Cross-species phenotype knowledge representation and processing

Tudor Groza

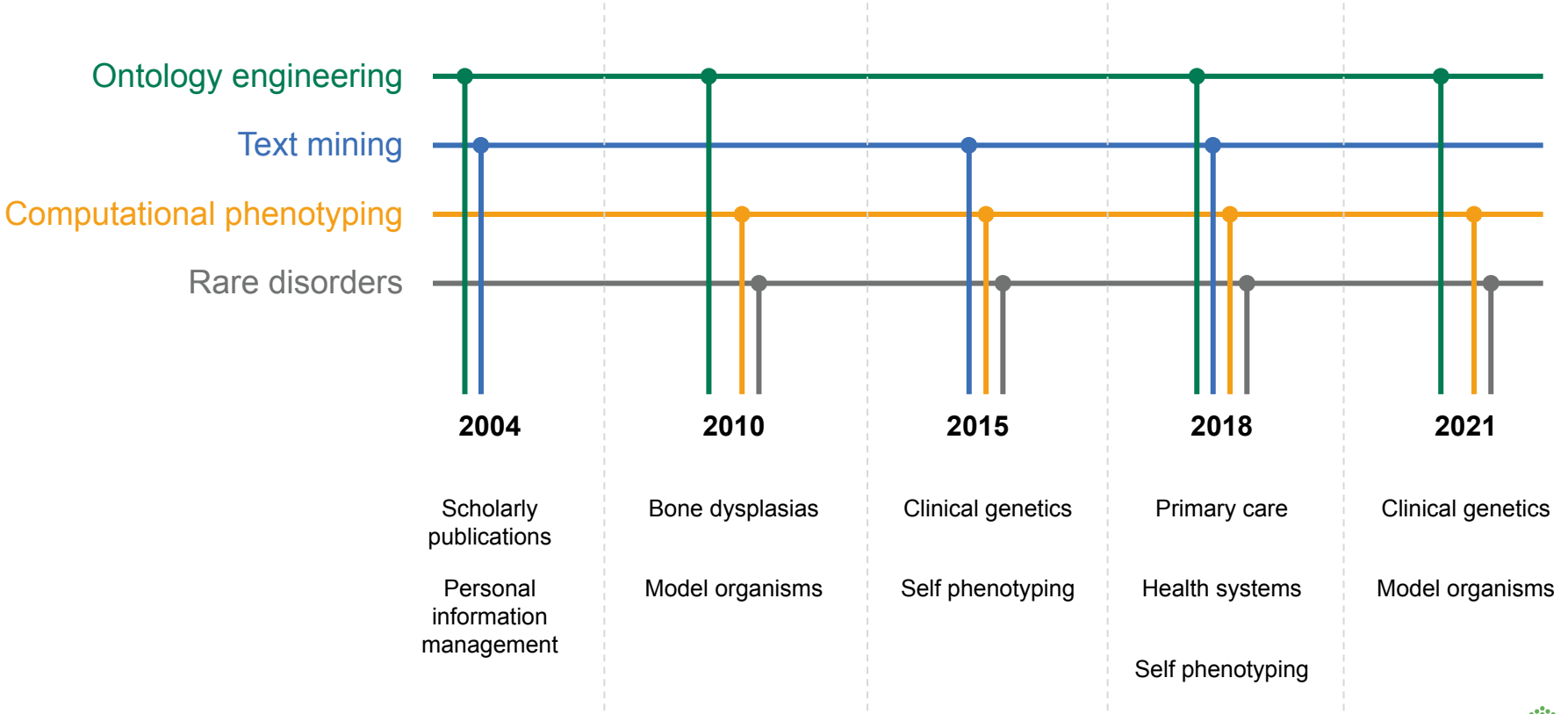
Phenomics Team Lead
EMBL-EBI

Open Science and FAIR Data for Neuroscience
Torino, 9 June 2022

monarch INITIATIVE

EMBL-EBI

Brief intro



EMBL-EBI

Research - Clinical practice continuum

monarch INITIATIVE

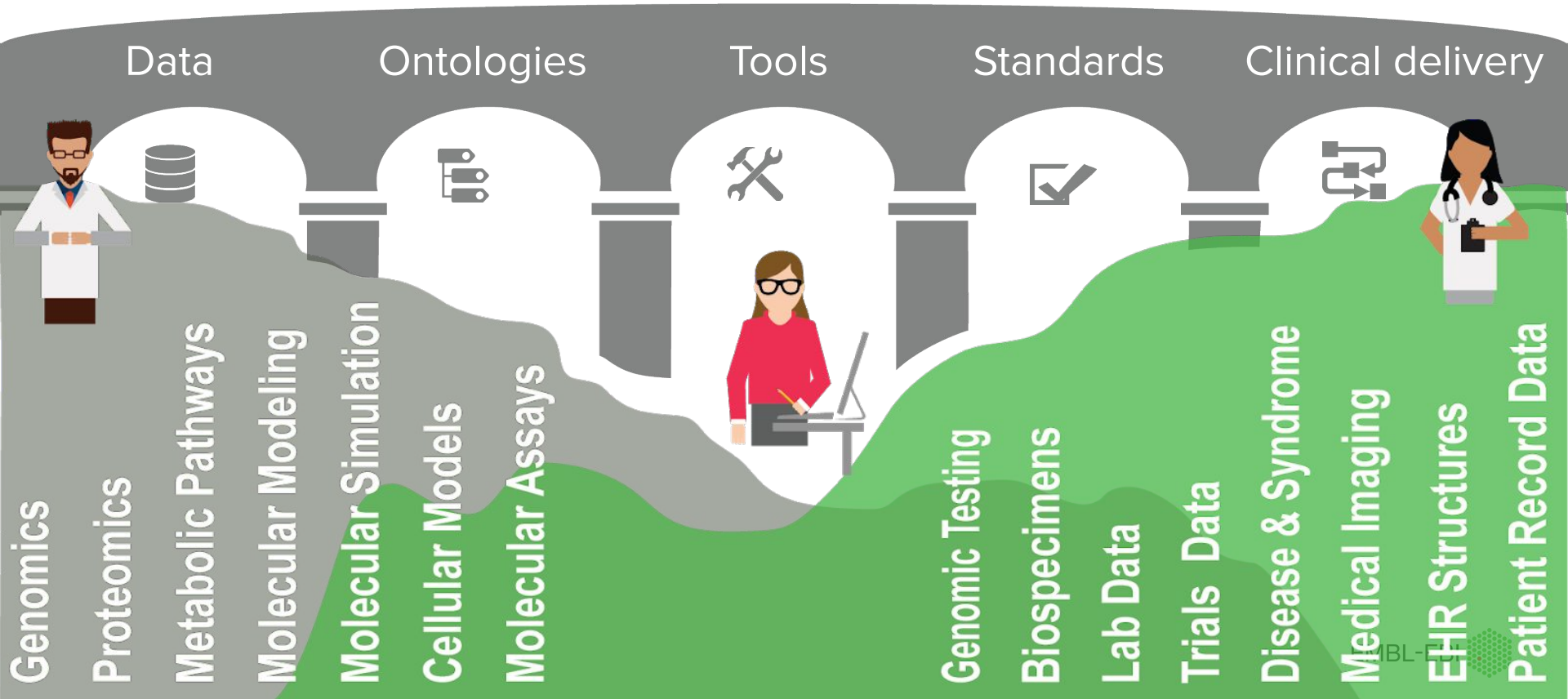
Unified Human Data

Development of ontologies and design patterns

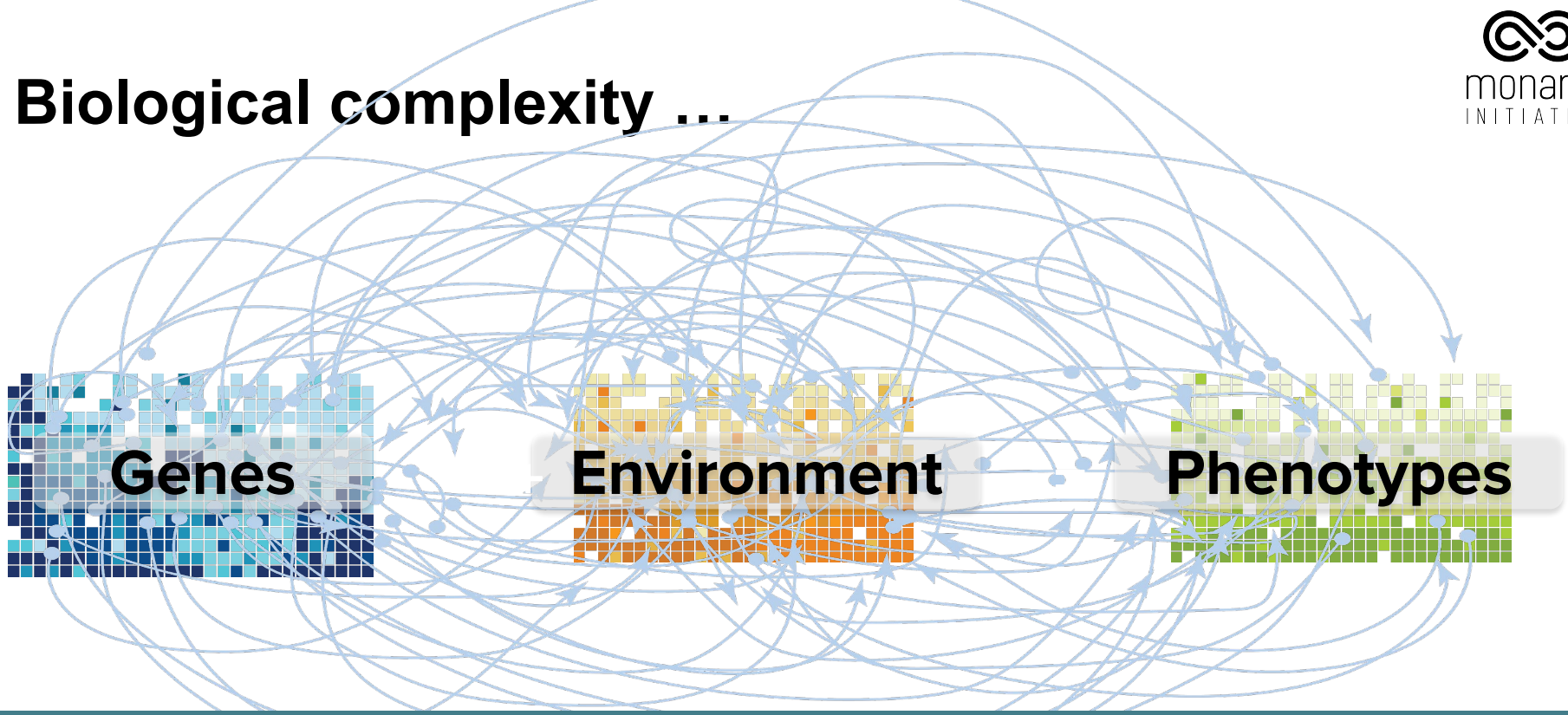
Tools, web services, visual widgets

E.g.,: Phenopackets

A bridge to the clinic



Biological complexity …



Standards for encoding and exchanging data must be up to these challenges.

And it's not just the bits …

G-P or D (disease)
causes
contributes to
is risk factor for
protects against
correlates with
is marker for
modulates
involved in
increases susceptibility to

G-G
regulates
negatively regulates (inhibits)
positively regulates (activates)
directly regulates
interacts with
co-localizes with
co-expressed with

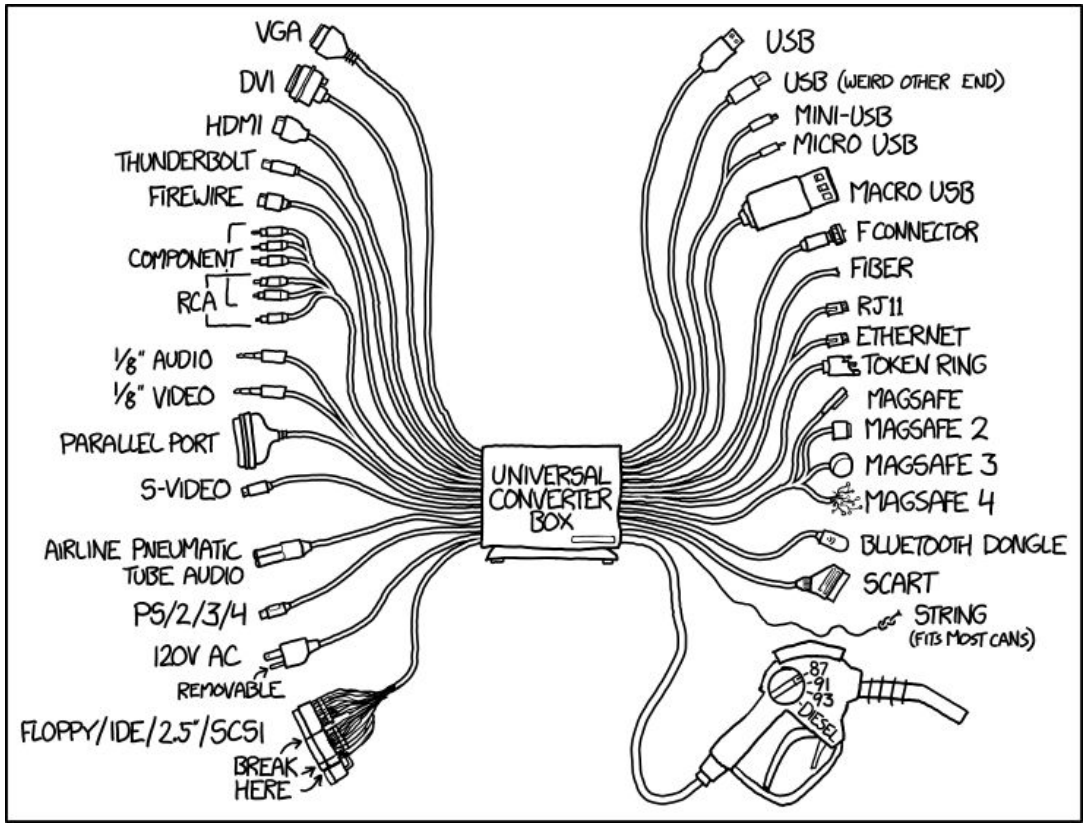
P/D - P/D
part of
hallmark of (P->D)

E-P
contributes to (E->P)
influences (E->P)
exacerbates (E->P)
manifest in (P->E)

G-E
expressed in
expressed during
contains

The relationships too must be captured

Semantics are the ultimate universal converter



http://xkcd.com/1406/

The role of cross-species phenotyping

“People are a lot like dogs”

http://honesttopaws.com/surprising-facts-about-dogs

- The dog’s retina has area centralis (analogous to the human macula) & fovea-like region, similar to humans; useful to study naturally occurring cone diseases

- Aged cats are natural models of Alzheimer’s Disease: they form Abeta oligomers, neurofibrillary tangles, and have neuronal loss

- Naked Mole Rats don't get cancer

- Armadillos are a natural host of M. leprae, the mycobacterium that causes leprosy (only one besides humans)

- Tree shrews’ glioblastomas are morphologically & genetically similar to humans (& more similar than mouse models)

- Great pond snails are models of inflammation-mediated memory dysfunction, and show evidence of spontaneous neural tissue regeneration after injury

- Silkworms are a model for uric acid metabolism. Decreases in plasma uric acid are correlated with clinical progression of Parkinson’s Disease

Other species aren’t just relevant; each has unique phenotypes

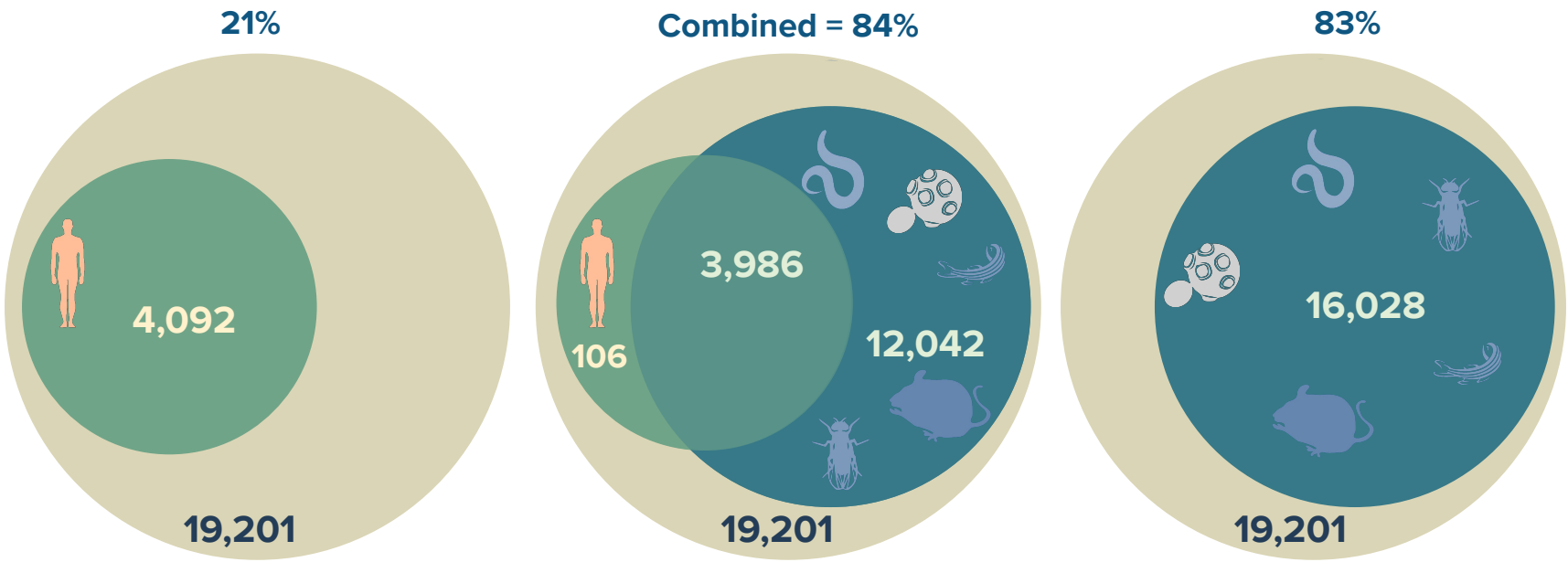
monarch INITIATIVE

EMBL-EBI

Model organisms matter to patients
More species = More coverage

monarch INITIATIVE

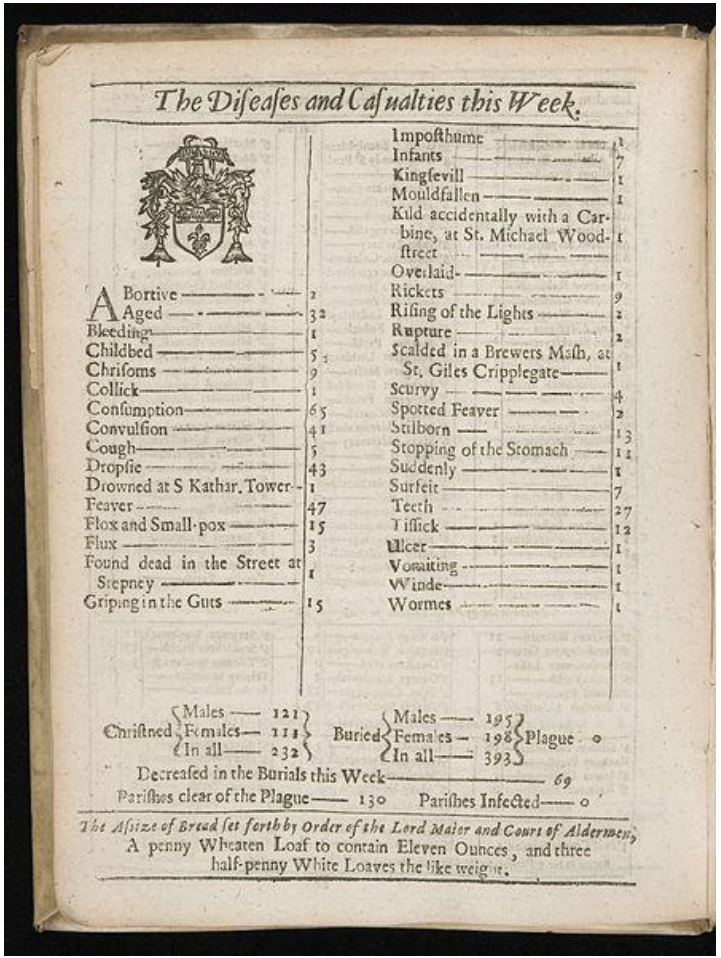
Model organisms provide key insight into phenotypic manifestations of human coding genes.



The inclusion of just five species boosts phenotypic coverage of genes by 63%

EMBL-EBI

What is an ontology?



The Diſeaſes and Caſualties this Week.

Abortive	2	Impoſthume	1
Aged	32	Infants	7
Bleeding	1	Kingſevill	1
Childbed	5	Mouldfallen	1
Chriſoms	9	Kild accidentally with a Carbine, at St. Michael Wood-ſtreet	1
Collick	1	Overlaid	1
Conſumption	65	Rickets	9
Convulſion	41	Riſing of the Lights	2
Cough	5	Rupture	2
Dropſie	43	Scalded in a Brewers Maſh, at St. Giles Cripplegate	1
Drowned at S Kathar. Tower	1	Scurvy	4
Feaver	47	Spotted Feaver	2
Flox and Small-pox	15	Stilborn	13
Flux	3	Stopping of the Stomach	11
Found dead in the Street at Stepney	1	Suddenly	1
Griping in the Guts	15	Surfeit	7
		Teeth	27
		Tiſſick	12
		Ulcer	1
		Vomiting	1
		Winde	1
		Wormes	1

Chriſtned: Males — 121, Females — 111, In all — 232
Buried: Males — 195, Females — 198, In all — 393
Plague — 0

Decreaſed in the Burials this Week — 69

Pariſhes clear of the Plague — 130 Pariſhes Infected — 0

The Aſſize of Bread ſet forth by Order of the Lord Maior and Court of Aldermen,
A penny Wheaten Loaf to contain Eleven Ounces, and three half-penny White Loaves the like weight.

Classification is not a new challenge …

London Bills of Mortality Feb 21 - 28 1664

Introduced to track deaths during the plague

Defined five kinds of infectious disease

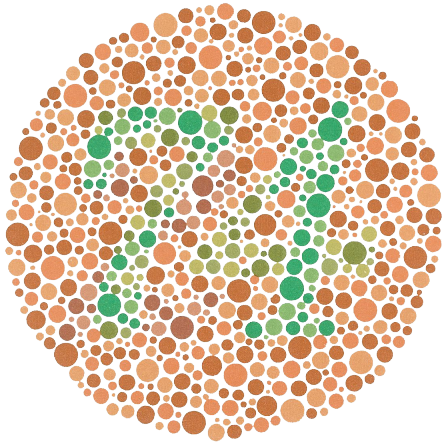
- Tuberculosis
- Small pox
- French pox
- Plague
- Measles

EMBL-EBI

Ambiguity and lack of precision in naming …

English is not a very precise language

- Same name for different concepts
- Different names for the same concept
- Changing names over time



Color blindness
Colour blindness
Abnormality of color vision
Colour vision defect, severe,
Abnormality of colour vision,
Loss in colour vision,
Colour vision defect,
Loss in color vision,
Color vision defect, severe
Abnormal colour vision,
Color vision defects,
Colour vision defects,
Abnormal color vision
Dyschromatopsia
Achromatopsia

cardiovascular disease

Search EFO | Search

http://www.ebi.ac.uk/efo/EFO_0000319 Copy

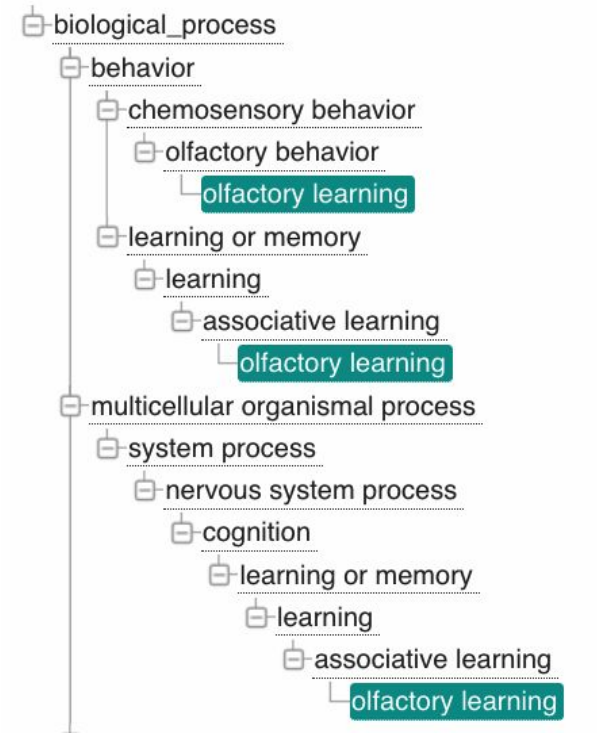
A disease involving the cardiovascular system. [https://orcid.org/0000-0002-6601-2165]

Synonyms: Cardiovascular Disorders | Circulatory system disease NOS (disorder) | Other ill-defined heart diseases | ASCVD | Other diseases of pericardium (disorder) | Cardiovascular system disease | Other pericardial disease NOS (disorder) | disorder of cardiovascular system | cardiovascular system disease or disorder | Other heart disease (disorder) | Other heart disease NOS (disorder) | Cardiovascular disease, unspecified | PAPILLARY MUSCLE DIS NEC | Disorder of circulatory system | Other diseases of pericardium | Disease affecting entire cardiovascular system (disorder) | Disorder of circulatory system, NOS | Other disorders of papillary muscle | Cardiovascular Disease (CVD) | [X]Other specified diseases of pericardium (disorder) | Other ill-defined heart disease | cardiovascular disease (CVD) | disease of subdivision of hemolymphoid system | Disorder of cardiovascular system (disorder) | Disease of cardiovascular system (disorder) | Ill-defined descriptions and complications of heart disease | cardiovascular disorder | Other heart disease NOS | disease or disorder of cardiovascular system | Cardiovascular disorder, NOS | CVD | Other sequelae of myocardial infarction, not elsewhere classified | [X]Cardiovascular disease, unspecified (disorder) | Other pericardial disease NOS | Other ill-defined heart disease NOS | Other forms of heart disease (disorder) | Other ill-defined heart disease NOS (disorder) | Unspecified circulatory system disorder | PERICARDIAL DISEASE NEC | CARDIOVASC DIS | CIRCULATORY DISEASE NOS | Other specified diseases of pericardium | Cardiovascular Diseases | Other diseases of endocardium (disorder) | Certain sequelae of myocardial infarction, not elsewhere classified | [X]Other specified diseases of pericardium | DISEASES OF THE CIRCULATORY SYSTEM | [X]Cardiovascular disease, unspecified | [X]Other ill-defined heart diseases | Disease of cardiovascular system, NOS | [X]Other forms of heart disease (disorder) | Disorder of the circulatory system | Other heart disease | Other ill-defined heart disease (disorder) | Other forms of heart disease | Other diseases of endocardium | [X]Other forms of heart disease | [X]Other ill-defined heart diseases (disorder) | Cardiovascular Disorder | circulatory system disease | disease of cardiovascular system | Other specified pericardial disease NOS | CVS disease | cardiovascular disease | cardiovascular system disease | Other specified pericardial disease NOS (disorder) | Diseases, Cardiovascular | Disorder of cardiovascular system | CVD, NOS | Disease, Cardiovascular | Cardiovascular disease, NOS | ILL-DEFINED HRT DIS NEC | Disease of cardiovascular system | Disease affecting entire cardiovascular system | Circulatory system disease NOS | OTHER SEQUELAE OF MI NEC

EMBL-EBI

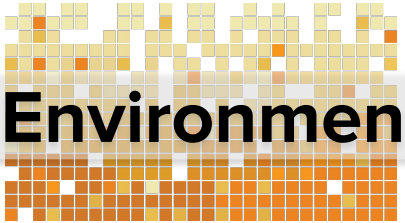
What is an ontology?

