Adventures in Translational Bioinformatics: Metadata, Annotation, and Integration:

Harry Hochheiser

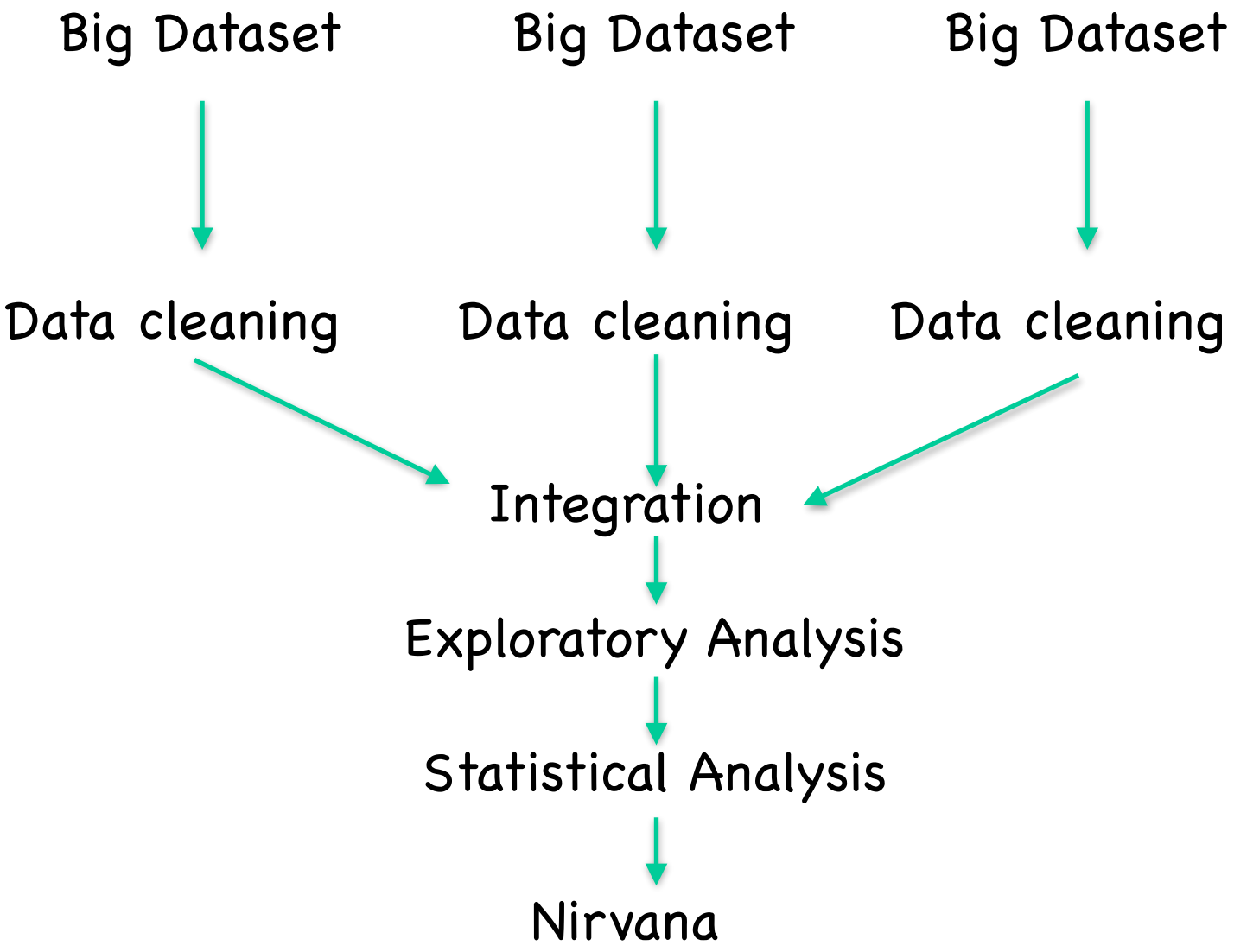
University of Pittsburgh

Department of Biomedical Informatics

harryh@pitt.edu

+1 410 648 9300

The "Big Data" story that I seem to hear so often.



Assumptions behind this story?

- Assume that you know
  - what the data are - what they represent
    - how they were collected
  - how they can be cleaned
  - how they can be integrated

- Metadata is key

Metadata

- Data about the data
- How was it collected?
- What were the sources?
- What has been done to it?

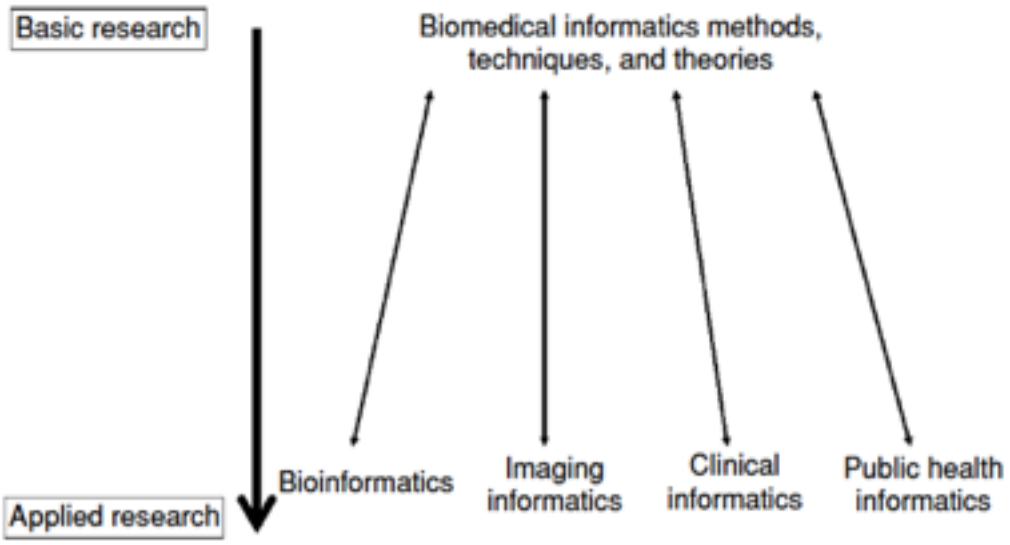
- Vitally important
- Often not available.
- Implications: Garbage In -> Garbage Out
- Why does this happen, and what can be done about it?

Outline

- Discussion of metadata challenges in biomedical research
- Metadata challenges
  - case study: FaceBase
- Ontologies to the rescue?
  - maybe..
- Computation with structured metadata via ontologies
- Challenges...

My perspective: Biomedical Informatics

- The use of computer systems for the improvement of biomedical research and clinical care



Shortliffe & Blois "The Computer Meets Medicine and Biology: Emergence of a Discipline" in Shortliffe & Cimino, eds., "Biomedical Informatics: Computer Applications in Health Care and Biomedicine

Friedman "A "fundamental theorem" of biomedical informatics. JAMIA 2009

Translational Research

Applying basic biological research to the generation of interventions that improve human health.

Model Systems

Genetics/Genomics

Human Disease

Grand Canyon NPS http://www.fotopedia.com/items/flickr-7553734530

Gulf(s) of Informatics: Translating across…

communities of practice - clinical vs. research
mouse vs. worm

data types - images, gene expression, clinical data, etc.

Harry Hochheiser, harryh@pitt.edu

Metadata is key for translational research

- In some fields, data may appear to be straightforward
  - (although, of course, appearances may be deceiving)
- Translational research
  - varied data formats
  - data sources
  - complex protocols
  - detailed analytics…

FaceBase

http://www.facebase.org, Hochheiser 2011

National Institute of Dental and Craniofacial Research

Five-year initial phase 2009-2014

"..systematically compile the biological instructions to construct the middle region of the human face and precisely define the genetics underlying its common developmental disorders, such as cleft lip and palate"

10 Projects + Data Management and Coordination Hub

- "One-stop access to craniofacial research data"
- "Allow scientists to more rapidly and effectively generate hypotheses and accelerate the pace of their research"
- Our task - build a site to present this data to the community.

- Can we develop tools that will help identify opportunities for data integration and sharing across projects, organisms, and modalities?
- Can we use these tools to promote data reuse and translational application of model system data?

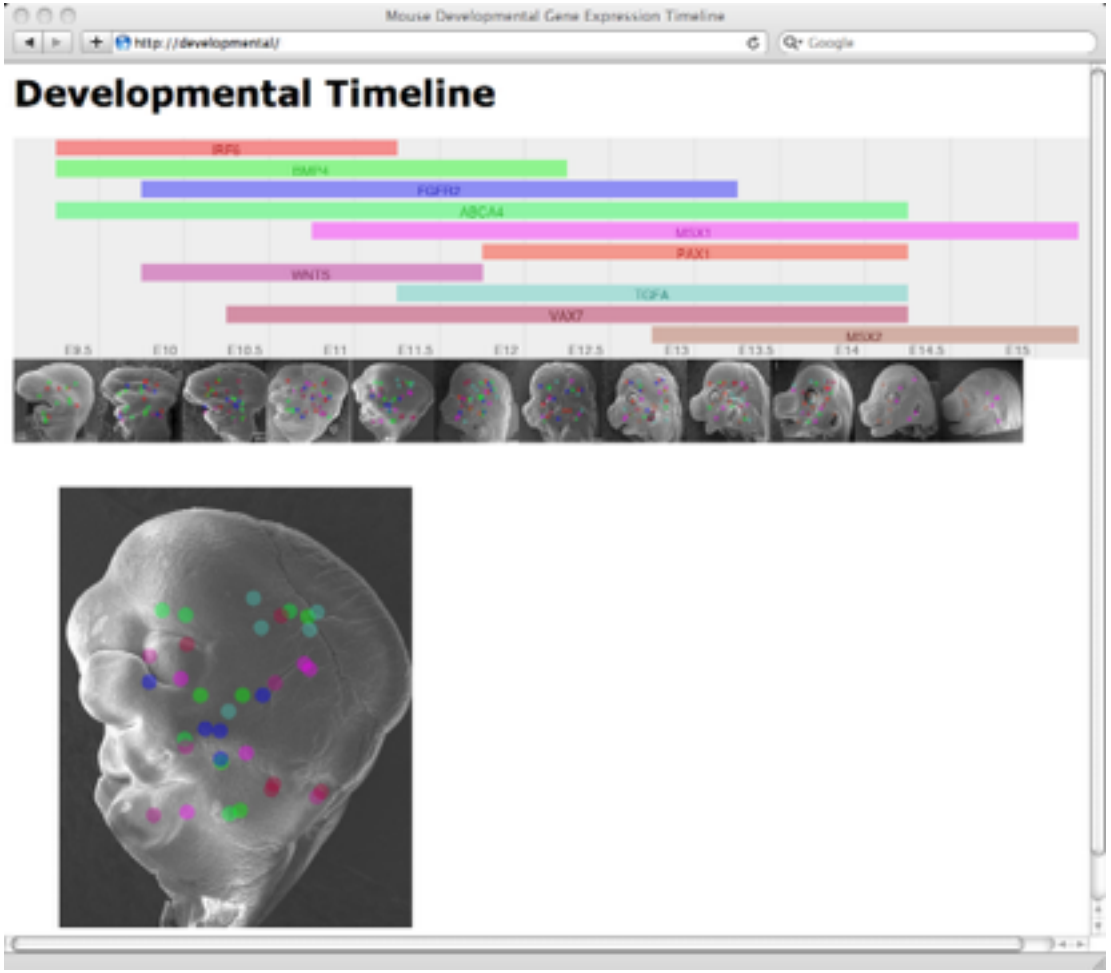
FaceBase Data

Data Diversity	Anatomy	miRNA	Phenotype	Images
	Genotypes	microarrays	RNA-Seq	Facial Images
Models	Human	Mouse	Zebrafish	
Developmental Stages	Embryos —> adults			

Goal: greater insight through data integration

Harry Hochheiser, harryh@pitt.edu

Gene Atlas



- Genes displayed on timeline, indicating when active
- Images display expression localization
- Large image for detailed views.
- Mouse-over coordinated links
- Integrate data from different groups, modalities, etc. to provide a comprehensive picture of gene expression localization through development of craniofacial region in mouse.

A translational Example

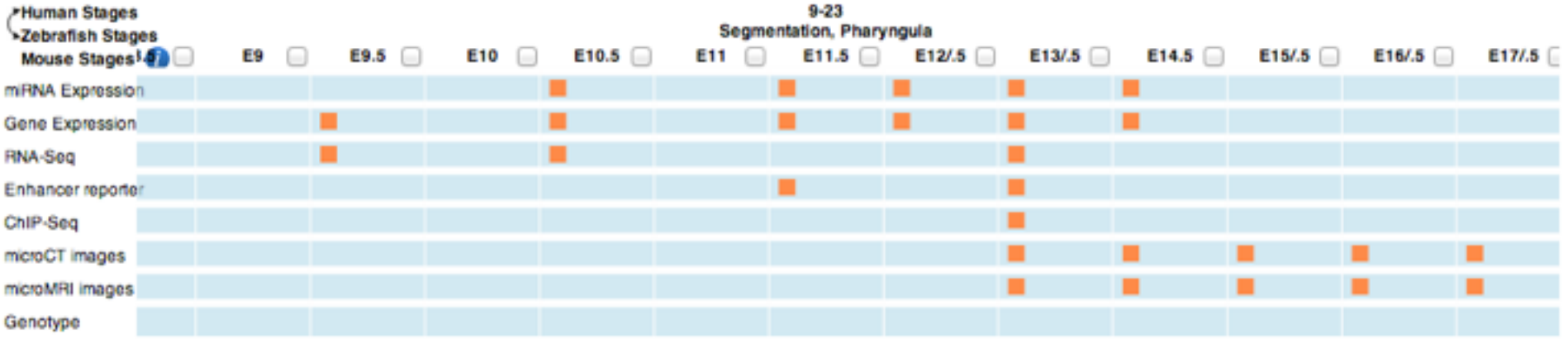
- LCM microarray identifies interesting gene expression in developing mouse snout
- What are the implications for humans
- Humans don't have snouts
- Genes may or may not be homologous
- How to map gene expression in developing mouse snout to relevant structures in developing/developed human???
  - Tests and therapeutic intervention.

What will it take to realize this vision?

- Gene expression data sets (microarray or RNA-Seq)
  - each from a specific
    - mouse strain
    - anatomic location
    - developmental time point
    - data collection platform
    - analysis pipeline
- What if you don't have one of these? What can you do?

