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

Deep Phenotyping and Clinical Research

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

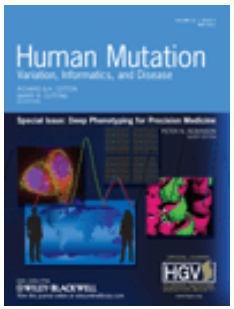
1. Reliably identify pathogenicity of variants in known disease genes
2. Quickly identify remaining Mendelian disease genes
3. Differential diagnosis and clinical decision support system
4. Characterize natural history of RDs and discover clinically actionable complications and risks
5. Basis to include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

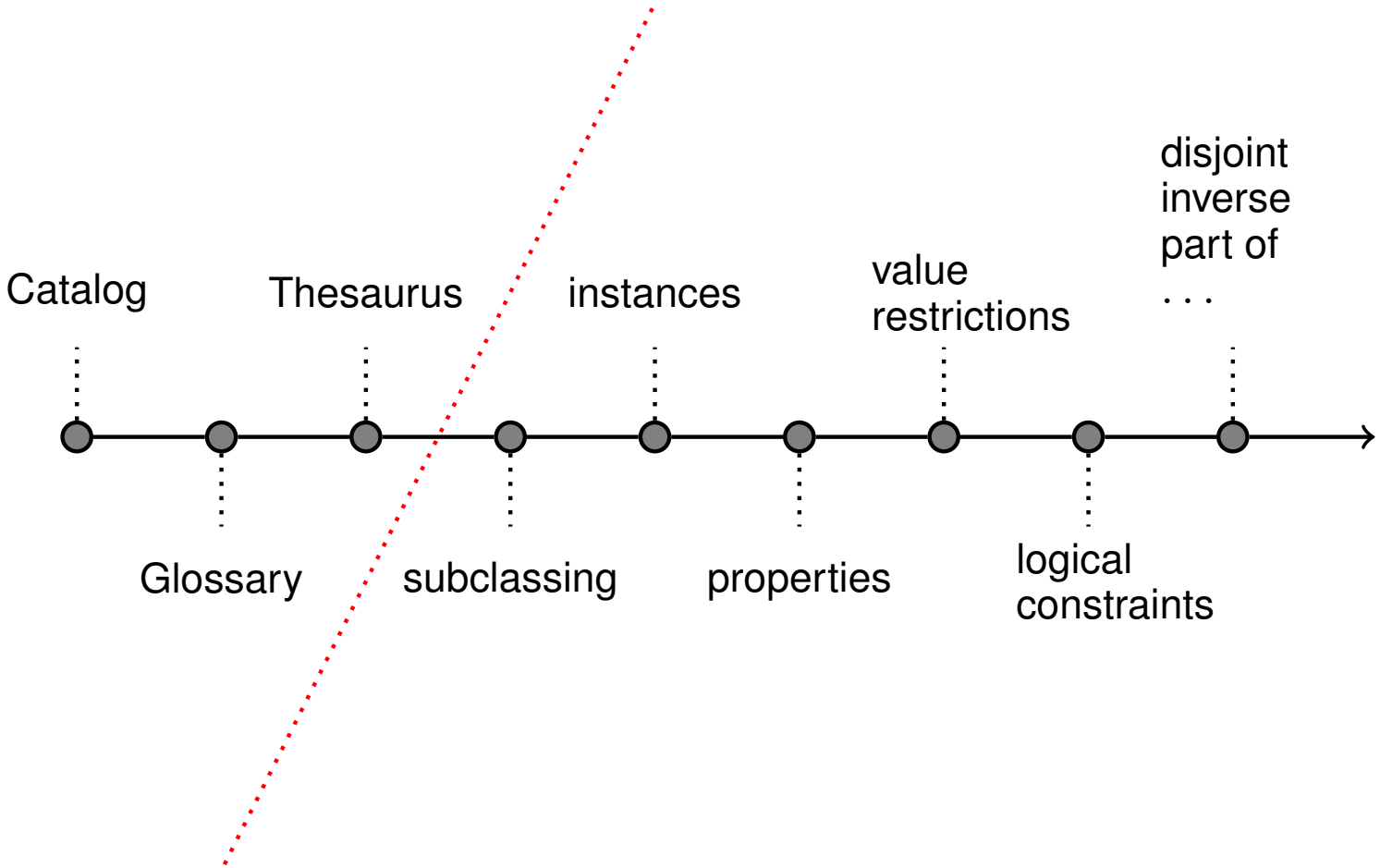
Special Issue of *Human Mutation* on Deep Phenotyping



What is an Ontology?

"An ontology is a specification of a conceptualization."

– Tom Gruber, 1993

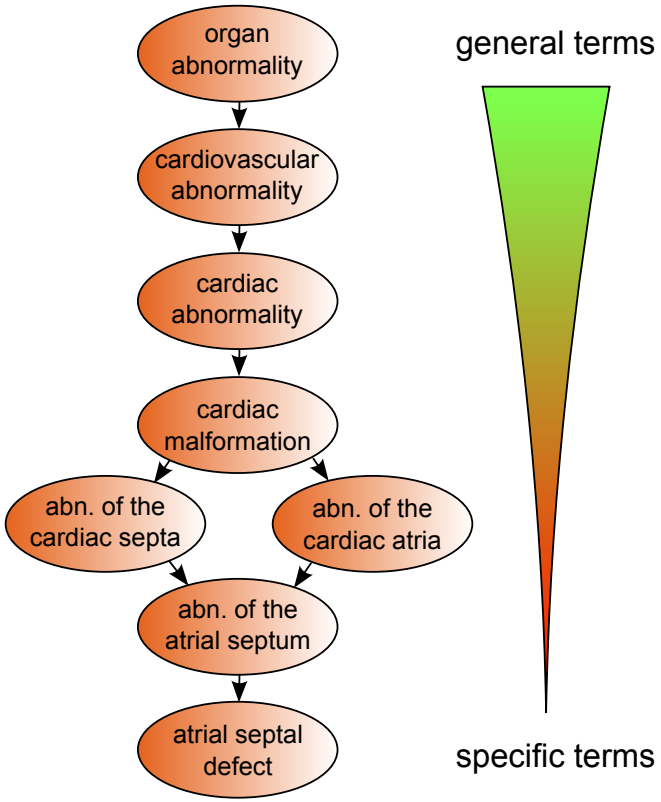


What is a phenotype ontology?