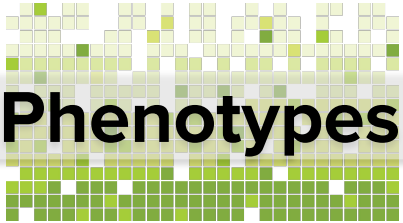
- **Representation of important things in a specific domain**
  - Describes types of entities and relations between them
- **An active, formal computational artifact**
  - A mathematical model based on a subset of first order logic
  - Tools can automatically process ontologies
- **A communication tool**
  - Provides a dictionary and shared understanding
  - Allows data sharing



Domain Ontologies

Starting point

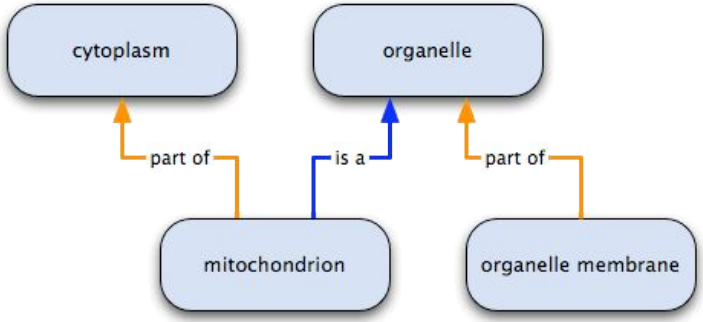


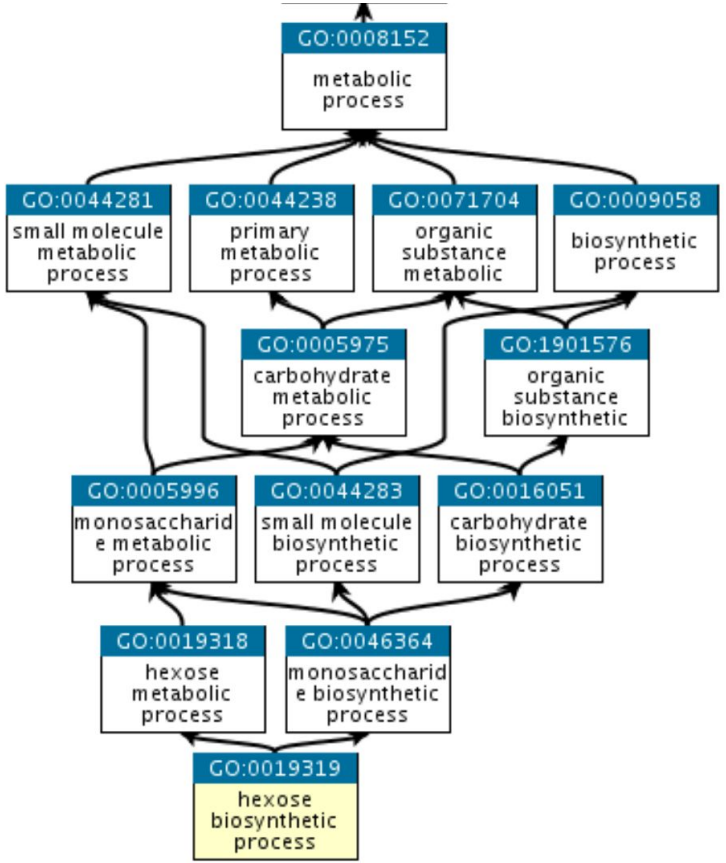


Disease = f<G, P, E>

GENEONTOLOGY
Unifying Biology

- Molecular function
- Cellular component
- Biological process



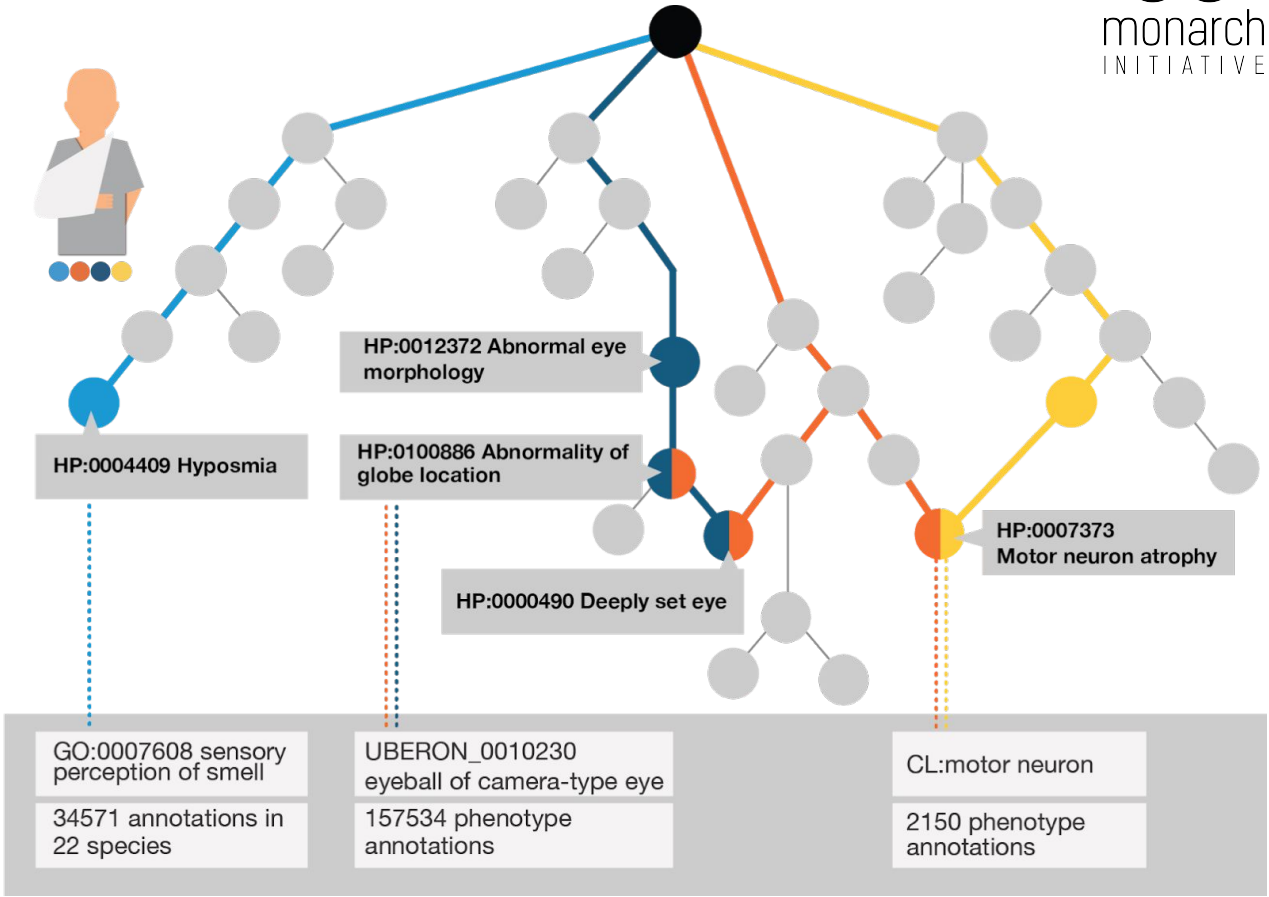


EMBL-EBI

Human Phenotype Ontology (HPO)

- **Phenotyping terminology**
  >14,500 terms

- **Widely adopted in rare disease genomic diagnostic tools**
  100,000 Genomes Project, SOLVE-RD, NIH-UDP, etc.



monarch INITIATIVE

human phenotype ontology

EMBL-EBI

OMIM®

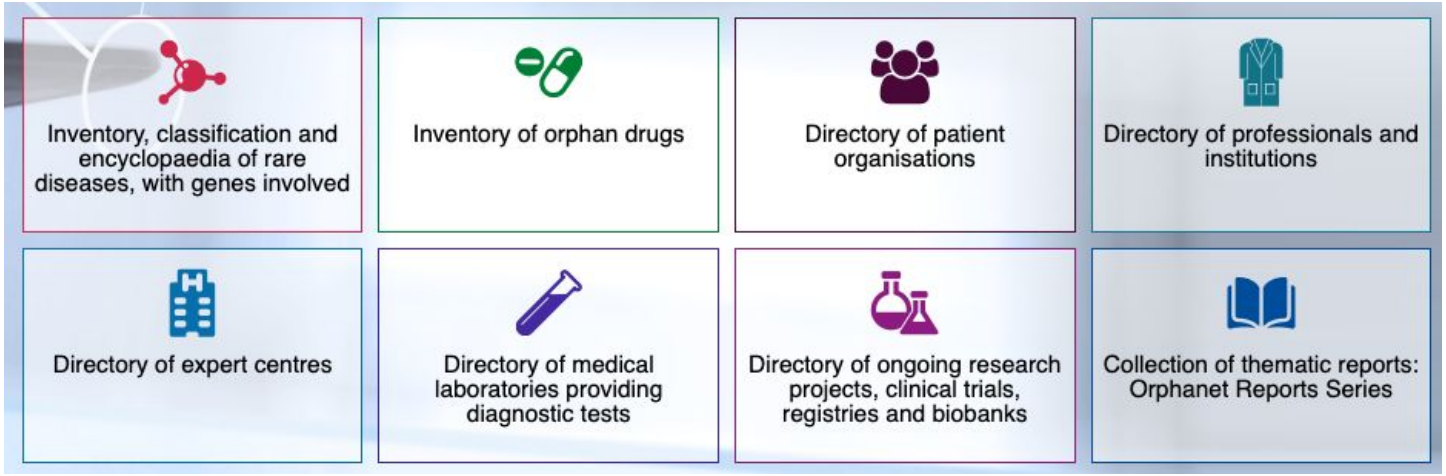
Online Mendelian Inheritance in Man®

An Online Catalog of Human Genes and Genetic Disorders

MIM Number Prefix	Autosomal	X Linked	Y Linked	Mitochondrial	Totals
Gene description *	15,913	751	51	37	16,752
Gene and phenotype, combined +	27	0	0	0	27
Phenotype description, molecular basis known #	5,994	369	5	34	6,402
Phenotype description or locus, molecular basis unknown %	1,401	112	4	0	1,517
Other, mainly phenotypes with suspected mendelian basis	1,650	102	3	0	1,755
Totals	24,985	1,334	63	71	26,453

EMBL-EBI

orphanet

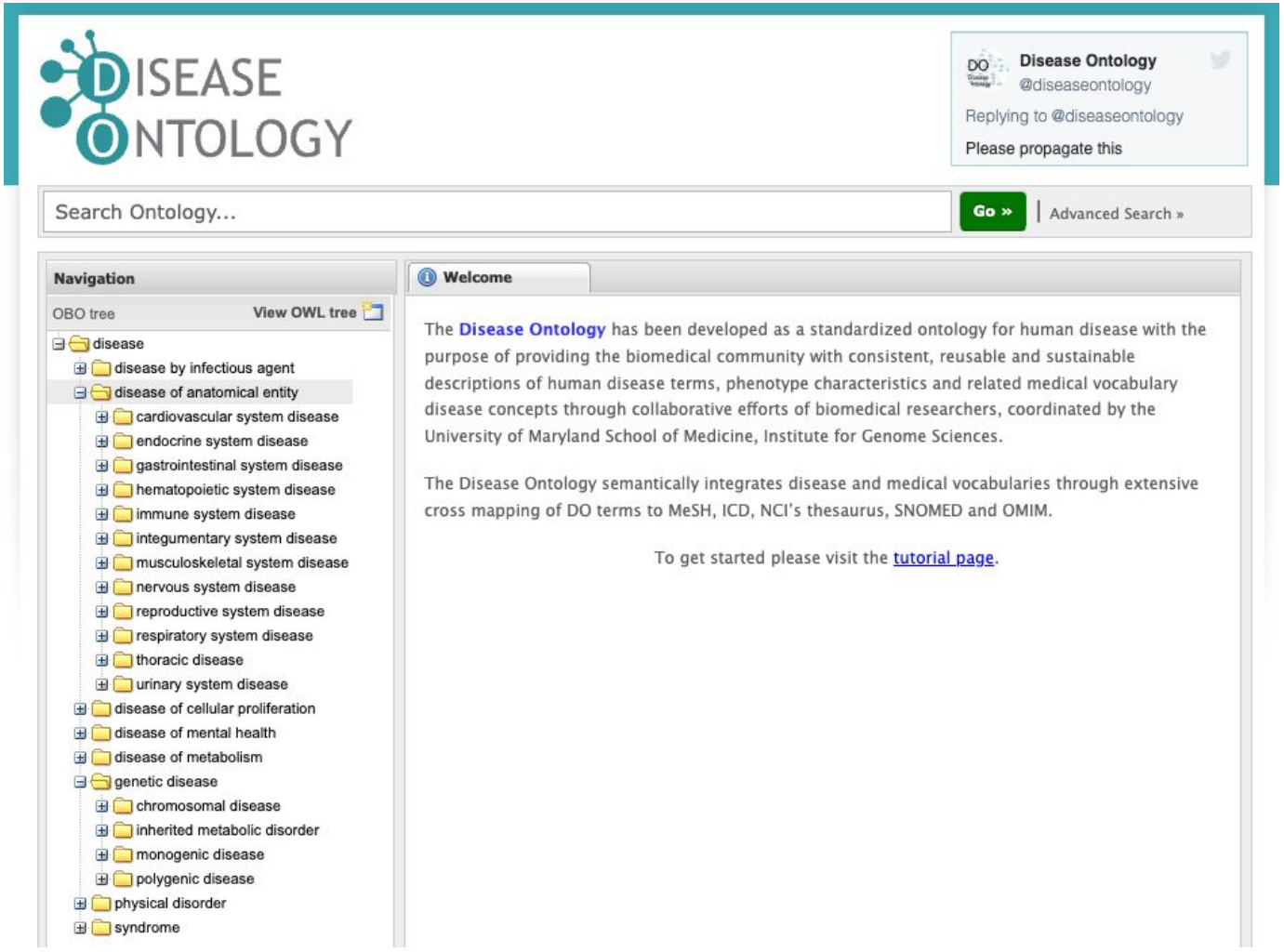

Inventory, classification and encyclopaedia of rare diseases, with genes involved
Inventory of orphan drugs
Directory of patient organisations
Directory of professionals and institutions
Directory of expert centres
Directory of medical laboratories providing diagnostic tests
Directory of ongoing research projects, clinical trials, registries and biobanks
Collection of thematic reports: Orphanet Reports Series

Browse Terms | Browse Properties

- CC BY 4.0
- age of onset
- clinical entity
  - disorder
    - Biological anomaly
    - Clinical syndrome
    - Disease
    - Malformation syndrome
    - Morphological anomaly
    - Particular clinical situation in a disease or syndrome
  - group of disorders
  - subtype of a disorder
- epidemiology
- genetic material
- geography
- inactive clinical entity
- inheritance

- clinical entity
  - disorder
  - group of disorders
    - Rare abdominal surgical disease
    - Rare allergic disease
    - Rare bone disease
    - Rare cardiac disease
    - Rare circulatory system disease
    - Rare developmental defect during embryogenesis
    - Rare disorder due to toxic effects
    - Rare disorder potentially indicated for transplant or complication after transplantation
    - Rare endocrine disease
    - Rare gastroenterologic disease
    - Rare genetic disease
    - Rare gynecologic or obstetric disease
    - Rare hematologic disease
    - Rare hepatic disease
    - Rare immune disease
    - Rare infectious disease
    - Rare infertility
    - Rare maxillo-facial surgical disease
    - Rare neoplastic disease
    - Rare neurologic disease
    - Rare odontologic disease
    - Rare ophthalmic disorder
    - Rare otorhinolaryngologic disease
    - Rare renal disease
    - Rare respiratory disease
    - Rare skin disease
    - Rare surgical thoracic disease
    - Rare systemic or rheumatologic disease
    - Rare systemic or rheumatological disease of childhood
    - Rare urogenital disease
  - subtype of a disorder

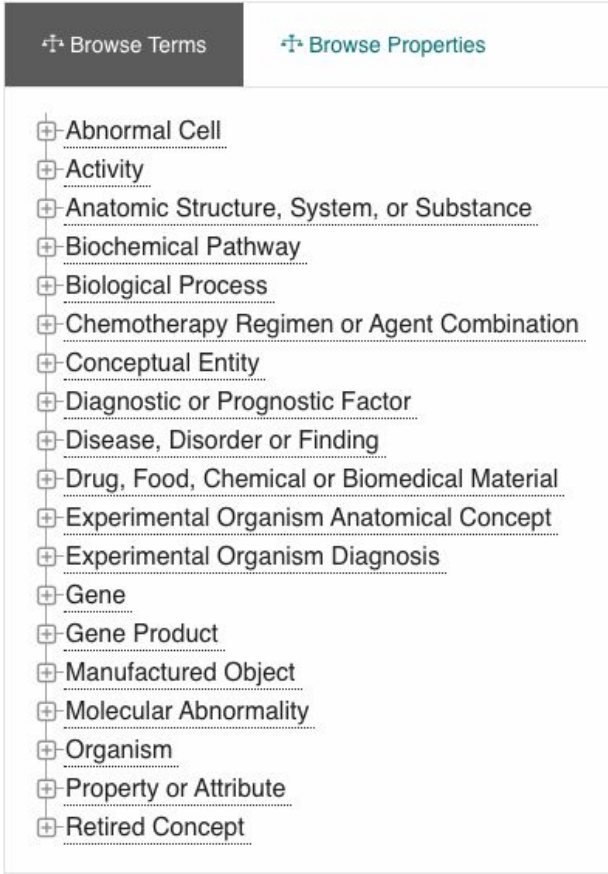
EMBL-EBI


DISEASE ONTOLOGY
Disease Ontology
@diseaseontology
Replying to @diseaseontology
Please propagate this
Search Ontology...
Go »
Advanced Search »
Navigation
OBO tree
View OWL tree
disease
disease by infectious agent
disease of anatomical entity
cardiovascular system disease
endocrine system disease
gastrointestinal system disease
hematopoietic system disease
immune system disease
integumentary system disease
musculoskeletal system disease
nervous system disease
reproductive system disease
respiratory system disease
thoracic disease
urinary system disease
disease of cellular proliferation
disease of mental health
disease of metabolism
genetic disease
chromosomal disease
inherited metabolic disorder
monogenic disease
polygenic disease
physical disorder
syndrome
Welcome
The Disease Ontology has been developed as a standardized ontology for human disease with the purpose of providing the biomedical community with consistent, reusable and sustainable descriptions of human disease terms, phenotype characteristics and related medical vocabulary disease concepts through collaborative efforts of biomedical researchers, coordinated by the University of Maryland School of Medicine, Institute for Genome Sciences.
The Disease Ontology semantically integrates disease and medical vocabularies through extensive cross mapping of DO terms to MeSH, ICD, NCI's thesaurus, SNOMED and OMIM.
To get started please visit the tutorial page.

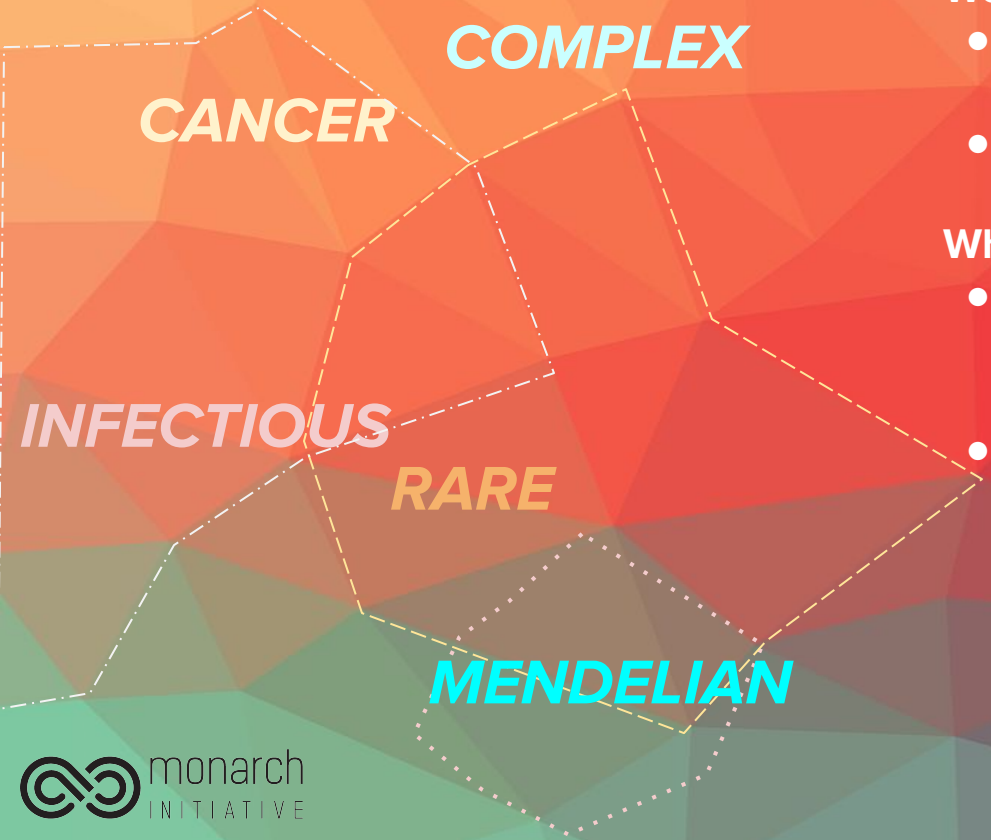
EMBL-EBI

NCI thesaurus

- Broad coverage of the cancer domain
- >100K concepts



What is the most clinically useful way to define and group diseases?

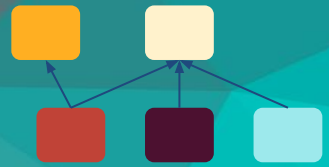


We needed:

- Disease concepts spanning multiple categories
- A systematic way of relating these concepts

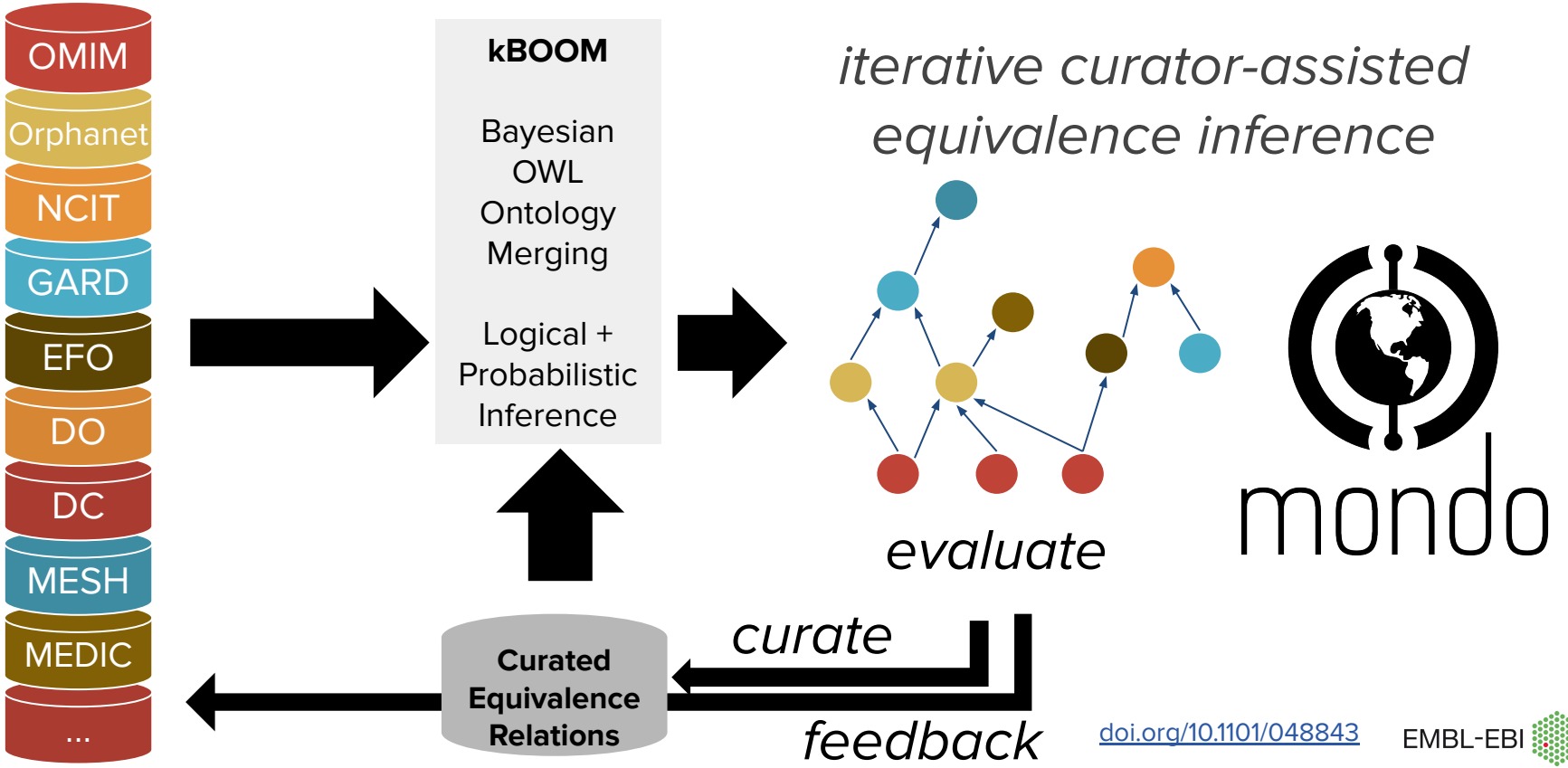
Why not just use mappings?