Developmental Timeline Viewer

https://www.facebase.org/timeline



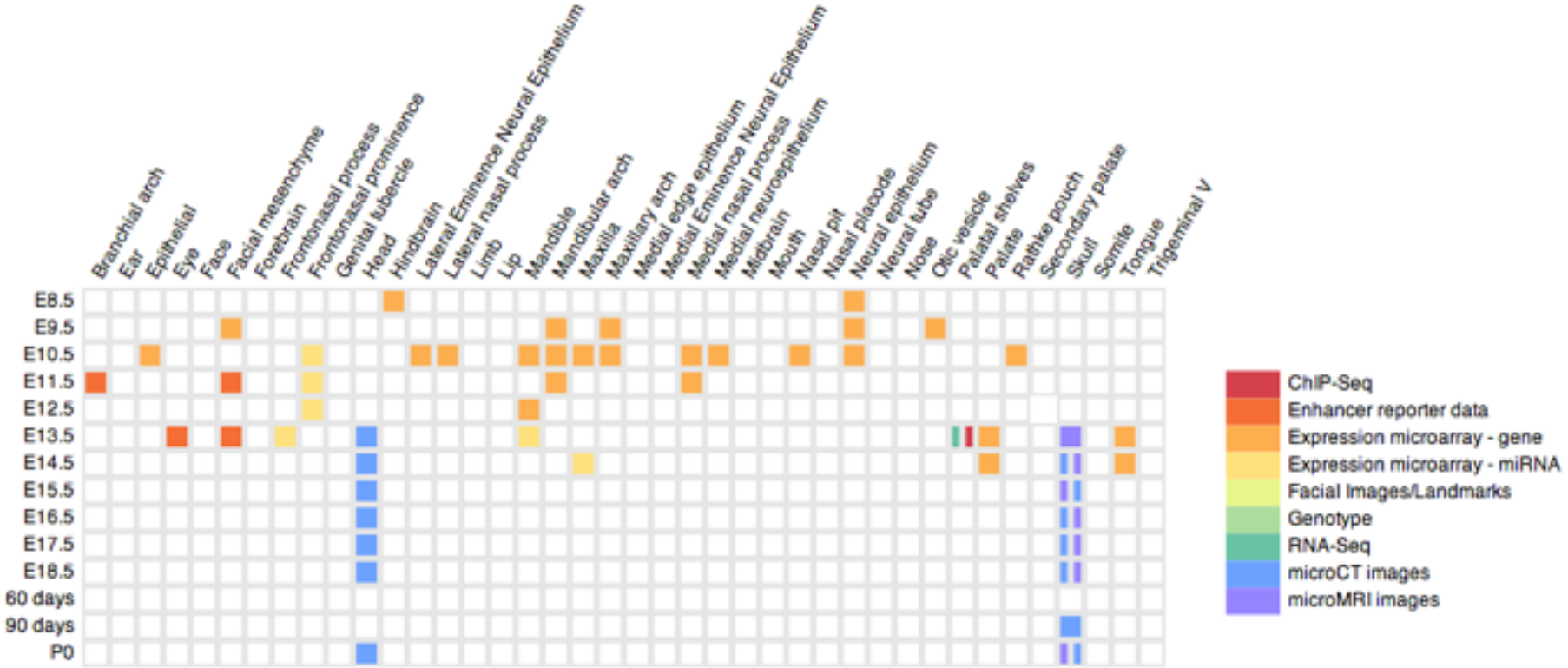
- Datasets on developmental timeline
- roughly aligned across organisms
- Lanes for different data modalities
- Filter by stage, data type...

Data Explorer

https://www.facebase.org/visualization

Support identification of data sets that might be "comparable"

- differing in at most one critical dimension



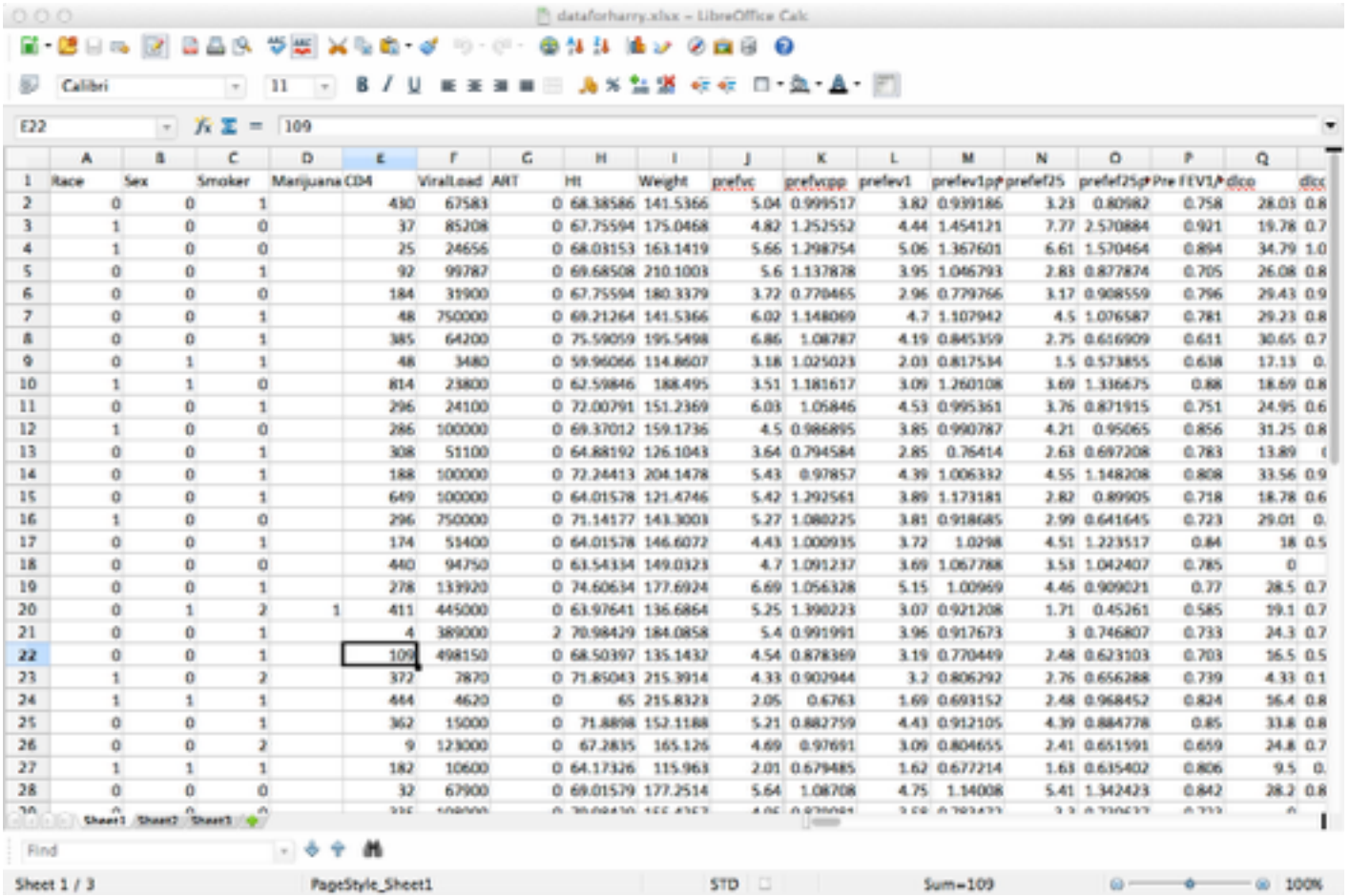
Challenges: Metadata

- Ideally, well-defined metadata fields/attributes
- Controlled vocabularies provide consistent terminology for each field
- Link to appropriate resources as needed: NCBI, MGI, etc..
- Additional attributes specific to each data type
- Consistent metadata supports search, navigation

Metadata in practice

- Ad-hoc formats, semantics
- Little or no agreement between projects
- Inconsistent terminology
- Why is this ?

The most widely-used biomedical research data management software tool?



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
1	Race	Sex	Smoker	Marijuana	CD4	ViralLoad	ART	Ht	Weight	prefvc	prefvcpp	prefev1	prefev1pp	prefef25	prefef25p	Pre FEV1/F	dlco	dlcc
2	0	0	1		490	67583	0	68.38586	141.5366	5.04	0.999517	3.82	0.939186	3.23	0.80982	0.758	28.03	0.8
3	1	0	0		37	85208	0	67.75594	175.0468	4.82	1.252552	4.44	1.454121	7.77	2.570884	0.921	19.78	0.7
4	1	0	0		25	24656	0	68.03153	163.1419	5.66	1.298754	5.06	1.367601	6.61	1.570464	0.894	34.79	1.0
5	0	0	1		92	99787	0	69.68508	210.1003	5.6	1.137878	3.95	1.046793	2.83	0.877874	0.705	26.08	0.8
6	0	0	0		184	31900	0	67.75594	180.3379	3.72	0.770465	2.96	0.779766	3.17	0.908559	0.796	29.43	0.9
7	0	0	1		48	750000	0	69.21264	141.5366	6.02	1.148069	4.7	1.107942	4.5	1.076587	0.781	29.23	0.8
8	0	0	1		385	64200	0	75.59059	195.5498	6.86	1.08787	4.19	0.845359	2.75	0.616909	0.611	30.65	0.7
9	0	1	1		48	3480	0	59.96066	114.8607	3.18	1.025023	2.03	0.817534	1.5	0.573855	0.638	17.13	0.
10	1	1	0		814	23800	0	62.59846	188.495	3.51	1.181617	3.09	1.260108	3.69	1.336675	0.88	18.69	0.8
11	0	0	1		296	24100	0	72.00791	151.2369	6.03	1.05846	4.53	0.995361	3.76	0.871915	0.751	24.95	0.6
12	1	0	0		286	100000	0	69.37012	159.1736	4.5	0.986895	3.85	0.990787	4.21	0.95065	0.856	31.25	0.8
13	0	0	1		308	51100	0	64.88192	126.1043	3.64	0.794584	2.85	0.76414	2.63	0.697208	0.783	13.89	(
14	0	0	1		188	100000	0	72.24413	204.1478	5.43	0.97857	4.39	1.006332	4.55	1.148208	0.808	33.56	0.9
15	0	0	1		649	100000	0	64.01578	121.4746	5.42	1.292561	3.89	1.173181	2.82	0.89905	0.718	18.78	0.6
16	1	0	0		296	750000	0	71.14177	143.3003	5.27	1.080225	3.81	0.918685	2.99	0.641645	0.723	29.01	0.
17	0	0	1		174	51400	0	64.01578	146.6072	4.43	1.000935	3.72	1.0298	4.51	1.223517	0.84	18	0.5
18	0	0	0		440	94750	0	63.54314	149.0323	4.7	1.091237	3.69	1.067788	3.53	1.042407	0.785	0	
19	0	0	1		278	133920	0	74.60634	177.6924	6.69	1.056328	5.15	1.00969	4.46	0.909021	0.77	28.5	0.7
20	0	1	2	1	411	445000	0	63.97641	136.6864	5.25	1.390223	3.07	0.921208	1.71	0.45261	0.585	19.1	0.7
21	0	0	1		4	389000	2	70.98429	184.0858	5.4	0.991991	3.96	0.917673	3	0.746807	0.733	24.3	0.7
22	0	0	1		109	498150	0	68.50397	135.1432	4.54	0.878369	3.19	0.770449	2.48	0.623103	0.703	16.5	0.5
23	1	0	2		372	2870	0	71.85043	215.3914	4.33	0.902944	3.2	0.806292	2.76	0.656288	0.739	4.33	0.1
24	1	1	1		444	4620	0	65	215.8323	2.05	0.6763	1.69	0.693152	2.48	0.968452	0.824	16.4	0.8
25	0	0	1		362	15000	0	71.8898	152.1188	5.21	0.882759	4.43	0.912105	4.39	0.884778	0.85	33.8	0.8
26	0	0	2		9	123000	0	67.2835	165.126	4.69	0.97691	3.09	0.804655	2.41	0.651591	0.659	24.8	0.7
27	1	1	1		182	10600	0	64.17326	115.963	2.01	0.679485	1.62	0.677214	1.63	0.615402	0.806	9.5	0.
28	0	0	0		32	67900	0	69.01579	177.2514	5.64	1.08708	4.75	1.14008	5.41	1.342423	0.842	28.2	0.8

I've made this claim to many groups… haven't had any disagreement yet.

What does this mean for the data pipeline?

	A	B	C	D	E	F	G	H	I	
1	Subject ID	File name	species	stage (embryonic day-dpc)	stage-other	strain/background	mutation	genotype	phenotype	litter
2	s1_6Feb13	s1_6Feb13.nii	mus muscululs:mouse	E18.5	null	C57Bl/6J	none	N/A	Normal	
3	s2_6Feb13	s2_6Feb13.nii	mus muscululs:mouse	E18.5	null	C57Bl/6J	Smad4 conditional	Myf5-Cre;Smad4fl/fl	Small tongue	JIL02
4	s3_6Feb13	s3_6Feb13.nii	mus muscululs:mouse	E18.5	null	C57Bl/6J	none	N/A	Normal	
5	s4_6Feb13	s4_6Feb13.nii	mus muscululs:mouse	E18.5	null	129/SvJ;C57B6J	Alk5 and Tgfbr2 conditional	Wnt1-Cre;Alk5fl/fl;Tgfbr2fl/+	Craniofacial abnormalities	JIL01
6	s5_6Feb13	s5_6Feb13.nii	mus muscululs:mouse	E18.5	null	C57Bl/6J	none	N/A	Normal	JIL02
7	s6_6Feb13	s6_6Feb13.nii	mus muscululs:mouse	E18.5	null	129/SvJ;C57B6J	Tgfbr2 conditional	K14-Cre;Tgfbr2fl/fl	Submucosal cleft palate, Cleft s	JIL01
8	s7_6Feb13	s7_6Feb13.nii	mus muscululs:mouse	E18.5	null	C57Bl/6J	none	N/A	Normal	JIL01

"What's the difference between mutation and genotype?"

"Uhh…. I'll have to get back to you on that."

"Strain name? Is "C57B6J" the same as "C57Bl/6J?"

