Semantic phenotype integration for clinical diagnostics

Peter Robinson

Curating the Clinical Genome 2019

These slides: bit.ly/hpo-ccg19

@monarchinit

30 May 2019

Washington, DC

CD2H: A National Center for Digital Health Informatics Innovation; 5U24TR002306-03
The Monarch Initiative: Linking Diseases to Model Organism Resources; R24OD011883

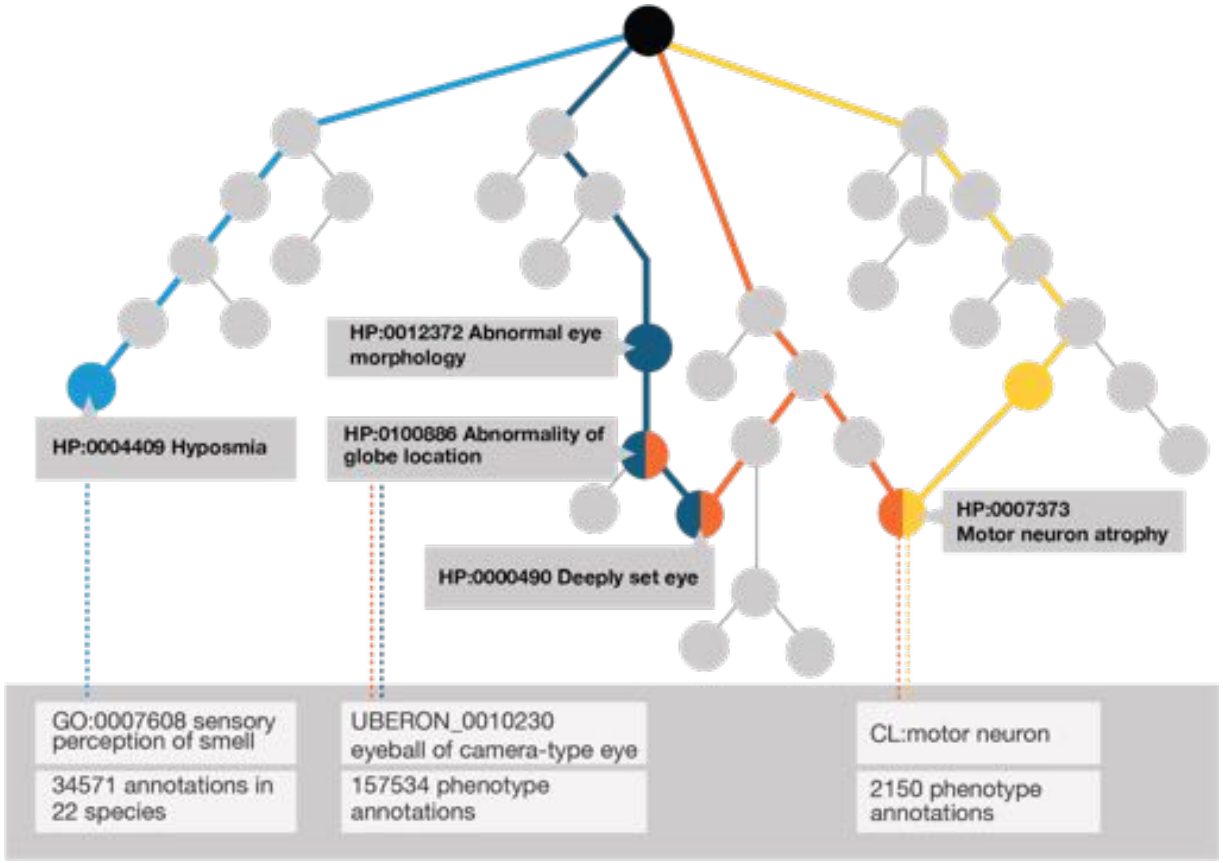
human phenotype ontology

monarch INITIATIVE

The Jackson Laboratory

Human Phenotype Ontology

- logically defined vocabulary for describing medically relevant abnormal phenotypes.
- *de facto* standard for computational phenotype analysis in genomics and rare disease



Robinson *Am J Hum Genet* 2008
Köhler *Am J Hum Genet* 2009
Robinson *Clin Genet.* 2010 J
Schulz *BMC Bioinformatics*. 2011
Köhler *Hum Mutat.* 2012
Bauer *Bioinformatics* 2012
Köhler *Nucleic Acids Res* 2014
Köhler *J Med Genet* 2014
Zemojtel *Sci Transl Med*. 2014
Robinson *PLoS Genet*. 2014
Robinson *Genome Res.* 2014
Groza *Am J Hum Genet* 2015
Groza *Database (Oxford)*. 2015
Smedley *Nat Protoc* 2015
Smedley *Genome Med.* 2015
Smedley *Am J Hum Genet.* 2016
Köhler *Nucleic Acids Res* 2017
Vasilevsky *Nat Genet* 2018
Köhler *Nucleic Acids Res* 2019