- Many terminologies / ontologies / lists include mappings
  - These can be used to cross-walk
- Problems:
  - Often mutually inconsistent
  - N^2 sets of mappings!
  - Not 1:1 equivalents



monarch INITIATIVE

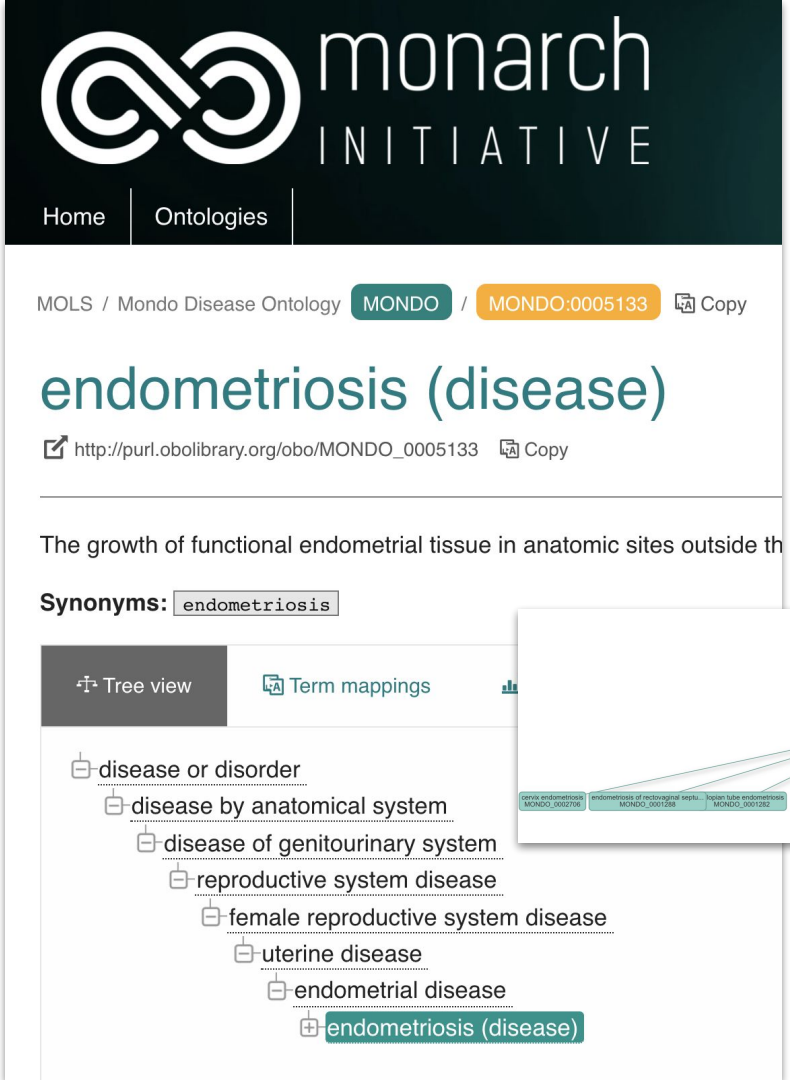
Evidence-based merging of equivalent disease concepts

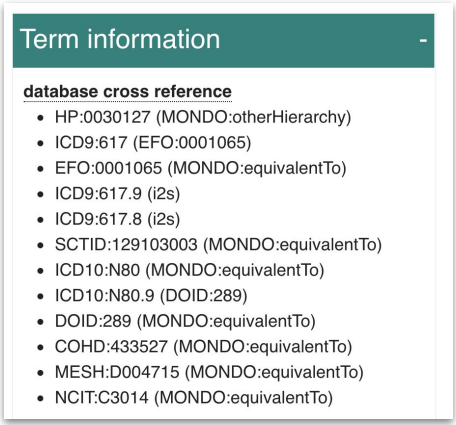


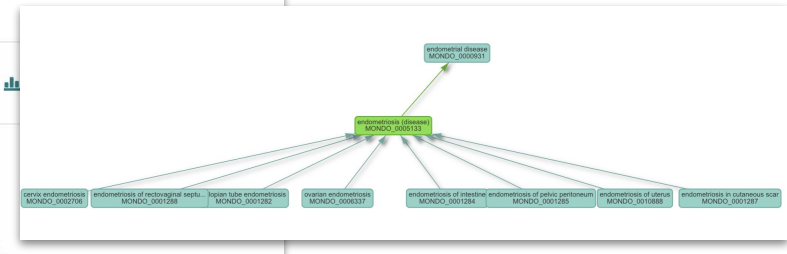
doi.org/10.1101/048843

Mondo

A logic-based structure to harmonize diseases and phenotypes across sources and species







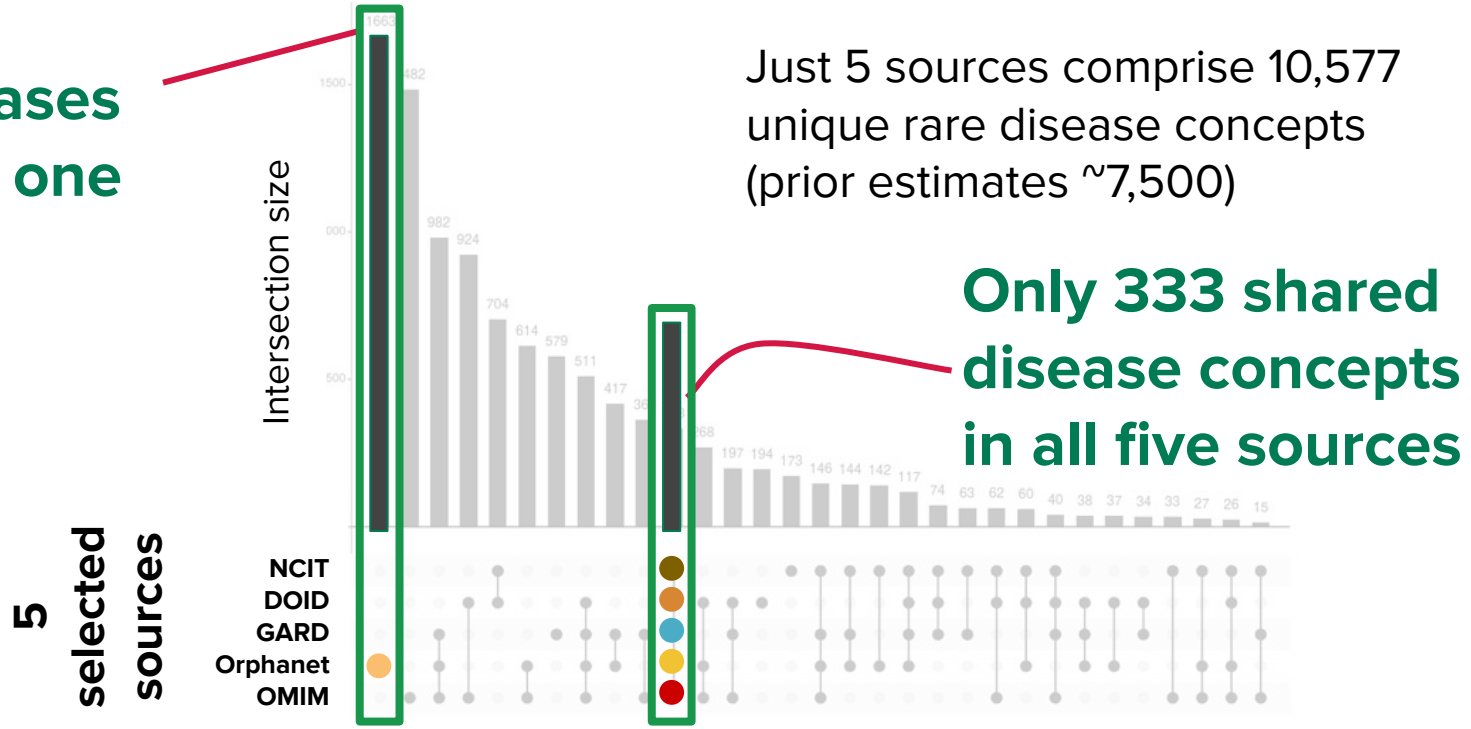
mondo

EMBL-EBI

If rare diseases are not counted, rare disease patients will not count

monarch INITIATIVE

Many diseases are in only one source



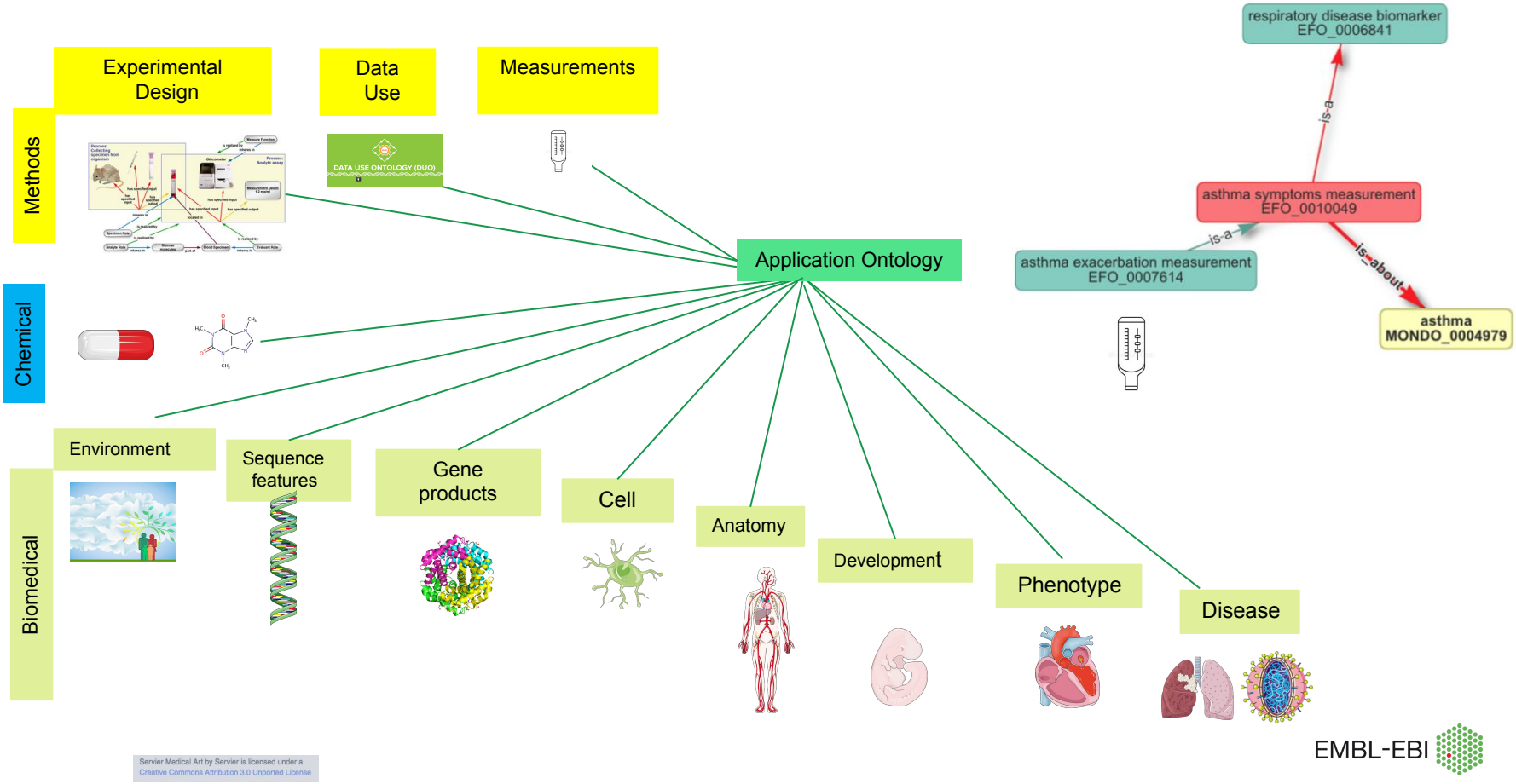
Just 5 sources comprise 10,577 unique rare disease concepts (prior estimates ~7,500)

Only 333 shared disease concepts in all five sources

Nature Reviews Drug Discovery (bit.ly/nature-rare-diseases)

EMBL-EBI

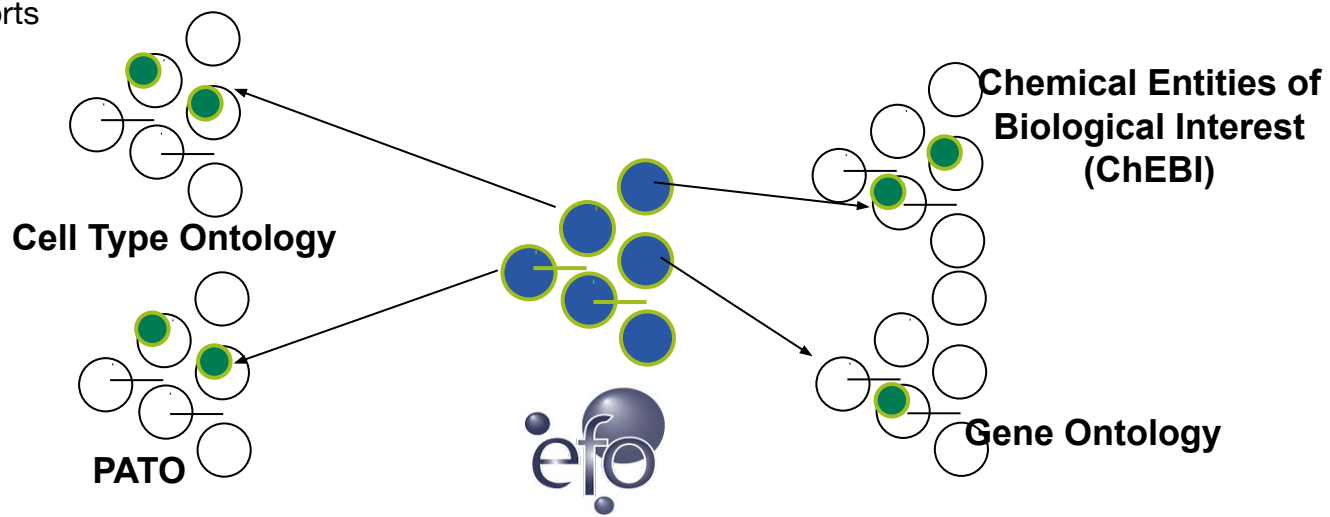
The complexity of the biomedical landscape



EMBL-EBI

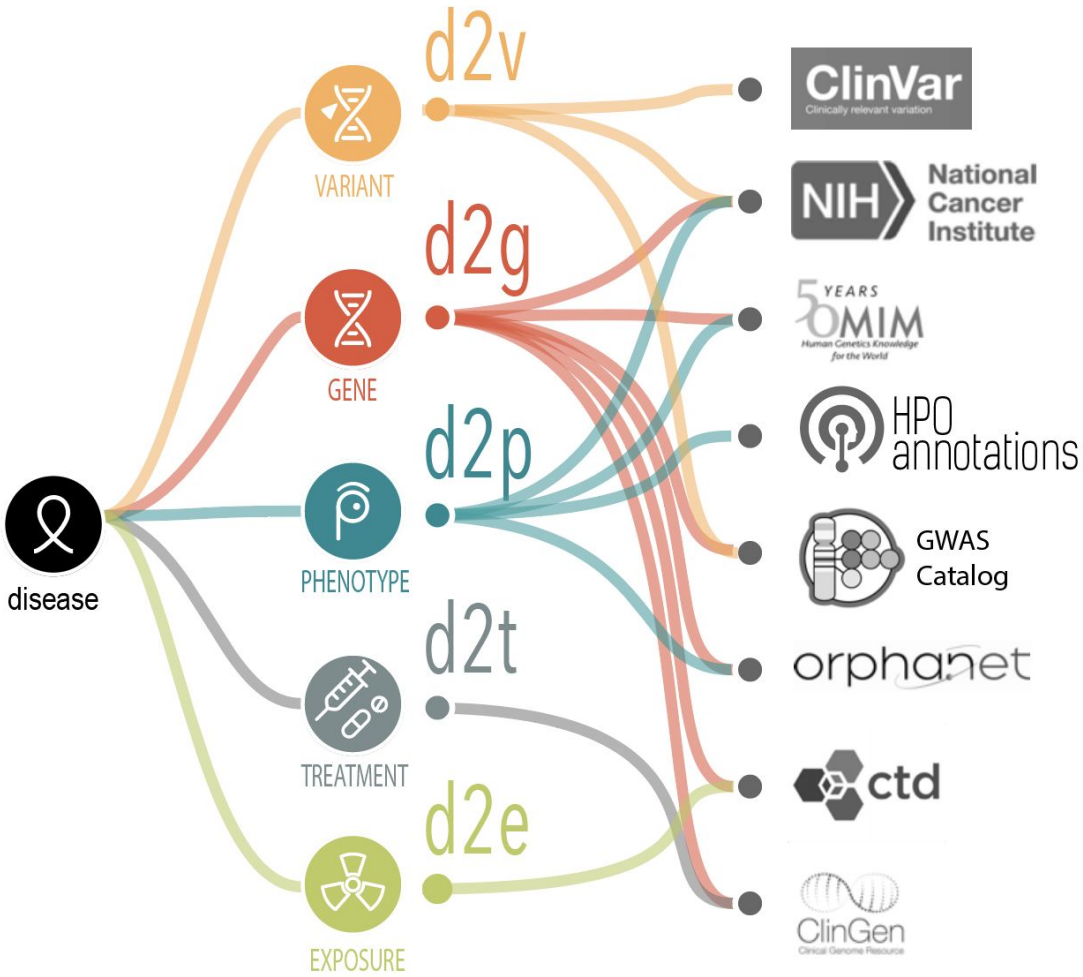
EFO - One ontology, many applications

- Experimental Factor Ontology is an **application ontology**, built for use in production services in OWL
  - Imports from >10 ontologies
  - Cross referenced to 25 additional ontologies
  - Extensive synonyms
  - Continuous integration build process, reasoning, manual error checking, multi-editor environment, imports



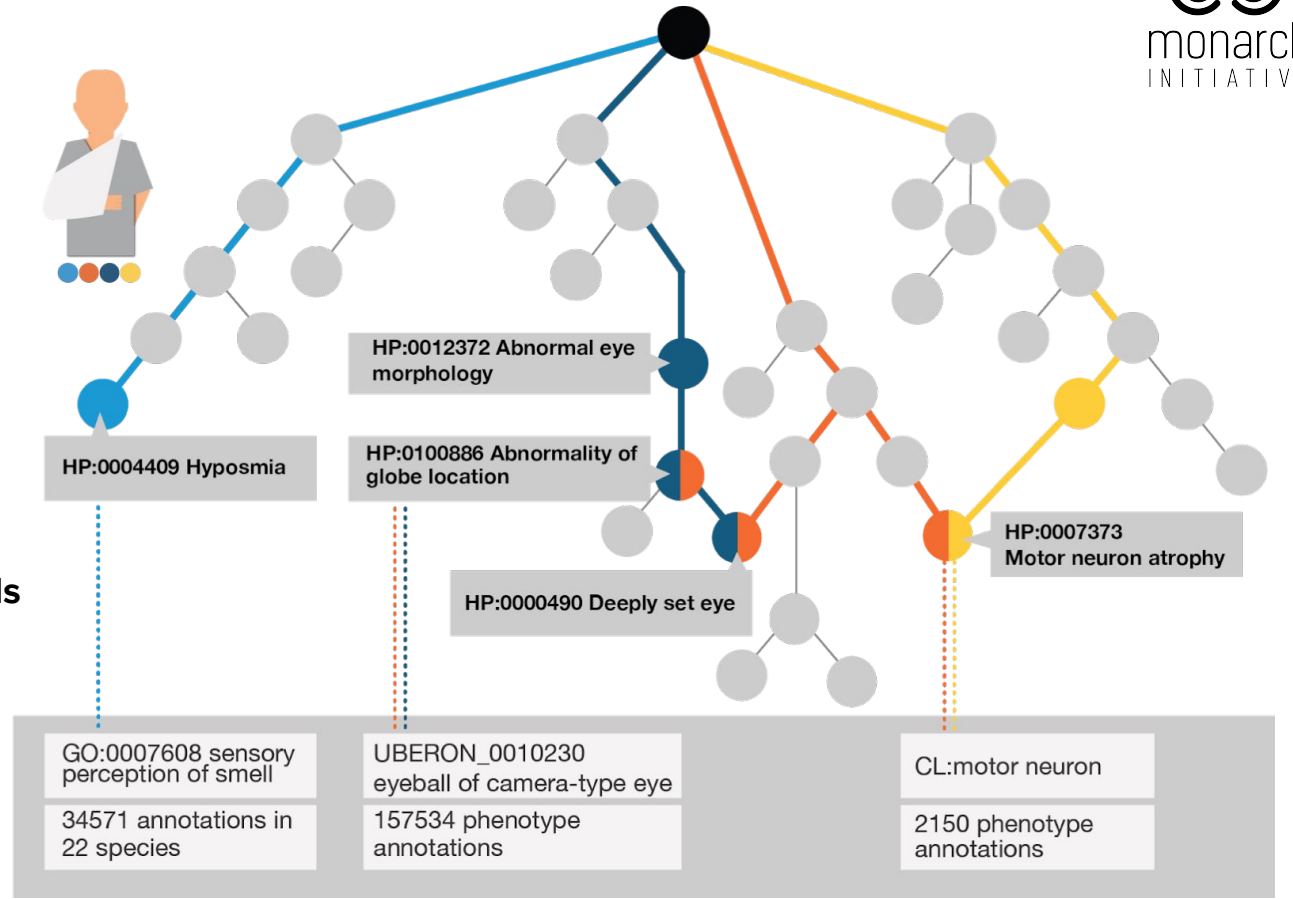
Knowledge bases

Different communities annotate different relationships, at different levels of granularity and using different vocabularies



Human Phenotype Ontology (HPO)

- **Computational disease models**
  >190,000 disease-phenotype annotations (associations)

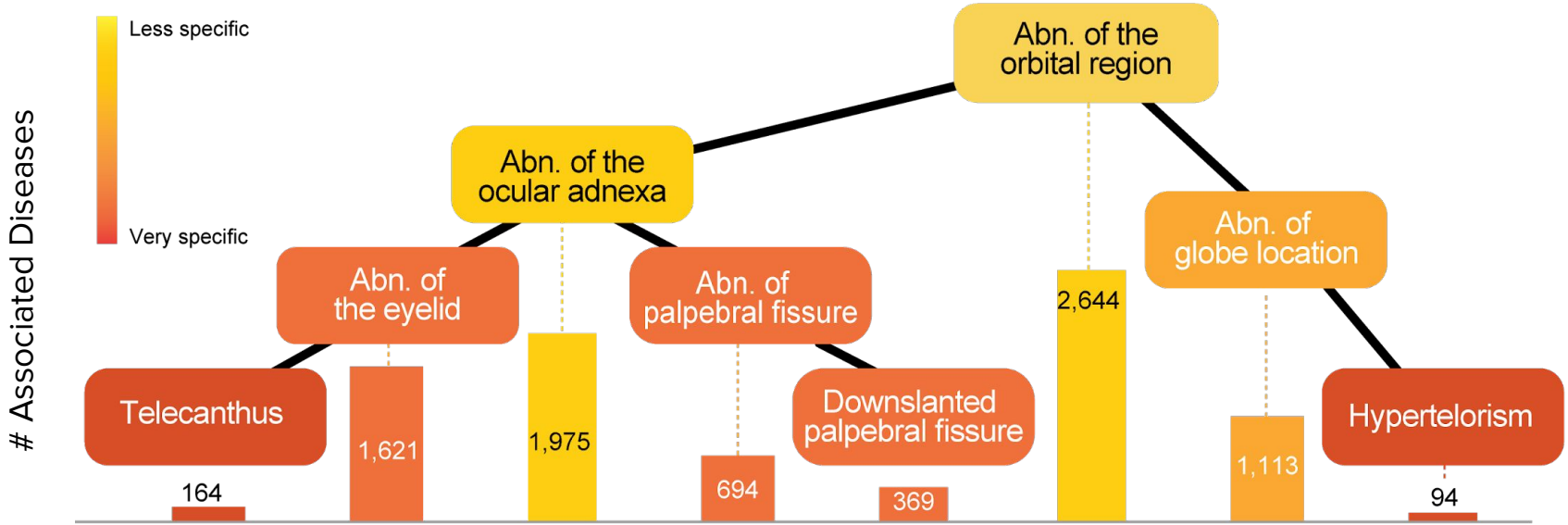


monarch INITIATIVE

human phenotype ontology

EMBL-EBI

Each disease has a gold standard phenotype profile



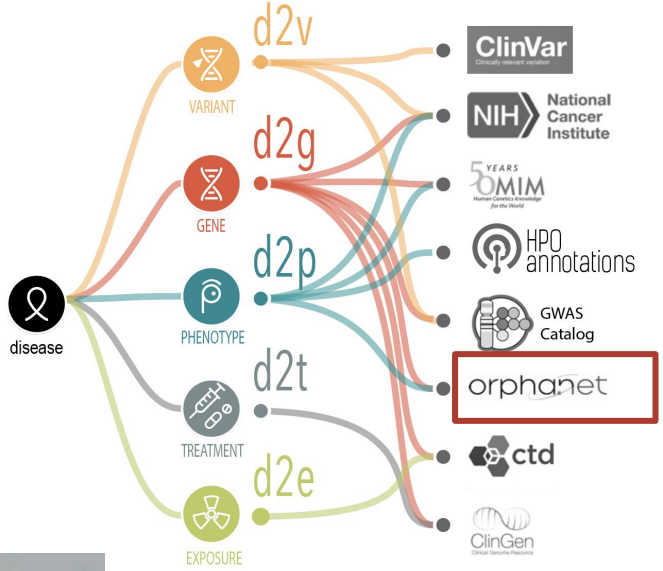
mondo

human phenotype ontology

EMBL-EBI

orphanet

- Frequency
- Functional consequences



EMBL-EBI

OMIM®

Online Mendelian Inheritance in Man®

An Online Catalog of Human Genes and Genetic Disorders

#100800
Table of Contents

MIM Entry

100800

ACHONDROPLASIA; ACH

INHERITANCE
- Autosomal dominant

GROWTH
Height
- Short-limb dwarfism identifiable at birth
- Mean male adult height, 131 cm
- Mean female height, 124 cm

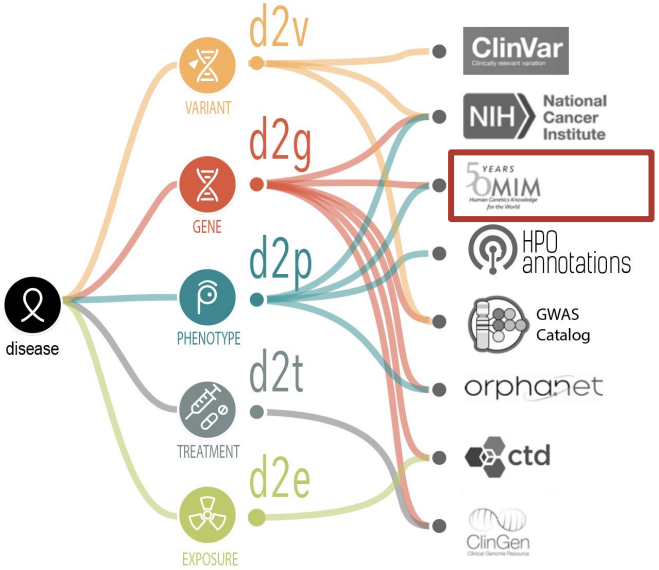
HEAD & NECK
Head
- Frontal bossing
- Megalencephaly

Face
- Midface hypoplasia

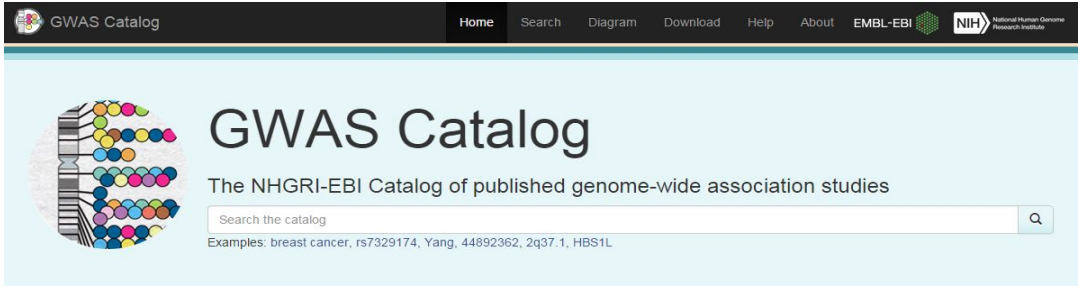
Ears
- Recurrent otitis media in infancy and childhood
- Conductive hearing loss