"We're lucky to have any information at all when it comes to these mice…"

"The official name of the strain is _____, but everybody calls it _____"

The problem is widespread

C57BL/6N Mutation in *Cytoplasmic FMRP interacting protein 2* Regulates Cocaine Response

Vivek Kumar,[1,2] Kyungin Kim,[1] Chryshanthi Joseph,[1] Saïd Kourrich,[3]* Seung-Hee Yoo,[1]* Hung Chung Huang,[1] Martha H. Vitaterna,[4] Fernando Pardo-Manuel de Villena,[5] Gary Churchill,[6] Antonello Bonci,[3,7] Joseph S. Takahashi[1,2]‡

Science, 20 December 2013

Relative to C57BL/6J …
"We found that C57BL/6N has a lower acute and sensitized response to cocaine and methamphetamine." - single variant difference

How easy would it be for two variants to be confused?

Research Reproducibility

On the reproducibility of science: unique identification of research resources in the biomedical literature

Nicole A. Vasilevsky[1], Matthew H. Brush[1], Holly Paddock[2], Laura Ponting[3], Shreejoy J. Tripathy[4], Gregory M. LaRocca[4], Melissa A. Haendel[1]

PeerJ:e148

238 articles from top journals in 5 fields

54% of resources are not uniquely identifiable

Policy: NIH plans to enhance reproducibility

Francis S. Collins & Lawrence A. Tabak

Nature, 27 January 2014

OPEN ACCESS Freely available online

PLOS | ONE

Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome

Daniel Garijo[1], Sarah Kinnings[2], Li Xie[3], Lei Xie[4], Yinliang Zhang[5], Philip E. Bourne[3]*, Yolanda Gil[6]¶

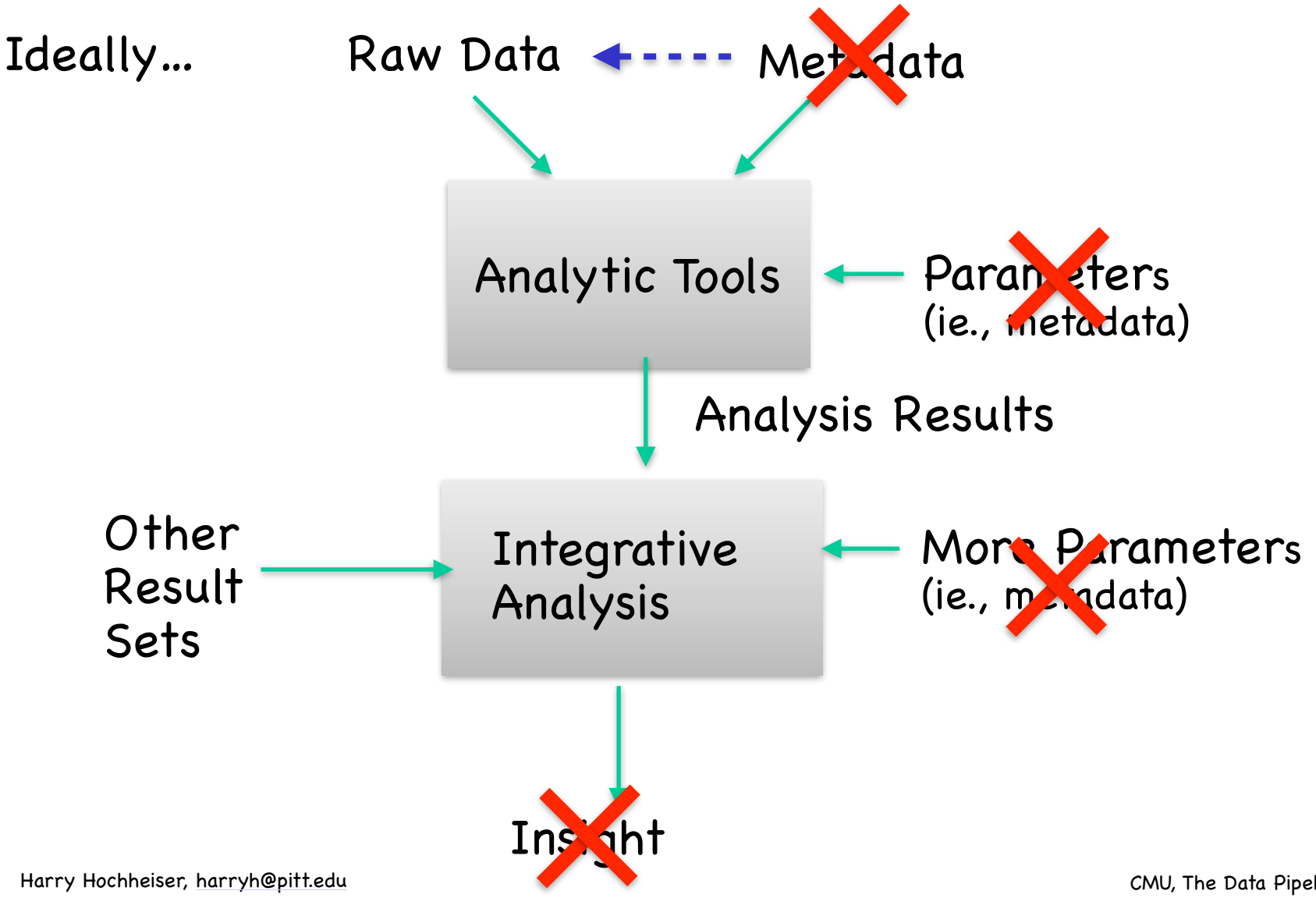
November 2013

aka "I cannot reproduce the work from my own laboratory" by Phil Bourne, now NIH Directory for Data Science
http://www.slideshare.net/pebourne/ebi121102013

Analytic metadata

- Well-curated metadata + unspecified analytics = ?
- Interpretation depends on analysis
  - normalizing microarrays
  - assembling genome sequences
- How to interpret results of unknown analysis?
- Gronenschild, et al. "The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements" DOI: 10.1371/journal.pone.0038234
  - Analytic results differ from OS 10.5 - 10.6

Where can things go wrong?



Biomedical Ontologies

Ontology: A set of categories, connected by meaningful relations:
"is part of", "connected to", etc. Def. thanks to Gary Merrill

- Nose is part of face
- Human nose is analogous to mouse snout.

- Knowledge representation
- Search/retrieval
- Contextualization
- Computation
  - Calculation
  - Inference

Ontological annotations

Metadata needed for translational science

The Gene Ontology (GO)

http://www.geneontology.org/GO.doc.shtml

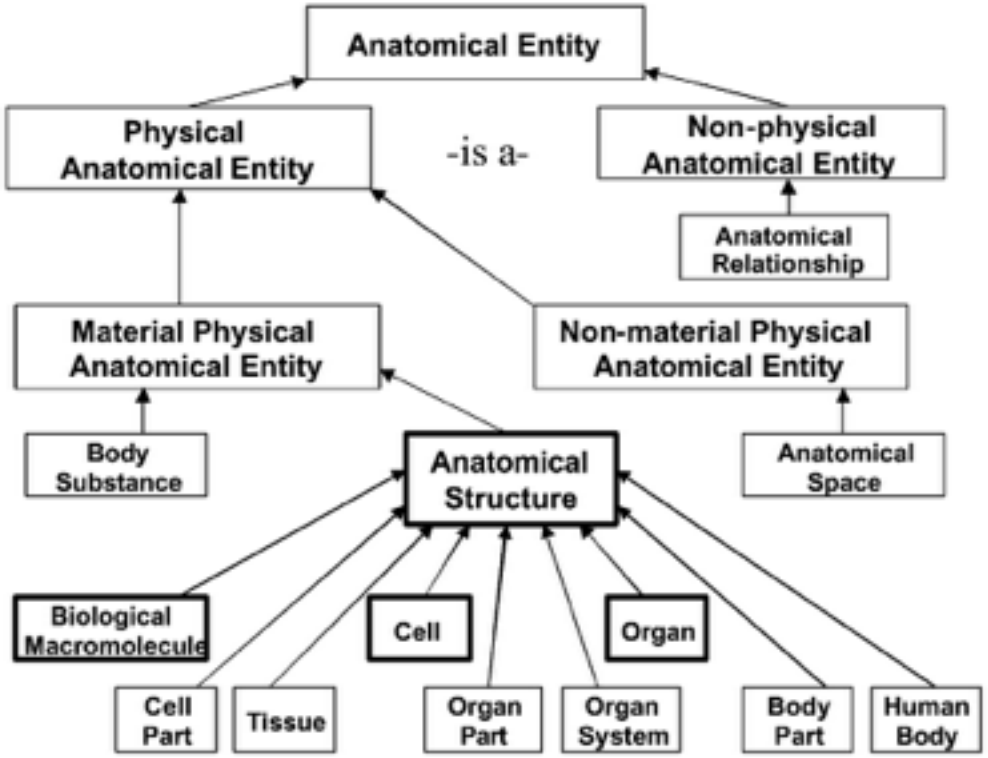
- 3 domains of gene product properties
  - Cellular component
  - Molecular function
  - Biological process
- How are they organized?
- Recurrent Theme: is a vs. part of
- Browse at AmiGO
  - http://amigo.geneontology.org/cgi-bin/amigo/go.cgi

Foundational Model of Anatomy

Rosse & Mejino 2003

http://sig.biostr.washington.edu/projects/fm/

- Human anatomy >75K classes, 168 relationship types
  120K terms, 2.1 million relationships



All anatomy is not equal…

	Human	Mouse	Zebrafish
Adult	FMA	Mouse Anatomy (MA)	Zebrafish Anatomy (ZFA)
Developmental	EHDAA Human Developmental Anatomy	EMAPA	

How to discuss analogous structures across organisms?

Necessary for using ontologies for translational comparisons…

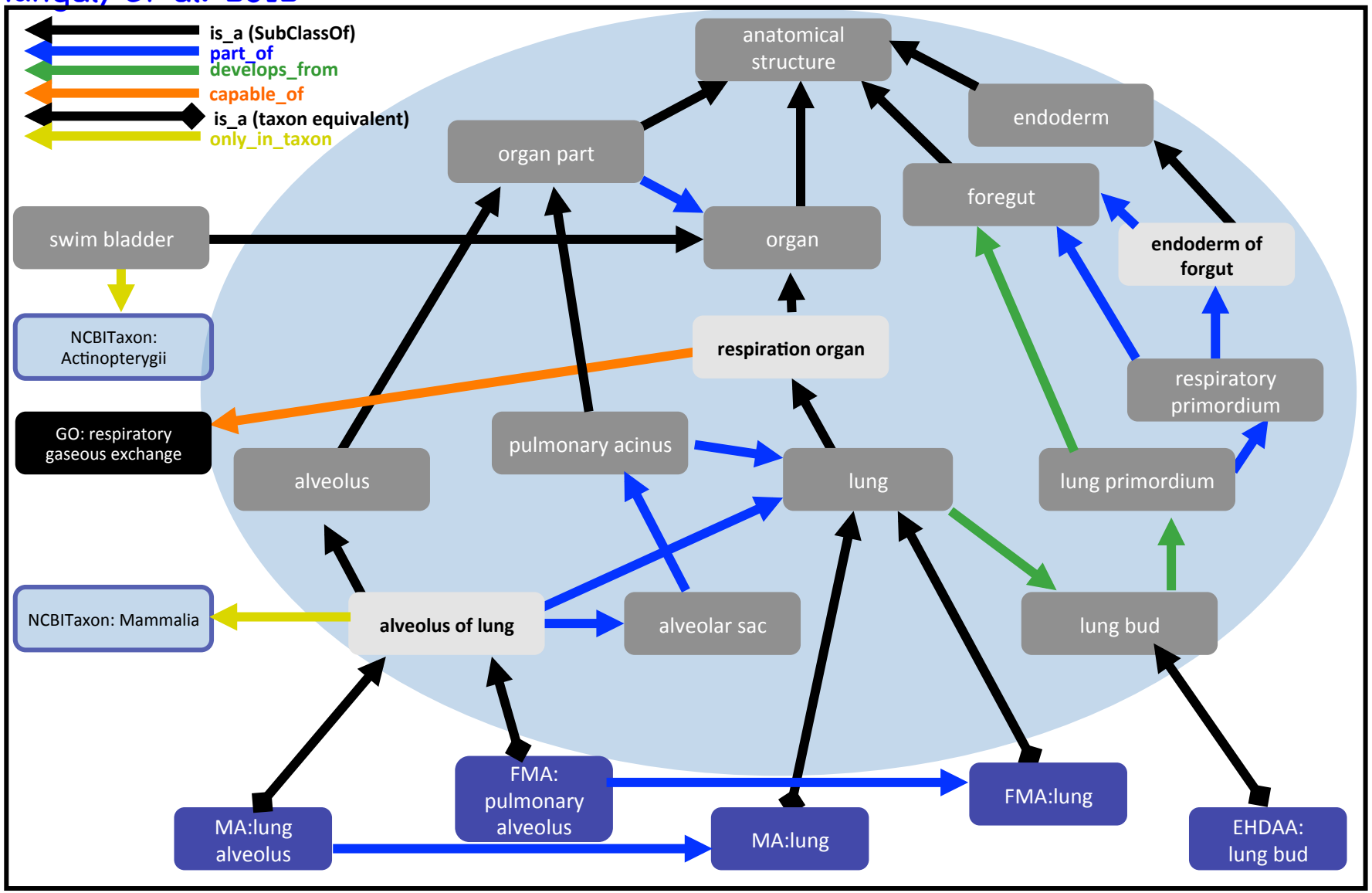
Two approaches for bridging the gap

Mapping: inter-ontology links

Unifying ontology: super structure with subclasses linking related items.