Integration of HPO and Improvement of Core Algorithms

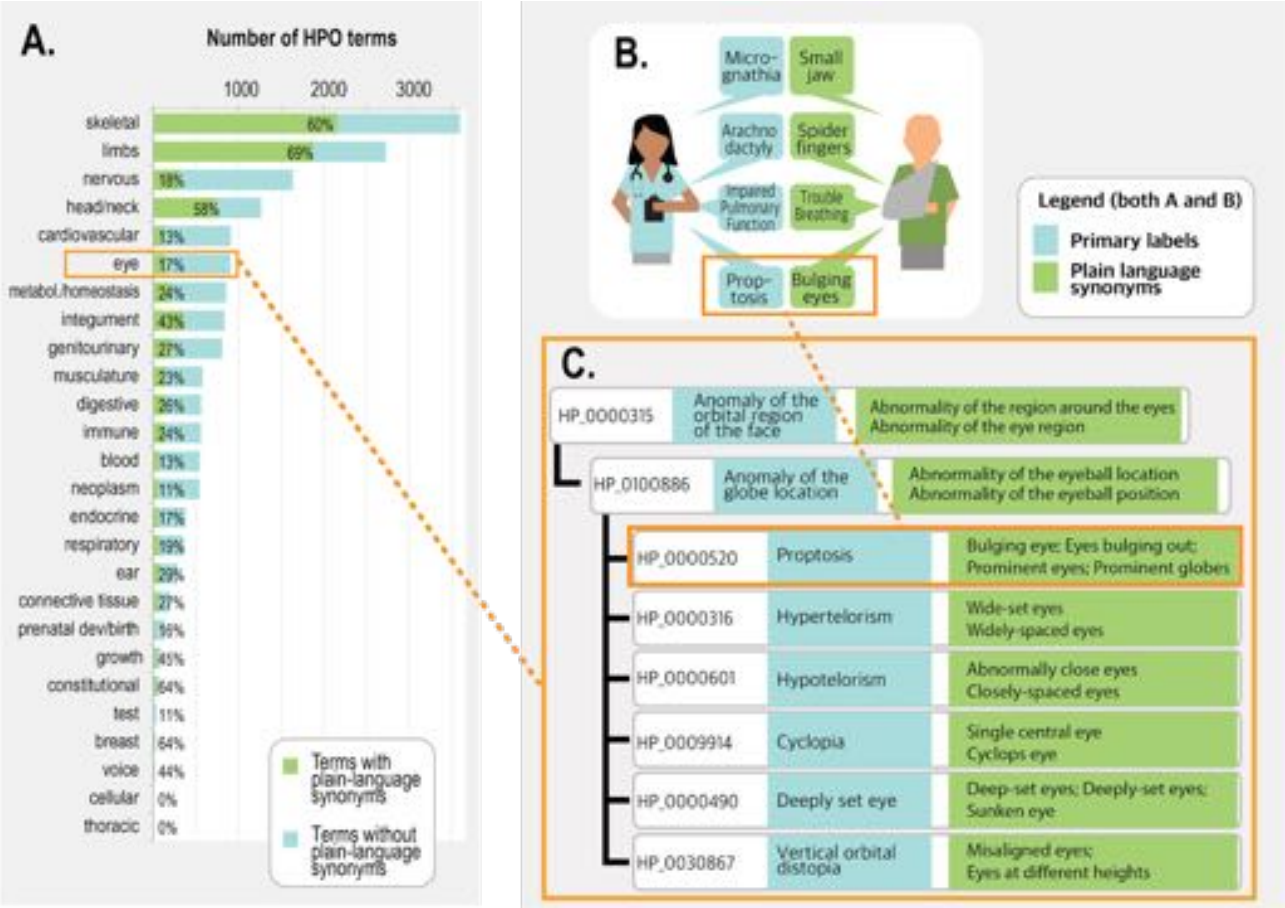
1) Plain language “translation” of HPO

2) Integration of HPO with EHR standards

3) A new format for phenotype data exchange

4) A clinically interpretable algorithm for phenotype-driven exome/genome analysis

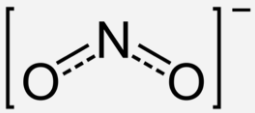
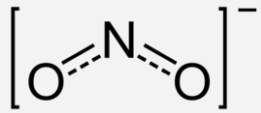
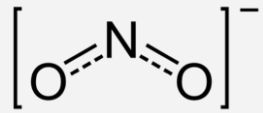
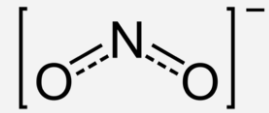
Plain-language medical vocabulary for precision diagnosis



4887 / 13823 HPO terms have lay synonyms

Vasilevsky, …, Haendel (2018) Plain-language medical vocabulary for precision diagnosis. *Nature Genet* **50**, 474–476

Multiple LOINC codes can be interpreted as same phenotypic outcome

LOINC: 20407-3	LOINC: 2657-5	LOINC: 50558-6	LOINC: 32710-6
			
Test strip	**Instrument**	**Test strip**	**Instrument**
mg/dL	**mg/dL**	☑ Yes ☐ No	☑ Yes ☐ No

HPO:0031812
Nitrituria

How can we infer phenotypic data from these lab codes?

LOINC2HPO provides first comprehensive way to roll up granular terms into meaningful buckets

LOINC® from Regenstrief | human phenotype ontology | monarch INITIATIVE | NATIONAL CENTER FOR DATA TO HEALTH

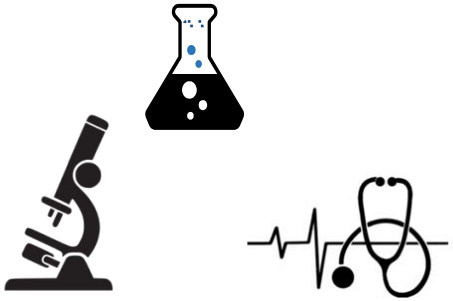
A universal standard for identifying medical laboratory observations in EHRs. e.g. LOINC 789-8 Erythrocytes [#/volume] in Blood by Automated Count

LOINC codes are widely used by healthcare providers and becoming mandatory by many regulatory agencies.

~86,000 LOINC codes exist

Integration of HPO with EHR standards such as LOINC

Electronic health records (EHRs) contain a vast amount of patient phenotype information

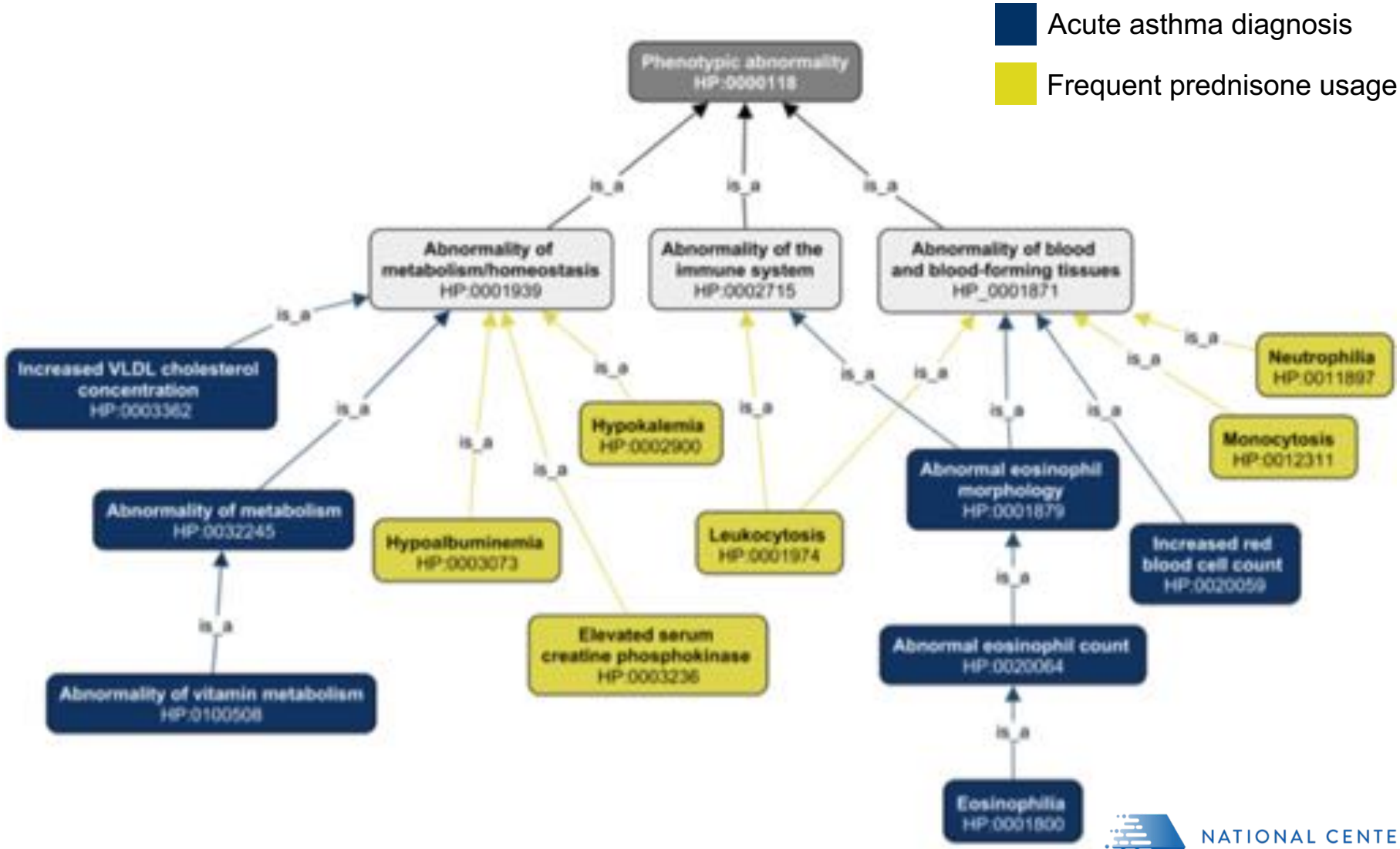


laboratory results are an untapped source of descriptive phenotypes, but are not directly amenable to semantic similarity algorithms used in clinical genetics