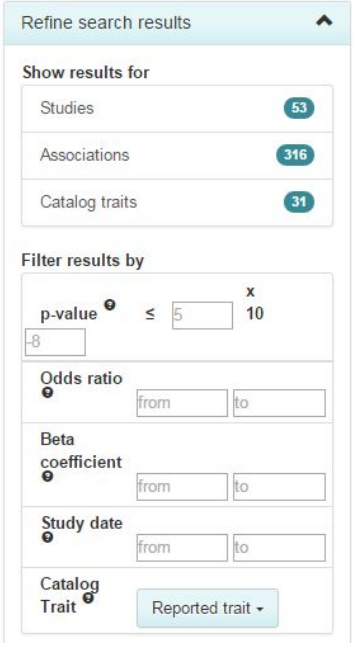
Nose
- Low nasal bridge

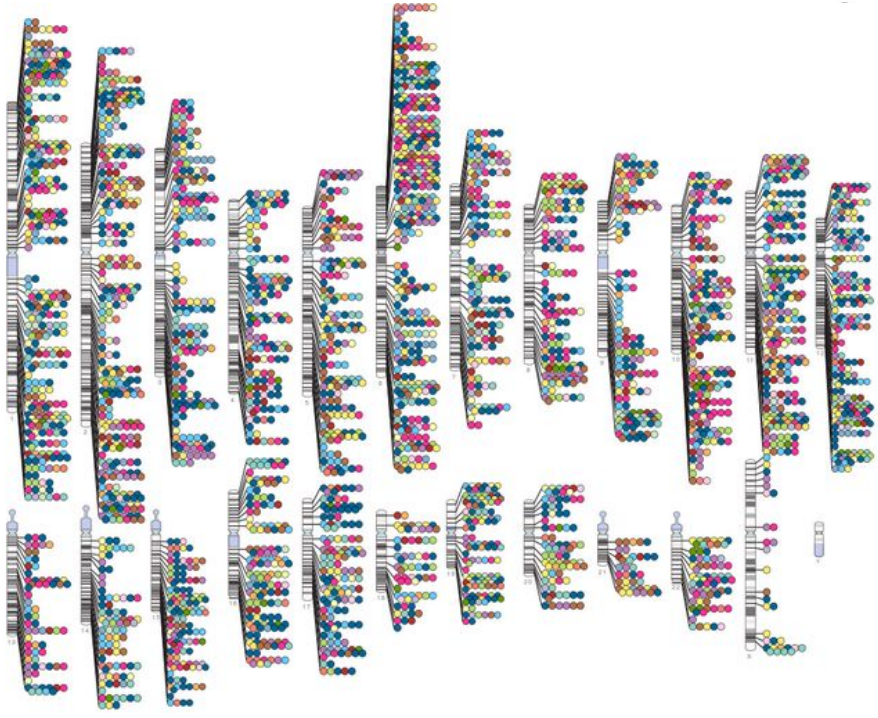
RESPIRATORY
Airways
- Upper airway obstruction

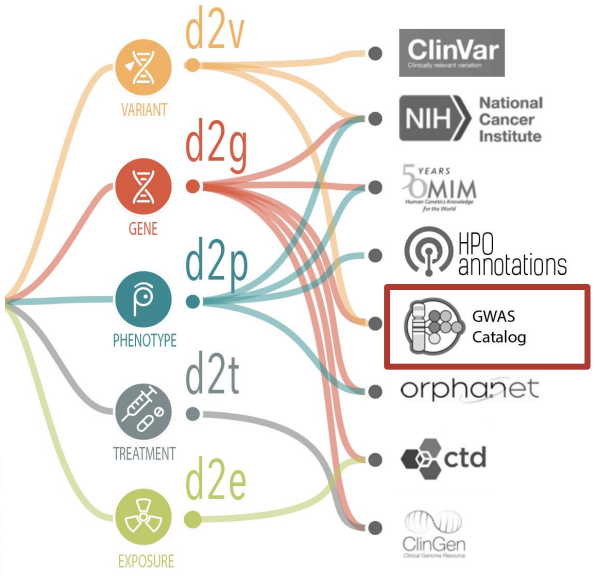


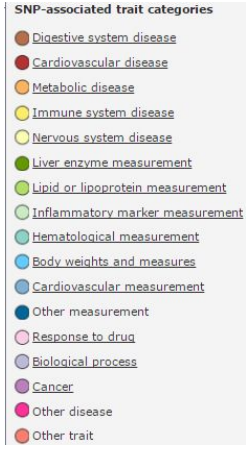
EMBL-EBI


GWAS Catalog
Home
Search
Diagram
Download
Help
About
EMBL-EBI
NIH
GWAS Catalog
The NHGRI-EBI Catalog of published genome-wide association studies
Search the catalog
Examples: breast cancer, rs7329174, Yang, 44892362, 2q37.1, HBS1L


Refine search results
Show results for
Studies 53
Associations 316
Catalog traits 31
Filter results by
p-value ≤ 5 x 10 -8
Odds ratio from to
Beta coefficient from to
Study date from to
Catalog Trait Reported trait

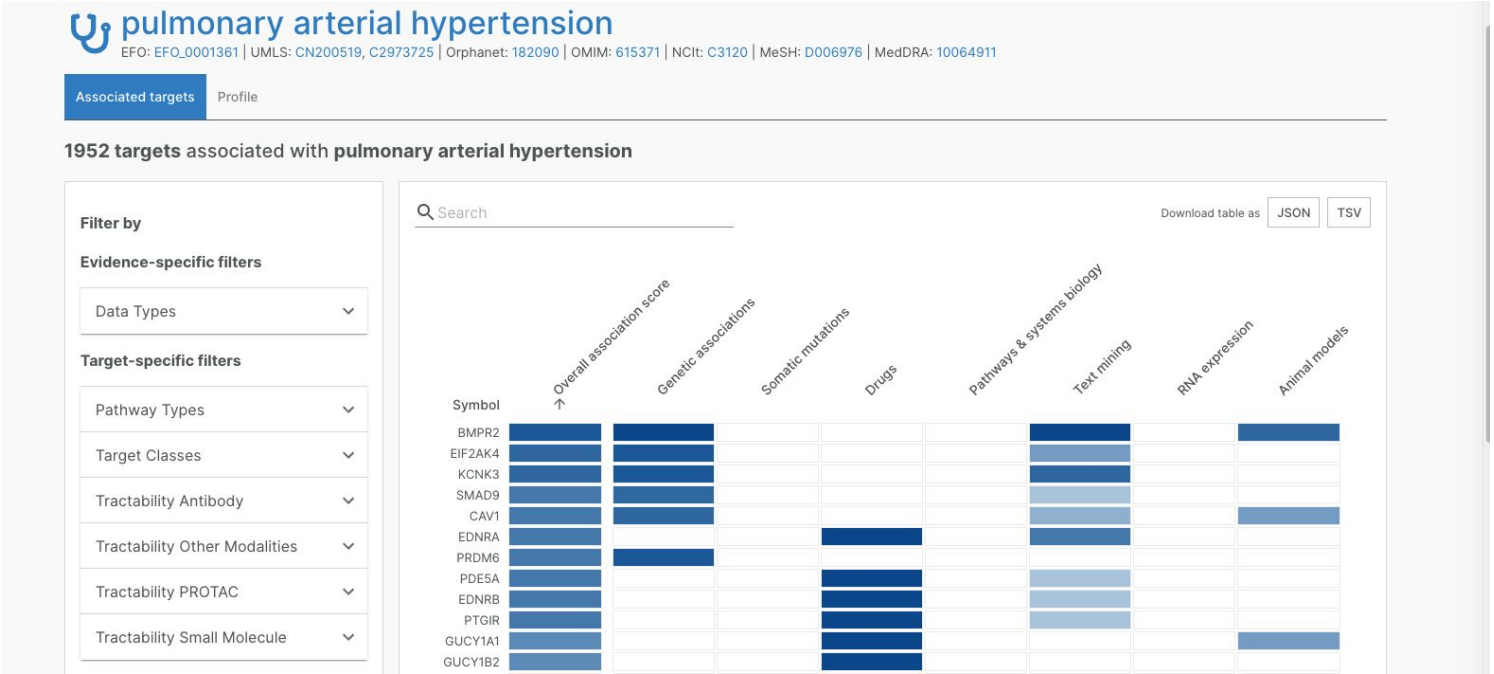



d2v
VARIANT
d2g
GENE
d2p
PHENOTYPE
d2t
TREATMENT
d2e
EXPOSURE
ClinVar
NIH National Cancer Institute
OMIM
HPO annotations
GWAS Catalog
orphanet
ctd
ClinGen


SNP-associated trait categories
Digestive system disease
Cardiovascular disease
Metabolic disease
Immune system disease
Nervous system disease
Liver enzyme measurement
Lipid or lipoprotein measurement
Inflammatory marker measurement
Hematological measurement
Body weights and measures
Cardiovascular measurement
Other measurement
Response to drug
Biological process
Cancer
Other disease
Other trait

EMBL-EBI

Open Targets platform



Open Targets

EMBL-EBI

pulmonary arterial hypertension

EFO: EFO_0001361 | UMLS: CN200519, C2973725 | Orphanet: 182090 | OMIM: 615371 | NCIt: C3120 | MeSH: D006976 | MedDRA: 10064911

Associated targets | Profile

Description

Pulmonary arterial hypertension (PAH) is a group of diseases characterized by elevated pulmonary arterial resistance leading to right heart failure. PAH is progressive and potentially fatal. PAH may be idiopathic and/ or familial, or induced by drug or toxin (drug-or toxin-induced PAH) or associated with ... [show more]

Synonyms

PPH1 | PHT - Pulmonary hypertension | pulmonary hypertension (disorder) | pulmonary hypertension, primary, 1 | Syndrome, Ayerza | Ayerza-Arrilaga Syndrome | PULM HYPERTENSION | PAH | pulmonary arterial hypertension | Ayerza Arrilaga Syndrome

... [show more]

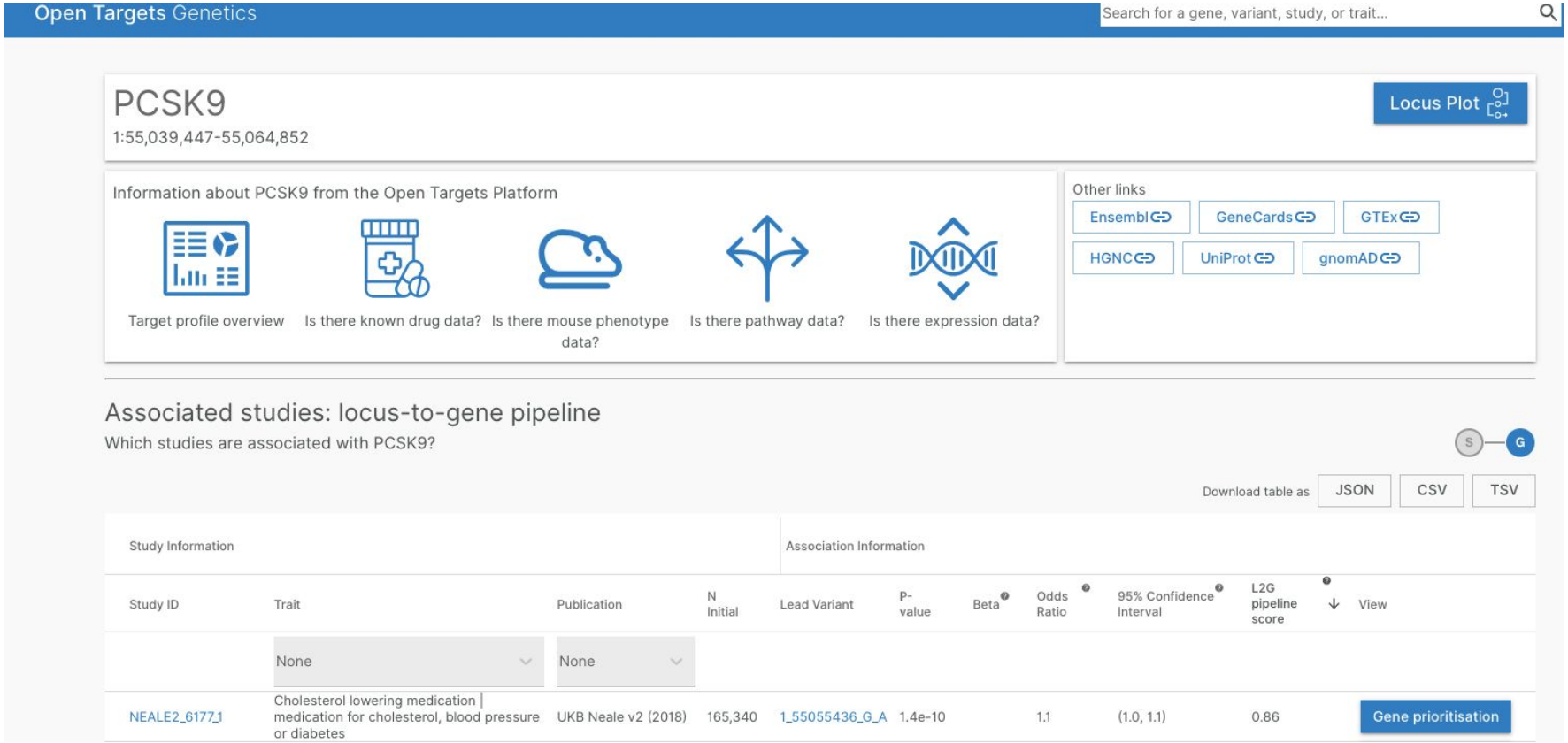
Ontology	Known Drugs	Clinical signs and symptoms	Bibliography
Belongs to 2 therapeutic areas	*73 drugs with 148 targets*	*no data*	*19,943 publications*

Ontology

*Ontology subgraph including children, ancestors and therapeutic areas of **pulmonary arterial hypertension**. Source: EFO.*

Open Targets

EMBL-EBI



Open Targets

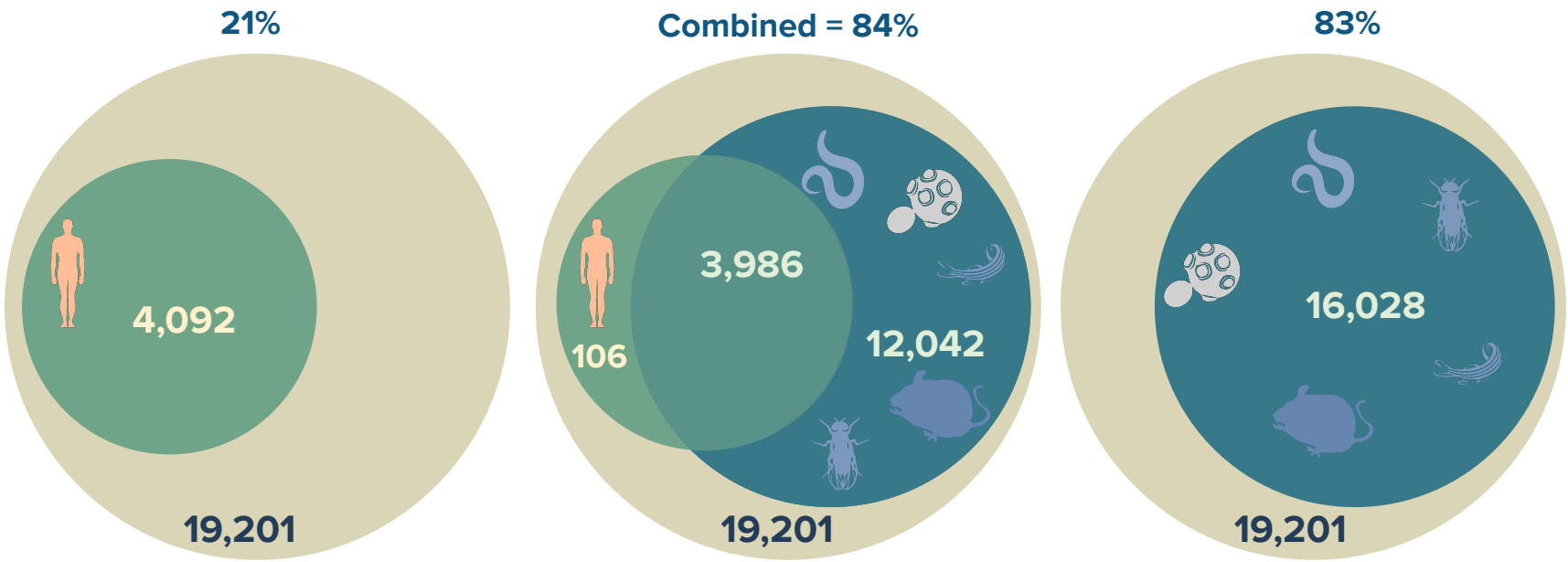
EMBL-EBI

Cross-species computational phenotyping

Model organisms matter to patients
More species = More coverage

monarch INITIATIVE

Model organisms provide key insight into phenotypic manifestations of human coding genes.



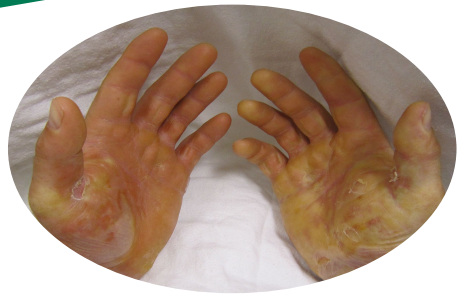
The inclusion of just five species boosts phenotypic coverage of genes by 63%

EMBL-EBI

Different communities use different languages

monarch INITIATIVE

Palmoplantar hyperkeratosis

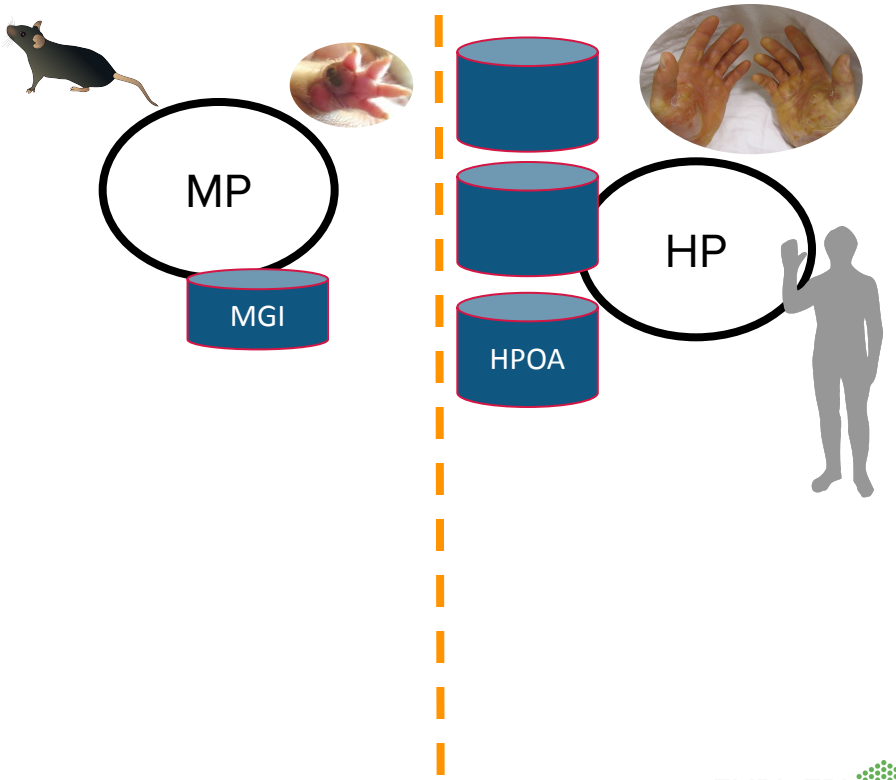


Thick hand skin

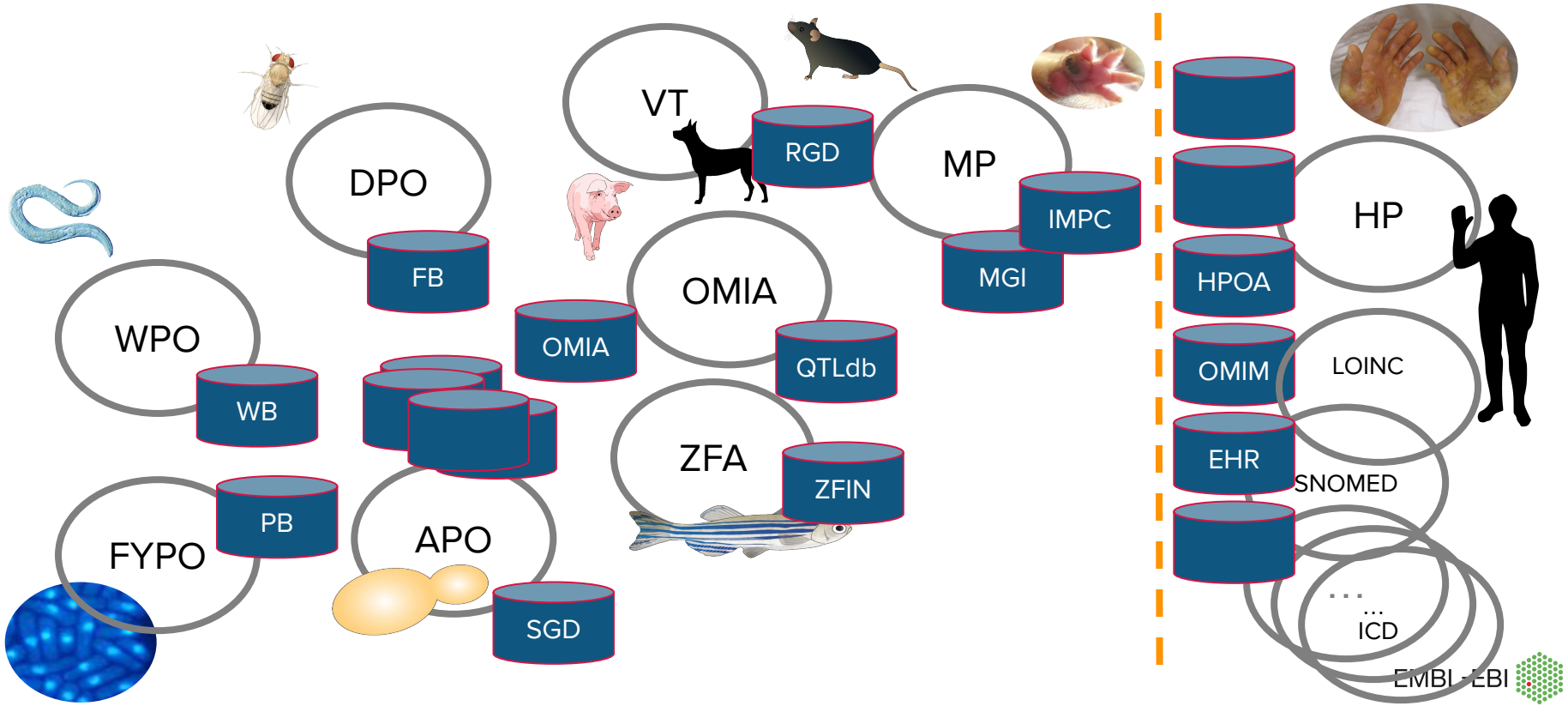
Ulcerated paws

EMBL-EBI

Challenge: Each data source uses their own vocabulary/ontology



Challenge: Each data source uses their own vocabulary/ontology

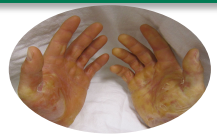


Logical decomposition of complex concepts allows interoperability

monarch INITIATIVE

Human phenotype

"Palmoplantar hyperkeratosis"



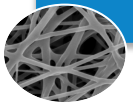
=

PATO

increased

GO

keratinization

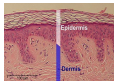


Mouse phenotype

"Ulcerated paws"

=

Uberon



Stratum corneum layer of skin

Autopod

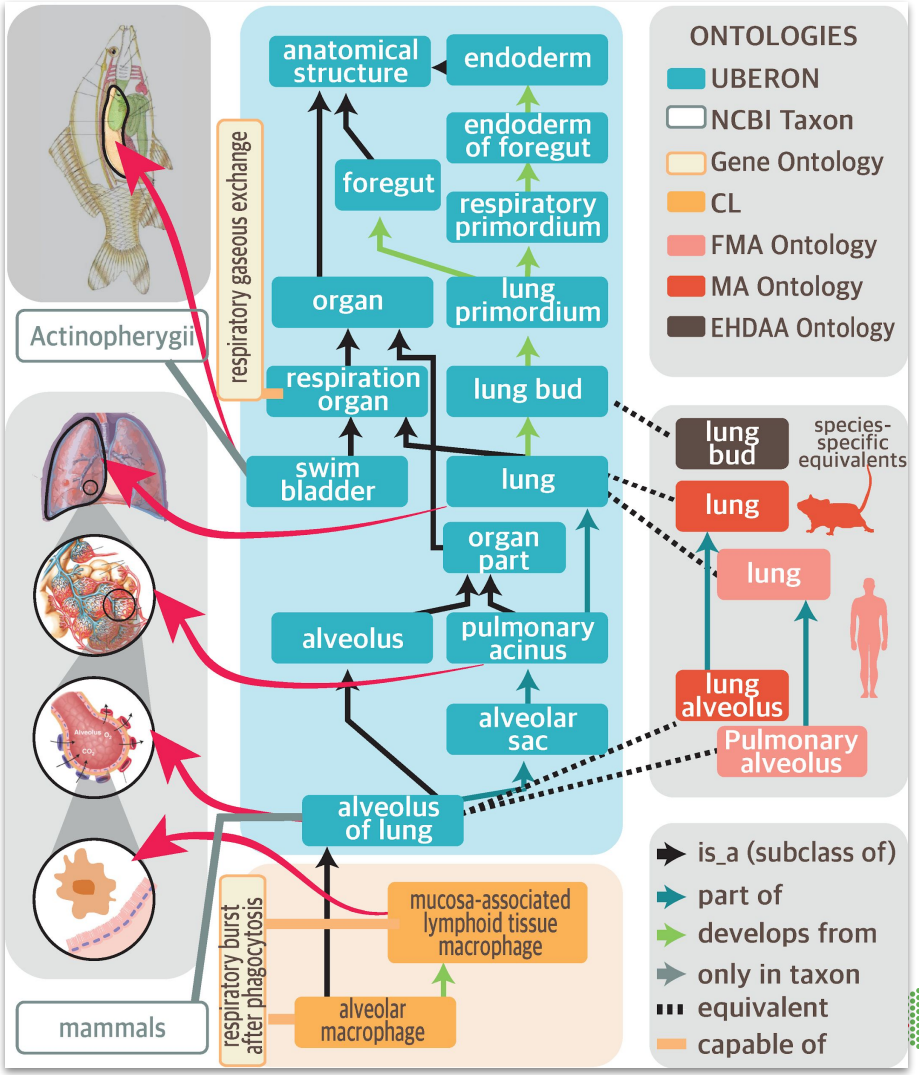
Species neutral ontologies, homologous concepts