Uberon, an integrative multi-species anatomy ontology

Mungall, et al. 2012

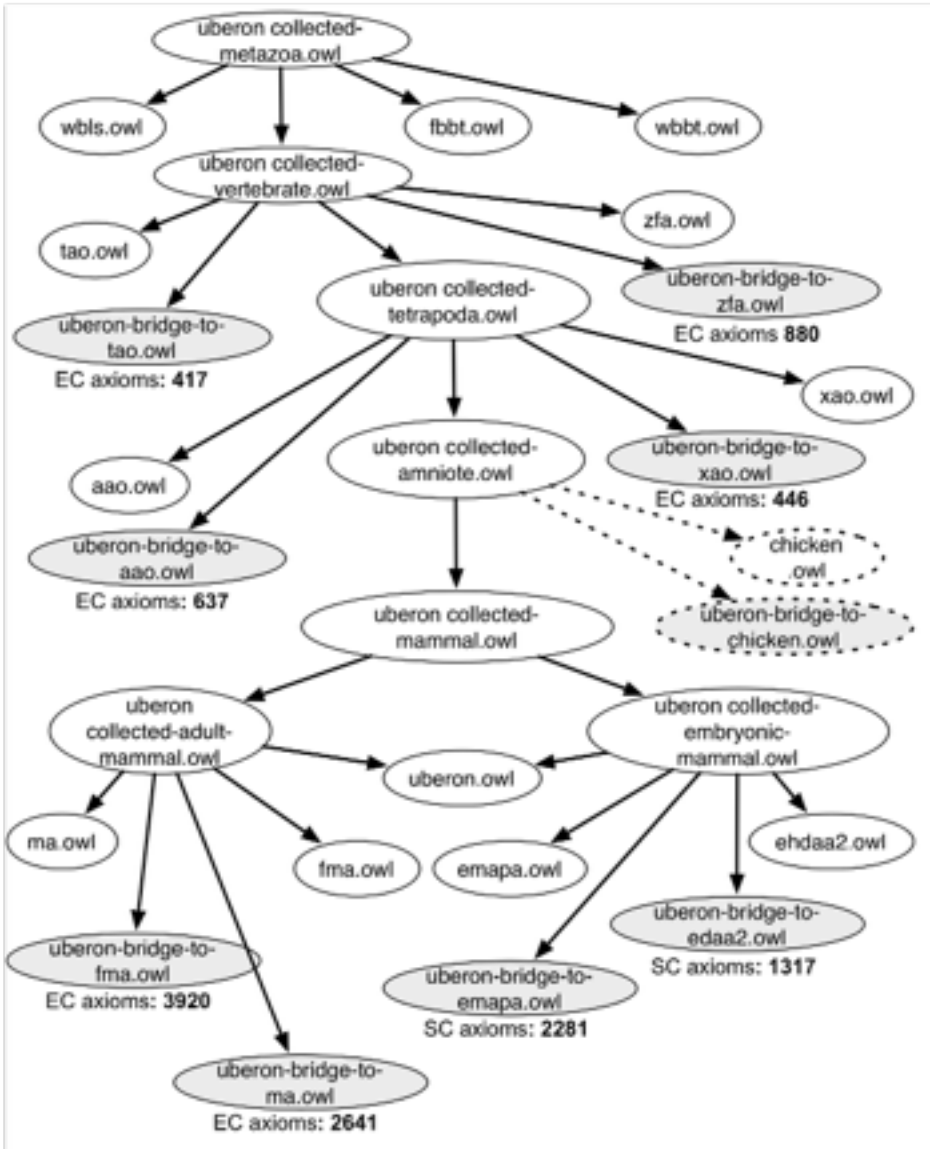


Harry Hochheiser, harryh@pitt.edu

UBERON

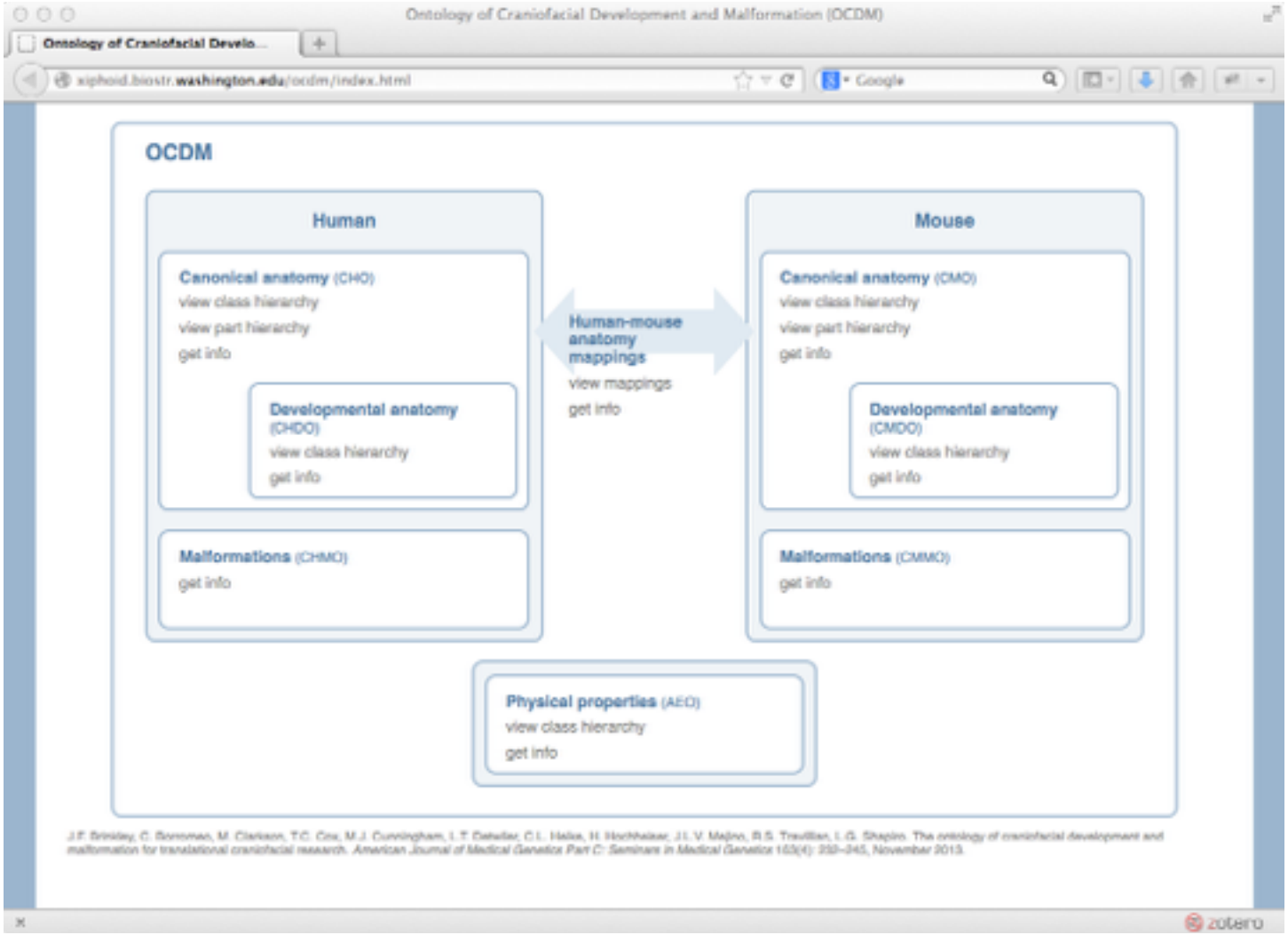
Mungall, et al. 2012

Combination of automated and manual effort



Mapping: the Ontology of Craniofacial Development and Malformation

http://xiphoid.biostr.washington.edu/ocdm/index.html Brinkley, et al. 2013



Harry Hochheiser, harryh@pitt.edu

Comparing the Methods

- Mappings
  - Potentially more lightweight
    - No higher-structure required
  - More flexible
  - No semantics – all structural
- Unifying
  - Clarity of structure
  - Homology vs. other similarity? - judgment call
  - Multiple information sources required?
- **Curation Required!**
  - Human "extensor retinaculum of wrist" does not correspond to mouse retina!

Objections

1. "Canonical"/expected anatomy is not all that we need to understand the genetics of development and disease

   we need malformations and disease to link to genetic variation

   ... more on this in a minute

2. ?????

   (this one is up to you to figure out..)

Beyond Anatomy

- Anatomy ontologies describe canonical organisms
  - Or, at least their parts. May not describe appearance, etc.
- Translational bioinformatics is often interested in specific characteristics
  - Mendelian traits
  - Malformations/abnormalities
    - FaceBase - cleft lip/palate
    - Absence of Structures

Phenotypes

- "observed" properties
  - Physical
  - Behavioral, etc.

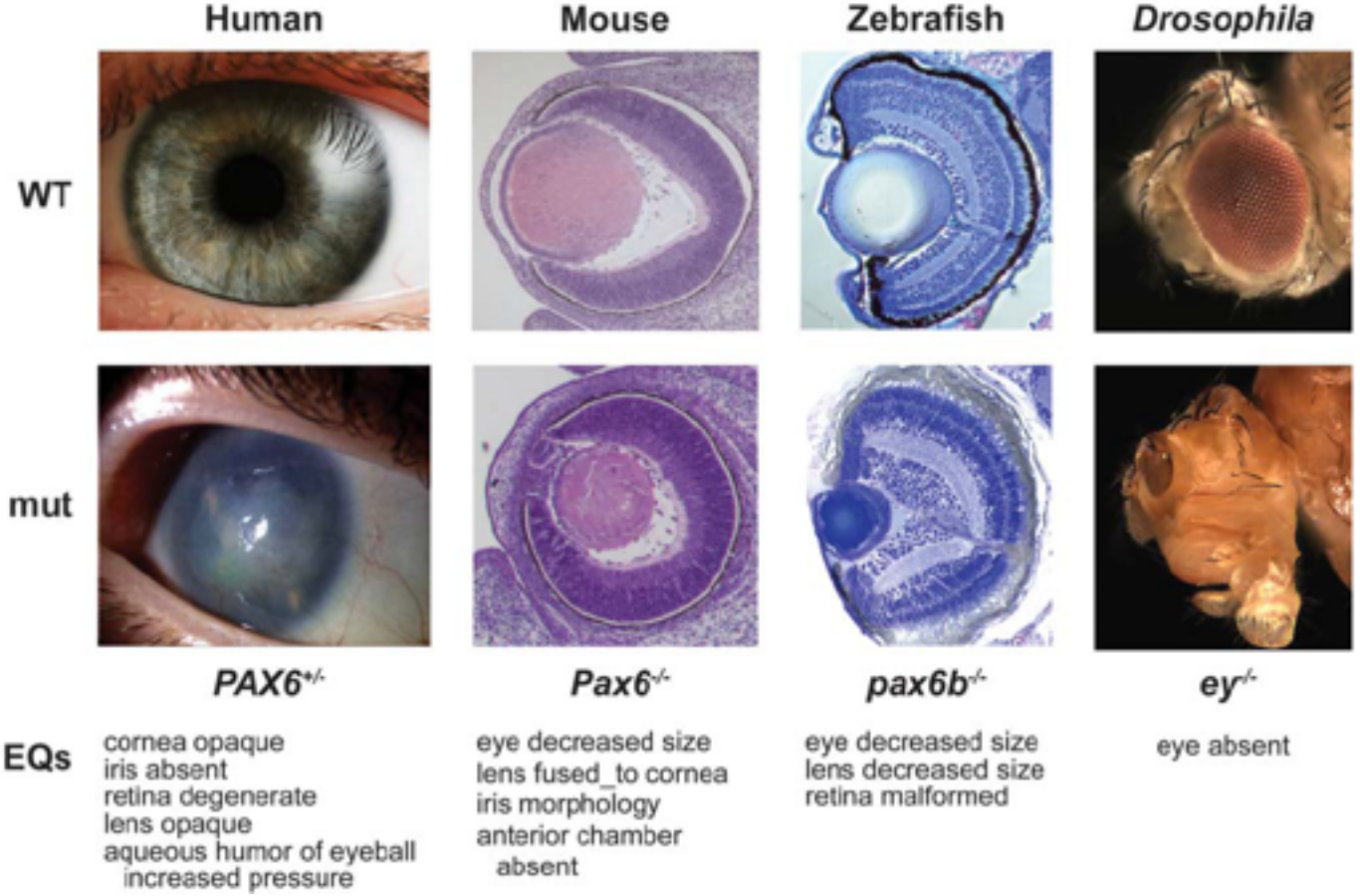
- Characterization of phenotypes is key to translation from model organism → human
  - Phenotype in model organism, might be a model for comparable phenotype in human ..
  - ..particularly if there is a homologous gene

Phenotype-genotype linkages

Washington, et al. 2009

Comparable phenotypes across organisms with homologous genes

-> translational research



Phenotype Ontologies

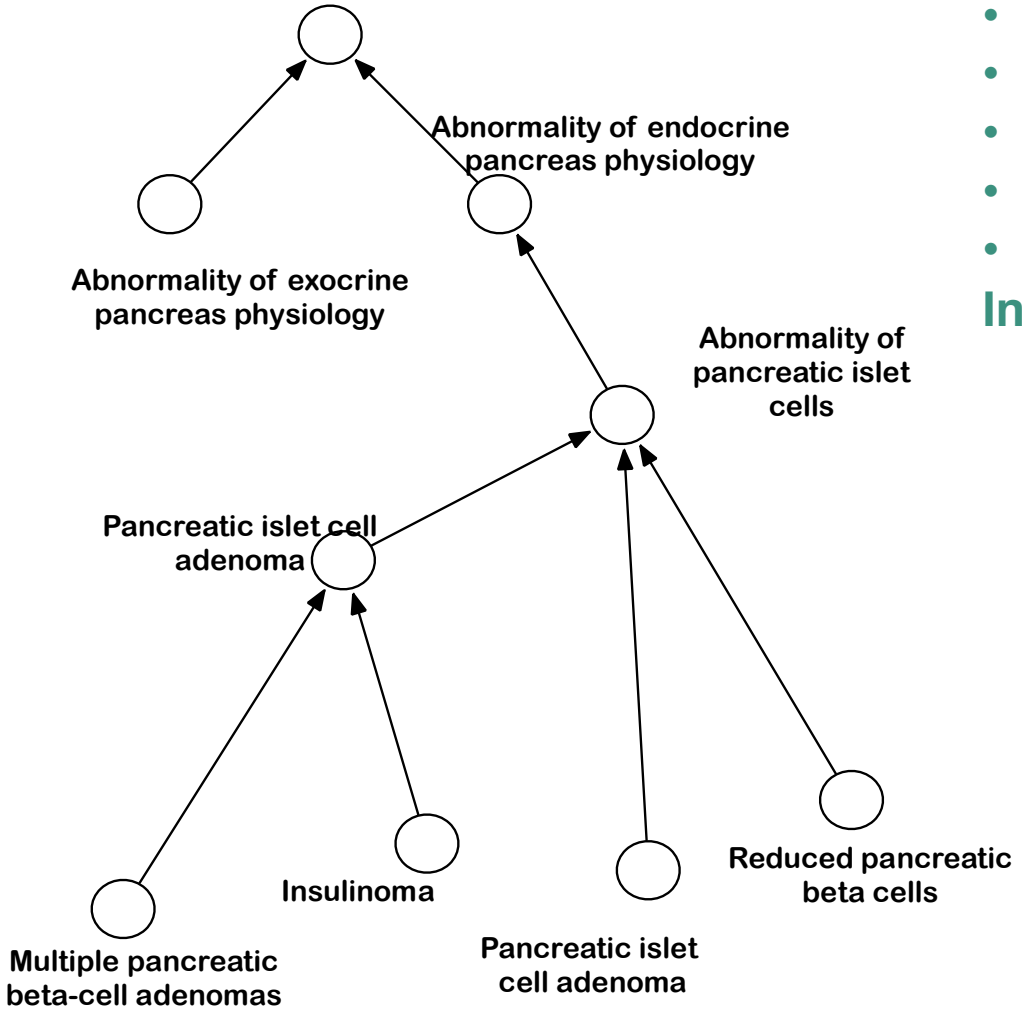
- Human Phenotype Ontology

  (Köhler, et al. 2014, www.human-phenotype-ontology.org)