Ontologies can act as a bridge that harmonizes disparate types of data

GOAL: Map EHR data to HPO to enable automatic extraction of detailed deep phenotypic profiles of laboratory results for downstream studies

HPO terms enriched in acute asthma and frequent prednisone groups



NATIONAL CENTER FOR DATA TO HEALTH

Via LOINC2HPO we independently discovered asthma biomarkers from lab tests

odds ratio(medical outcome) ~ $\exp(\beta_0 + \beta * x)$

- Significant associations with acute asthma diagnosis:

Eosinophilia, Increased red blood cell count, Increased VLDL cholesterol concentration, Abnormality of vitamin metabolism

- Significant associations with frequent prednisone prescription:

Hypoalbuminemia, Neutrophilia, Monocytosis, Leukocytosis, Hypokalemia, Elevated serum creatine phosphokinase

npj | Digital Medicine

Article | OPEN | Published: 02 May 2019

Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery

Xingmin Aaron Zhang, Amy Yates, [...] Peter N. Robinson

npj Digital Medicine **2**, Article number: 32 (2019) | Download Citation

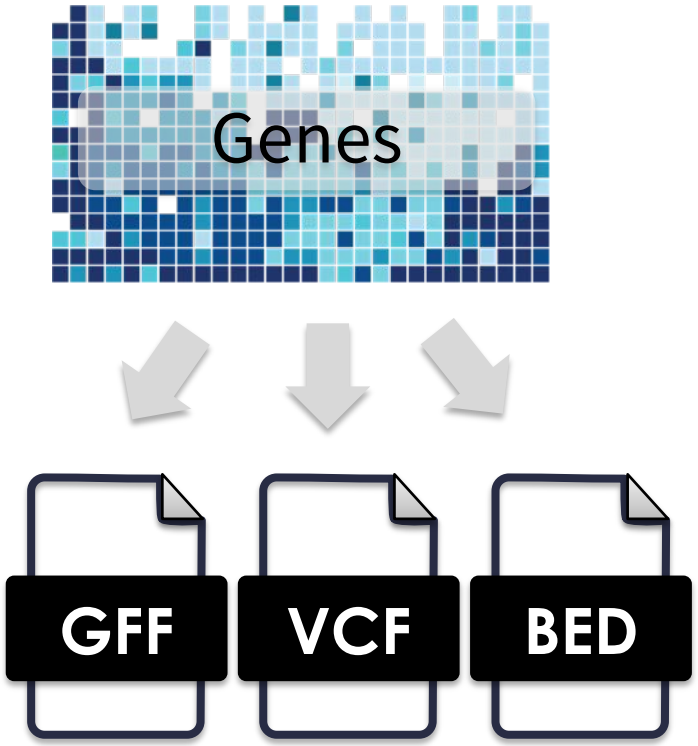
bit.ly/loinc2hpo-npj

asthma
Price DB et al. 2015, Lancet Respir Med
Allen S et al, 2009, Thorax
Jolliffe D et al, 2017, Lancet Respir Med
Thuesen BH, 2010, Allergy
Yiallouros PK, 2012 Clin Exp Allergy
Ramaraju K et al, 2013, Lung India
Ko SH et al, 2018, Lipids Health Dis
Chen YC, et al, 2013 Diabetes Metab Syndr
Al-Shawwa B et al, 2006, J Asthma
Cottrell L et al, 2011, Am J Respir Crit Care Med

Prednisone
Aplasca EC et al, 1986, J Am Diet Assoc
Dale DC et al, 1975, J Clin Invest
Shoenfeld Y et al, 1981, Am J Med
Veltri KT et al, 2015, P T
Smithson J et al, 2009, Australian Pharmacist

NATIONAL CENTER FOR DATA TO HEALTH

Exchanging phenotype information: Phenopackets



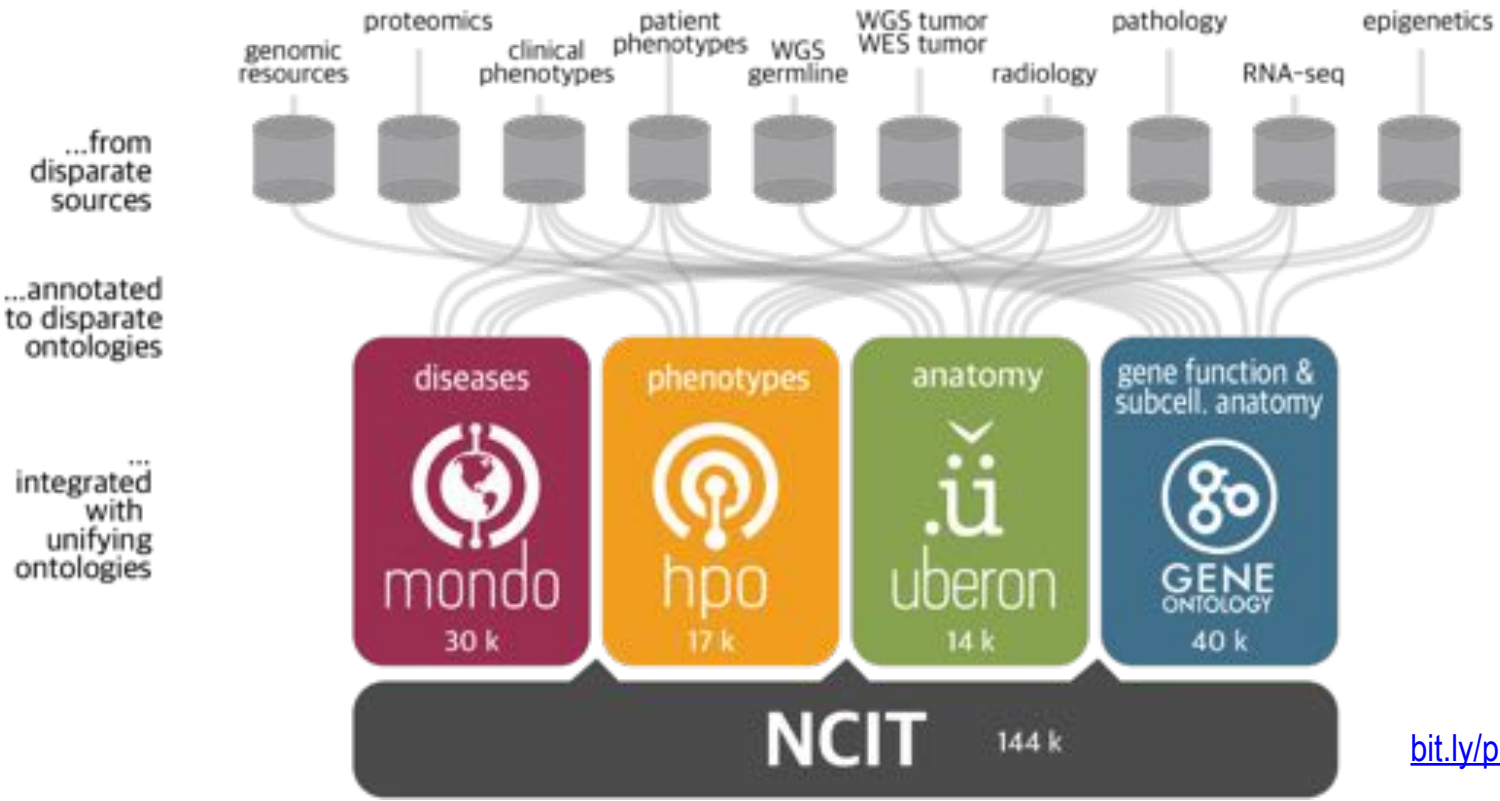
Phenopackets

Phenopackets are aligned with GA4GH mission: Interoperability of disease and phenotype data