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



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

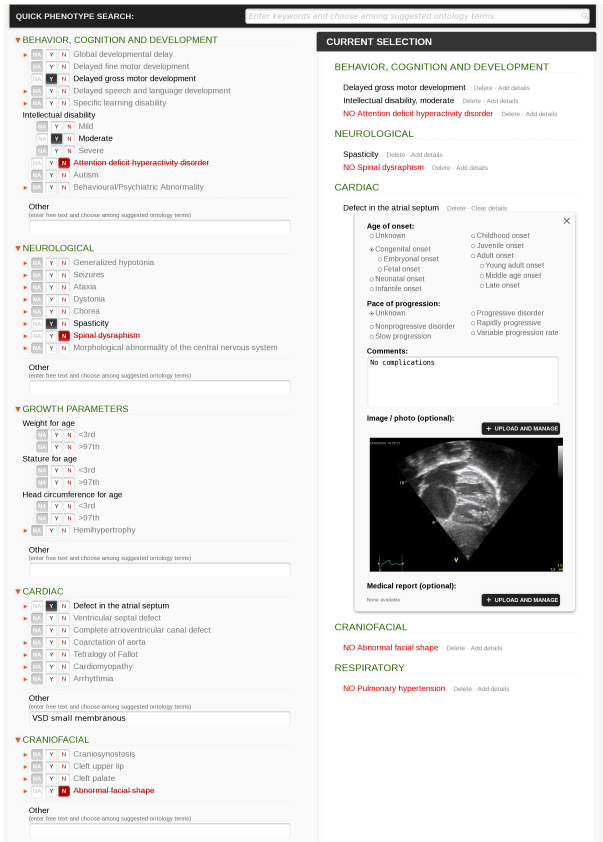
- http://www.human-phenotype-ontology.org
- Robinson PN et al. (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*. **83**:610–5.

Uptake in community

Databases & Bioinformatics Resources Using HPO

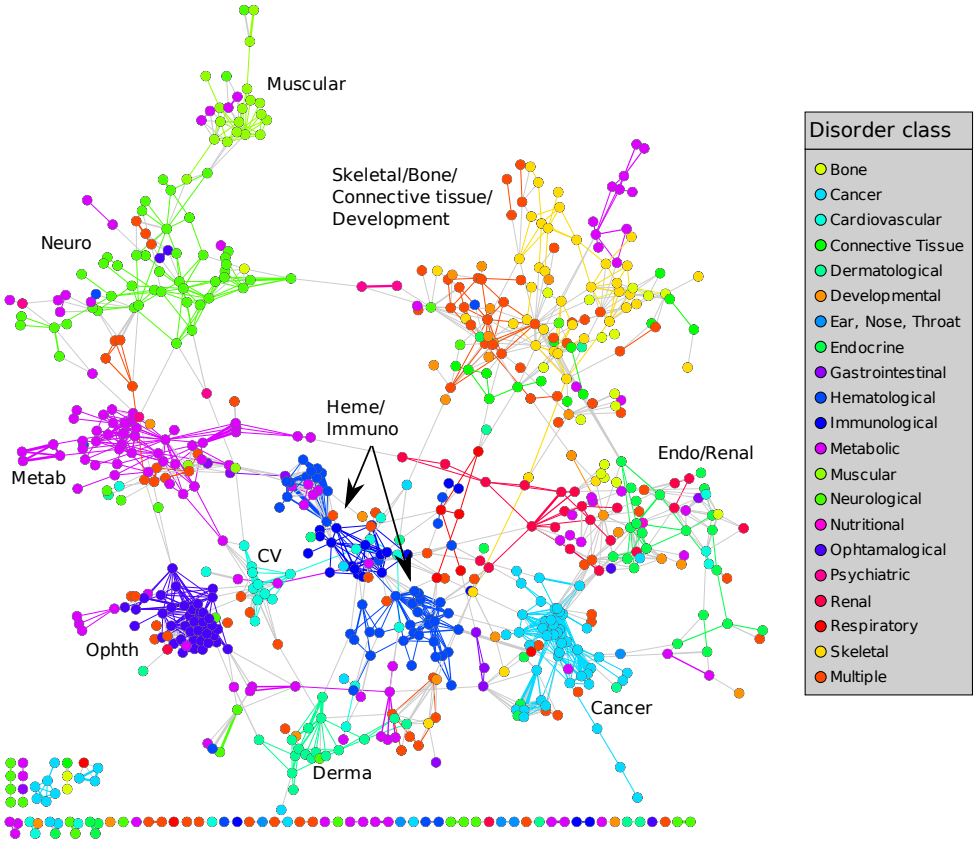
DECIPHER (Sanger Institute)
DDD (Sanger Institute)
ECARUCA
FORGE (Genome Canada)
GWAS Central
IRDiRC
ISCA
NCBI Genetic Testing Registry
NIH Undiagnosed diseases program
PhenomeNET
RIKEN
. . .