- Strict is_a hierarchy
- Mode of inheritance: autosomal dominant, somatic mutation, etc.
- Onset and clinical course – age of onset speed, pace of progression, variability
- Phenotypic Abnormality – abnormality by anatomic region
- Also, Disease ontology (DO) Schriml et al. 2011

Human Phenotype Ontology

M. Haendel, used with permission



Used to annotate:

- **Patients**
- **Disorders**
- **Genotypes**
- **Genes**
- **Sequence variants**

In human

Mammalian Phenotype Ontology

Smith, et al. 2004

- Heavily biased towards mouse

Mammalian Phenotype Browser

Term Detail

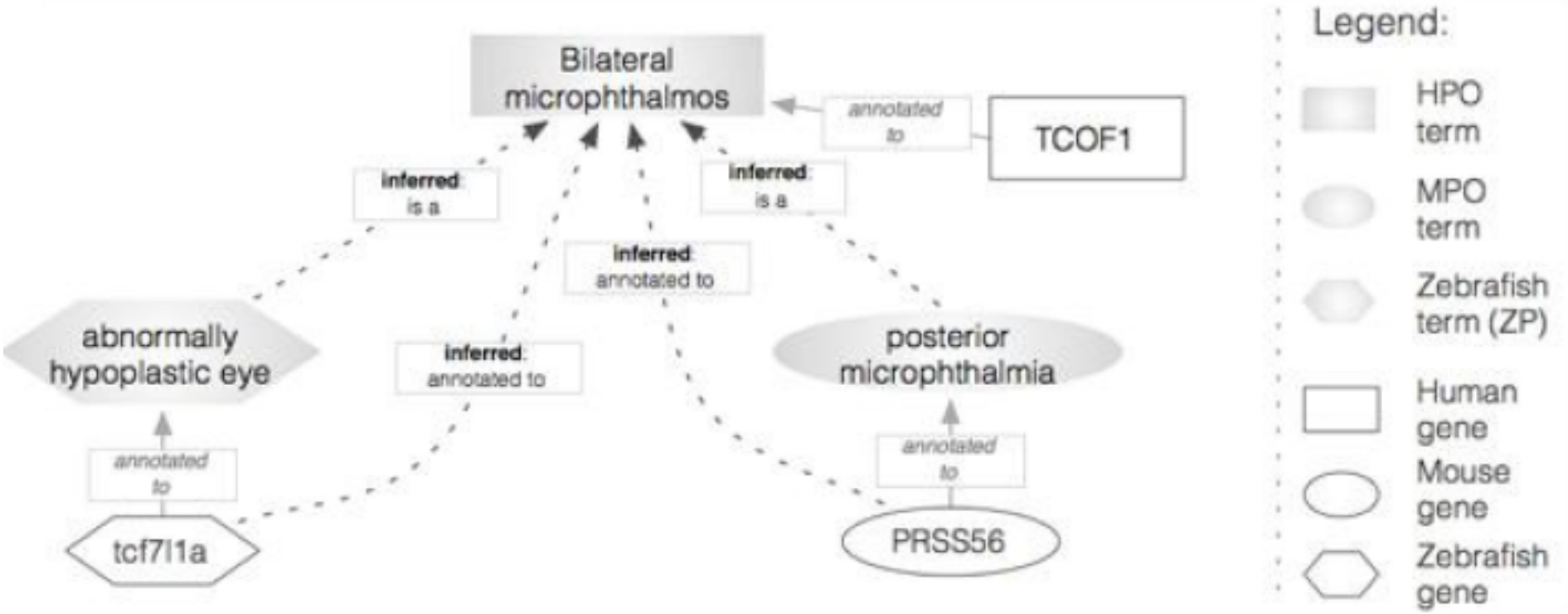
MP term: **lethality-embryonic/perinatal**
Synonym: **survival**
MP id: **MP:0005374**
Number of paths to term: **1**

denotes an 'is-a' relationship
denotes a 'part-of' relationship

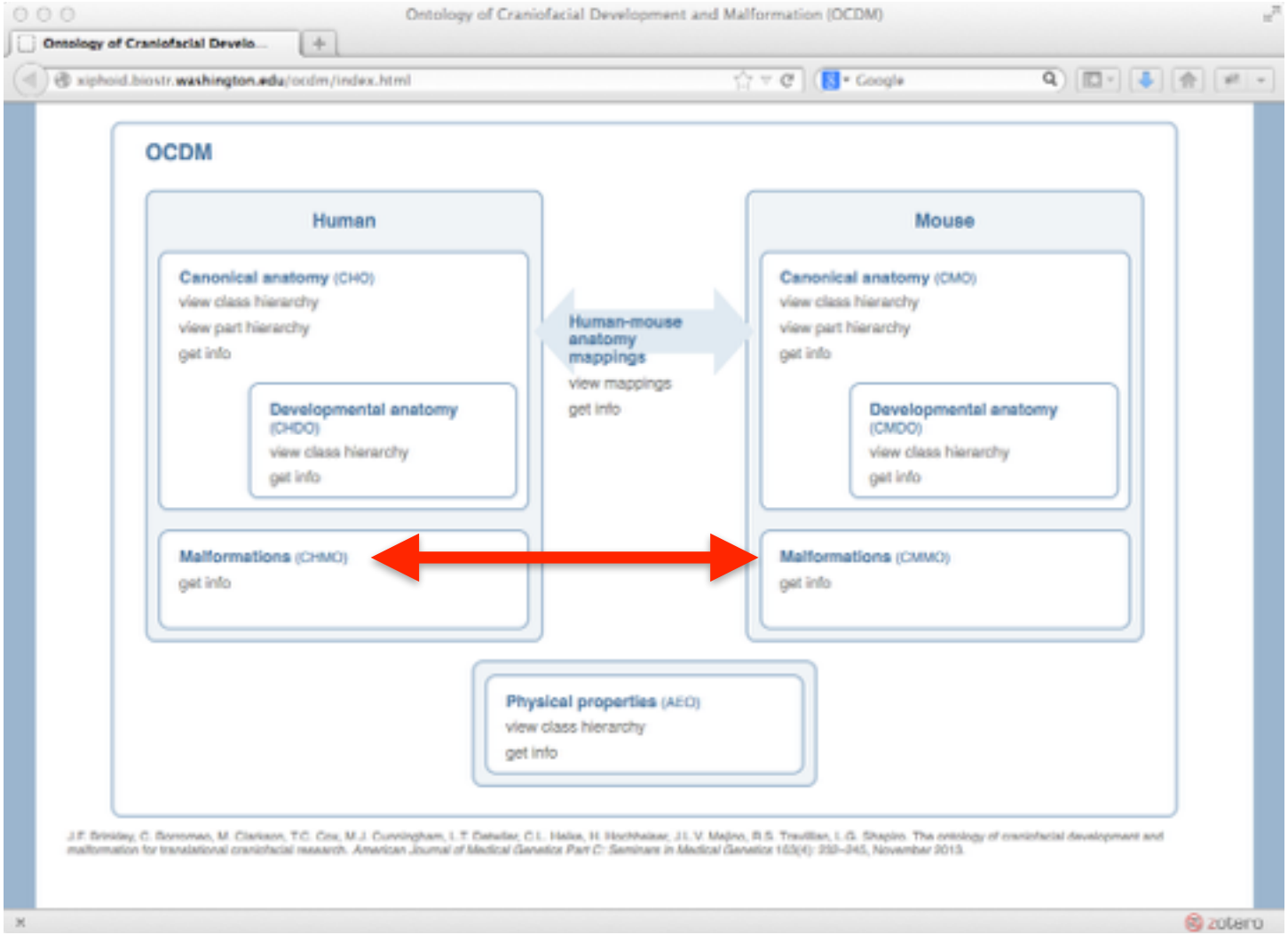
Phenotype Ontology
- adipose tissue phenotype +
- behavior/neurological phenotype +
- cardiovascular system phenotype +
- cellular phenotype +
- central nervous system phenotype +
- craniofacial phenotype +
- digestive/alimentary phenotype +
- embryogenesis phenotype +
- endocrine/exocrine gland phenotype +
- growth/size phenotype +
- hearing/ear phenotype +
- hematopoietic system phenotype +
- homeostasis/metabolism phenotype +
- immune system phenotype +
- lethality-embryonic/perinatal [MP:0005374] *(1482 genotypes, 1596 annotations)*
  - embryonic lethality
  - perinatal lethality +
- lethality-postnatal +
- life span-post-weaning/aging +
- limbs/digits/tail phenotype +
- liver/biliary system phenotype +
- muscle phenotype +
- normal phenotype +
- obsolete +
- other phenotype +
- peripheral nervous system phenotype +
- renal/urinary system phenotype +
- reproductive system phenotype +
- respiratory system phenotype +
- skeleton phenotype +
- skin/coat/nails phenotype +
- taste/olfaction phenotype +
- touch/vibrissae phenotype +
- tumorigenesis +
- vision/eye phenotype +

Uberpheno: cross-species phenotype ontology

Köhler, et al. 2013



Mapping – OCDM redux.



Mapping between anatomy and phenotype

Mungall, et al. 2010

"Logical equivalences"

- Define Phenotypes as intersection of
  - Anatomic entities (FMA,MA. Etc..)
  - Phenotypic Qualities

<Q> that inheres_in some <E>

MP: 008152 decreased diameter of femur is equivalent to

Decreased diameter that inheres_in some femur

So what does all of this give us?

Rich anatomy ontologies

+

Phenotype ontologies

+

cross-species links..

serious ability to create rich models of relationships between translational data types - animal models, human diseases, etc.

OWL and Reasoning

Representation of biomedical ontologies - OWL

OWL reasoners - consistency checks

Inference.

A part_of B & B part_of C -> A part_of C

Ability to ask some interesting questions:

Which genes are differentially expressed in developmental anatomic regions that develop into the palate?

Reasoning via transitivity

C. Mungall, with permission

tooth SubClassOf develops_from some tooth bud

tooth bud SubClassOf develops_from some tooth placode

develops_from is transitive

->**tooth SubClassOf develops_from some tooth placode**

… as does any subclass of tooth.

Computational methods for extracting insight from ontologically-annotated data

Use structure of ontologies and related data to compute, generate new hypotheses, etc.

PAX6 example - human, mouse, fish phenotypes linked by knowledge of common orthologous gene...

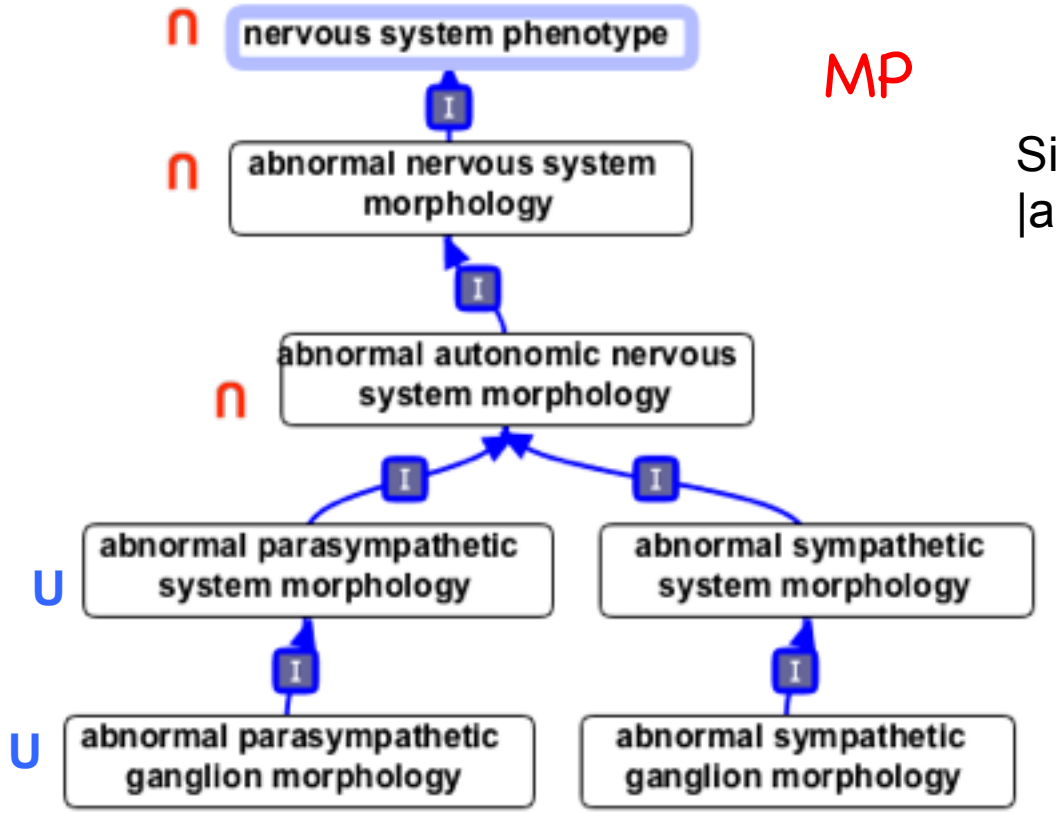
Can we use phenotypic annotations to find models of human diseases, even when a gene is not known?

Broad idea

- Use relationships in ontologies to infer relationships/ develop new theories
- Probabilistic and similarity measures
- Start with one ontology..
- Then combine...

Graph Similarity for the Identification of Candidate Genes

C. Mungall, with permission



MP

- What genes are **phenotypically** similar to Phox2a?