Criteria for terminology use in GA4GH/Phenopackets

Open
Good content coverage (cancer, rare, and complex disease)
Interoperable with other standards
Relevant for genomic medicine (e.g. variant interpretation)

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

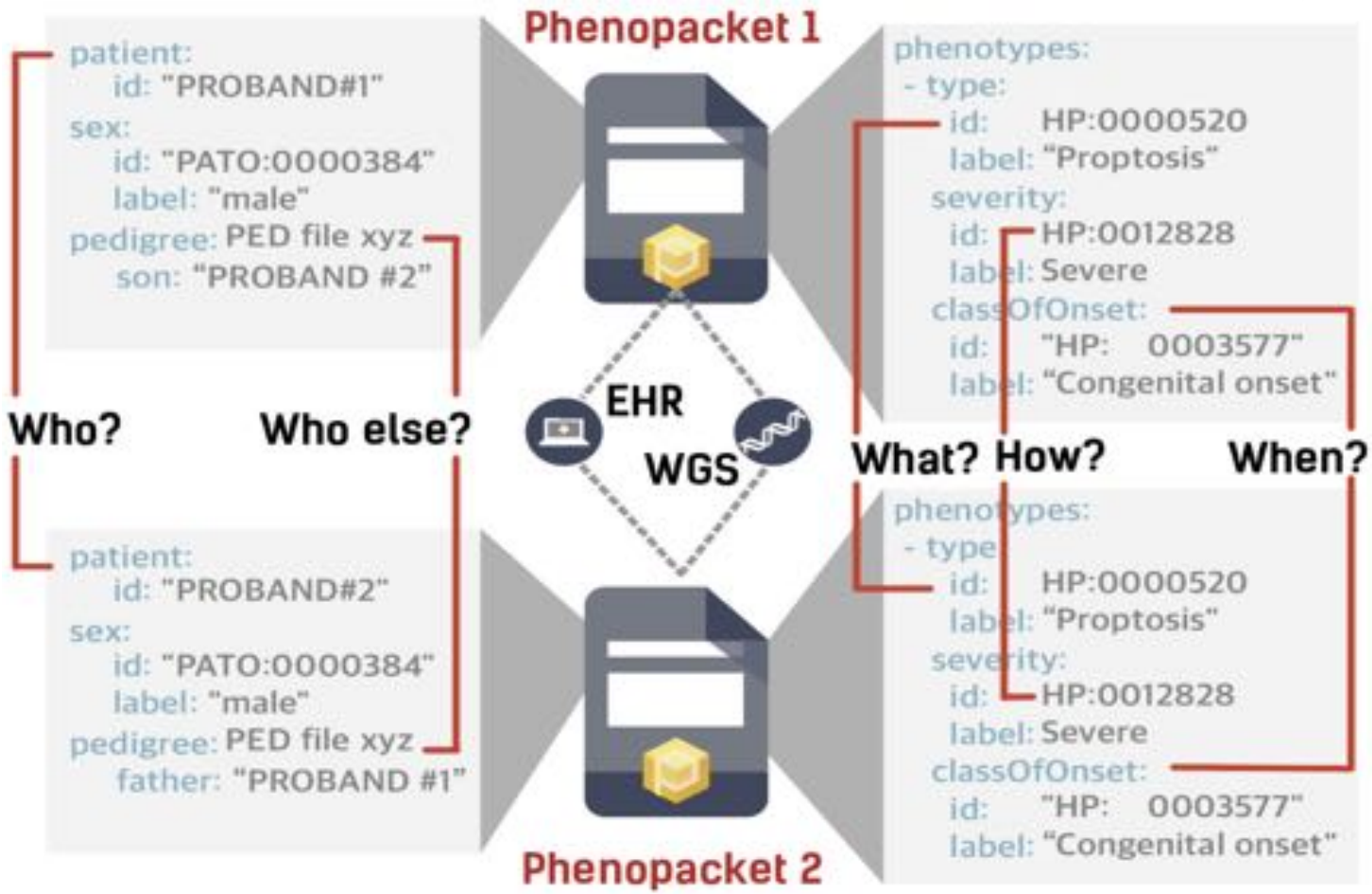


bit.ly/phenopackets-basel

Phenopackets

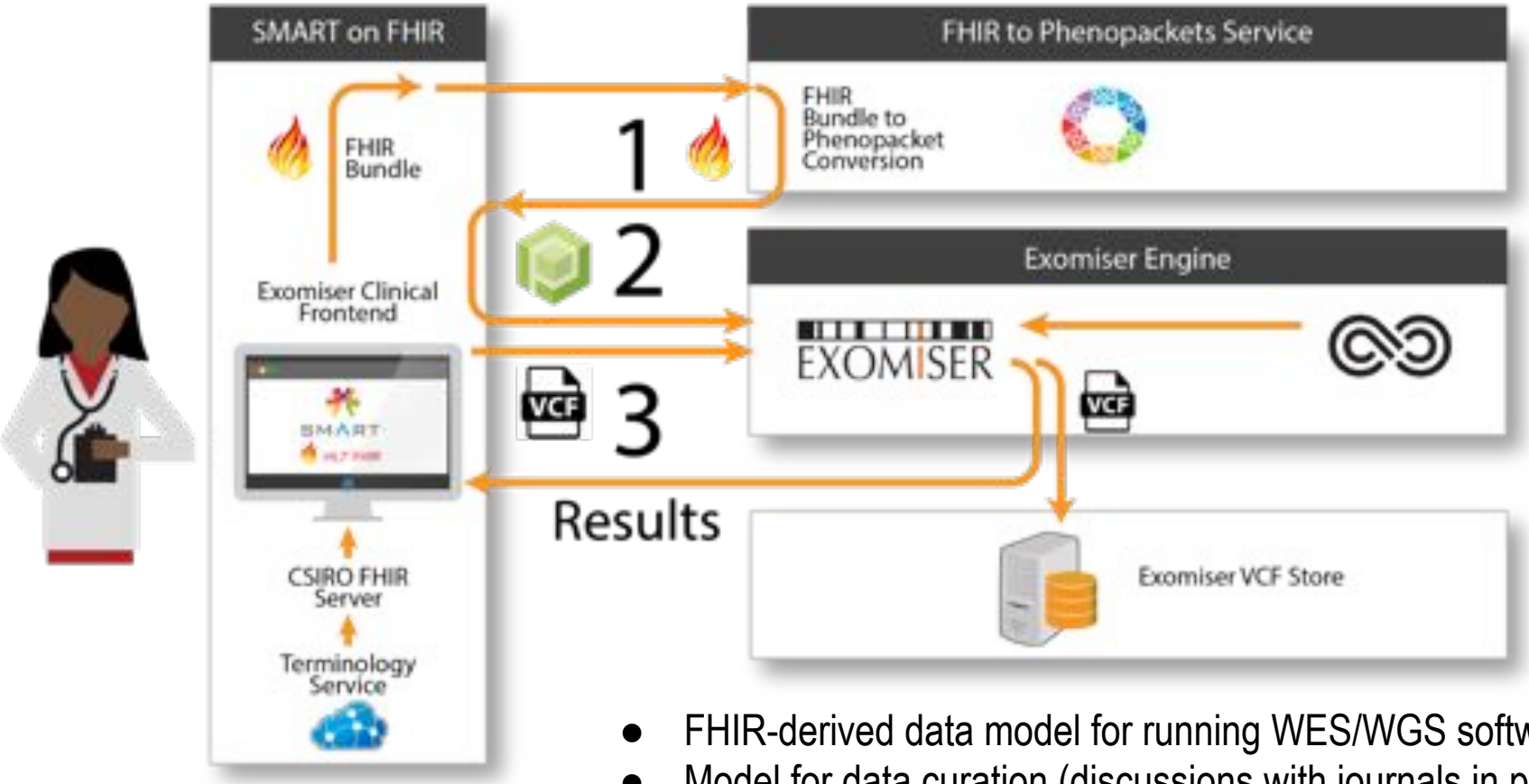
A Data Exchange and Analysis Model for Rare disease and Cancer

now obtaining GA4GH certification!