Close integration with other important efforts



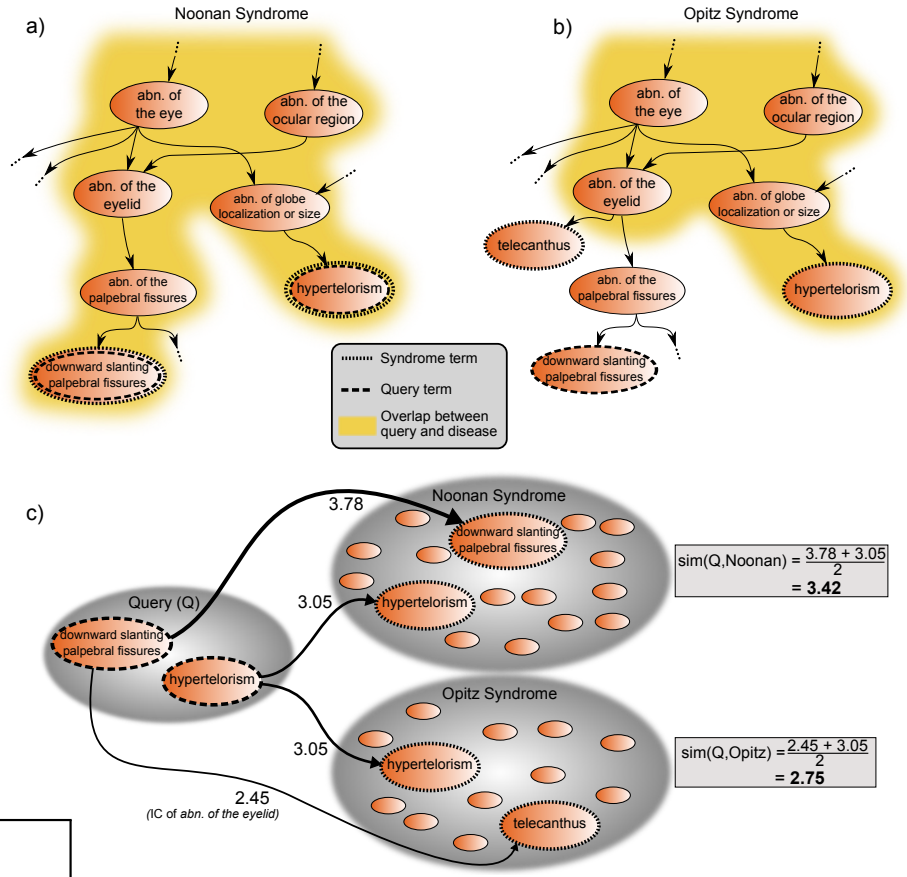
Phenotips (Brudno group, U Toronto)