EMBL-EBI

Uberon

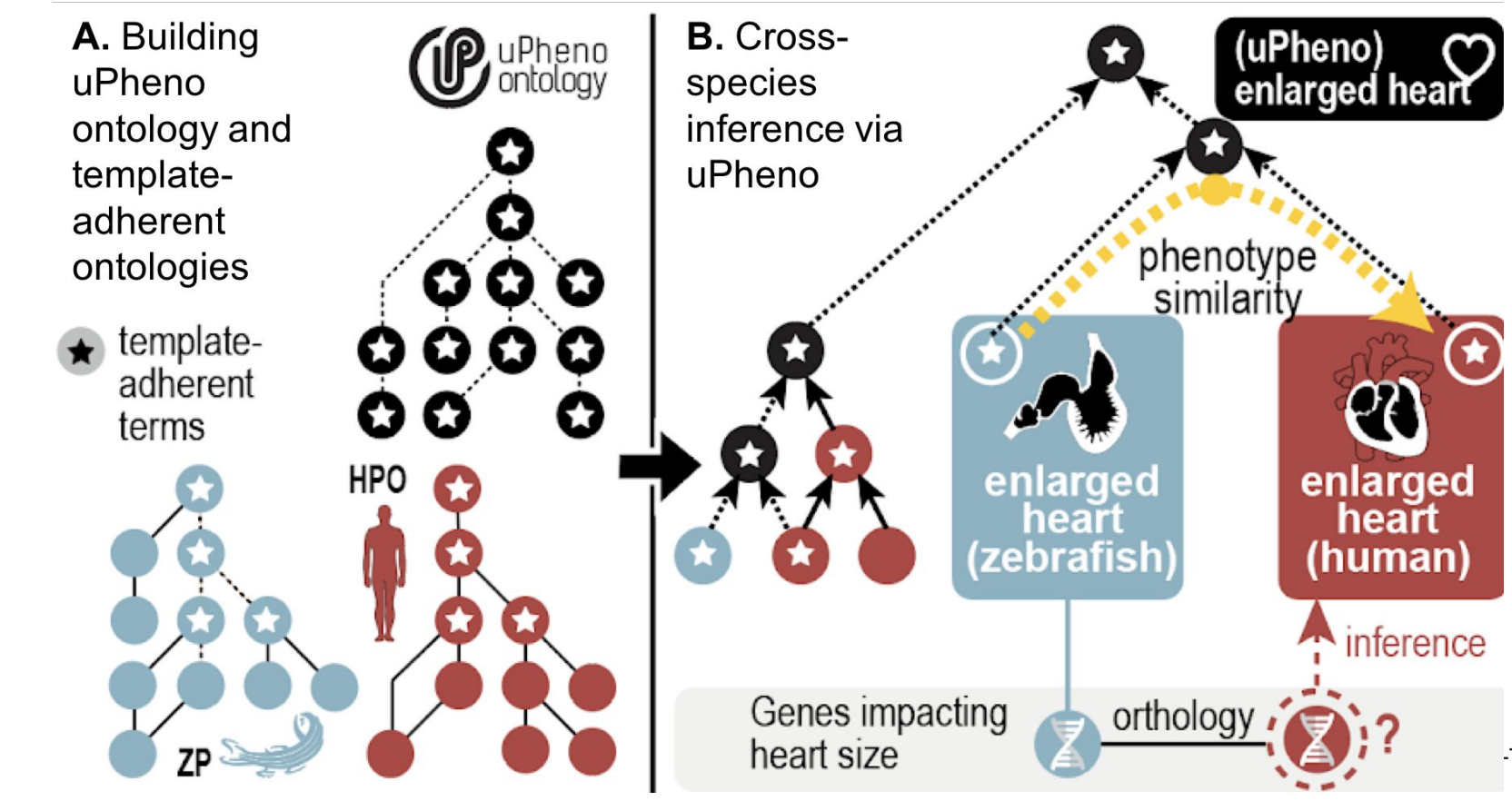
multi-species knowledge graph

- Integrate multiple anatomy ontologies into a unified, interoperable, cross-species one
- Can readily generate different views for different taxa, domains (e.g., respiratory system), or contexts (e.g., data collection forms)



monarch INITIATIVE

Template-driven ontology development and harmonization



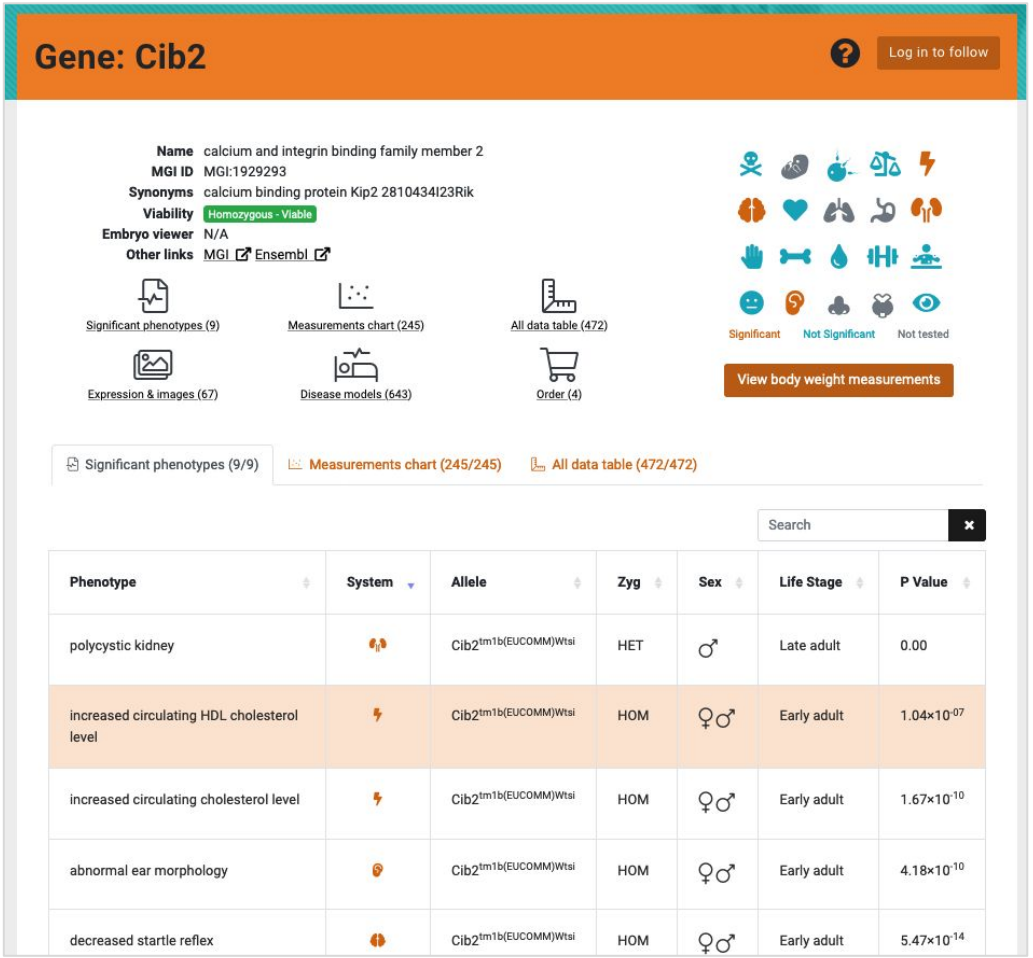
IMPC

International Mouse Phenotyping Consortium

8093 total knockout genes phenotyped

- Number of phenotyped lines: **8741**
- Statistically Significant Calls: **93235**
- Data Release Version: **16.0**
- Published: **21 April 2022**

View Release Full Report

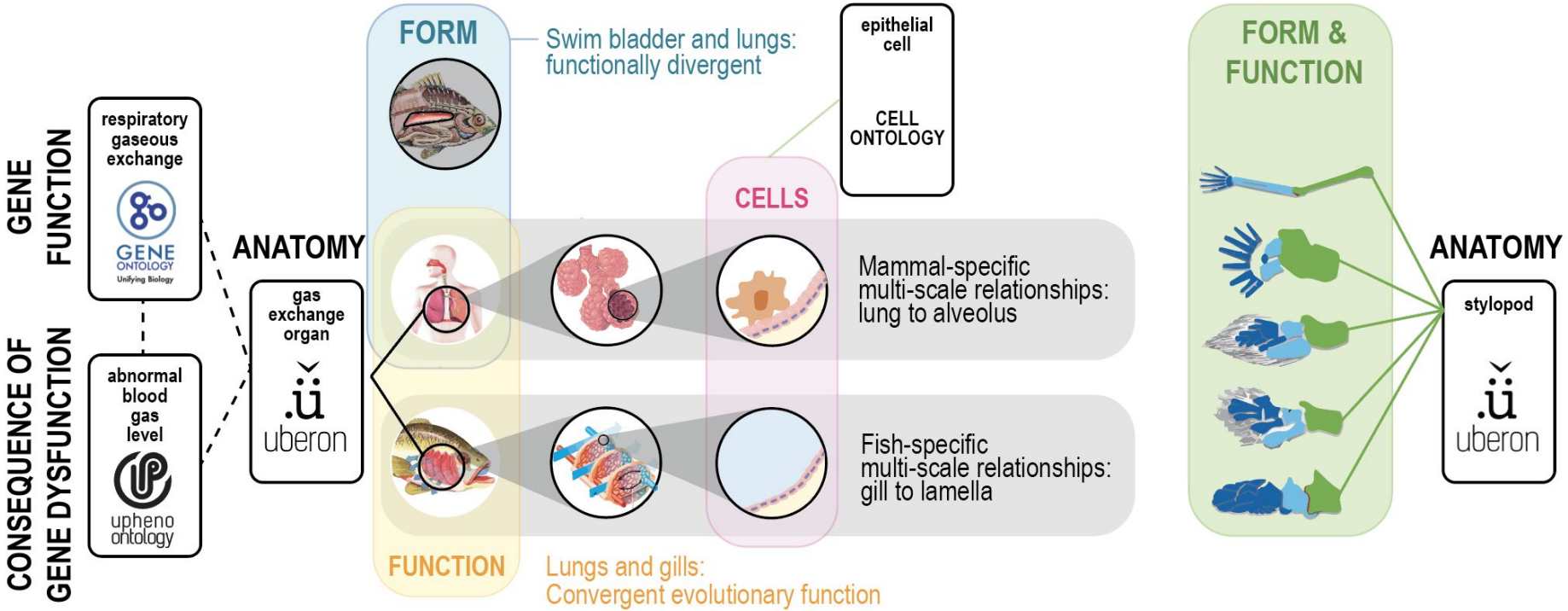


EMBL-EBI

BREAK

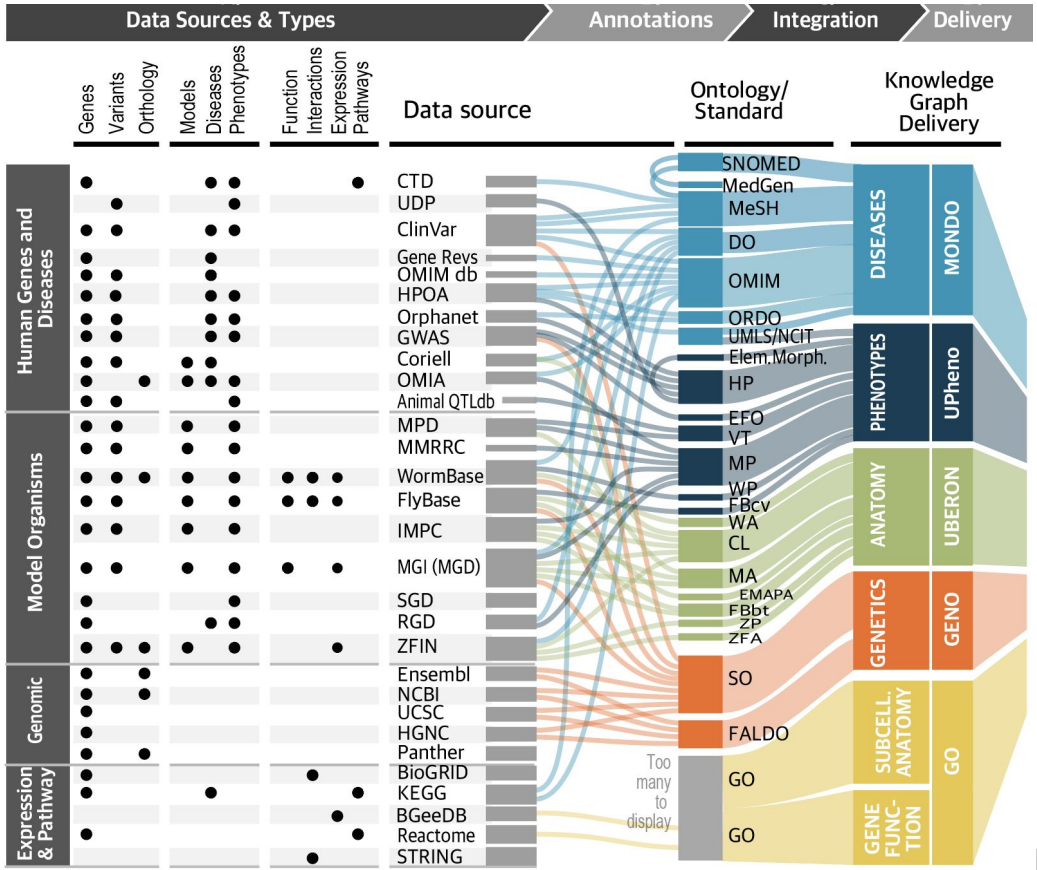
Putting it all together

We need computable means to relate form, function, and dysfunction in order to interpret the genome (for diagnostics or otherwise)



EMBL-EBI

Monarch Knowledge Graph

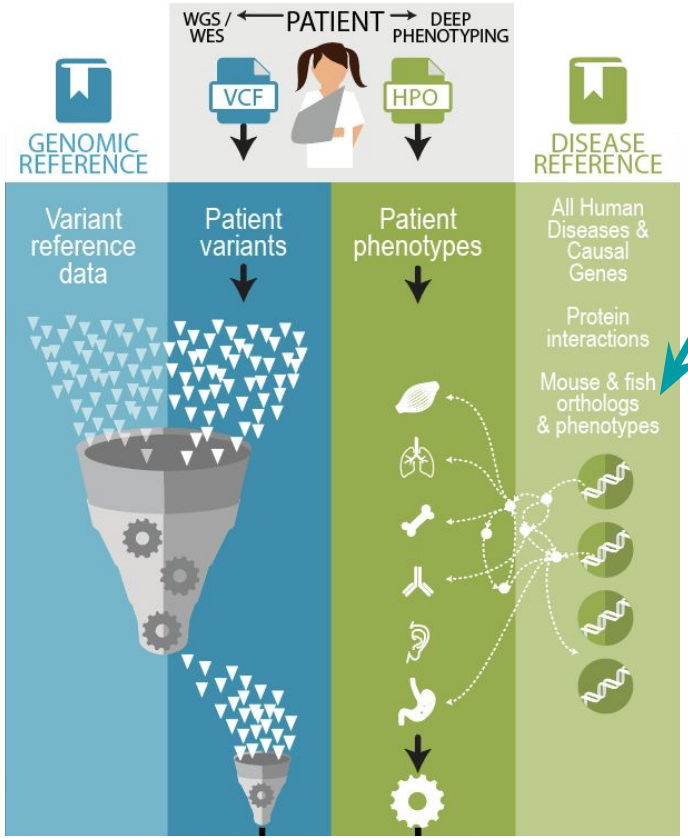


Genotype-phenotype associations across a variety of species

- **Gene to Phenotype Associations**: 818,690
  From approx. 50 species, including mouse, worm, yeast, American mink, Japanese rice fish, various species of livestock, and many species in the Drosophila group
- **Causal Gene to Disease Associations**: 9,197
  From human and mouse data
- **Non-causal Gene to Disease Associations**: 30,220
  From more than 70 species

https://api.monarchinitiative.org

Graph Machine Learning on Monarch KG

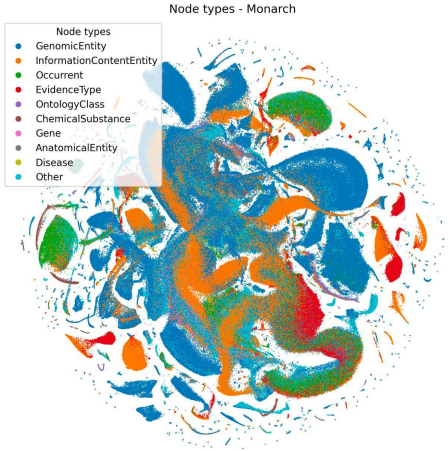


- Several use cases, (e.g., **variant prioritization**,), would benefit from characterizing understudied human proteins
- The Monarch KG contains **rich data** from other a variety of species that can help characterize these proteins
- Graph ML can **effectively capture** the available data
- Deep learning (graph convolutional networks) to characterize understudied proteins by leveraging other species in Monarch KG
- **One application:** feed data from more species into Exomiser (which currently uses human and mouse data from Monarch KG) to allow it to use these species data to prioritize variants

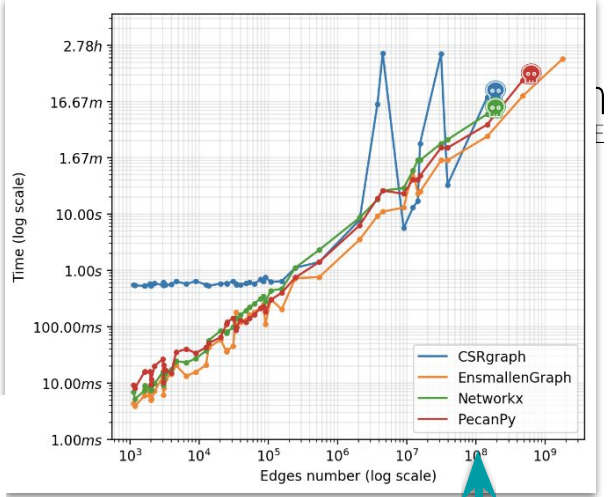
More data means bigger KGs, and slower ML

Enter EnsmallenGraph/Embiggen:

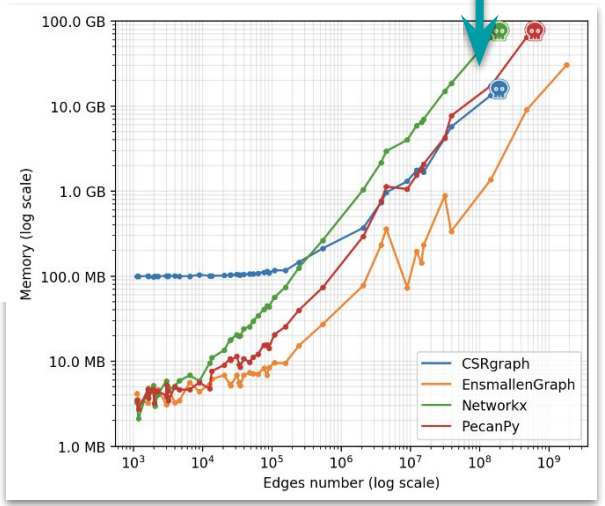
- Monarch KG has >10e7 nodes/edges (and getting bigger) - existing software struggles with it
- EnsmallenGraph/Embiggen: Performant graph machine learning
- Faster loading, lower memory footprint (10x smaller)
- Scales to **billions** of nodes
- Algorithms: graph convolutional networks, node2vec, TransE (and friends), many more
- Novel graph ML algorithms in development
- Other possible graph ML experiments on Monarch KG:
  - learn “vector” for drug -> disease in latent space, find candidate drugs for untreatable diseases
  - find causal genes for rare/orphan diseases



Embedding of Monarch KG in ~3h (node2vec)



Billions of edges

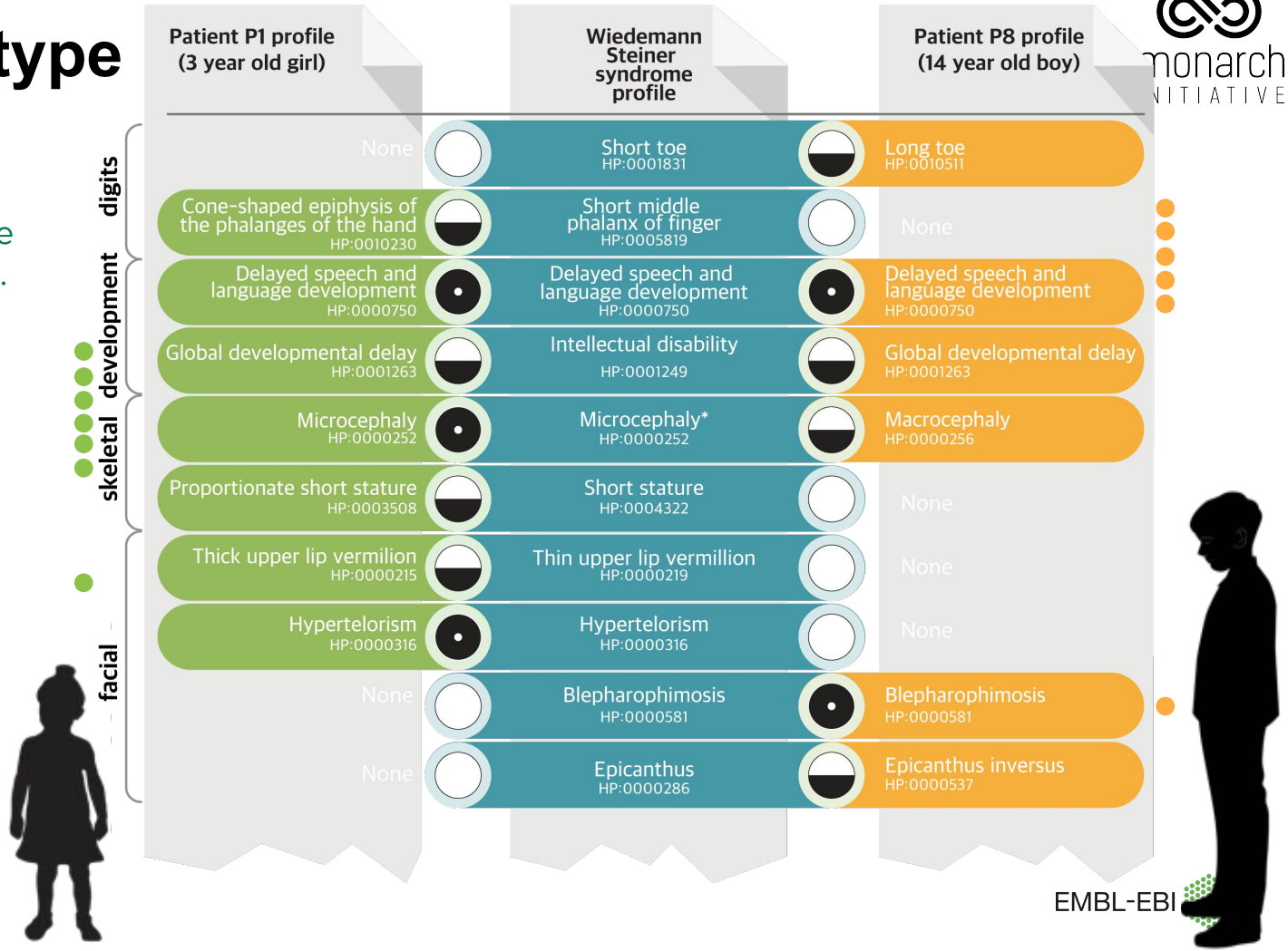


Fuzzy Phenotype Matching

Not same variant, but same disease and gene, KMT2A.