Anatomy of a Phenopacket: Phenotype information; Biosample (histologic diagnosis, phenotypes); Genotypes, exome/genome file references; Pedigree, cohort; Links to EHR formats such as FHIR

Phenopacket use cases



- FHIR-derived data model for running WES/WGS software
- Model for data curation (discussions with journals in progress)
- Reporting phenotypes to labs
- Common data model for algorithm development
- Making oncology path reports more easily computable

Clinically interpretable bioinformatics?

- Current WES/WGS detection rate: ~40-50%
- But prevailing algorithms don't help us realize that there are no good candidates. #1 may still be poor.

- Some phenotypic features may support a diagnosis while others do not
- Prevailing algorithms don't help users interpret the relative contribution of individual HPO terms to the computational ranking

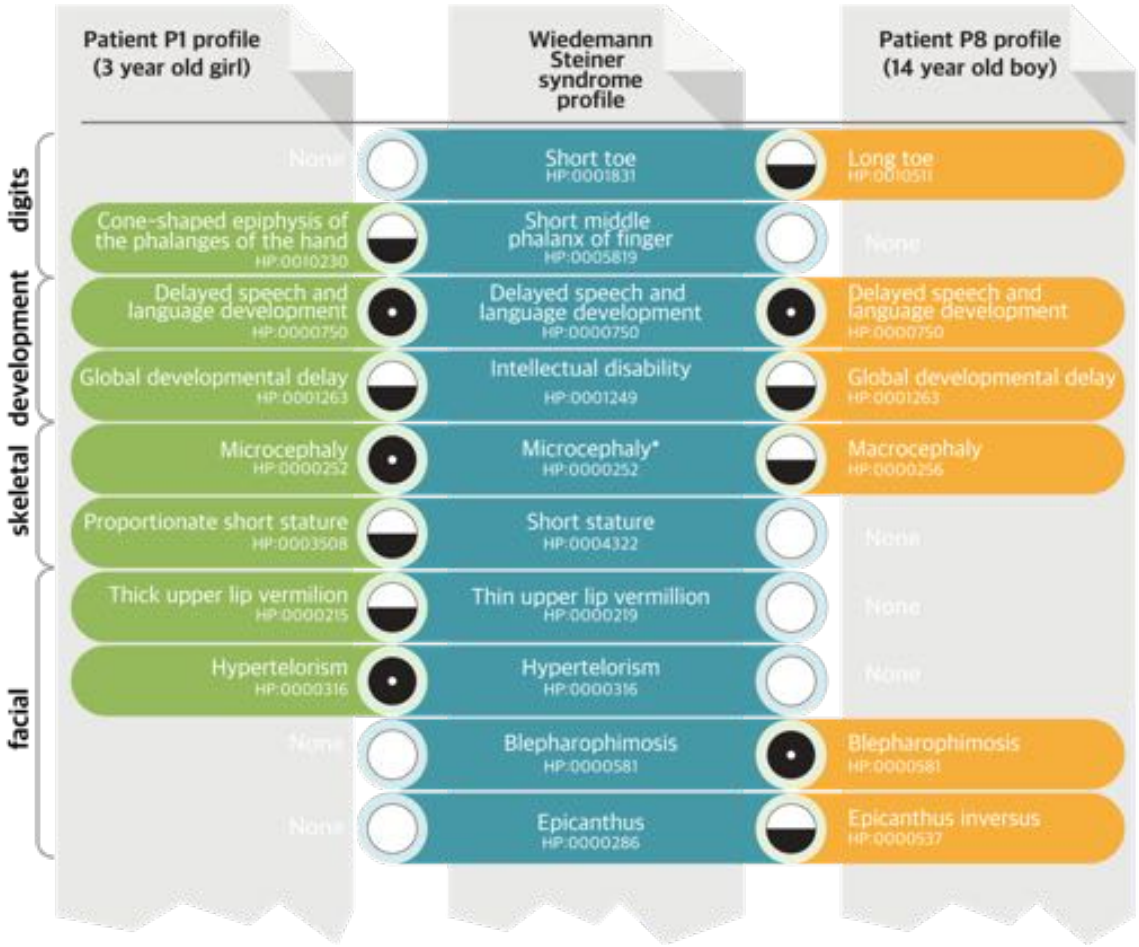
New algorithm: LIRICAL

Likelihood **R**atio **I**nterpretation of **C**linical **A**bnorma**L**ities

- Frames prioritization in a novel likelihood ratio based method for phenotypes and optionally genotypes
- Provides estimate of posttest probability
- Displays contribution of each feature to the final ranking
- Avoids arbitrary thresholds for imperfectly matching genotypes
- Outperforms existing algorithms for deeply phenotyped cases