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

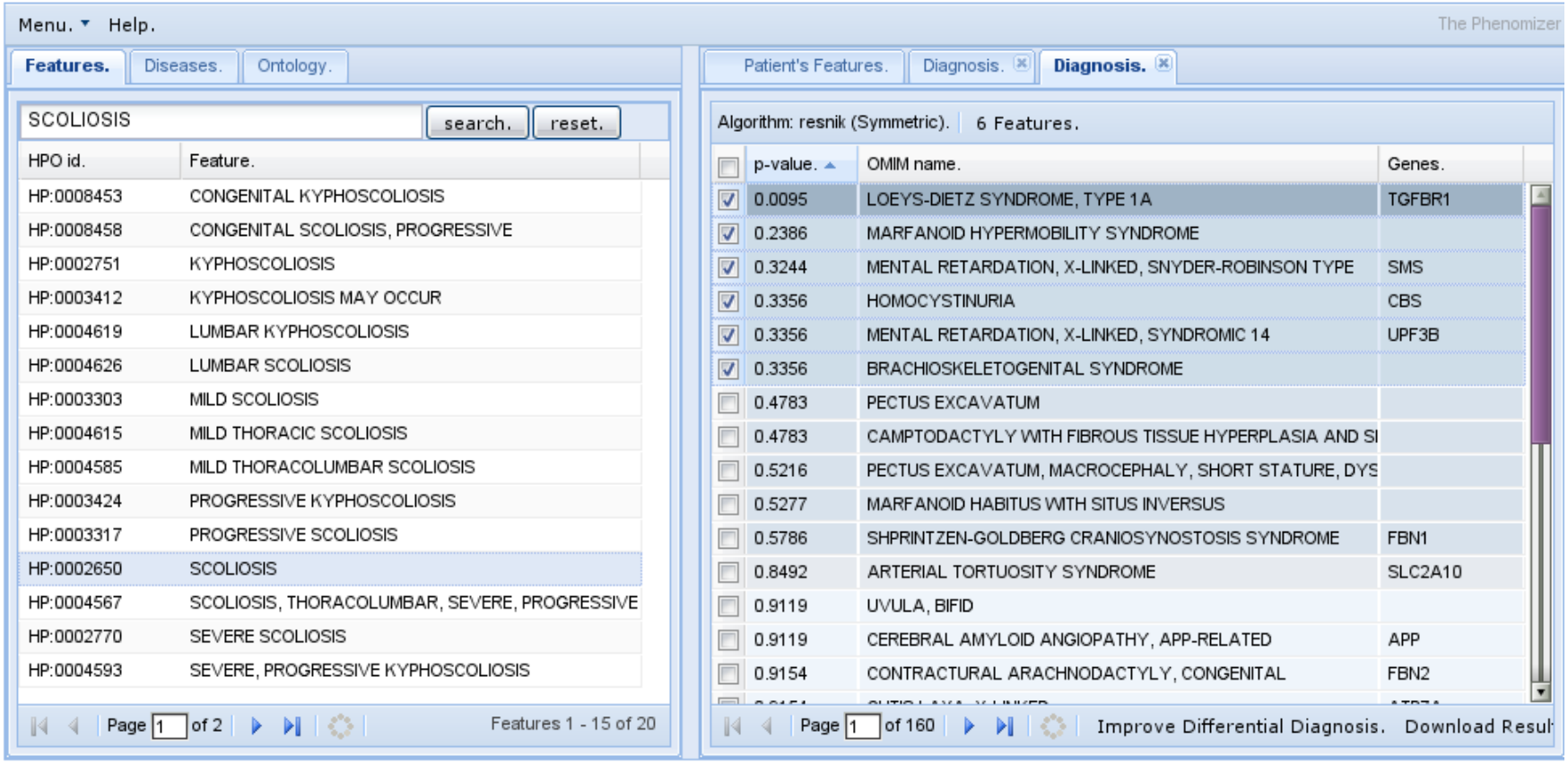


Q: Query terms

d: Disease terms

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

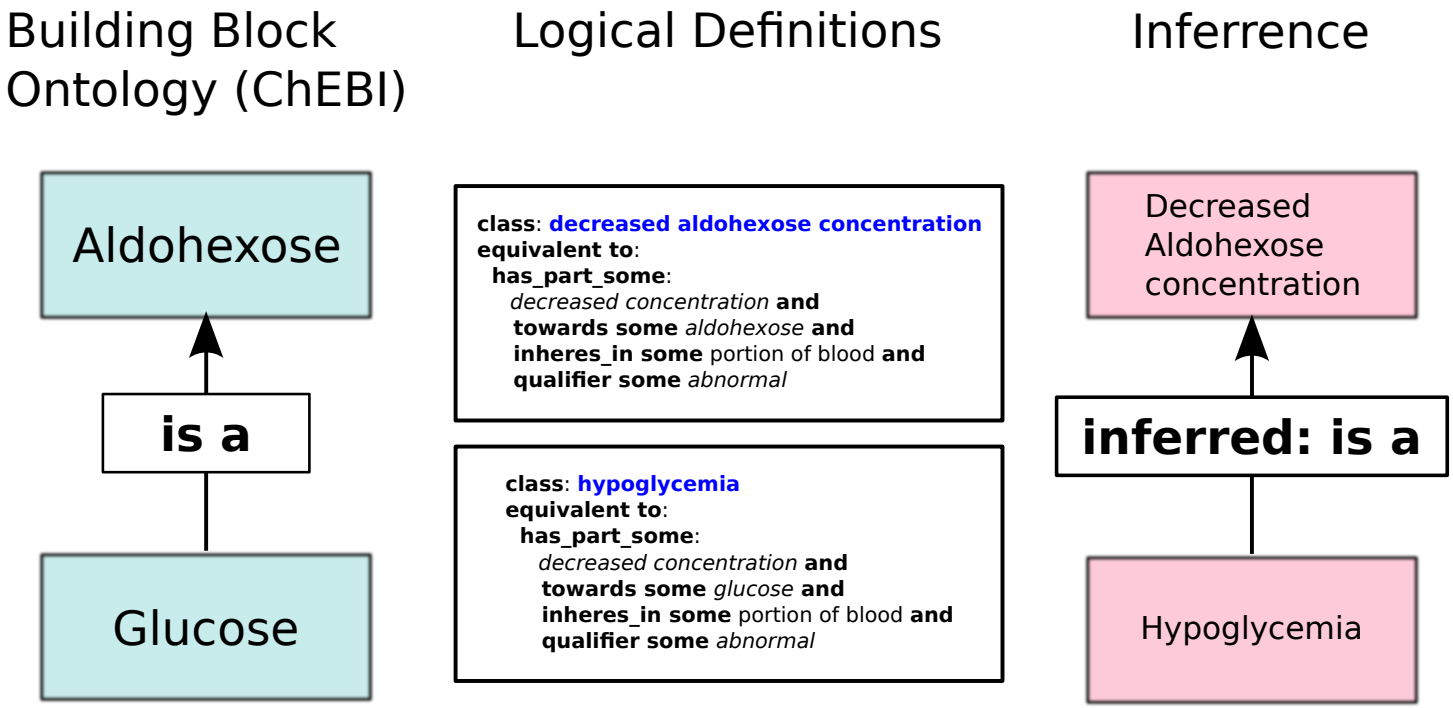
The Phenomizer



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

http://compbio.charite.de/Phenomizer

Reasoning over Phenotypes



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.