Legend

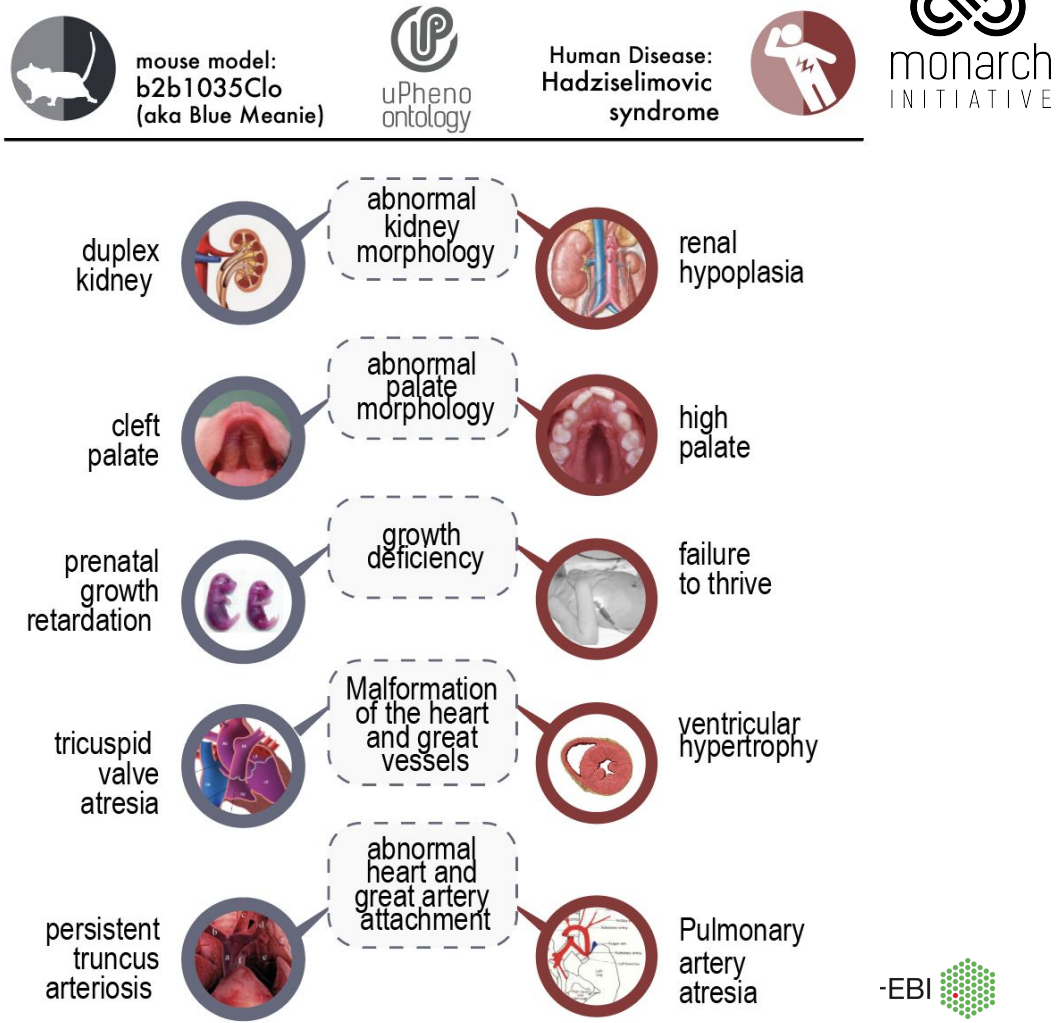
- Perfect Match
- Fuzzy Match
- No Match



DOI: 10.1126/scitranslmed.3009262

EMBL-EBI

Fuzzy matching across species improves diagnostics

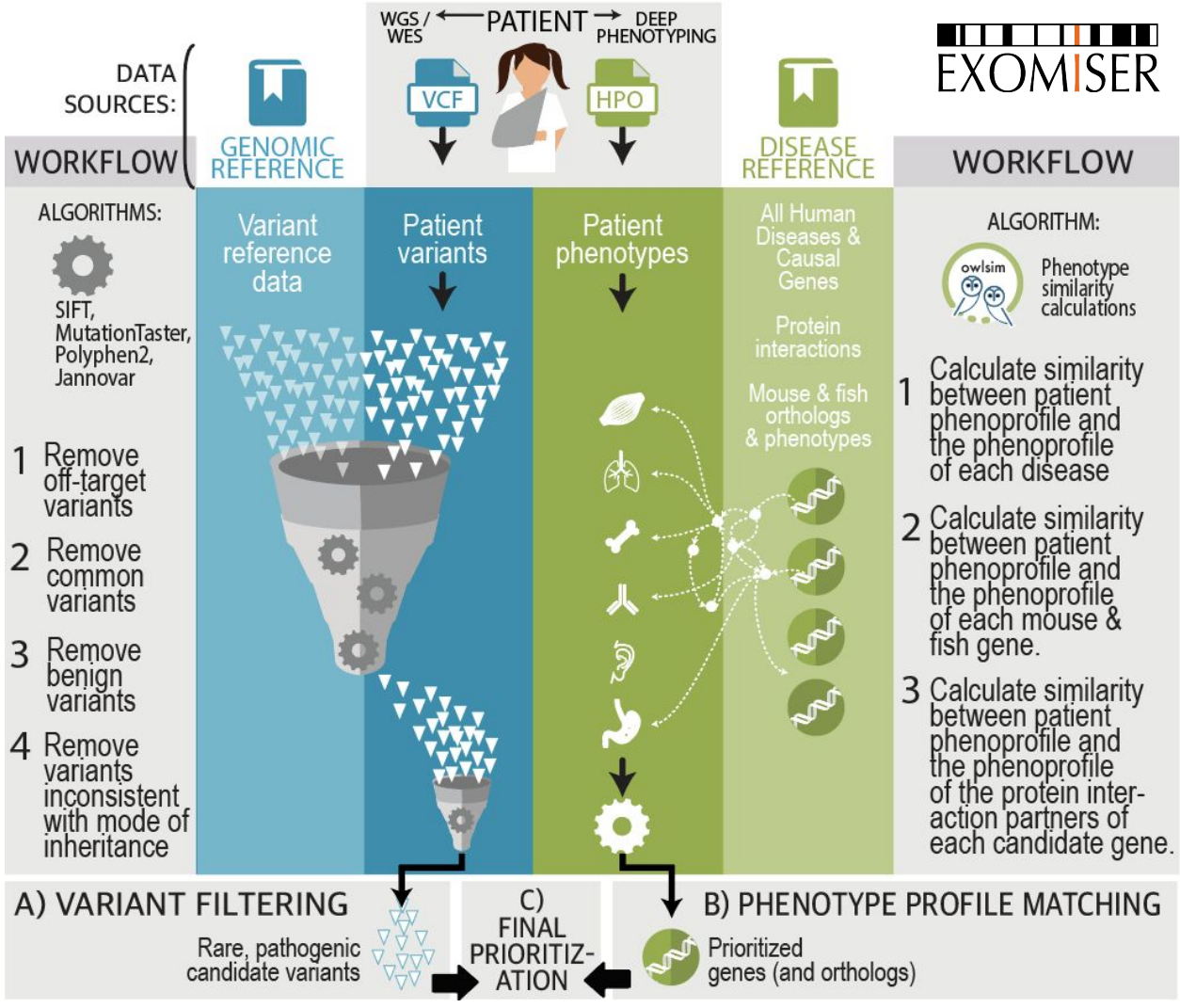


upheno ontology

-EBI

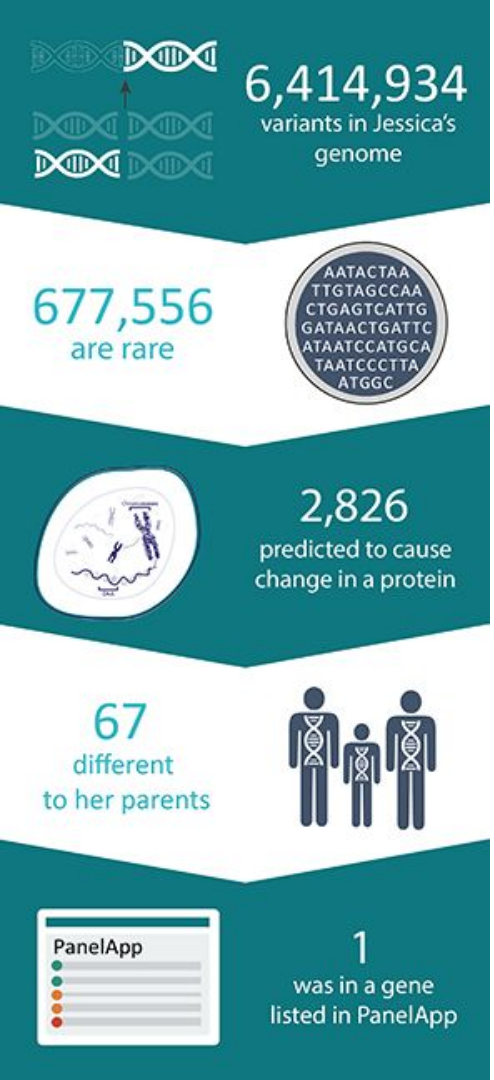
Use Cases

Combining genomic and phenomic data improves variant prioritization for diagnosis



monarch INITIATIVE

doi: 10.1038/gim.2015.137



Jessica

monarch INITIATIVE

- Jessica (age 4) has a rare condition which causes epilepsy, affects her movement and developmental delay. Standard genetics tests negative.
- De novo deletion in *SLC2A1* identified as the cause of her Glut 1 deficiency syndrome
- Exomiser ranked this variant first
- Now being successfully treated with a ketogenic, low-carb diet
- Low risk for future pregnancies

Genomics england

Exomiser ranking 94% in top 3 candidates using human + model organism data

Aiding diagnosis using phenotype matching to model organisms - IMPC data key when no human data exists

CDKN2A Exomiser Score: 0.975 Phenotype Score: 0.755 Variant Score: 0.992

EXOMISER

Phenotype matches:
Phenotypic similarity 0.755 to mouse mutant involving CDKN2A.
Best Phenotype Matches:
HP:0000519, Congenital cataract - MP:0001304, cataract
HP:0000164, Abnormality of the dentition -

Proximity score 0.510 in interactome to TP63 and phenotypic similarity 0.600 to Hay-Wells syndrome associated with TP63.
Best Phenotype Matches:
HP:0000519, Congenital cataract -
HP:0000164, Abnormality of the dentition - HP:0000687, Widely spaced teeth

Proximity score 0.510 in interactome to TP63 and phenotypic similarity 0.620 to mouse mutant of TP63.
Best Phenotype Matches:
HP:0000519, Congenital cataract - MP:0006000, abnormal corneal epithelium morphology
HP:0000164, Abnormality of the dentition - MP:0000764, abnormal tongue epithelium morphology

Known diseases:
OMIM:155601 Melanoma, cutaneous malignant, 2 (susceptibility)
OMIM:155755 Melanoma and neural system tumor syndrome - autosomal dominant
OMIM:606719 Pancreatic cancer/melanoma syndrome - autosomal dominant
ORPHA:1333 Familial pancreatic carcinoma
ORPHA:618 Familial melanoma

Gene scores under compatible inheritance modes:

AUTOSOMAL_DOMINANT Exomiser Score: 0.975 Phenotype Score: 0.755 Variant Score: 0.992

Variants contributing to score:
MISSENSE chr9:g.21971040C>T [0/1:0/1] rs199888003 (variation viewer)
Variant score: 0.992 CONTRIBUTING VARIANT
Transcripts:
CDKN2A:ENST00000361570:c.484G>A:p.(Ala162Thr)
CDKN2A:ENST00000530628:c.361G>A:p.(Ala121Thr)
CDKN2A:ENST00000579755:c.361G>A:p.(Ala121Thr)
CDKN2A:ENST00000304494:c.318G>A:p.(=)
CDKN2A:ENST00000446177:c.318G>A:p.(=)
CDKN2A:ENST00000479692:c.165G>A:p.(=)

Pathogenicity Data:
Best Score: 1.0

Frequency Data:
Local: 0.0379%

Significant Phenotypes

IMPC

Phenotype: All

Show 10 entries

System	Phenotype	Allele	Zyg	Sex	Life Stage	P Value	Data
	abnormal lens morphology	Cdkn2a$^{tm1a(EUCOMM)Wtsi}$	HOM	♂	postnatal	2.61E-7	
	cataract	Cdkn2a$^{tm1a(EUCOMM)Wtsi}$	HOM	♂	postnatal	2.61E-7	
	abnormal eye morphology	Cdkn2a$^{tm1a(EUCOMM)Wtsi}$	HOM	♂	postnatal	6.67E-5	

Genomics england

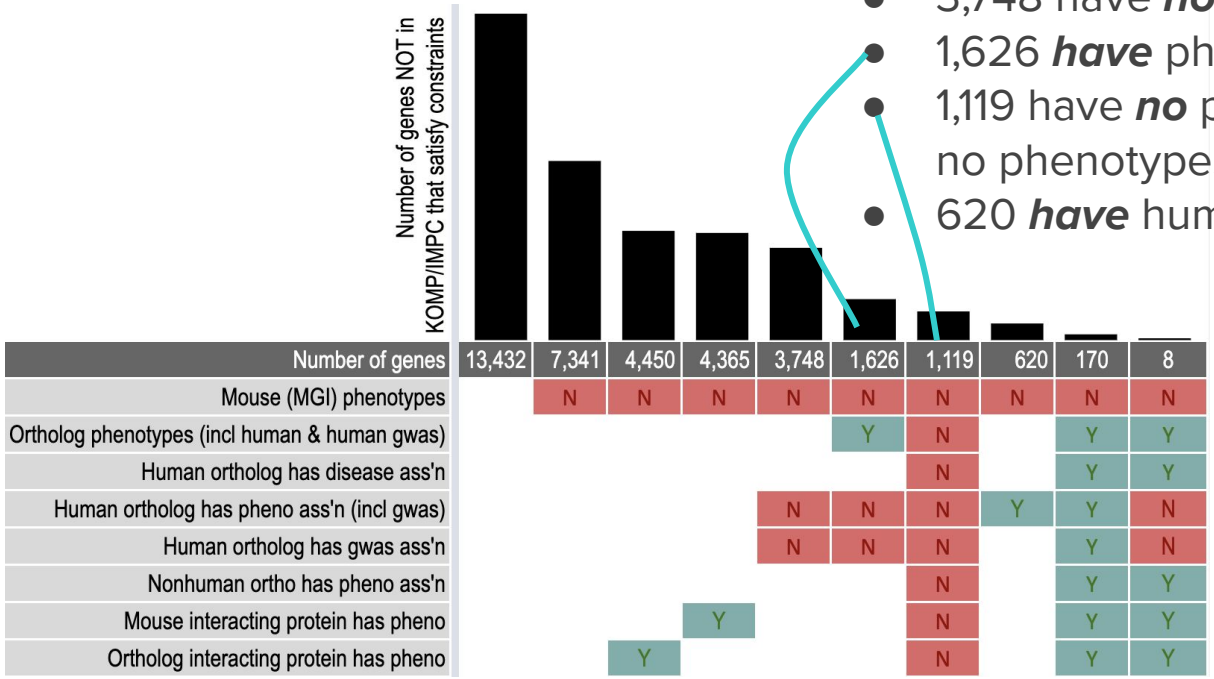
CDKN2A candidate for cataracts GeL patient

https://dx.doi.org/10.1038%2Fng.3901

Using phenotype data to inform gene selection for KOMP

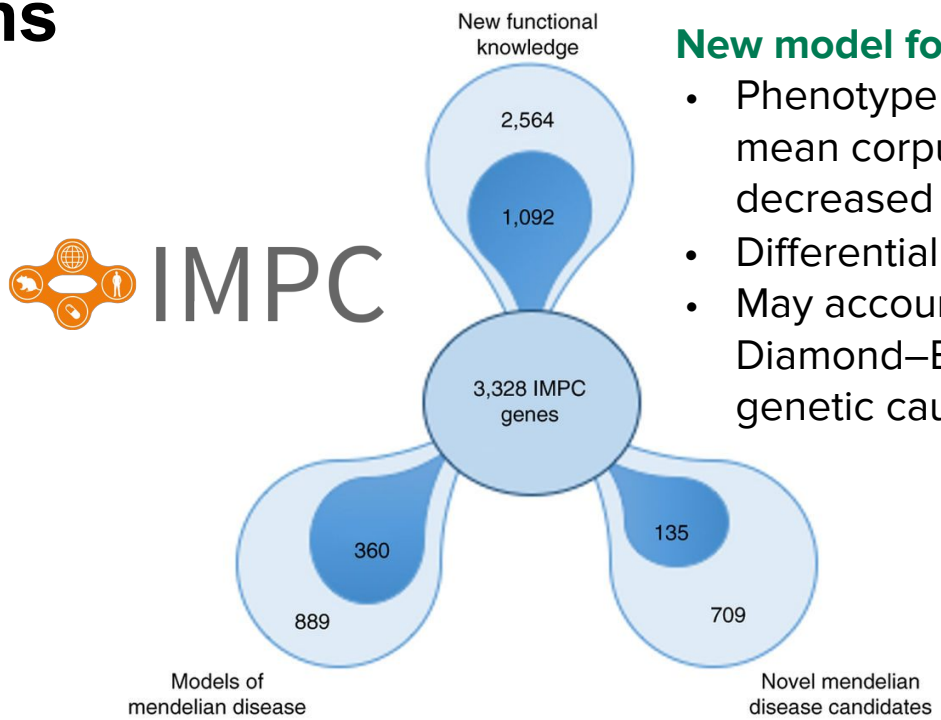
7,341 have ***no*** phenotypes in mouse (MGI). Of these:

- 4,450 ***have*** phenotypes of interacting proteins in other species
- 4,365 ***have*** phenotypes of interacting proteins in mice
- 3,748 have ***no*** phenotypes in human*
- 1,626 ***have*** phenotypes in other species
- 1,119 have ***no*** phenotypes in ***an***y species AND no phenotypes in any interaction partners
- 620 ***have*** human phenotype or disease ass'n



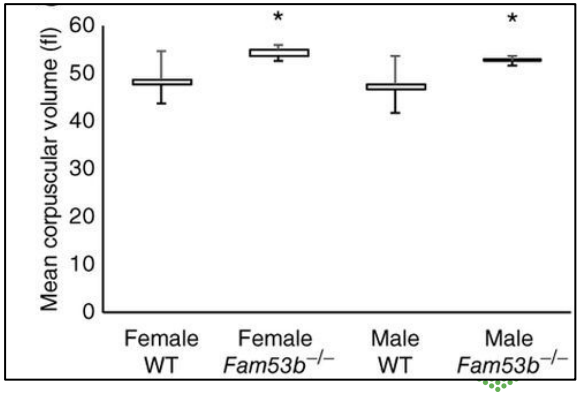
bit.ly/komp-priority-genes

Identifying candidate genes using model organisms



New model for Diamond–Blackfan anemia

- Phenotype profile similarity: increased mean corpuscular hemoglobin and decreased erythrocyte cell numbers
- Differential expression
- May account for 46% of people with Diamond–Blackfan anemia with unknown genetic causes



135 new candidate genes for Mendelian disorders

https://dx.doi.org/10.1038%2Fng.3901

Standardisation

Global Alliance for Genomics & Health

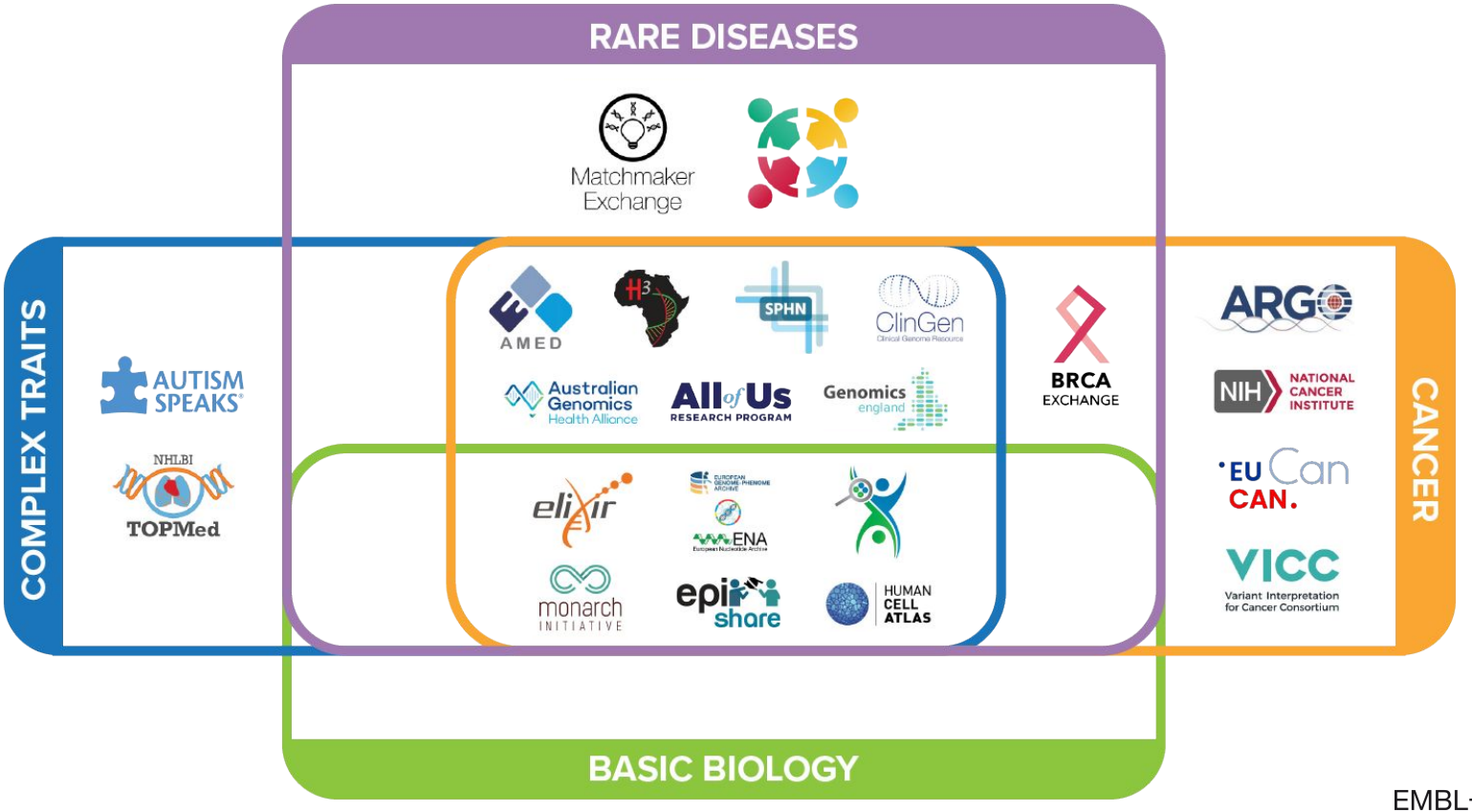
The Mission of the GA4GH is to accelerate progress in genomic science and human health by developing standards and framing policy for responsible genomic and health-related data sharing.

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

EMBL-EBI

GA4GH Driver projects

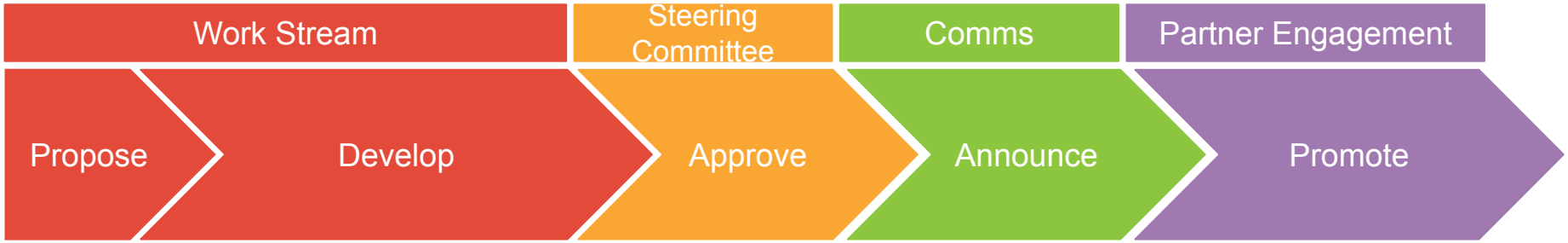
Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.



EMBL-EBI

Product pipeline

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



- Driver Projects
- Regulatory & Ethics
- Data Security

- Product Review Committee
- Regulatory & Ethics
- Data Security

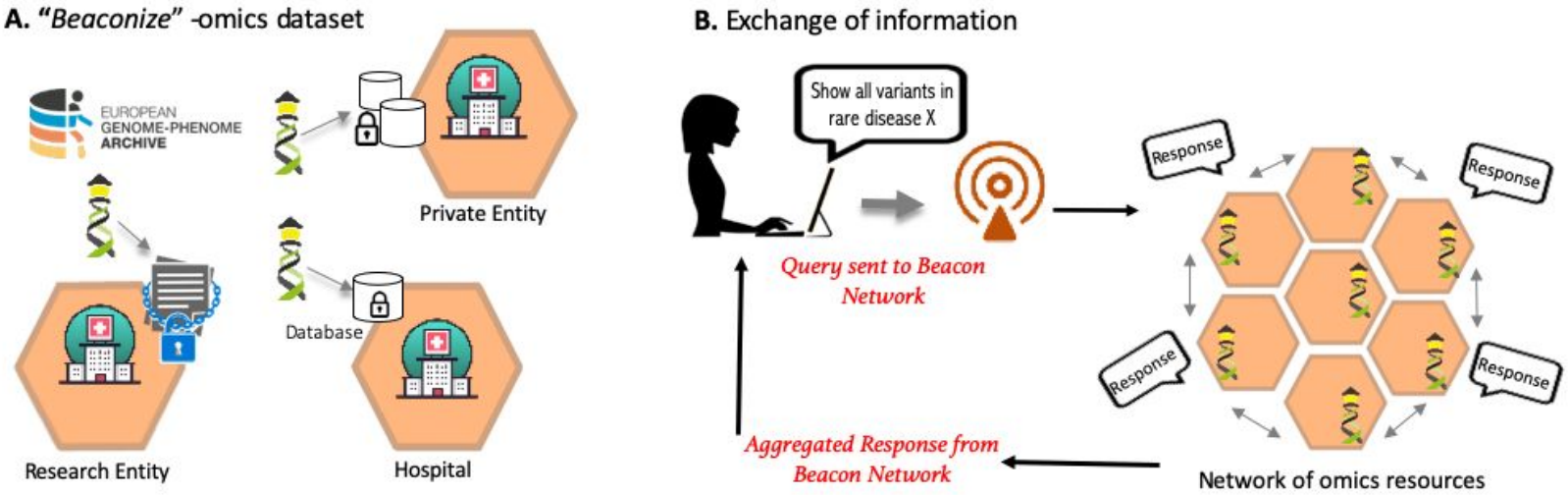
- Work Stream
- Partner Engagement

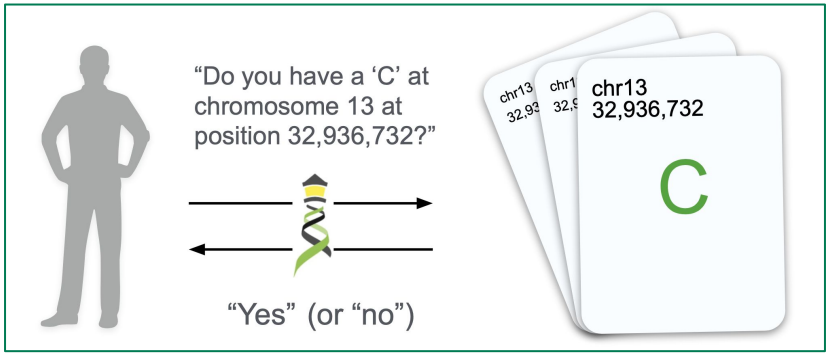
- Work Stream
- Comms

EMBL-EBI

GA4GH Beacon Standard

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





EMBL-EBI

GA4GH Beacon Standard

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

Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?

Beacon *v2* API

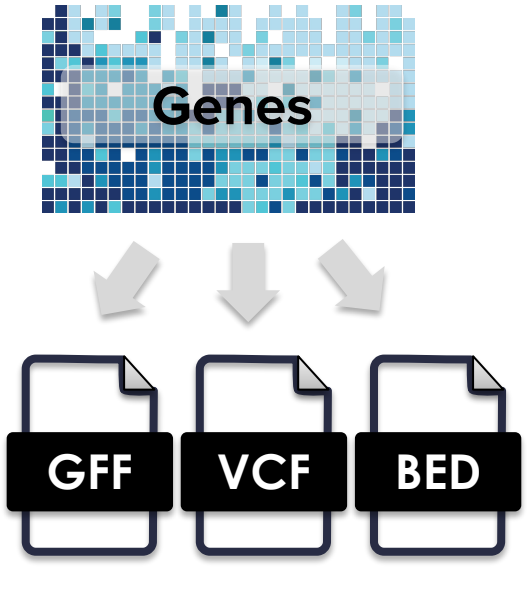
The Beacon API v2 proposal opens the way for the design of a simple but powerful "**genomics API**".

EMBL-EBI

Standard exchange formats exist for sequence/genomes but not for phenotypes

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

monarch INITIATIVE



We need a standard way to share case-level phenotypic information that is not free text, a candidate diagnosis proxy, nor full EHR data exported via PDF

EMBL-EBI

Phenopackets can help us do better

- Craniosynostosis
- Brachydactyly
- Proptosis
- Broad thumb...

How are these linked to a patient?
To genomic info?
To samples?
To parents and siblings?

Were they NOT observed?

When were they first observed?

How severe are these?

Are some more severe than others?

obal Alliance
Genomics & Health
. Innovate. Accelerate.