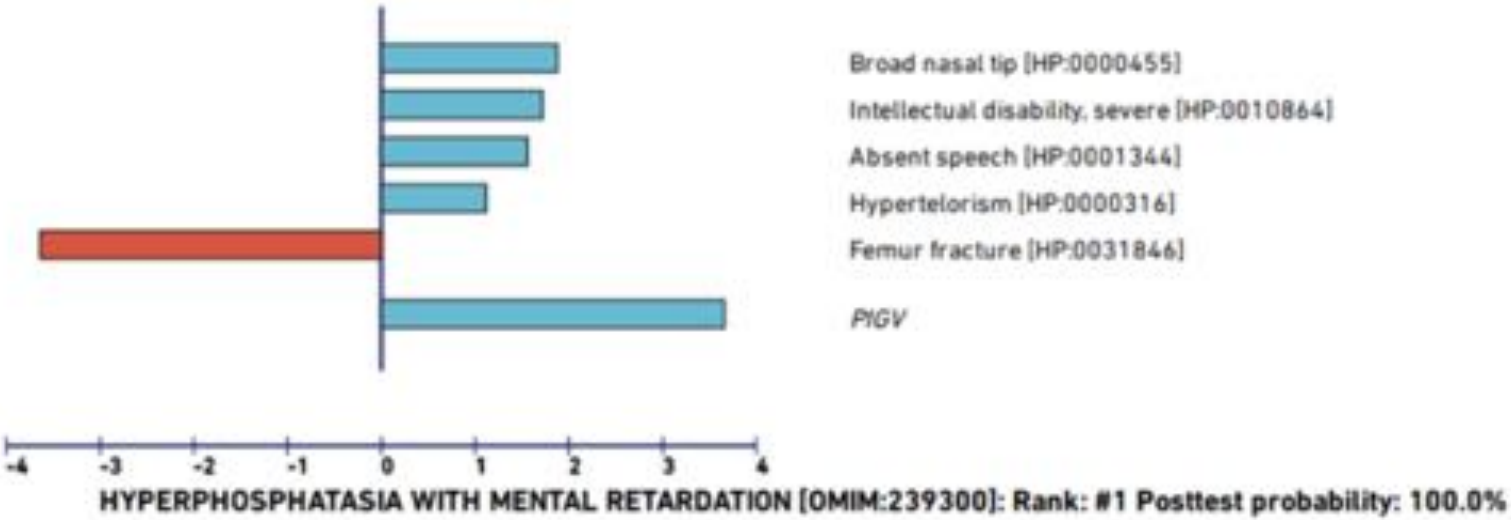
COI: patent application

Phenotype matching



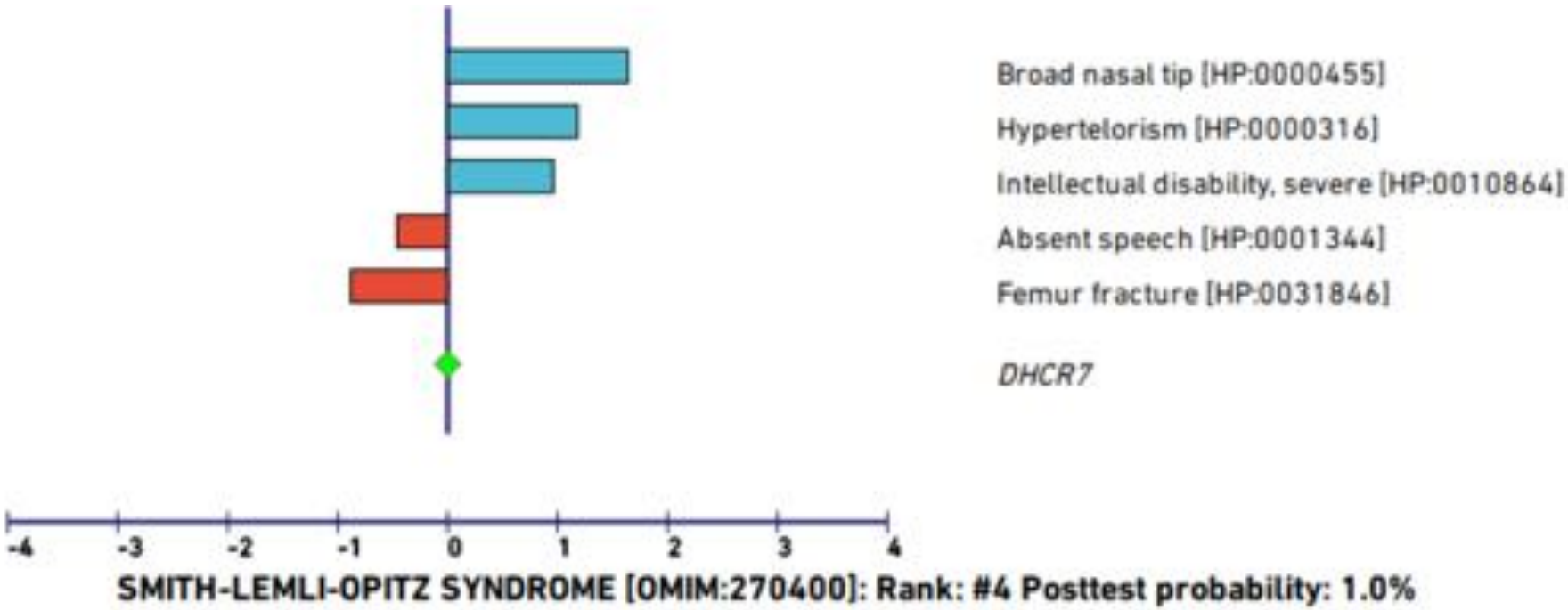
- Patient phenotypic features rarely exactly-match database descriptions of disease
- Patients may have unrelated (or related but never before observed) abnormalities
- Physicians may describe abnormalities at different granularity (often reflected in HPO hierarchy)

LIRICAL's Best Match: Mabry syndrome (Posttest probability: 100%)



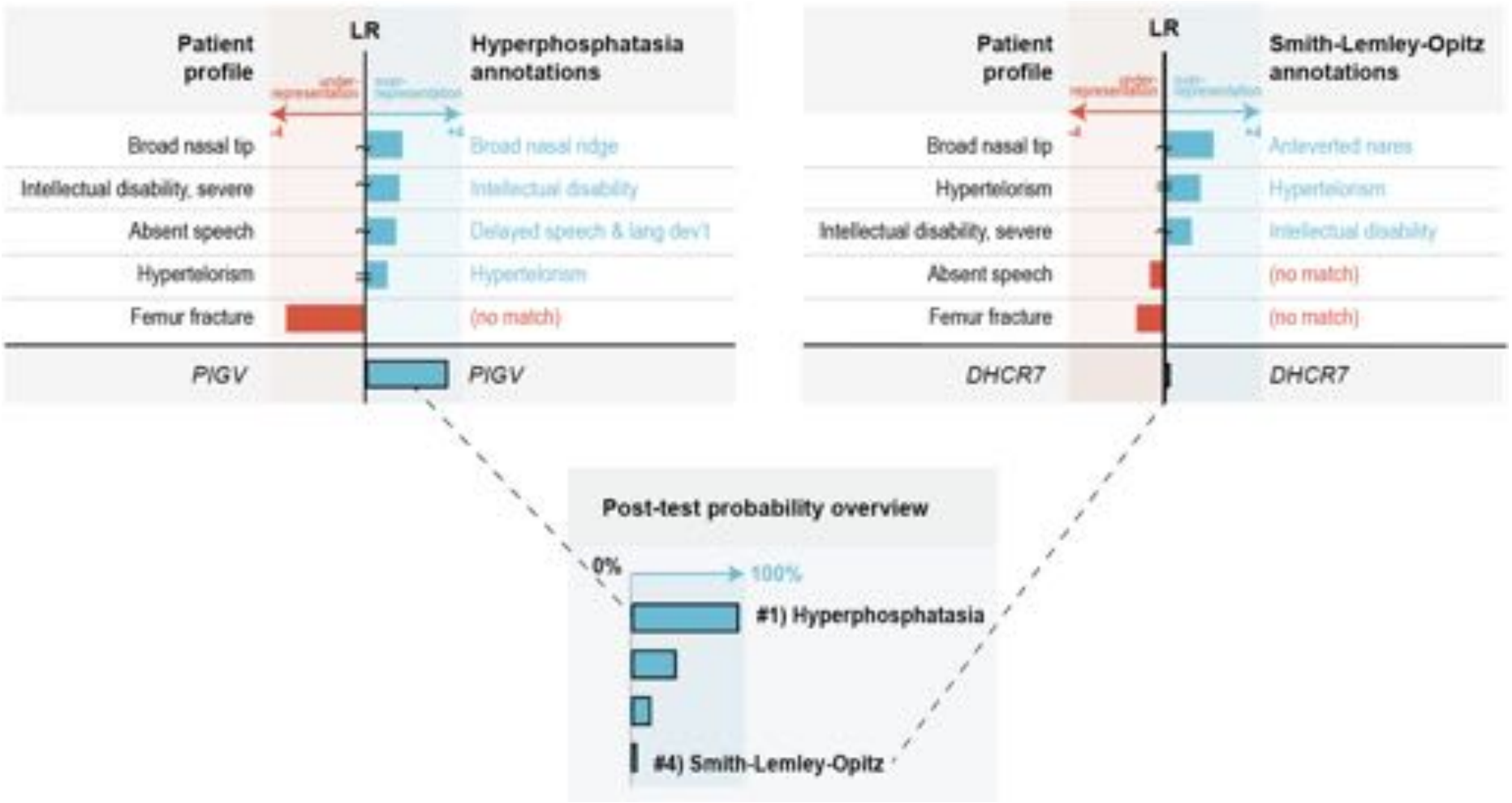
- 4 of th 5 features are characteristic of Hyperphosphatasia with mental retardation syndrome 1 (HPMRS1; Mabry syndrome)
- Each of the two variants in *PIGV* NM_017837.3:c.615C>G and c.854A>G is assessed as predicted pathogenic, and since the expected number of rare predicted pathogenic variants in *PIGV* is low, the likelihood ratio is $10^{3.66}$.
- In this case, a clinician might downweight the finding of ***Femur fracture*** as being unrelated and decide to follow up on the differential of HPMRS1.