SimJ(a,b) =
|a ∩ b| / |a ∪ b|

SimJ(Phox2a,Sox10) = 3/7 = 0.42
SimJ(Phox2a,Phox2b) = 1

Phox2a

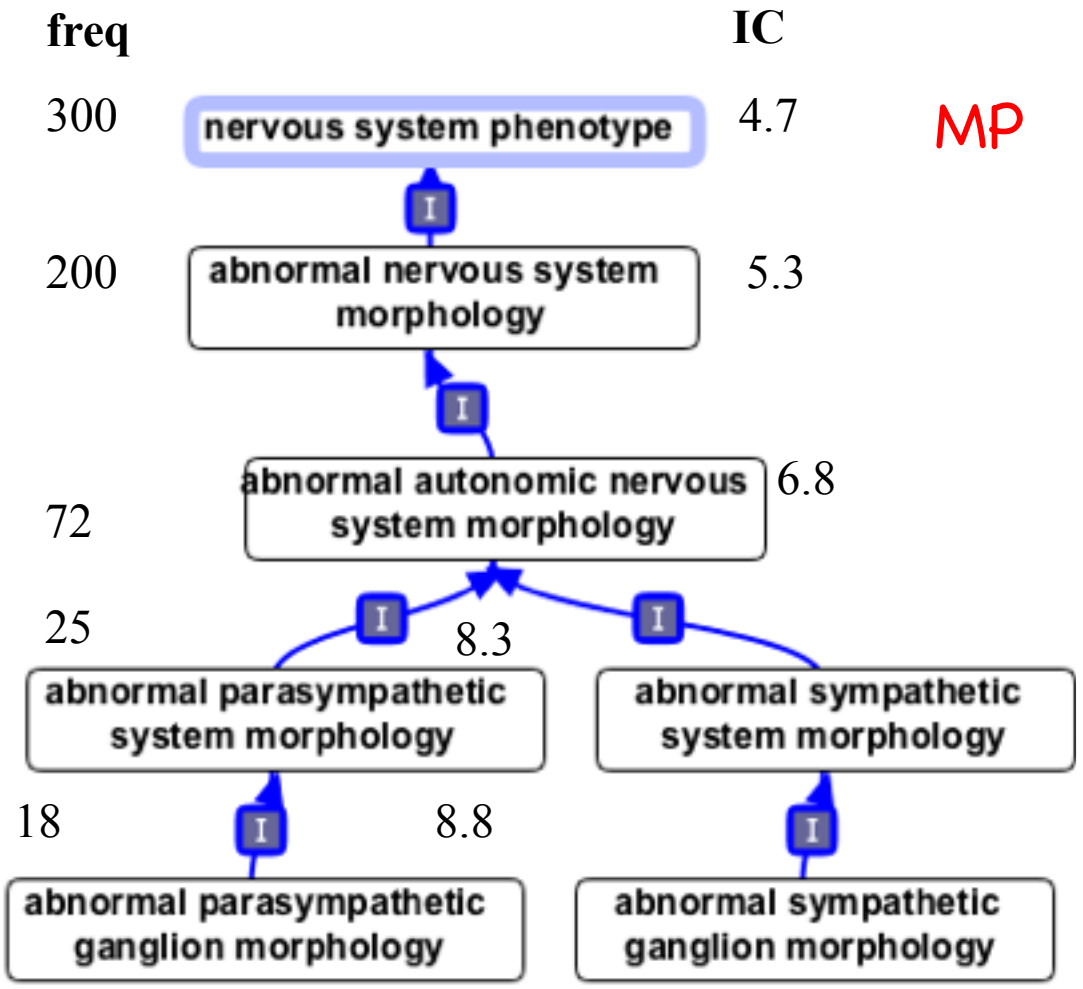
Phox2b

Sox10

Phox2b

Information Content

C. Mungall, with permission



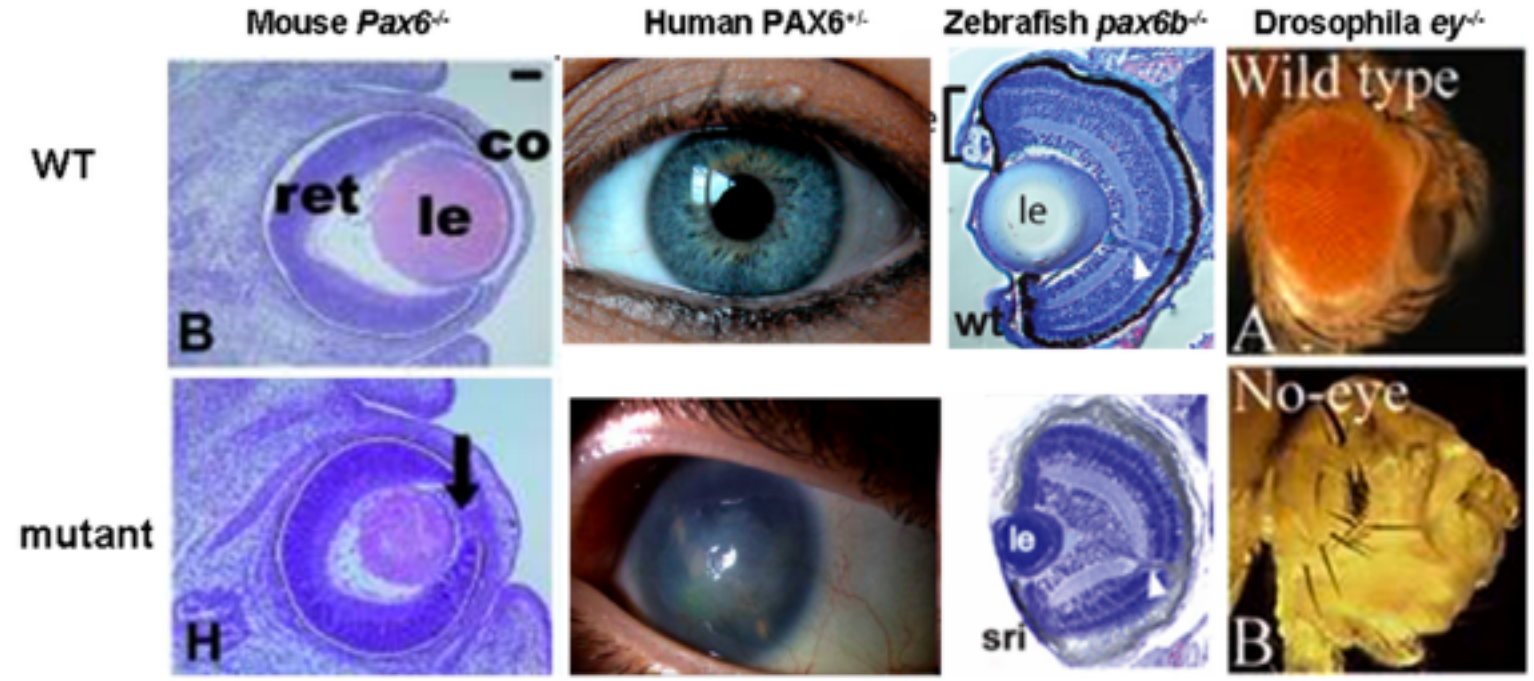
IC(t) = -log(p(t))

p(t) = |annot(t)|/|annot|

MaxIC(Phox2a,Sox10) = 6.8
MaxIC(Phox2a,Phox2b) = 8.8

Cross-species comparison

C. Mungall, with permission



EQs

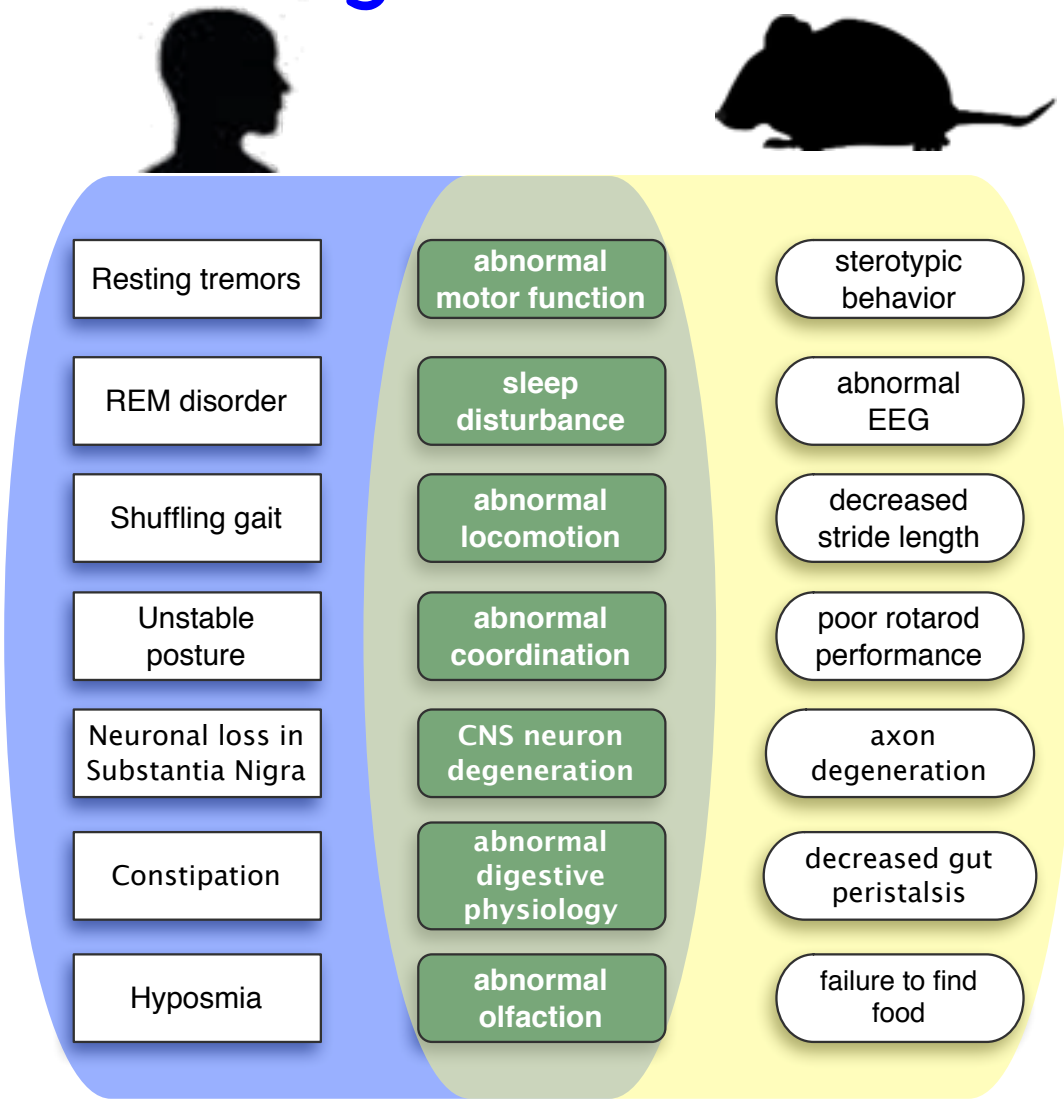
Mouse *Pax6*$^{-/-}$	Human PAX6$^{+/-}$	Zebrafish *pax6b*$^{-/-}$	Drosophila *ey*$^{-/-}$
eye decreased size cornea opaque lens fused_to corneal epithelium anterior chamber decreased size	eye decreased size cornea opaque iris decreased size or absent retina malformed lens opaque anterior chamber decreased size	eye decreased size lens decreased size retina malformed	eye absent

SimJ: 0.42
MaxIC: 13.4

SimJ: 0.32
MaxIC: 12.1

SimJ: 0.17
MaxIC: 6.2

OWLsim: Phenotype similarity across patients or organisms

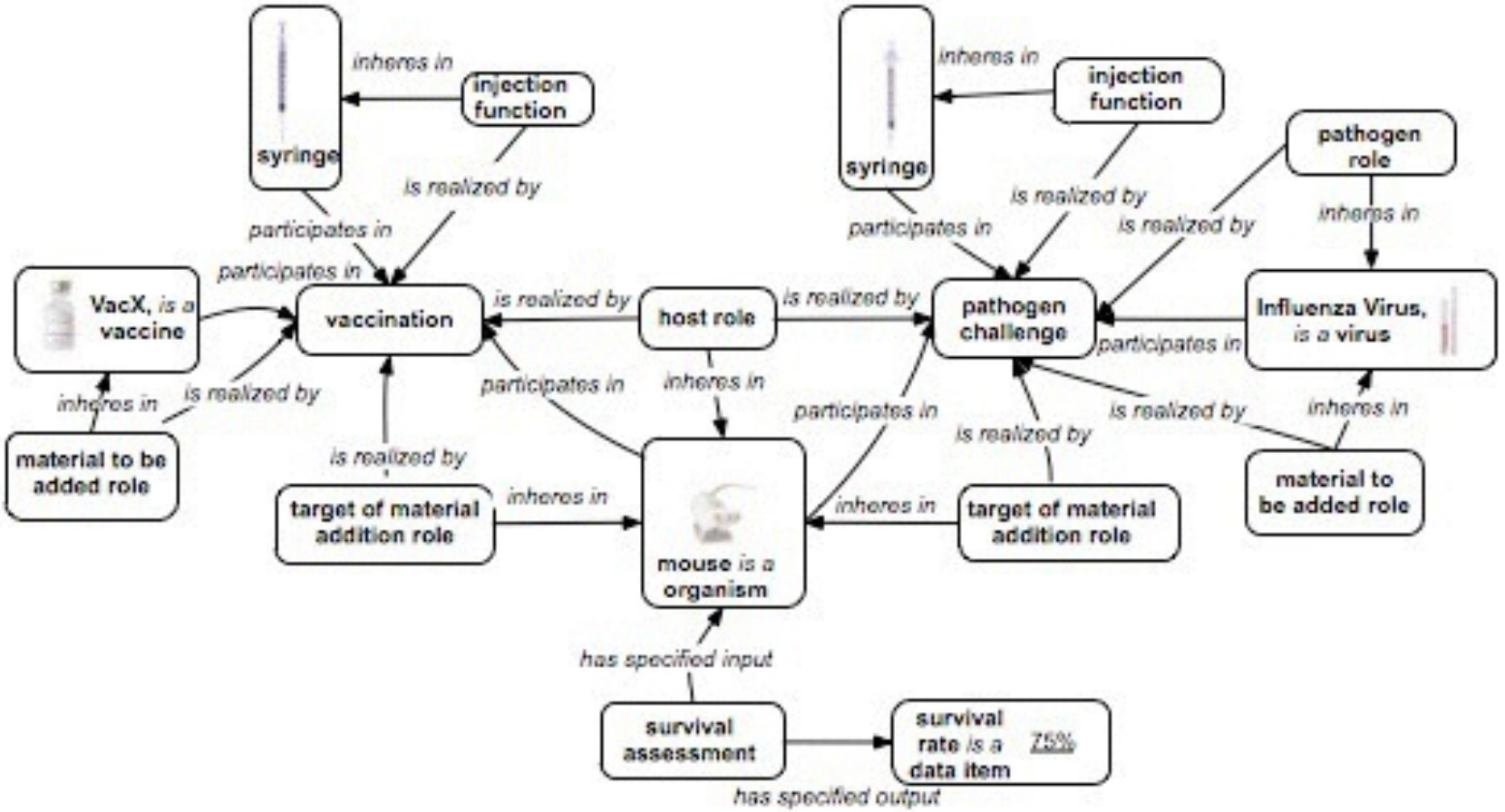


https://code.google.com/p/owltools/wiki/OwlSim

Harry Hochheiser, harryh@pitt.edu

Experimental data: Ontology for Biomedical Investigations

Brinkman, et al. 2010



Vaccine Protection Investigation

Objections

1. "Canonical"/expected anatomy is not all that we need to understand the genetics of development and disease

   we need malformations and disease to link to genetic variation

   ... more on this in a minute

2. ?????