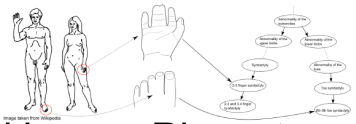
A semantic web of the human phenotype

Anatomy

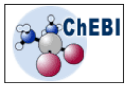
Gene Ontology

Cell Type

Proteins



Human Phenotype

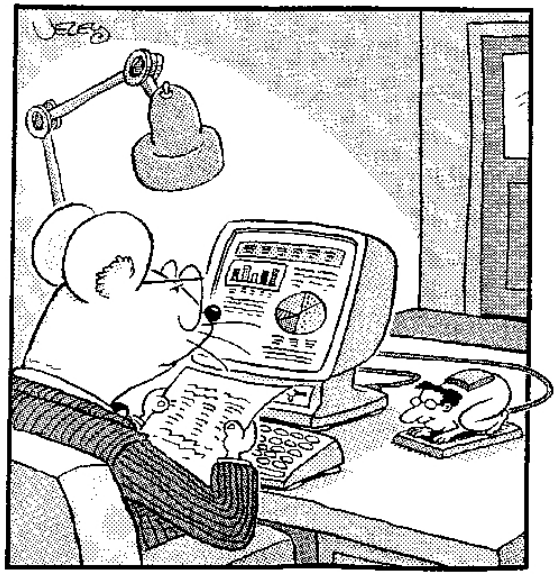


Small molecules

Embryology

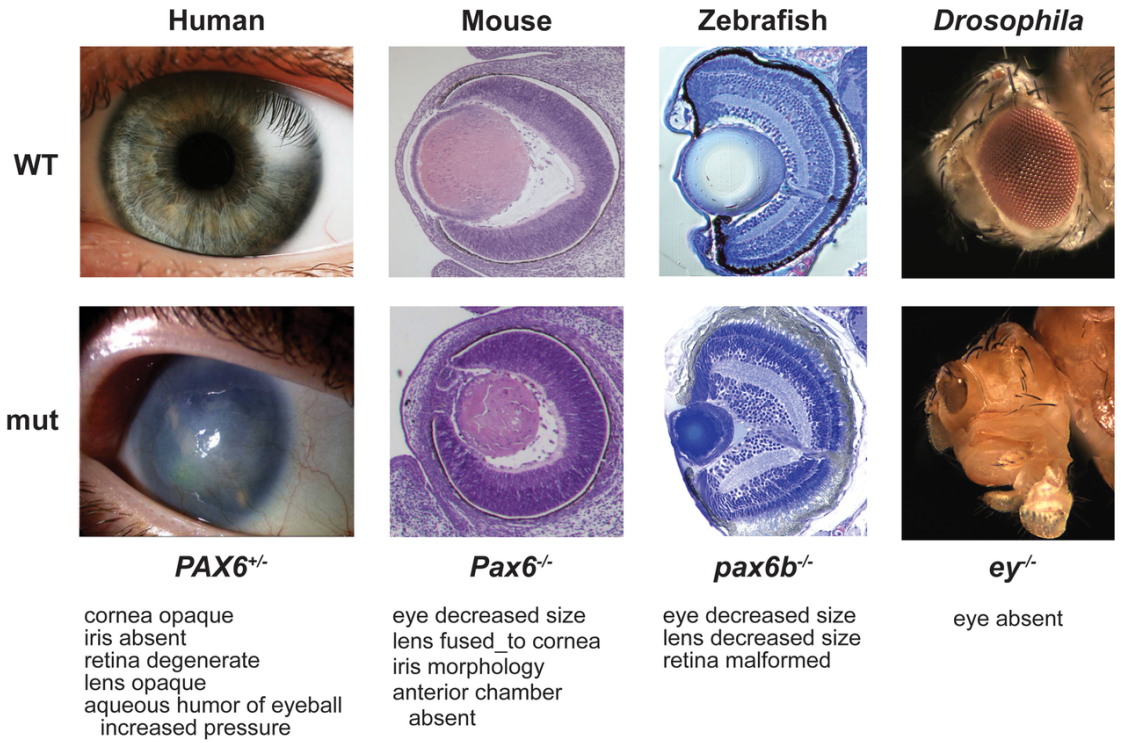
OBO Foundry

Pathology



- 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

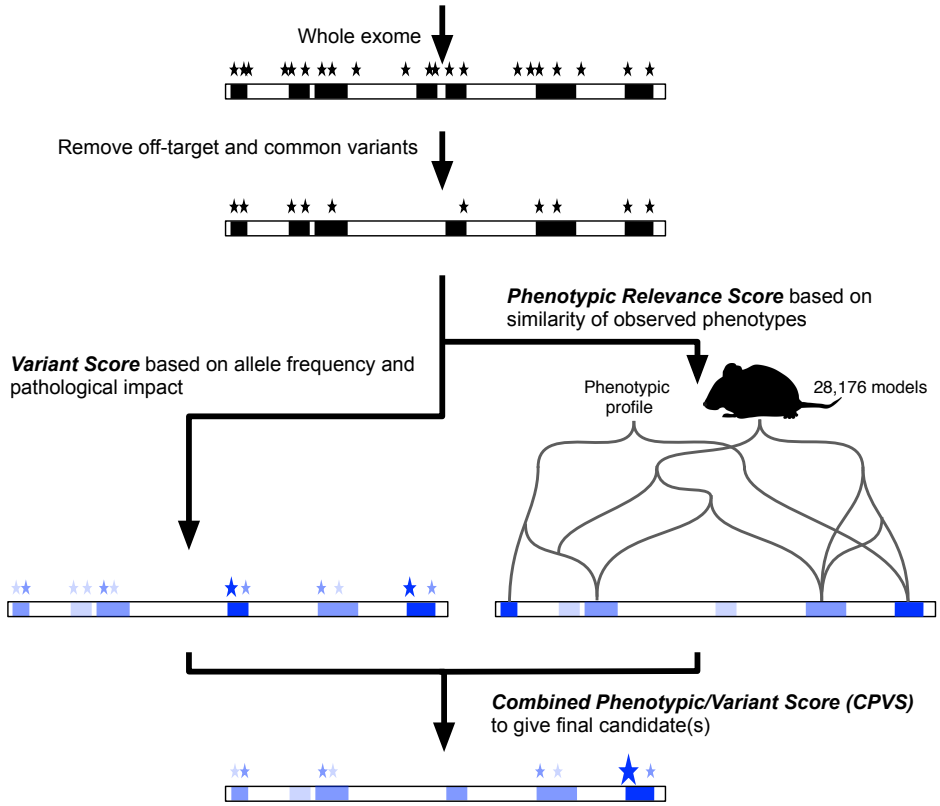
Finding the needle: 3,000,000,000 bases $\rightarrow$ 1 mutation

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

- Each genome: Lots of "private" variants with $\sim$ 100 genuine loss of function variants with $\sim$20 genes completely inactivated.
- $\therefore$ sequence-based prioritization alone will struggle to identify the disease-associated mutation

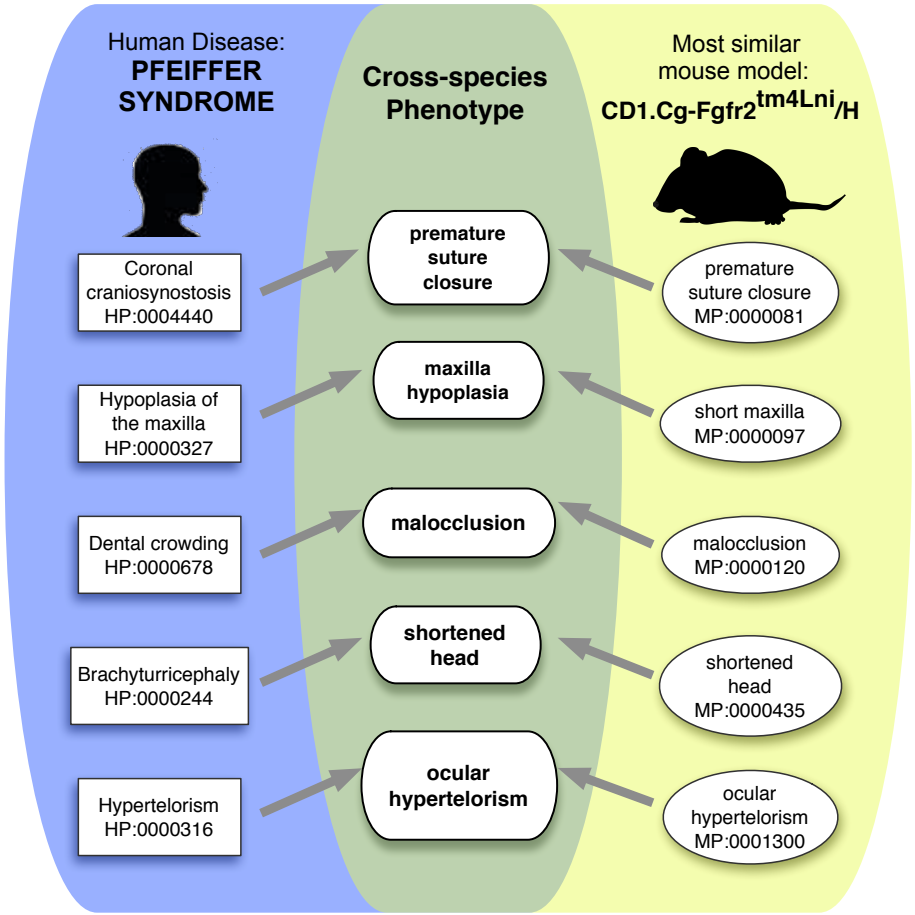
Exome Sequencing

- at least 30,000 variants per exome
- We all have $\sim$ 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.