LIRICAL: Posterior probabilities improve interpretation of rankings



- The fourth disease in the list is Smith Lemli Opitz (SLO) syndrome, with a posttest probability of 1% (only a single heterozygous predicted pathogenic variant in *DHCR7* was found, but SLO is an autosomal recessive disease, so the genotype likelihood ratio is about 1; the phenotype match is worse that for HPMRS1).
- The interpretation is that there is a low probability for this differential (which is much lower yet for following entries), and so users should focus attention on differentials above rank #4.

LIRICAL: Posterior probabilities improve interpretation of rankings



LIRICAL: Defining the likelihood ratio

$$\mathrm{LR}(x) = \frac{\text{sensitivity}}{1 - \text{specificity}} = \frac{P(x|D_j)}{P(x|\neg D_j)}$$

- The likelihood ratio is commonly used to compute the accuracy of tests

For LIRICAL, we adapted this framework to phenotype-driven genomic diagnostics

$$\mathrm{LR}(h_i) = \frac{P(h_i|D_j)}{P(h_i|\neg D_j)}$$

- $P(h_i|D_j)$ is the probability of a person with disease D_j of having phenotypic abnormality h_i (HPO term i)
- Definition: Probability of a test result x in a patient with the target disease (D_j) divided by the probability of that result in a person without the disease.
- For instance, $P(h_i|D_j)$ = 34/57=59.7% that a person with Machado-Joseph disease has ***External ophthalmoplegia (HP:0000544)*** according to PMID:19659750 and available as an HPO annotation.
- LIRICAL uses various heuristics to account for partial matches using the ontology structure of the HPO

LIRICAL: Defining the likelihood ratio

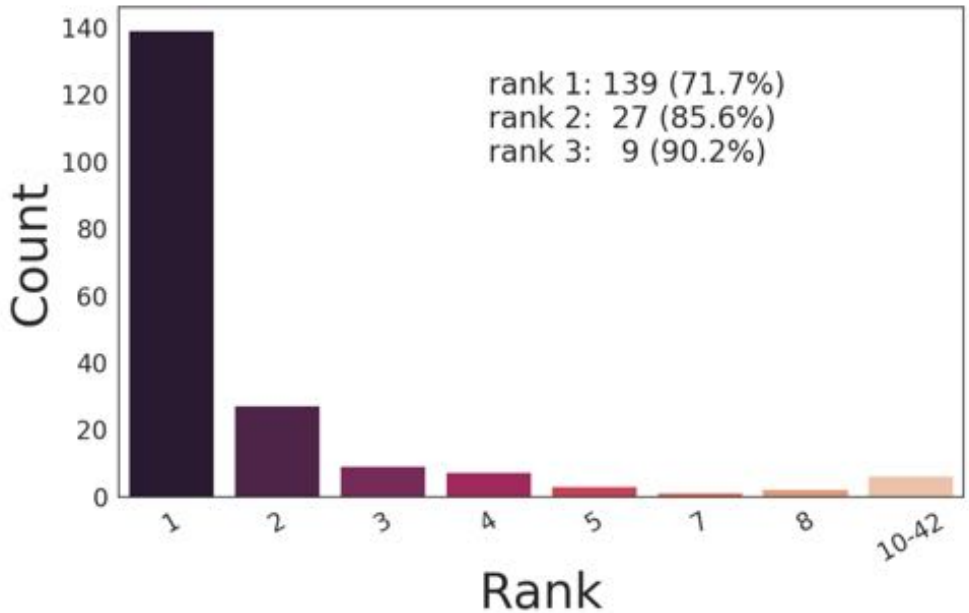
$$P(h_i|\neg D_j)$$

But how do we calculate the probability of HPO term *i* in an individual without disease D_j?

- Unknown population prevalence of the vast majority of the abnormalities represented in HPO
- Our model posits that the individual being tested has some Mendelian disease i, and so we can take the average frequency of the feature (f_i) amongst the N=~7000 diseases of the HPO disease-phenotype annotation database.

$$P(h_i|\neg D_j) = \frac{1}{(N-1)} \sum_{k \neq j} f_{i,D_k} \approx \frac{1}{N} \sum_{k=1}^{N} f_{i,D_k}$$

LIRICAL performs well on published case reports

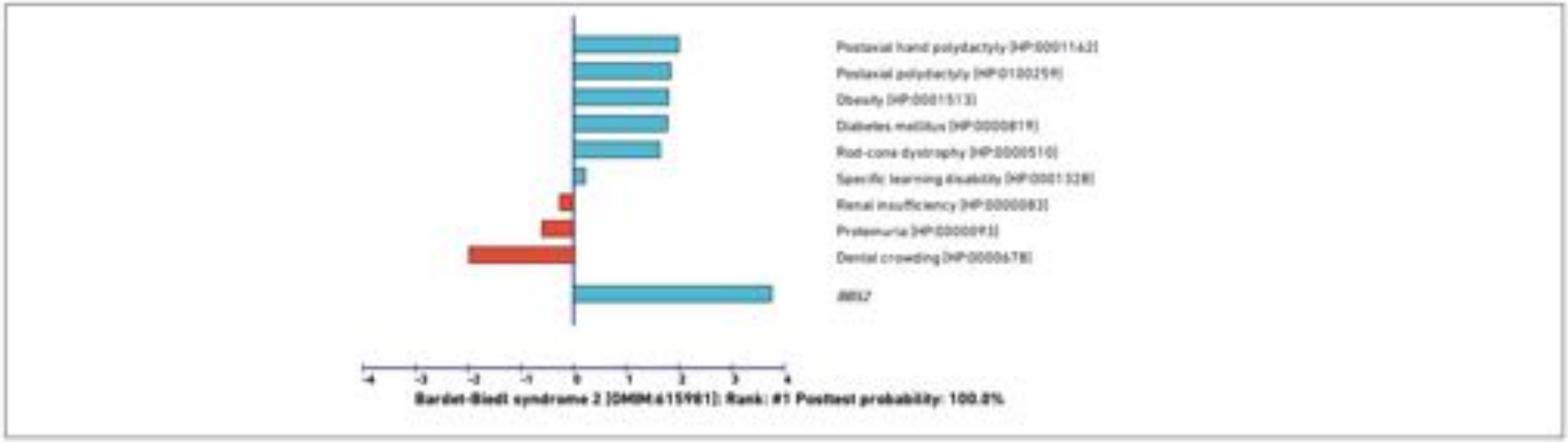


- Simulations on 194 phenopackets from **published case reports** representing 190 Mendelian diseases
- Over 90% were in the top 3 ranks, over 97% in the top 10
- Substantially better performance than related in house tools
- Intuitively, LIRICAL allows really strong matches to help more than previous semantic similarity based approaches that essentially average each match score.