   (this one is up to you to figure out..)

Usability!

Usability of biomedical ontologies for effective biomedical data science

To realize the potential of this vision, we must develop tools

that both

increase the quantity, quality, and variety of ontologically annotated data..

and

help biomedical researchers leverage the power of this data to generate novel insights.

Annotating data

- Extraction of terms from text
  - Textpresso, BioPortal annotator, etc.

- Support for structured experimental data models
  - Investigation-study-assay (ISA) tools
- Ontomaton - spreadsheet-integrated ontology annotation

- Manual curation - accurate, but expensive

Data Interpretation: The Monarch Initiative

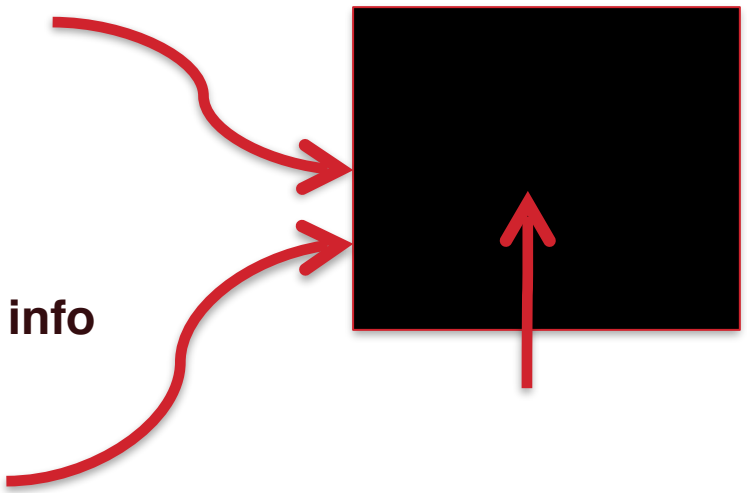
OHSU, UCSD, LBNL, U. Pitt

Phenotypes

P_1
P_2
P_3
…

Genotype info

G_1
G_2
G_3
G_4
…



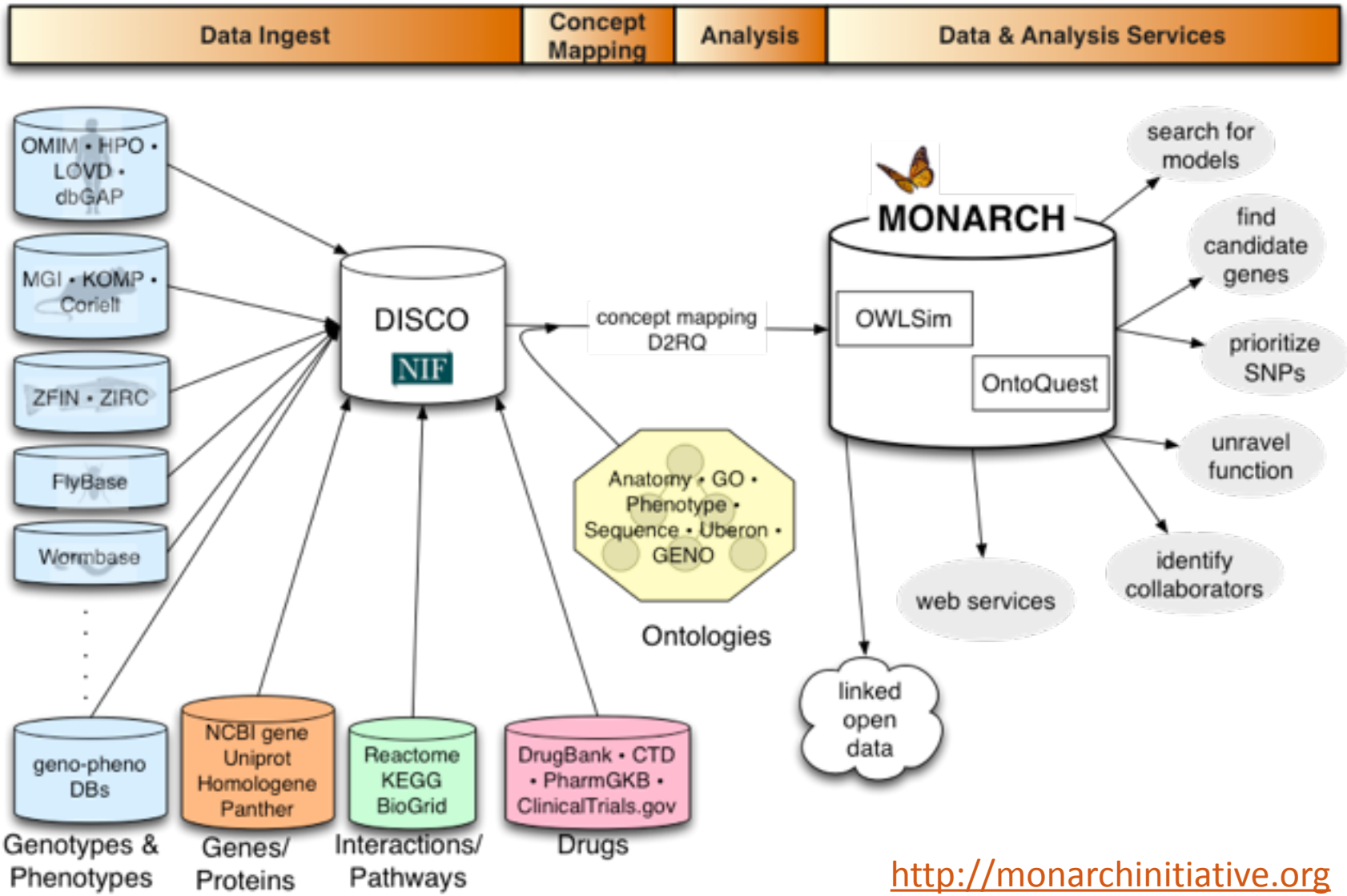
Pathogenicity, frequency, protein interactions, gene expression, gene networks, epigenomics, metabolomics….

Prioritized Candidates, Models, functional validation

M_1
M_2
M_3
M_4
…

- **What's in the box?**
- **How are candidates identified?**
- **How do they compare?**

The Monarch System



http://monarchinitiative.org

Monarch phenotype data

Species	Source	Unique	Disease/
Mouse	*MGI*	*53,573*	*406,618*
Zebrafish	*ZFIN*	*14,703*	*75,698*
C. elegans	*Wormbase*	*116,106*	*411,154*
Fruit fly	*Flybase*	*98,596*	*265,329*
Human	*OMIM*	*26,372*	*27,798*
Human	*Orphanet*	*2,872*	*5,095*
Human	*ClinVar*	*62,437*	*178,424*

Also in the system: Rat; IMPC; GO annotations; Coriell cell lines; OMIA; MPD; Yeast; CTD; GWAS; Panther, Homologene orthologs; BioGrid interactions; Drugbank; AutDB; Allen Brain …157 sources to date

Coming soon: Animal QTLs for pig, cattle, chicken, sheep, trout, dog, horse

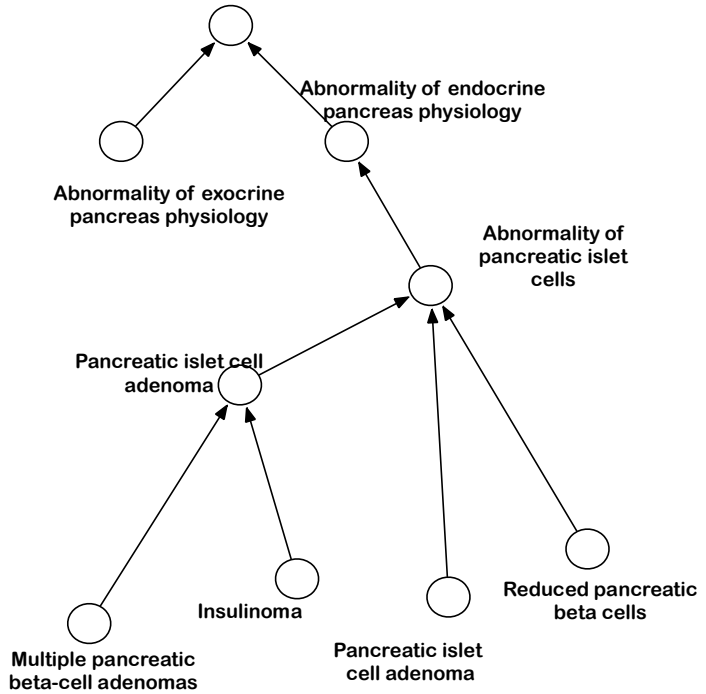
Interpretation challenges

How to make sense of the OWLSim calculation results?

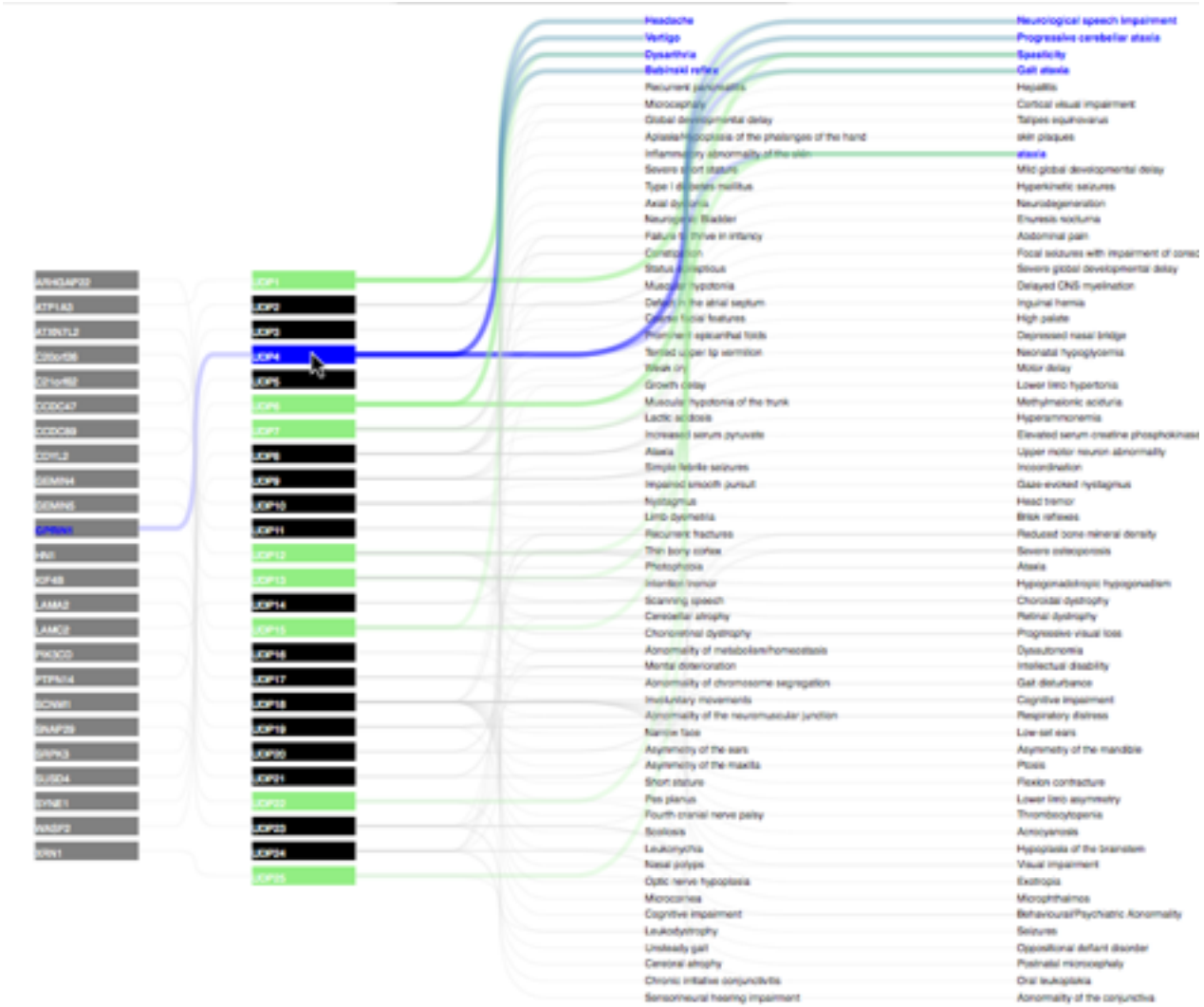
Why are phenotypic profiles similar?

When are small local differences important or not?