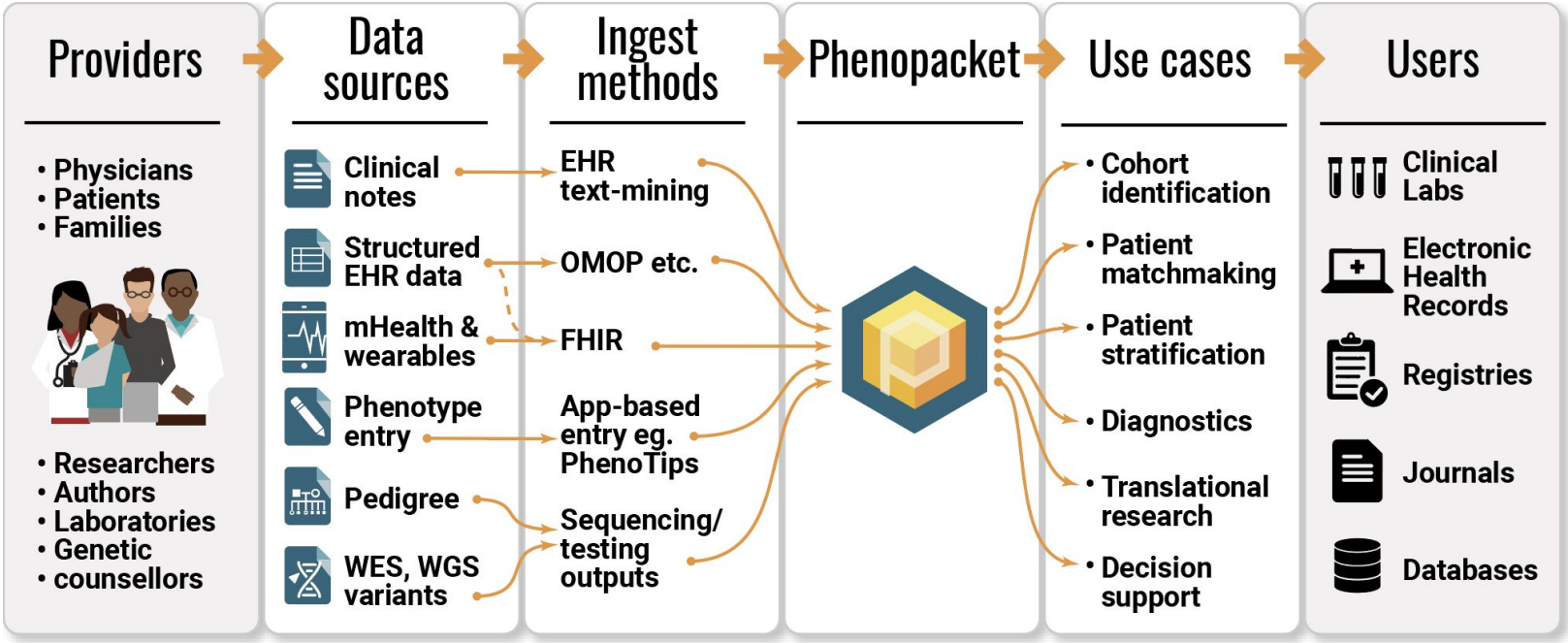
monarch INITIATIVE

EMBL-EBI

The Phenopacket Ecosystem: Users and use cases

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

monarch INITIATIVE



phenopackets.org

https://github.com/phenopackets

EMBL-EBI

Phenopacket structure - phenotypes

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

```
subject:
  id: "PROBAND#1"
  ageAtCollection:
    age: "P3M"
  sex: "MALE“
htsFiles:
- uri: "file:///data/file.vcf.gz"
  htsFormat: "VCF"
  genomeAssembly: "GRCh38"
```

who?

WGS data

```
phenotypes:
- type:
    id: "HP:0000520"
    label: "Proptosis"
  severity:
    id: "HP:0012828"
    label: "Severe"
  classOfOnset:
    id: "HP: 0003577"
    label: "Congenital onset"
```

what?

how?

when?

EMBL-EBI

A concrete example

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

monarch INITIATIVE

A 6-month-old girl conceived by in vitro fertilization (IVF) (own oocytes and anonymous donor sperm) was admitted to the hospital because of leukocoria and strabismus. Past medical history and physical examination were unremarkable except for clinodactyly of the right fifth finger. Indirect ophthalmoscopic examination and examination under anesthesia was performed by ophthalmologists. Orbital ultrasound and magnetic resonance imaging (MRI) scans showed a 14 × 13 × 11 mm left eye tumor located in the lower-external retinal side. Retinal detachment was also detected. Diagnosis of retinoblastoma was made and, based on International Classification for Intraocular Retinoblastoma, a grade E was established.

Gargallo, P., Oltra, S., Balaguer, J. et al. Retinoblastoma and mosaic 13q deletion: a case report. Int J Retin Vitr 7, 50 (2021). https://doi.org/10.1186/s40942-021-00321-9

Note: this example focuses only on the *textual part* of an article although the knowledge (and provenance) can be externalised from / associated with *images* as well

EMBL-EBI

Phenopacket structure - diagnosis

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

monarch INITIATIVE

```
diseases:
- term:
    id: "NCIT:C7541"
    label: "Retinoblastoma"
  onset:
    age:
      iso8601duration: "P4M"
  diseaseStage:
  - id: "LOINC:LA24739-7"
    label: "Group E"
  primarySite:
    id: "UBERON:0004548"
    label: "left eye"
```

Various clinical terminologies used to capture accurately domain-specific concepts:

- **NCIT** - National Cancer Institute Thesaurus - de facto standard to capture cancer-related knowledge
- **LOINC** - de facto standard to capture lab tests and measurements
- **UBERON** - cross-species ontology for modeling anatomy

EMBL-EBI

Phenopacket structure - phenotype

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

monarch INITIATIVE

```
phenotypicFeatures:
- type:
    id: "HP:0030084"
    label: "Clinodactyly"
  modifiers:
  - id: "HP:0012832"
    label: "Bilateral"
  onset:
    age:
      iso8601duration: "P3M"
```

```
phenotypicFeatures:
- type:
    id: "HP:0000486"
    label: "Strabismus"
  modifiers:
  - id: "HP:0012833"
    label: "Unilateral"
  onset:
    age:
      iso8601duration: "P5M15D"
```

```
phenotypicFeatures:
- type:
    id: "HP:0000541"
    label: "Exsudative retinal detachment"
  modifiers:
  - id: "HP:0012231"
    label: "Unilateral"
  onset:
    age:
      iso8601duration: "P6M"
```

- Accurate phenotype representation - including locality, onset, degree of severity, etc:
- **HPO** - Human Phenotype Ontology - de facto standard to capture human phenotypes (see UK 100k Genomes Project)

EMBL-EBI

Phenopacket structure - treatment

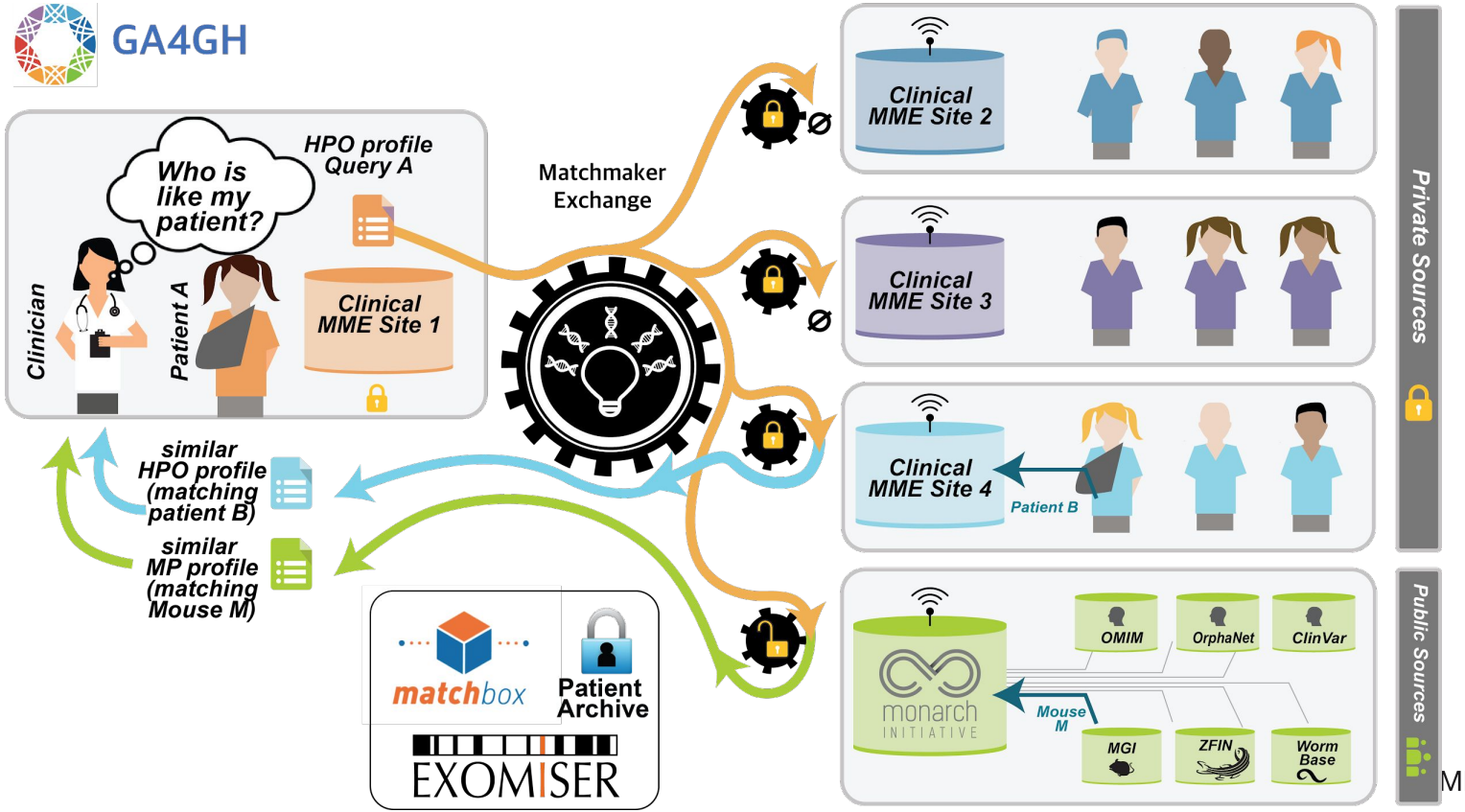
Comprehensive acquisition of medical actions using well established terminologies and drug repositories:

- **DrugCentral** - online drug compendium
- **NCIT** - National Cancer Institute Thesaurus
- **UO** - ontology capturing standardised units of measurement
- **HPO** - here, capturing adverse events as phenotypes

MatchMaker Exchange: match rare disease patients to model organisms

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

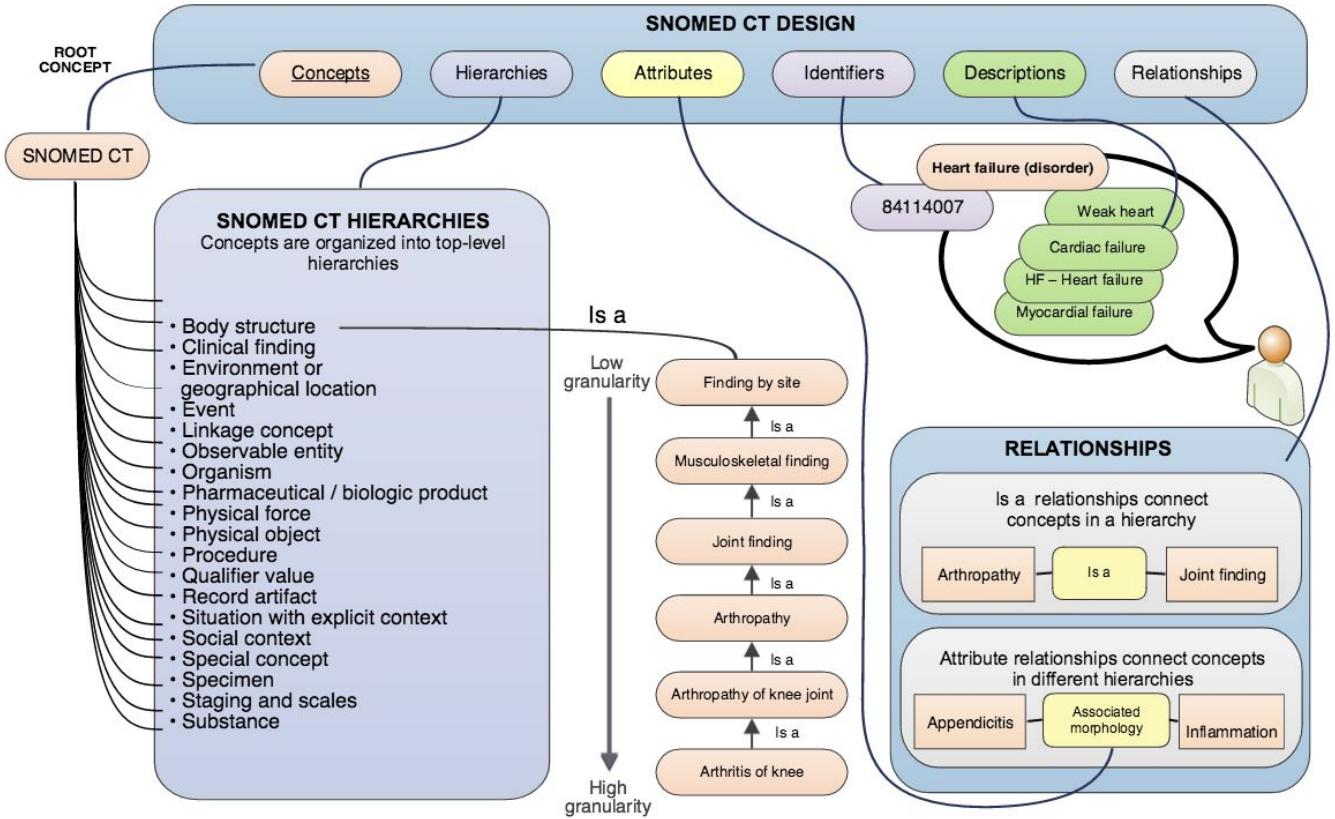
monarch INITIATIVE



MBL-EBI

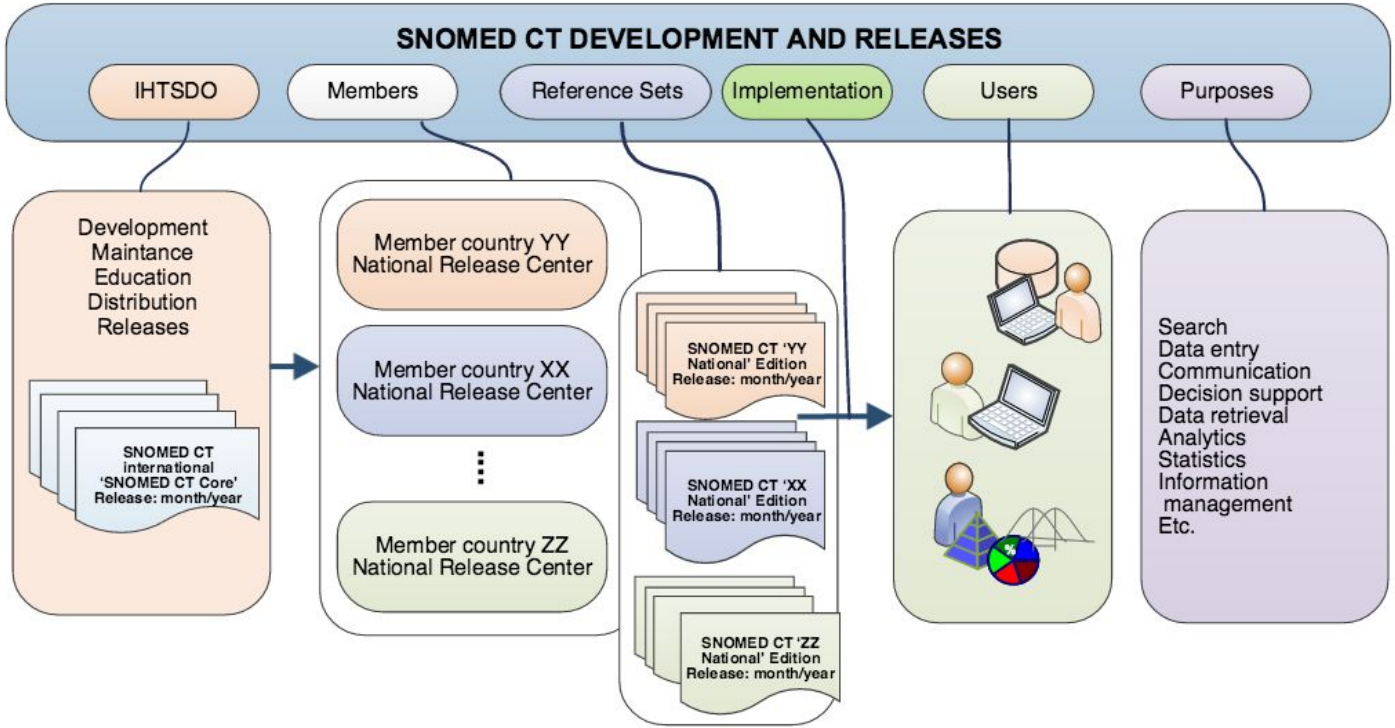
Clinical standards

SNOMED CT
The global language of healthcare



EMBL-EBI

SNOMED CT
The global language of healthcare

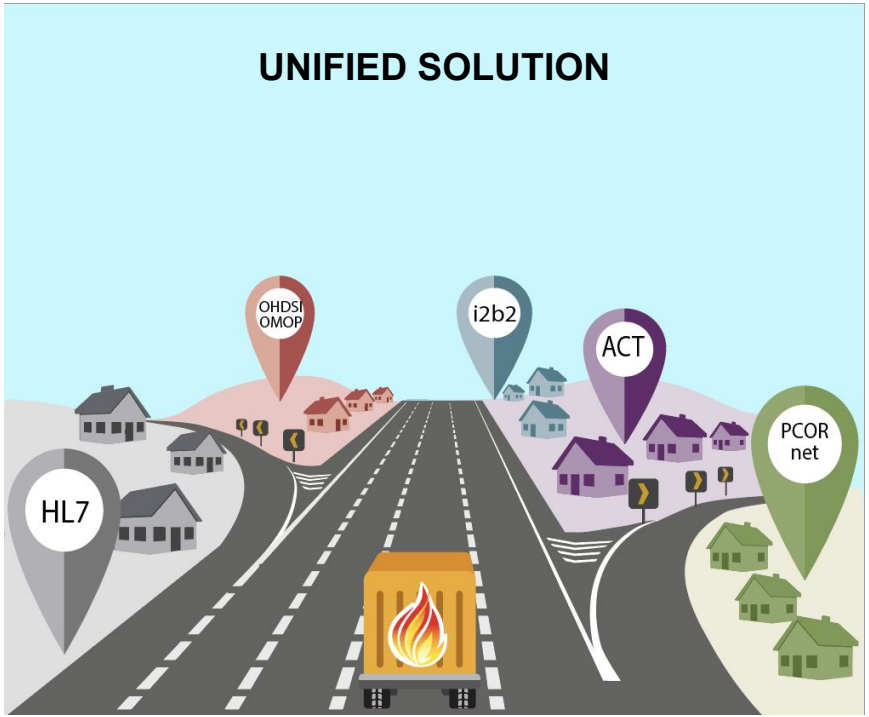


EMBL-EBI

FHIR: Fast Healthcare Interoperability Resources
A HL7 Standard for transmitting healthcare data

EMBL-EBI

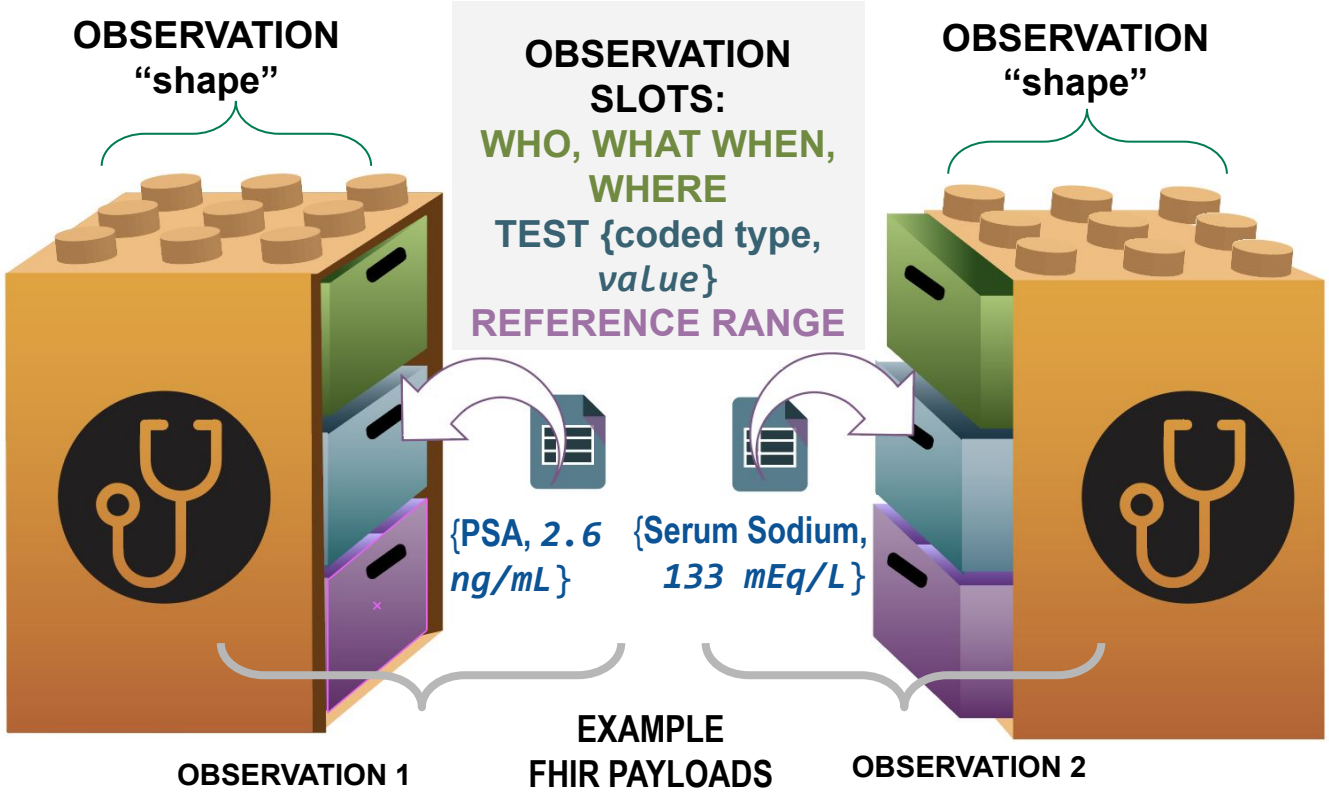
Transforming the clinical data landscape with FHIR



EMBL-EBI

How does FHIR help?

EXAMPLE FHIR "shapes" (Resources)



EMBL-EBI

Phenopackets-FHIR Implementation Guide

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

monarch INITIATIVE

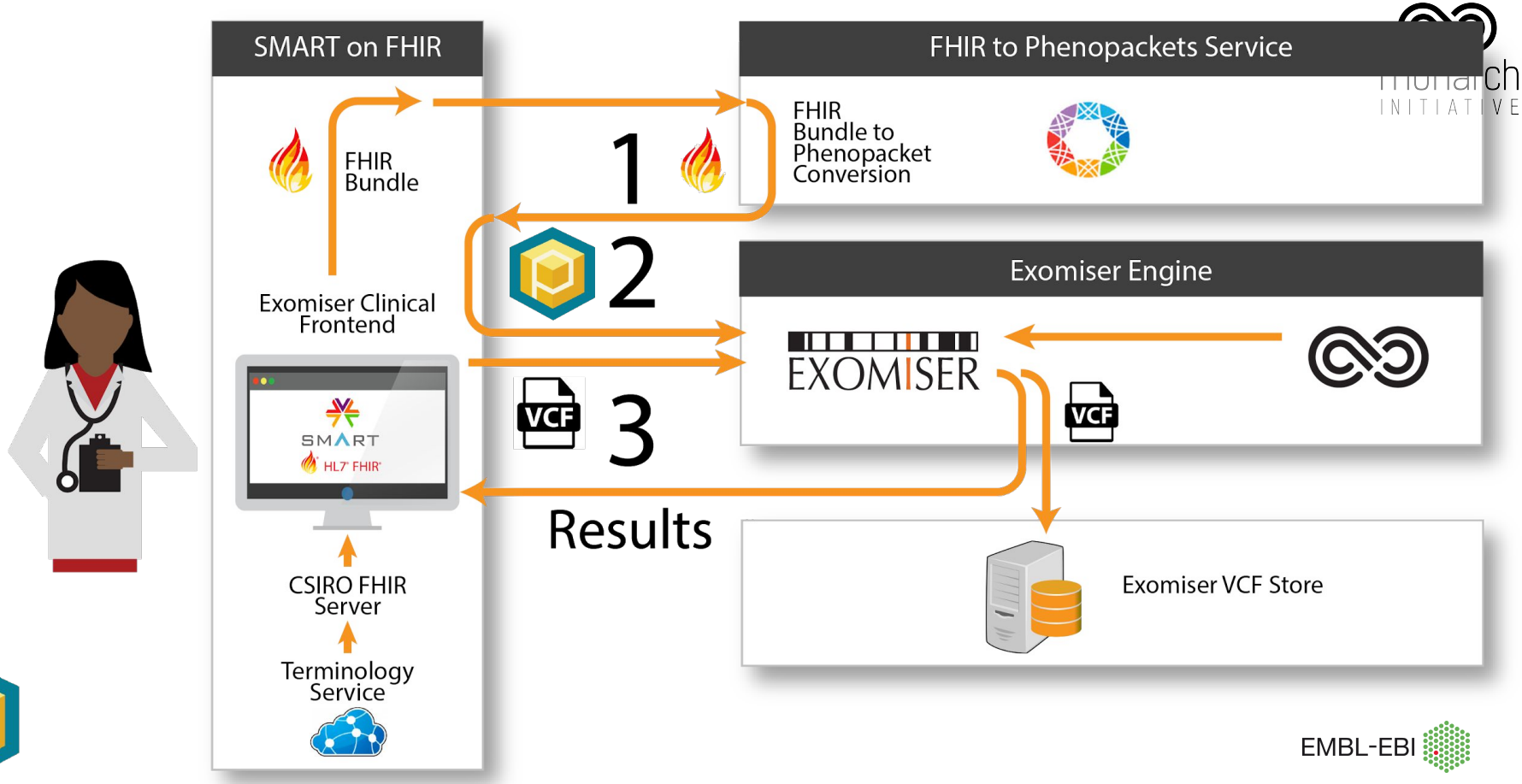
Overall Goal:

Increase availability of high-quality standardized phenotypic information for genomic research and genomic medicine

- Represent the contents of the clinical phenotype profile of a patient, as defined in the Phenopackets schema (including association with pedigree and variant/genome information), in the FHIR structure.
- Allow non-lossy exchange of clinical information between Phenopackets and FHIR structures.
- Include references to existing and/or novel FHIR resources and profiles.

EMBL-EBI

Example System Architecture



Phenopackets-FHIR IG: Objectives

- Leverage what our team has already developed for FHIR IG
- Map Phenopackets schema elements to FHIR Resources
- EHR interoperability
- Evaluation in different infrastructure systems:
  - biobanks, registries, journals

EMBL-EBI

Phenopackets-FHIR IG: Approach

1. Community organization: governance, communications
2. Workflow for collaborative review of current Phenopackets-FHIR mappings
3. Collaborative development of Phenopackets FHIR extensions
4. Pilot Testing
    a. Engaging candidate pilot test sites: GRIN participants
    b. Designing and implement testing plan: mappings, testing message round trip
    c. Pilot use cases: rare diseases, oncology.

