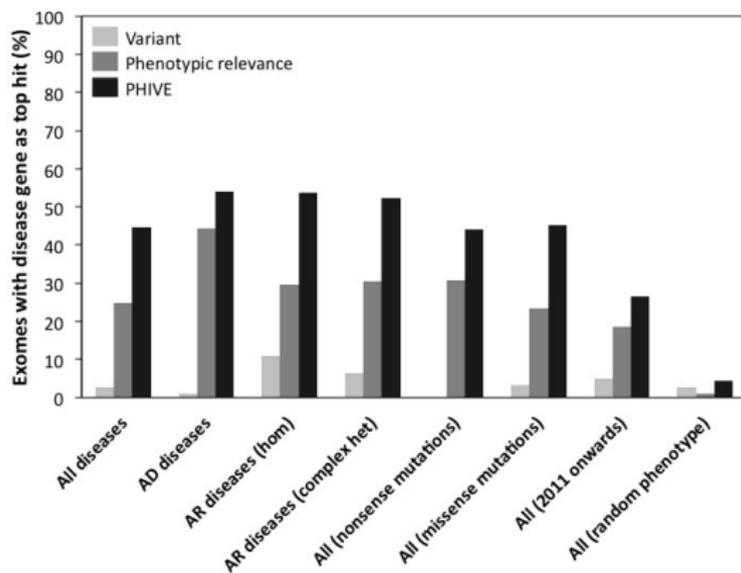
- Users enter VCF file and phenotype data and get back ranked list of candidates



- Phenotype data and variant data **synergistically** improve interpretation of exome data for gene discovery



- With ESP and 1000G Frequency data



- With frequency data only from ESP to remove any potential bias due to the non-causative variants also coming from the 1000 Genomes Project

# Conclusions

The Exomiser: Annotate and Filter Variants

The Exomiser is a Java program that functionally annotates variants from whole-exome sequencing data starting from a VCF file (version 4). The functional annotation code is based on [Annotovar](#) and uses [UCSC](#), [KnownGene](#) transcript definitions and hg19 genomic coordinates.

Variants are prioritized according to user-defined criteria on variant frequency, pathogenicity, quality, inheritance pattern, and model organism phenotype data. Predicted pathogenicity data was extracted from the [dbSNP](#) resource. Cross-species phenotype comparisons come from our [DerezoGen](#) tool powered by the [OASim](#) algorithm.

The Exomiser was developed by the Computational Biology and Bioinformatics group at the [Institute for Medical Genetics and Human Genetics](#) of the [Charité - Universitätsmedizin Berlin](#), the Mouse Informatics Group at the [Sanger Institute](#) and the [Lewin group](#) at the Lawrence Berkeley National Labs.

**Upload exome sequencing results in VCF format (single sample only)** Required

Associated Mendelian disease  
or Clinical phenotype(s) (HPO terms)

Minimum variant call quality e.g. 30.0

Maximum minor allele frequency (%) Required

Remove all dbSNP variants  No  Yes

Remove non-pathogenic variants  Yes  No

Inheritance model  Autosomal dominant  Autosomal recessive

- Large-scale validation of PHIVE analysis using 100,000 exomes containing known mutations demonstrated an improvement of up to 54.1 fold over purely variant-based methods with the correct gene recalled as the top hit in up to 83% of samples.
- Robinson PN, Köhler S, Oellrich A, Sanger Mouse Genetics Project, Wang K, Mungall C, Washington N, Bauer S, Seelow D, Krawitz P, Gillesen C, Haendel M, Smedley D (2013) Improved exome prioritization of disease genes through cross species phenotype comparison. *Genome Research*, early access, pmid: 24162188
- <https://www.sanger.ac.uk/resources/databases/exomiser/>

# Any Questions?