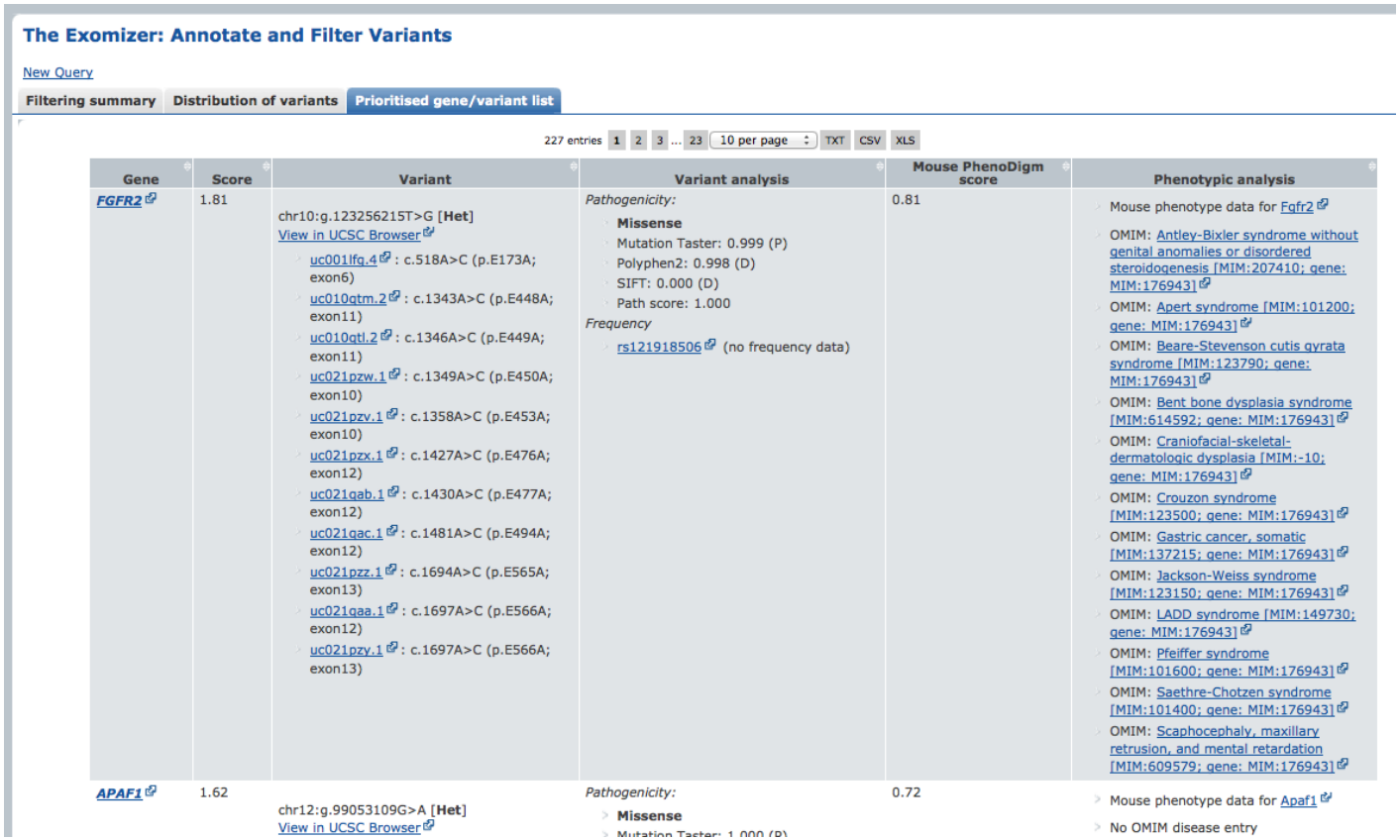
The Exomiser

PHIVE



- PHIVE: *PHenotypic Interpretation of Variants in Exomes*

Exomizer



The Exomizer: Annotate and Filter Variants

New Query

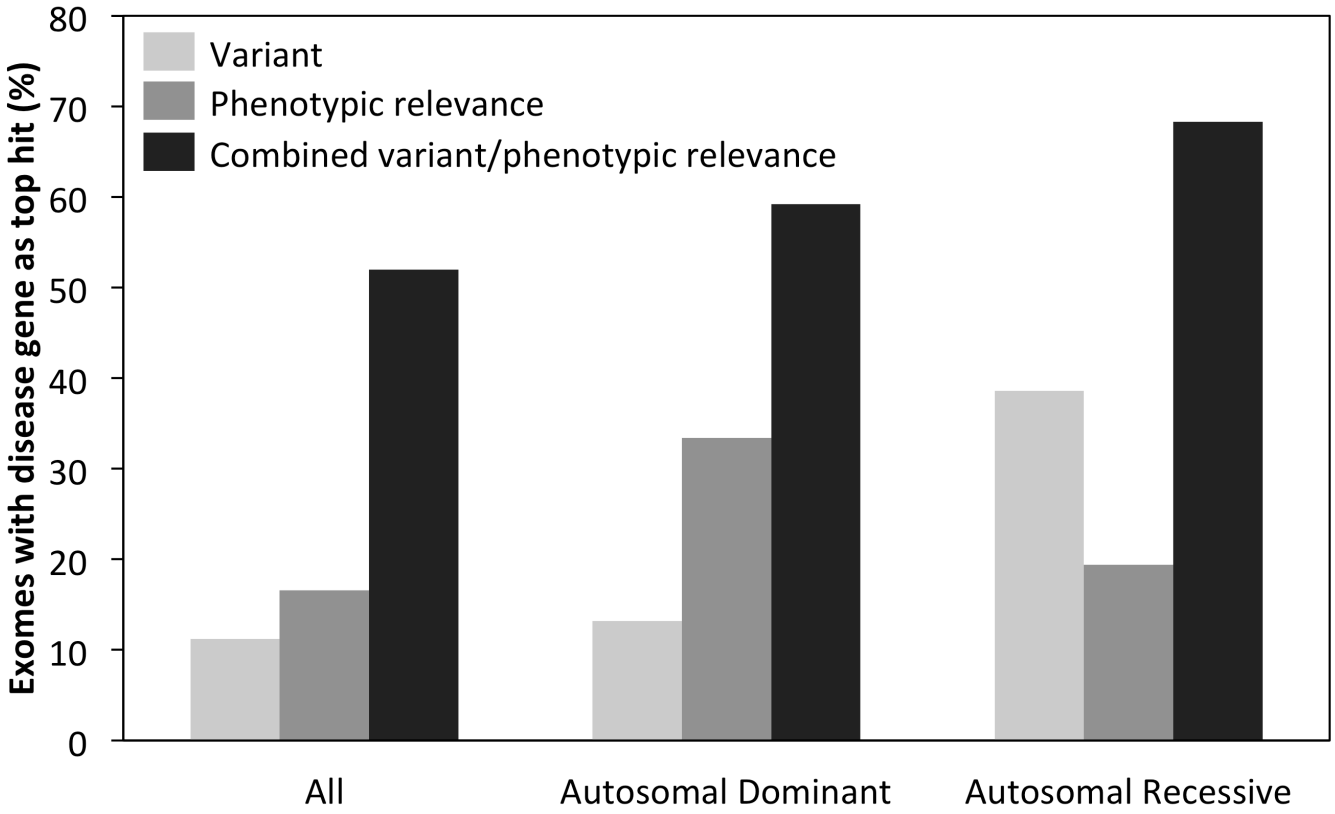
Filtering summary | Distribution of variants | Prioritised gene/variant list

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

Gene	Score	Variant	Variant analysis	Mouse PhenoDigm score	Phenotypic analysis
FGFR2	1.81	chr10:g.123256215T>G [Het] View in UCSC Browser uc001lfq.4 : c.518A>C (p.E173A; exon6) uc010qtm.2 : c.1343A>C (p.E448A; exon11) uc010qtl.2 : c.1346A>C (p.E449A; exon11) uc021pzw.1 : c.1349A>C (p.E450A; exon10) uc021pzv.1 : c.1358A>C (p.E453A; exon10) uc021pzx.1 : c.1427A>C (p.E476A; exon12) uc021qab.1 : c.1430A>C (p.E477A; exon12) uc021qac.1 : c.1481A>C (p.E494A; exon12) uc021pzz.1 : c.1694A>C (p.E565A; exon13) uc021qaa.1 : c.1697A>C (p.E566A; exon12) uc021pzy.1 : c.1697A>C (p.E566A; exon13)	*Pathogenicity:* **Missense** Mutation Taster: 0.999 (P) Polyphen2: 0.998 (D) SIFT: 0.000 (D) Path score: 1.000 *Frequency* rs121918506 (no frequency data)	0.81	Mouse phenotype data for Fgfr2 OMIM: Antley-Bixler syndrome without genital anomalies or disordered steroidogenesis [MIM:207410; gene: MIM:176943] OMIM: Apert syndrome [MIM:101200; gene: MIM:176943] OMIM: Beare-Stevenson cutis gyrata syndrome [MIM:123790; 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-Weiss 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]
APAF1	1.62	chr12:g.99053109G>A [Het] View in UCSC Browser	*Pathogenicity:* **Missense**	0.72	Mouse phenotype data for Apaf1 No OMIM disease entry

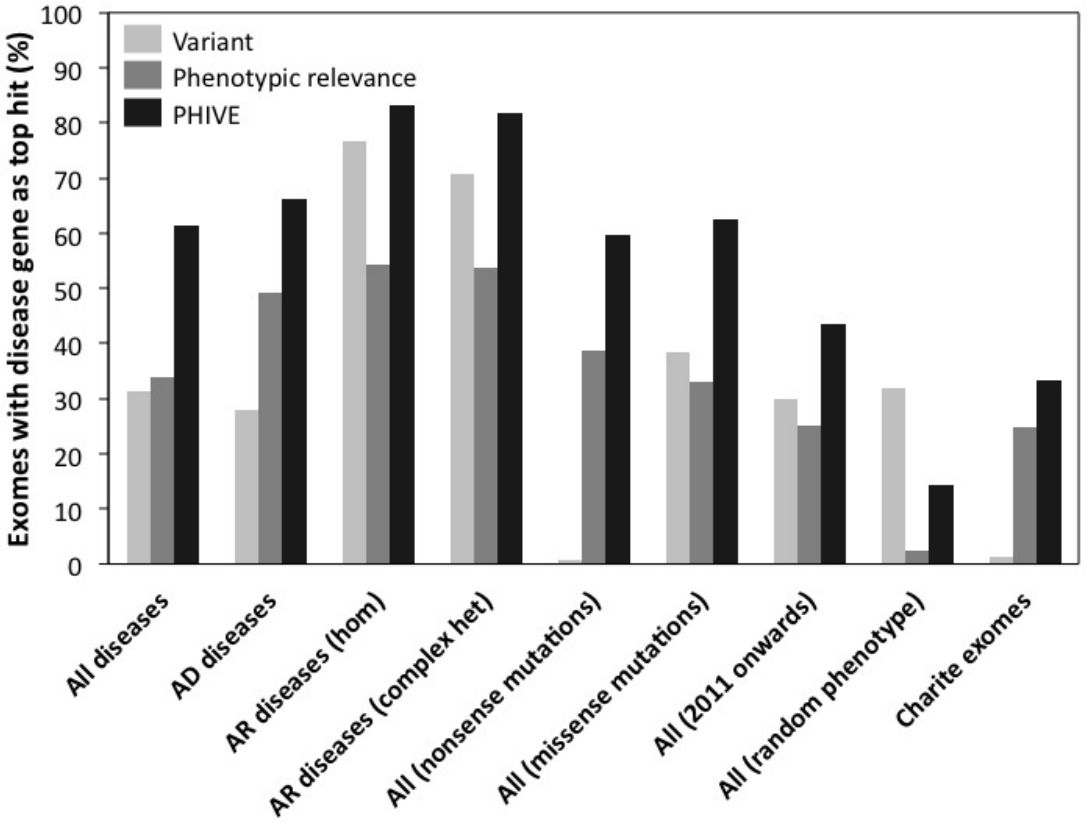
- Users enter VCF file and phenotype data and get back ranked list of candidates