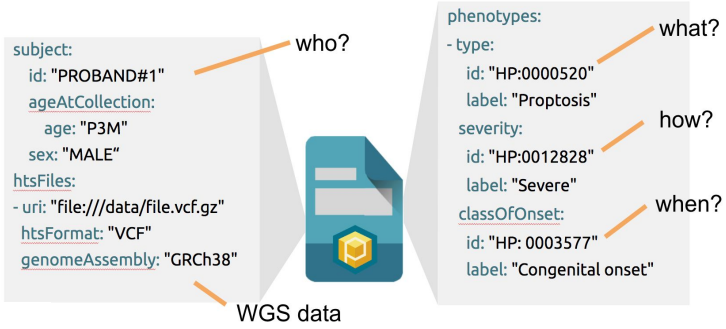
Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

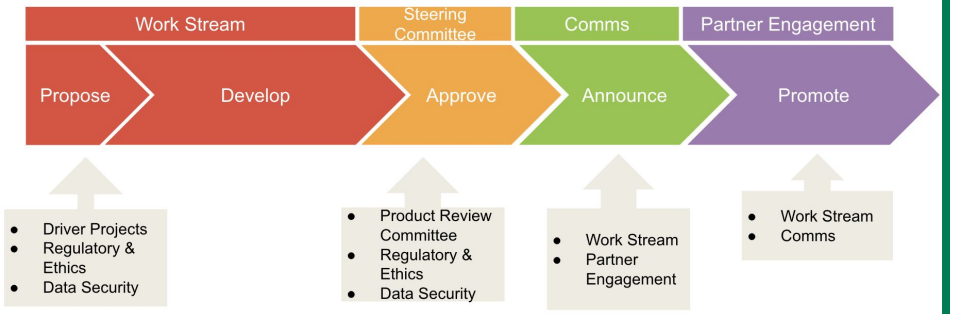
monarch INITIATIVE

EMBL-EBI

Turning Phenopackets into FHIR is a big endeavour

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

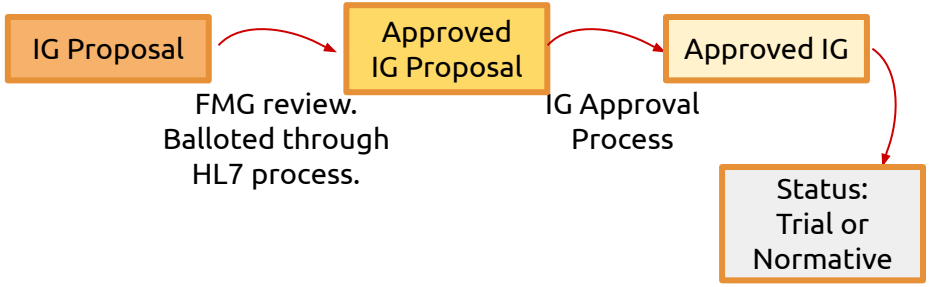




monarch INITIATIVE

Structure

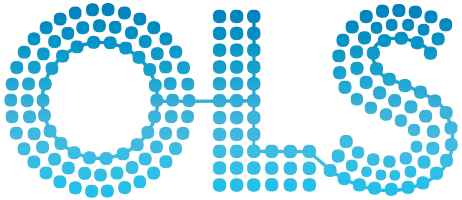
Name	Flags	Card.	Type
Patient	N		DomainResource
identifier	Σ	0..*	Identifier
active	?! Σ	0..1	boolean
name	Σ	0..*	HumanName
telecom	Σ	0..*	ContactPoint
gender	Σ	0..1	code
birthDate	Σ	0..1	date
deceased[x]	?! Σ	0..1	
deceasedBoolean			boolean
deceasedDateTime			dateTime
address	Σ	0..*	Address
maritalStatus		0..1	CodeableConcept
multipleBirth[x]		0..1	
multipleBirthBoolean			boolean
multipleBirthInteger			integer

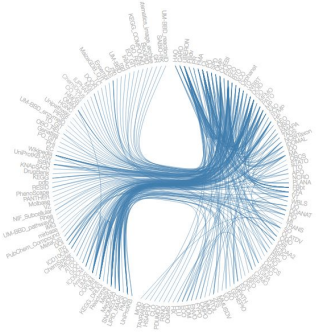


EMBL-EBI

Tooling

Semantics as Service tools



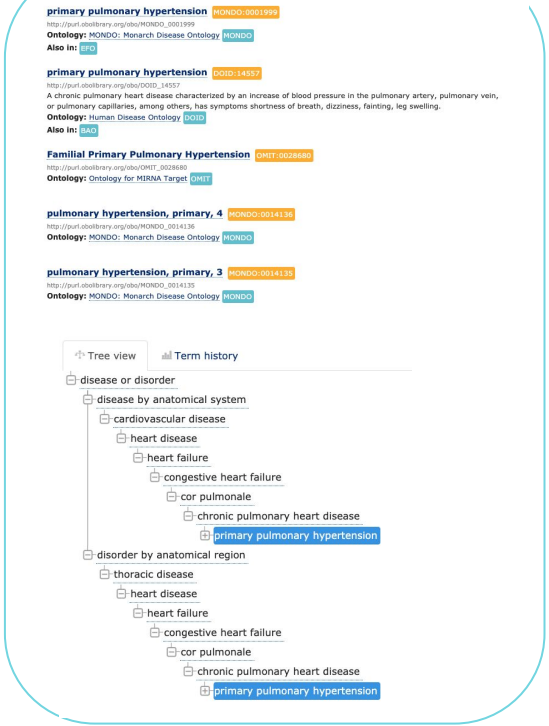


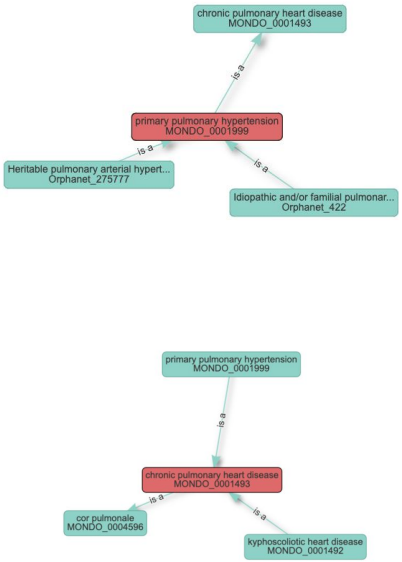
Ontology Lookup Service	Cross Ontology Mapping	OntoString
Find, visualise	Maps between ontologies Provenance Cross referencing to concept equivalence	Data Annotation Service Ontology term request broker Embeds other services Configurable for annotation scenarios

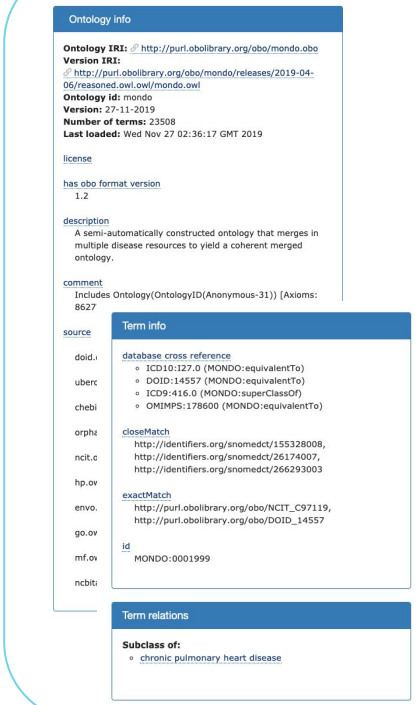
FAIR semantic Interoperability

EMBL-EBI

Ontology Lookup Service





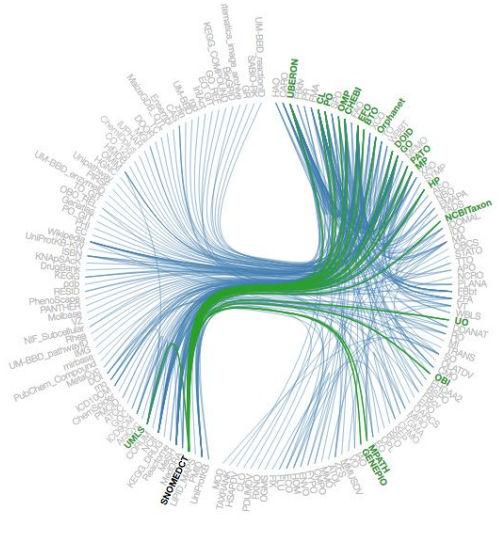


Ontology Xref Service

OxO | Documentation | About

Welcome to the EMBL-EBI Ontology Xref Service (OxO).

OxO is a service for finding mappings (or cross-references) between terms from ontologies, vocabularies and coding standards. OxO imports mappings from a variety of sources including the Ontology Lookup Service and a subset of mappings provided by the UMLS. We're still developing the service so please get in touch if you have any feedback.



Mapping results

Complete!

Download

Select a term to see more information. The evidence colums tells you how many times we have seen this mapping and the distance is a how many hops across other mappings you need to go to find this mapping. Distance 1 is a direct mapping, the greater the distance the less likely it is that a mapping holds true. Max distance is set to 2.

Showing 1 to 27 of 27 entries

Search:

Input	Mapped id	Id source	Evidence	Distance
MeSH:D009202 (Cardiomyopathies)	ICD10CM:I42.9 (Cardiomyopathy (primary) (secondary) NOS)	ICD10CM	3	2
MeSH:D009202 (Cardiomyopathies)	SNOMEDCT:89600008	SNOMEDCT	3	2
MeSH:D009202 (Cardiomyopathies)	SNOMEDCT:57809008	SNOMEDCT	4	2
MeSH:D009202 (Cardiomyopathies)	SNOMEDCT:195035002 (Secondary cardiomyopathy: [dilated] or [NOS])	SNOMEDCT	3	2
MeSH:D009202 (Cardiomyopathies)	SNOMEDCT:89461002	SNOMEDCT	3	2
MeSH:D009202 (Cardiomyopathies)	EFO:0000316 (cardiomyopathy)	EFO	1	1
MeSH:D009202 (Cardiomyopathies)	UMLS:C0878544 (Cardiomyopathies)	UMLS	1	1
MeSH:D009202 (Cardiomyopathies)	SNOMEDCT:195037005	SNOMEDCT	3	2
MeSH:D009202 (Cardiomyopathies)	MedDRA:10007635 (Cardiomyopathies)	MedDRA	1	2
MeSH:D009202 (Cardiomyopathies)	ICD10CM:I51.5 (Myocardial degeneration)	ICD10CM	3	2
MeSH:D009202 (Cardiomyopathies)	NCIt:C53654	NCIt	3	2
MeSH:D009202 (Cardiomyopathies)	UMLS:C0033141 (Cardiomyopathies, Primary)	UMLS	1	1
MeSH:D009202 (Cardiomyopathies)	NCIt:C34630	NCIt	4	2

Query

Use the text box to find possible ontology mappings for free text terms in the ZOOMA repository of curated annotation knowledge. You can add one term (e.g. 'Homo sapiens') per line. If you also have a type for your term (e.g. 'organism'), put this after the term, separated by a tab. If you are new to ZOOMA, take a look at our getting started guide.

Show me some examples...

```
Bright nuclei
Agammaglobulinemia 2  phenotype
Reduction in IR-induced 53BP1 foci in HeLa  cell
Impaired cell migration with increased protrusive activity  phenotype
C57Black/6  strain
nuclei stay close together
Retinal cone dystrophy 3B  disease
segregation problems/chromatin bridges/lagging chromosomes/multiple DNA masses
Segawa syndrome autosomal recessive  phenotype
BRCA1  gene
Deafness, autosomal dominant 17  phenotype
cooked broccoli  compound
```

Datasources

ZOOMA maps text to ontology terms based on curated mappings from selected datasources (more preferred), and by searching ontologies directly (less preferred). Here, you can select which curated datasources to use, optionally ranked in order of preference. You can also select which ontologies to search directly.

monarch INITIATIVE Tools About Documentation Help Examples

monarch INITIATIVE

All | Explore Monarch for phenotypes, diseases, genes and more..

e.g. Marfan Syndrome | Multicystic kidney dysplasia | SHH

DISEASE

- Overview
- Neighbors
- Phenotype 89
- Gene (causal) 1
- Gene (correlated) 2
- Variant 1544
- Model 79
- Pathway 6
- Publication 473
- Genotype 9

Marfan syndrome MONDO:0007947

Marfan's syndrome, MFS, Marfan syndrome, MARFAN SYNDROME; MFS

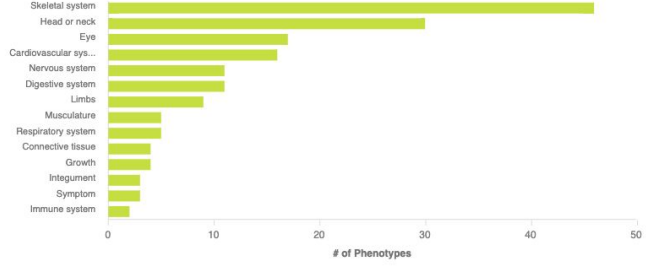
Overview

Marfan syndrome is a disorder of the connective tissue. Connective tissue provides strength and flexibility to structures throughout the body such as bones, ligaments, muscles, walls of blood vessels, and heart valves. Marfan syndrome affects most organs and tissues, especially the skeleton, lungs, eyes, heart, and the large blood vessel that distributes blood from the heart to the rest of the body (the aorta). It is caused by mutations in the FBN1 gene, which provides instructions for making a protein called fibrillin-1. Marfan syndrome is inherited in an autosomal dominant pattern. At least 25% of cases are due to a new (de novo) mutation. Treatment is based on the signs and symptoms in each person.

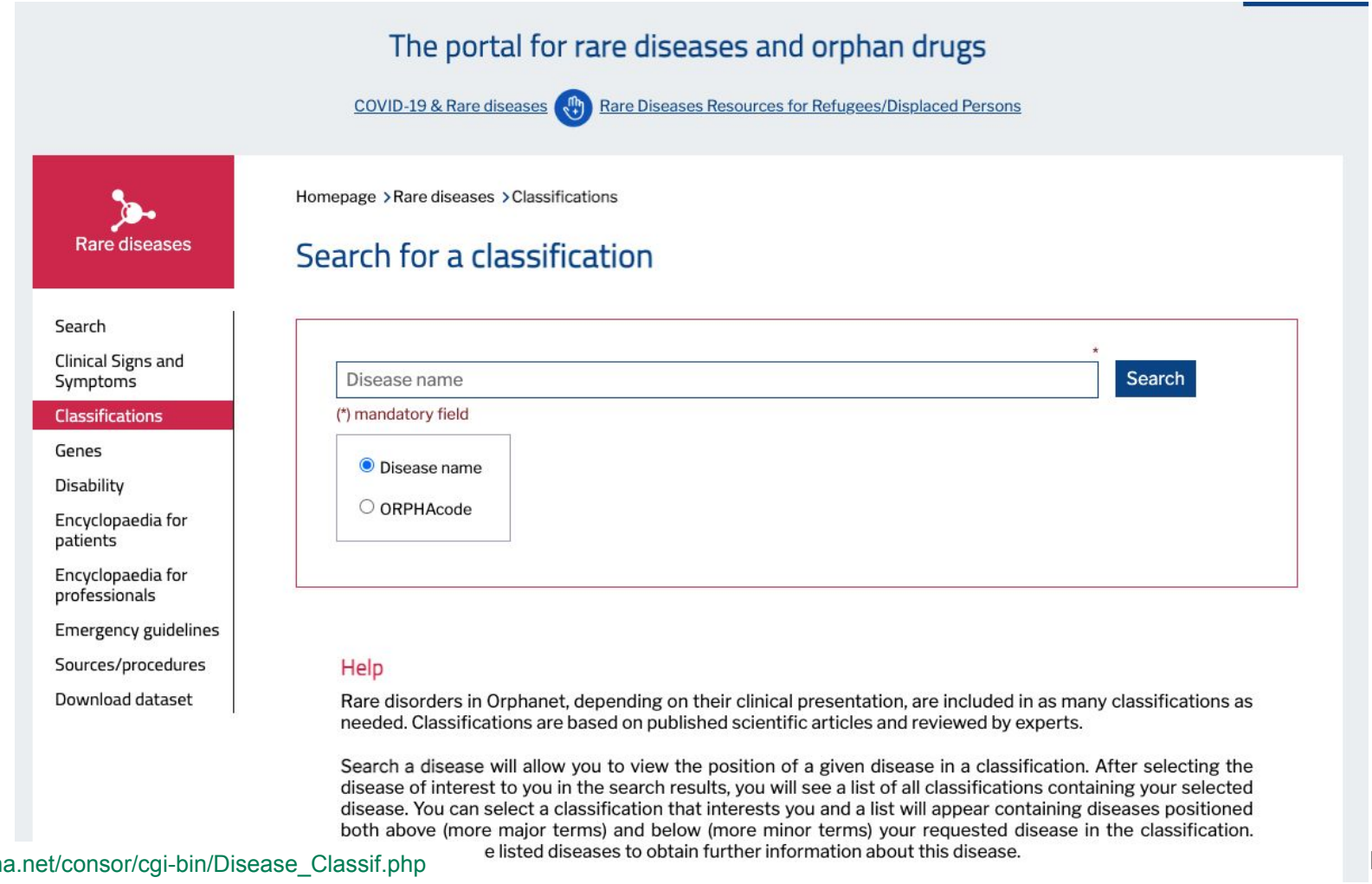
Key Features

Heritability: Autosomal dominant inheritance

Associated Phenotypes




IMPC
International Mouse Phenotyping Consortium
ABOUT THE IMPC
DATA
HUMAN DISEASES
PUBLICATIONS
NEWS
BLOG
The fourth PATHBIO course will take place in Barcelona from 4-12 July 2022. This hybrid event offers both hands-on and virtual participation
GENES
PHENOTYPES
HELP, NEWS, BLOG
Search All 8093 Knockout Data...
What you need to know about IMPC data
What you need to know about IMPC data
We generate our own data, it is not aggregated from publications
8093 total knockout genes phenotyped
Number of phenotyped lines: 8741
Statistically Significant Calls: 93235
Data Release Version: 16.0

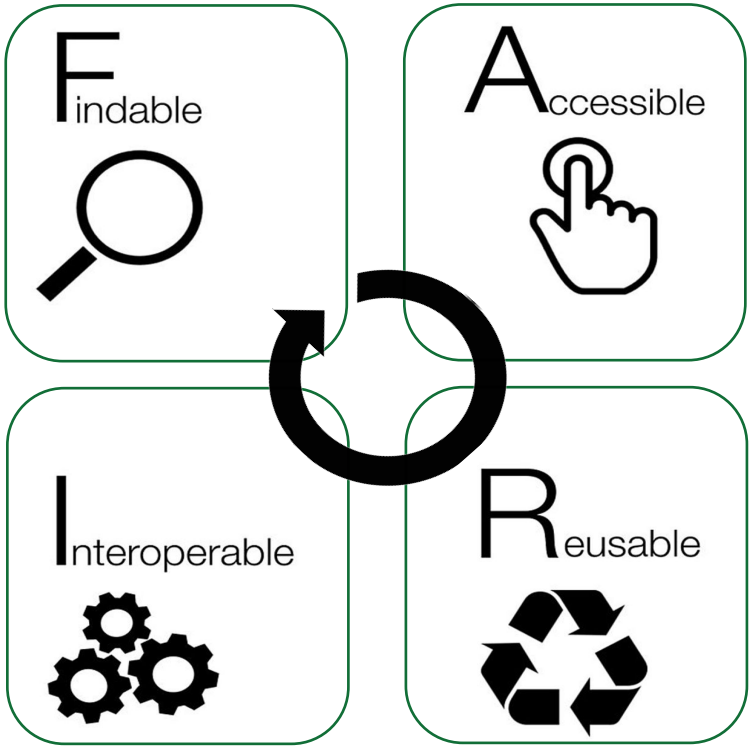


www.orpha.net/consor/cgi-bin/Disease_Classif.php

EBI

FAIR Data

FAIR Principles



www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of 'long-term care' of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes 'good data management' is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility.

This article describes four foundational principles—Findability, Accessibility, Interoperability, and Reusability—that serve to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publishing. Importantly, it is our intent that the principles apply not only to 'data' in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects[2]—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other's data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly

Correspondence and requests for materials should be addressed to B.M. (email: barend.mons@dtls.nl).
#A full list of authors and their affiliations appears at the end of the paper.

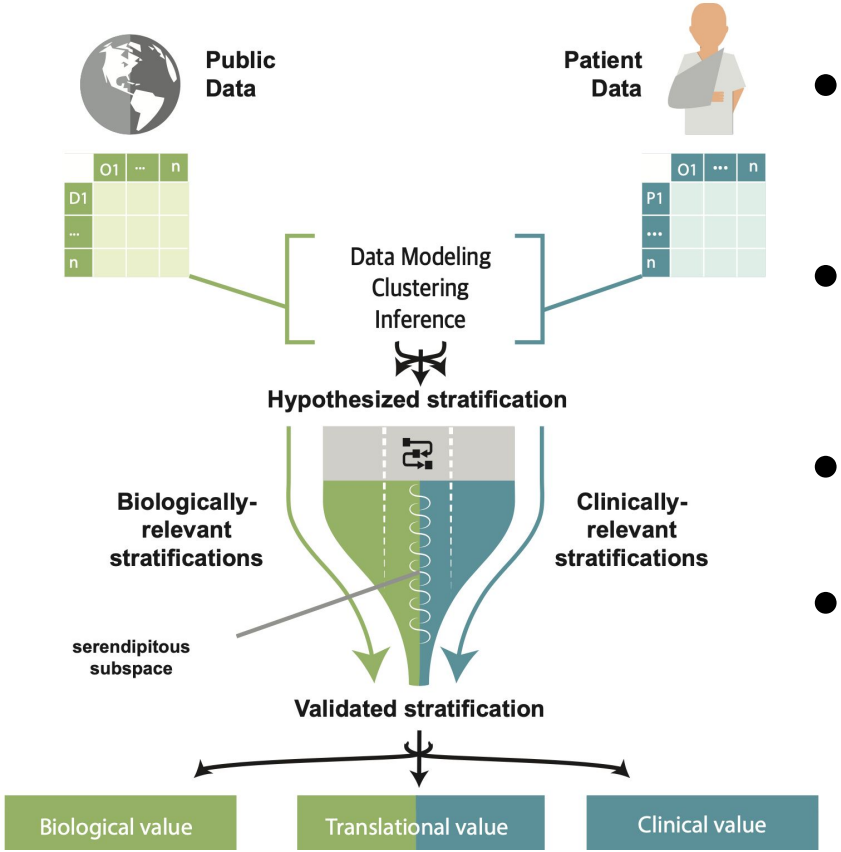
EMBL-EBI

What does FAIR data mean practically?

- Use of appropriate identifiers, versioning and deprecation
- Use and reuse of community-developed and maintained ontologies
- Use of appropriate retrieval mechanisms (preferably REST)

EMBL-EBI

Takeaways



- Semantics can help cross the “chasm of semantic despair” and support more meaningful patient classification
- Realizing standardized and computable phenotypic data akin to genomic data has revolutionized diagnostics and discovery
- Dynamic interplay between public data and clinical/patient-level data
- Combining clinical and basic research data supports new hypotheses, mechanism discovery, and better treatment management

Acknowledgements

Melissa Haendel (PI)
Peter Robinson (PI)
Chris Mungall (PI)
Damian Smedley (Site PI)
David Osumi-Sutherland (Site PI)
Julie McMurry
Jules Jacobsen
Kent Shefchek
Monica Munoz-Torres
Nomi Harris
Sebastian Koeller
Deepak Unni
Harshad Hedge
Justin Reese
Kevin Schaper
Lauren Chan
Matt Brush
Nicole Vasilevsky
Sabrina Toro
Seth Carbon
Shawn O'Neil
Sierra Moxon
Tudor Oprea

The Monarch Initiative is funded by NIH OD R24 OD011883

Additional support to TISLab: NHGRI RM1 HG010860, NCATS U24 TR002306

EMBL-EBI

Acknowledgements

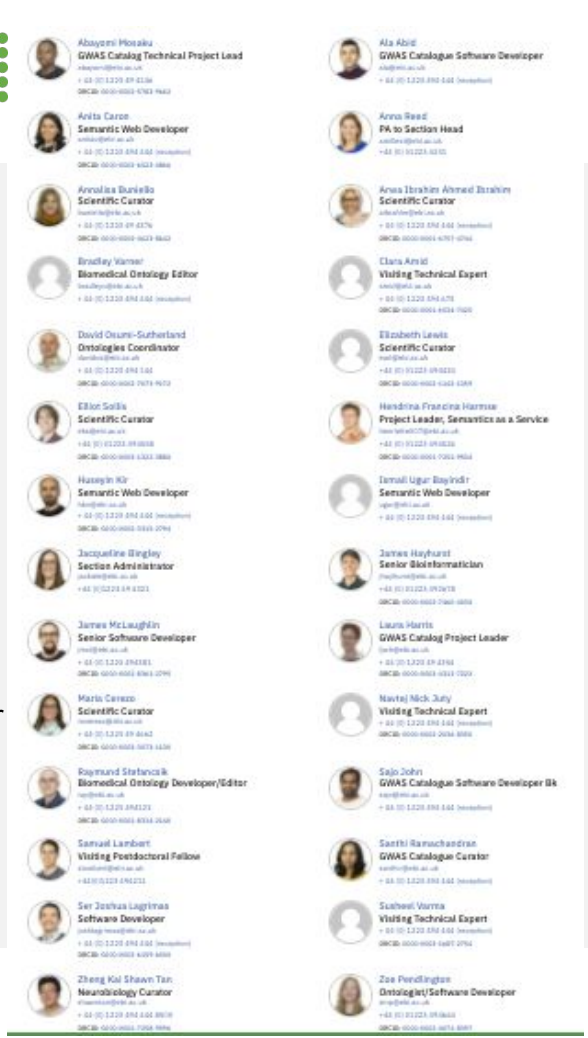
EMBL-EBI

People

- Parkinson Team and colleagues EMBL-EBI
- David Osumi-Sutherland, Henriette Harmse and colleagues – Ontologies and Ontology Tools
- Laura Harris, Yomi Mosaku and colleagues – GWAS Catalog

Funders:

- NIH Common Fund Award to EMBL-EBI, QMUL, MRC Harwell **3UM1HG006370-10S2**
- NIH NCI ITCR Award **5U24CA253539-02**
- NHGRI GWAS Catalog - **5U41HG007823-08**
- NHGRI Monarch Initiative - **2R24OD011883-10A1**
- NHGRI Phenomics First - **7RM1HG010860-02**
- Health Data Research UK - PGS Catalog
- EOSCLife has received funding from the European Union's Horizon 2020 programme under grant agreement number 824087
- Intervene has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016775
- OpenTargets funding for GWAS and Ontologies
- CZI for Human Cell Atlas



Acknowledgements

IMPC

National Institutes of Health (USA)

Toronto Centre for Phenogenomics (Canada)

Medical Research Council & MRC Harwell (UK)

The Wellcome Trust Sanger Institute (UK)

Wellcome Trust

Helmholtz Zentrum Munich (Germany)

Institute Clinique de la Souris (France)

UC Davis

European Bioinformatics Institute

The Jackson Laboratory

Children's Hospital Oakland Research Institute

Consiglio Nazionale delle Ricerche (Italy)

European Commission (EU)

Infrafrontier (EU)

Australian Phenomics Network (Australia)

RIKEN BioResource Center (Japan)

Genome Canada

Model Animal Research Center (Nanjing)

Baylor College of Medicine

Charles River Laboratories

Korea Mouse Phenotyping Center

Universitat Autònoma de Barcelona


Robert Wilson


Zina Perova


Hamed Haseli Mashhadi


Mauricio Martínez Jiménez


Violeta Muñoz Fuentes


Tushar Mandloi


Federico López Gómez


Osman Güneş

National Institutes of Health [NIH U54 - HG006370]
European Molecular Biology Laboratory Core Funding

EMBL-EBI