LIRICAL

LIkelihood Ratio Interpretation of Clinical AbnormaLities

Position	Pathogenicity score	Max. pop. frequency	Genotype	ClinVar	Annotation
chr16:56519631A>T	1	0%	HETEROZYGOUS	n/a	• uc002ejd.2: c.1930T>A p.(Tyr644Asn) MISSENSE_VARIANT
chr16:56530925G>A	0.996	0.026%	HETEROZYGOUS	n/a	• uc002ejd.2: c.1864C>T p.(Arg622*) STOP_GAINED
chr16:56548501C>T	n/a	100%	HOMOZYGOUS_ALT	n/a	• uc010ccg.2: c.209G>A p.(Ser70Asn) MISSENSE_VARIANT • uc002ejd.2: c.209A>A p.(=) SYNONYMOUS_VARIANT
Genotype score LR:	Mode of inheritance: autosomal recessive. Observed weighted pathogenic variant count: 2.00. $\lambda_{disease}$=2. $\lambda_{background}$=0.0096. P(G|D)=0.2709. P(G|~D)=0.0000. $\log_{10}$(LR): 3.76.				
Phenotype analysis:	**E**:Postaxial hand polydactyly(HP:0001162)[103.068]; **D+G**:Postaxial foot polydactyly(HP:0001830)–Postaxial polydactyly(HP:0100259)[69.851]; **E**:Obesity(HP:0001513)[64.266]; **E**:Diabetes mellitus(HP:0000819)[62.116]; **E**:Rod-cone dystrophy(HP:0000510)[65.074]; **G–D**:Specific learning disability(HP:0001328)–Neurodevelopmental abnormality(HP:0012759)[1.732]; **G–D**:Renal insufficiency(HP:0000083)–Abnormality of the genitourinary system(HP:0000119)[0.545]; **G–D**:Proteinuria(HP:0000093)–Abnormality of metabolism/homeostasis(HP:0001939)[0.243]; **NM**:Dental crowding(HP:0000678)[0.010]				



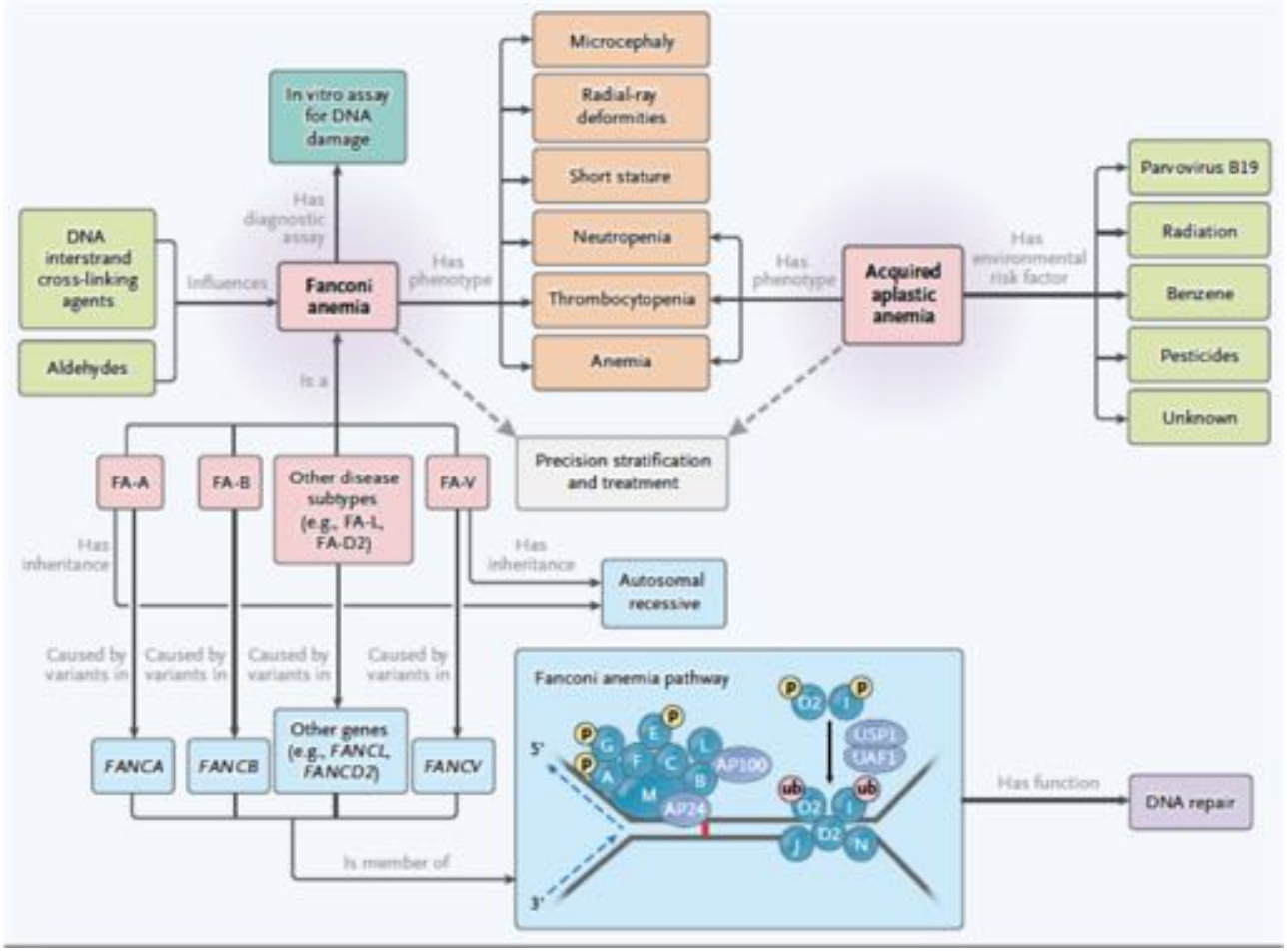
Complete LIRICAL output for one differential diagnosis

manuscript in preparation

https://github.com/TheJacksonLaboratory/LIRICAL

HPO + Monarch Initiative:

Semantic classification for precision medicine



- *Curation and harmonization of **disease definitions:** phenotypic features, genetic variation, and environmental factors*
- **Key Ontologies: HPO, uberon, MONDO, ECTO, close affiliation with GO**
- **Data integration across translational sources (human, model organisms, veterinary)**
- **Medical Action Ontology (MAxO)**: See Poster by Leigh Carmody

Haendel, Chute, Robinson (2018) Classification, Ontology, and Precision Medicine. *N Eng J Med* **379**:1452-1462

bit.ly/nejm-ontologies-medicine

Thanks!

Jackson Laboratory for Genomic Medicine
Hannah Blau
Leigh Carmody
Daniel Danis
Peter Hansen
Michael Gargano
Guy Karlebach
Vida Ravanmehr
Aaron Zhang

monarch INITIATIVE

OHSU
Melissa Haendel
Matt Brush
Kent Shefchek
Julie McMurry
Tom Conlin
Nicole Vasilevsky
Dan Keith
Maureen Hoatlin

Charité
Sebastian Kohler

Genomics England/Queen Mary
Damian Smedley
Jules Jacobson
Tomasz Konopka
Pilar Cacheiro

Lawrence Berkeley
Chris Mungall
Suzanna Lewis
Eric Douglass
Seth Carbon
(Nicole Washington)

Pryzm Health
Tudor Groza
Craig McNamara

RTI
Jim Balhoff

EBI
Helen Parkinson
David Osumi-Sutherland
Nicolas Matentzoglu

Funding

European Commission

CD2H NATIONAL CENTER FOR DATA TO HEALTH

NIH National Institutes of Health

The Jackson Laboratory