Exomiser



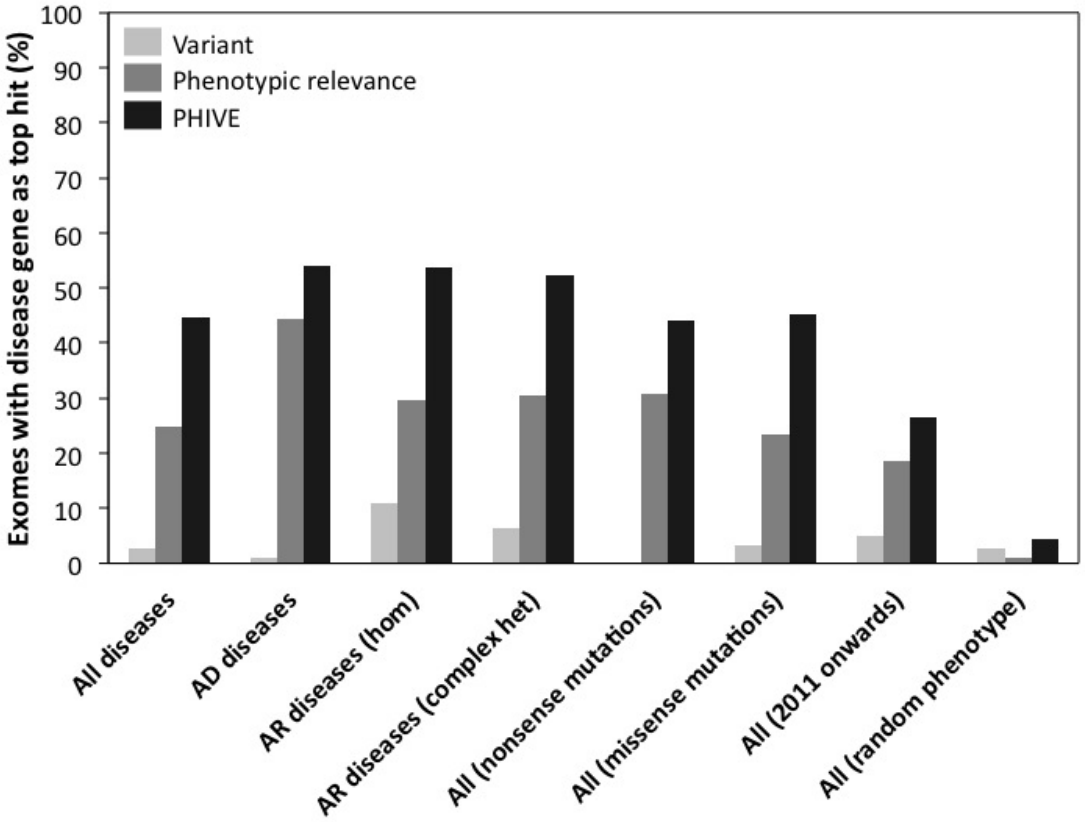
- Phenotype data and variant data **synergistically** improve interpretation of exome data for gene discovery

Exomiser



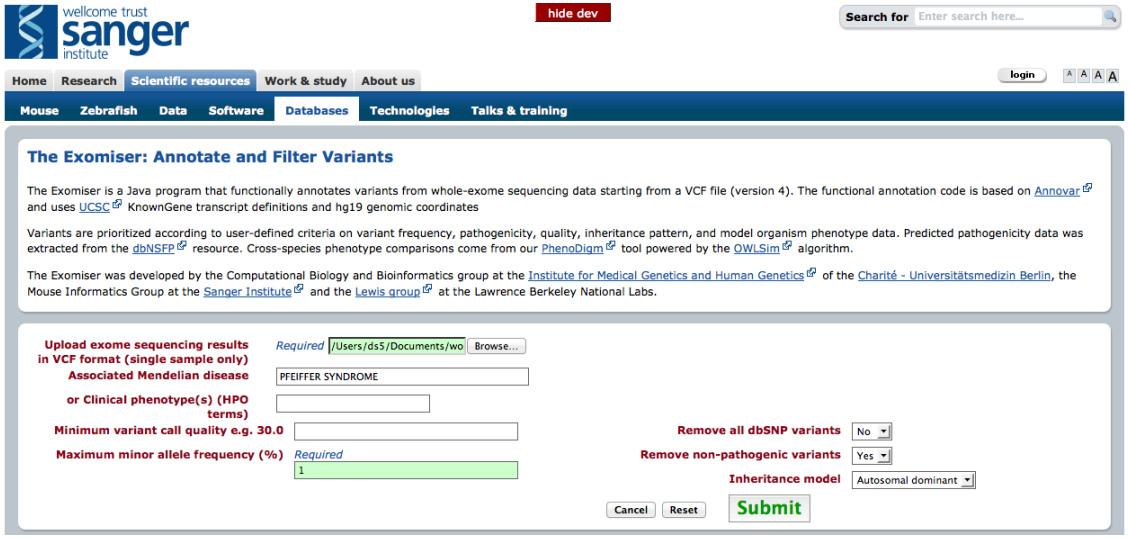
- With ESP and 1000G Frequency data

Exomiser



- 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



- 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, Gilessen 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?