Multiple phenotypes, multiple models

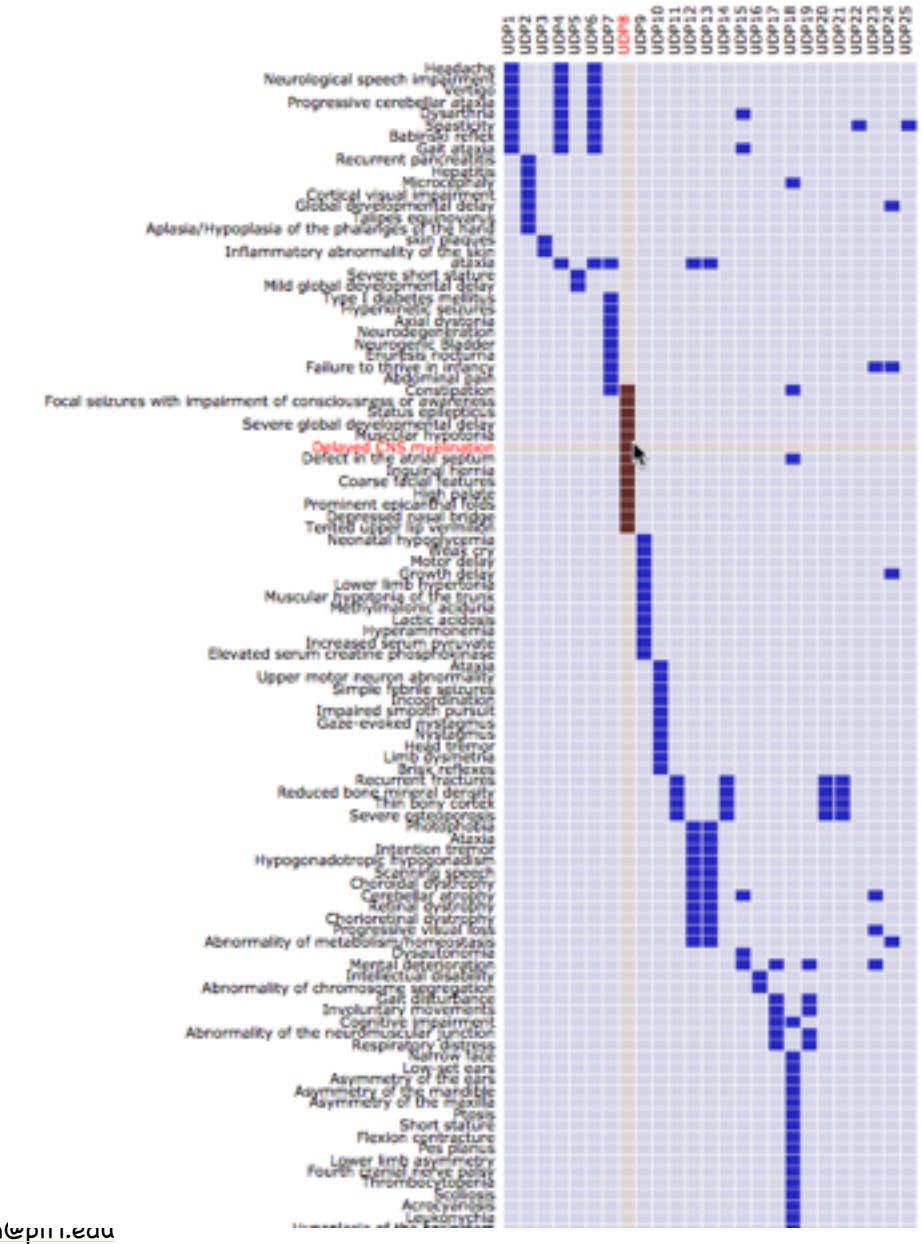


NIH Intramural Undiagnosed Disease Program - comparison of phenotype profiles



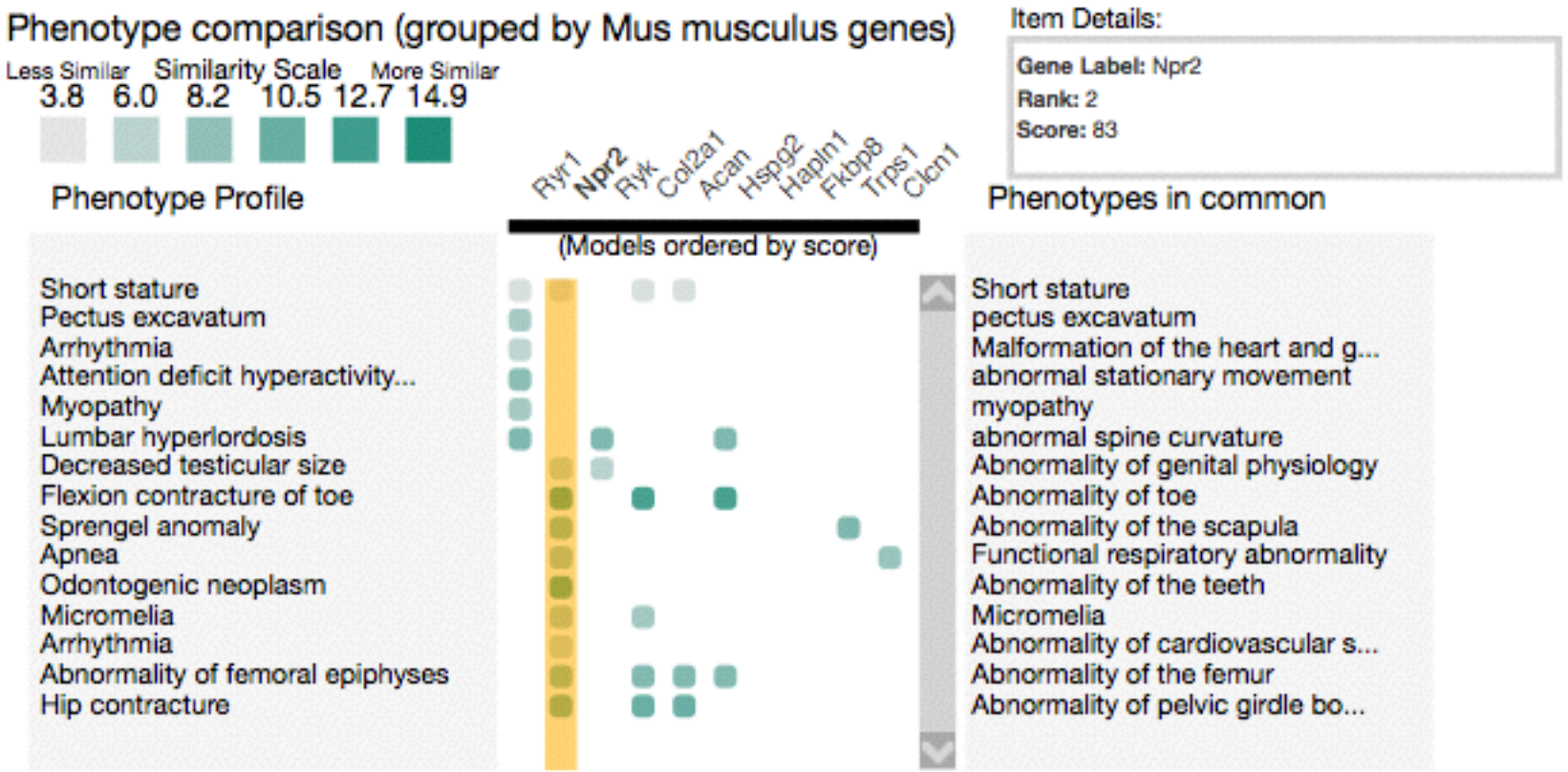
Harry Hochheiser, harryh@pitt.edu

UDP phenotype profile comparison…



Monarch Model Viewer

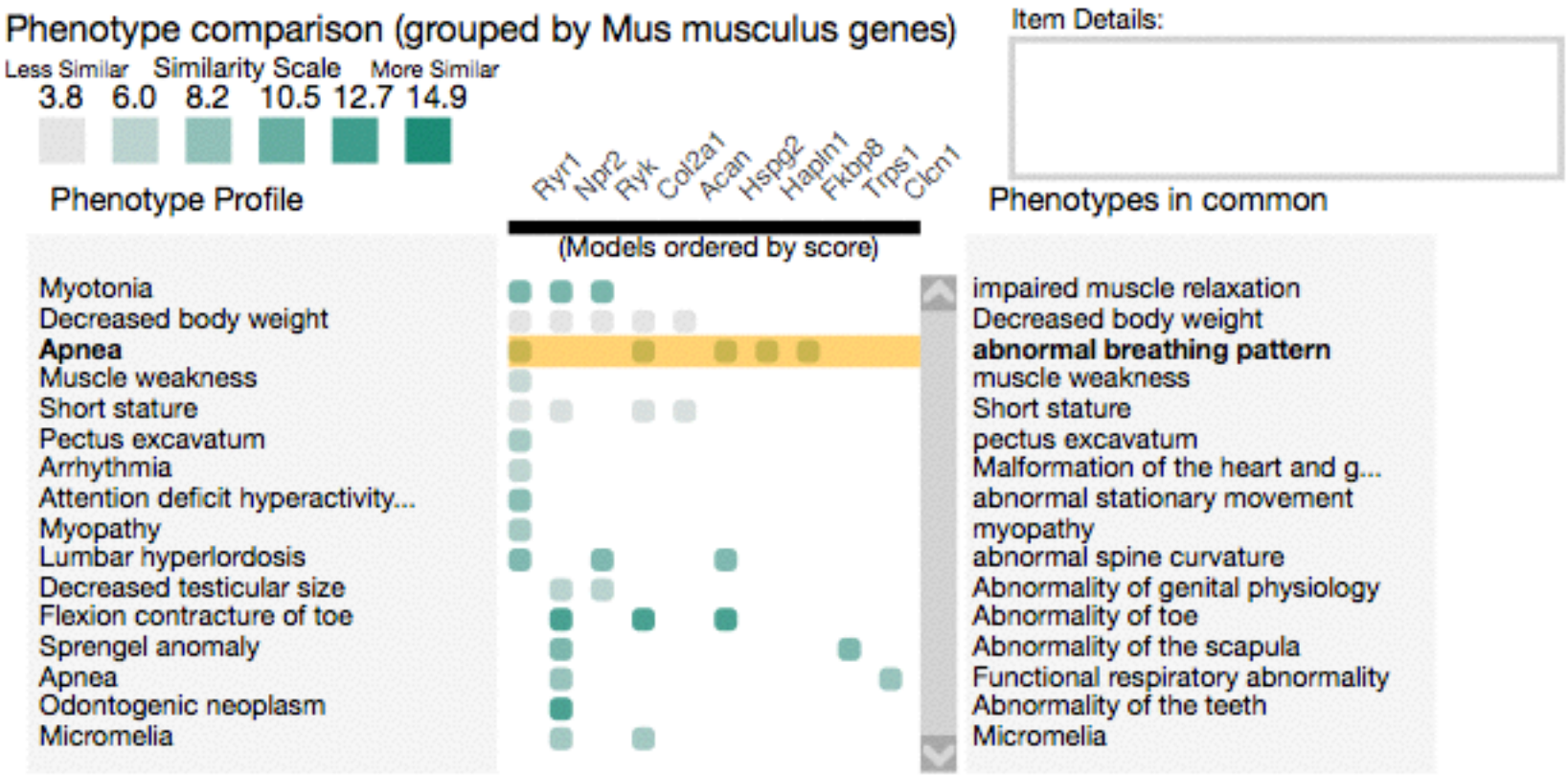
Schwartz-Jampel Syndrome (OMIM 255800)
a genetic disorder associated with the HSPG2 gene



HSPG2 is not the most similar model…….

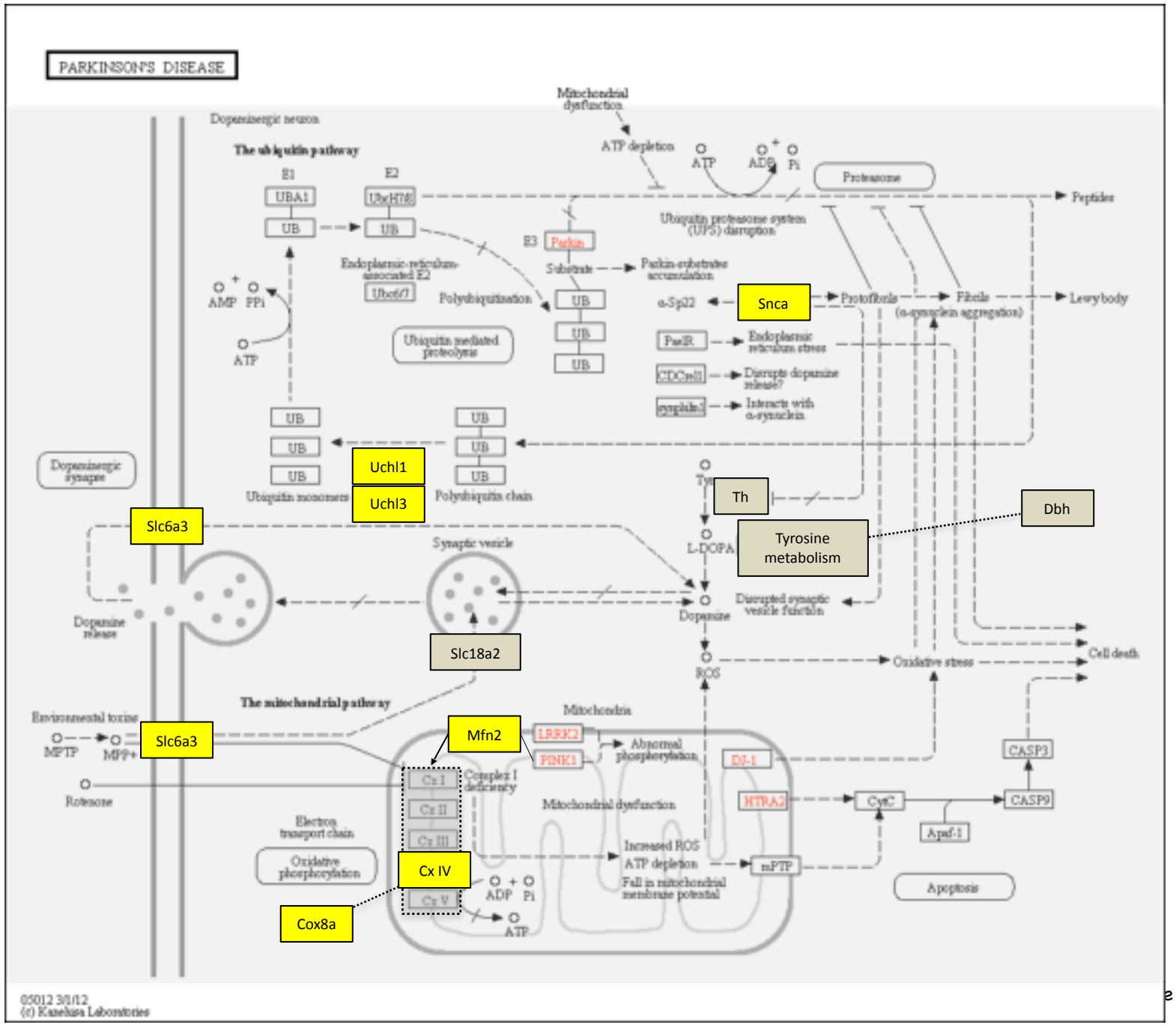
Harry Hochheiser, harryh@pitt.edu

Monarch Model Viewer



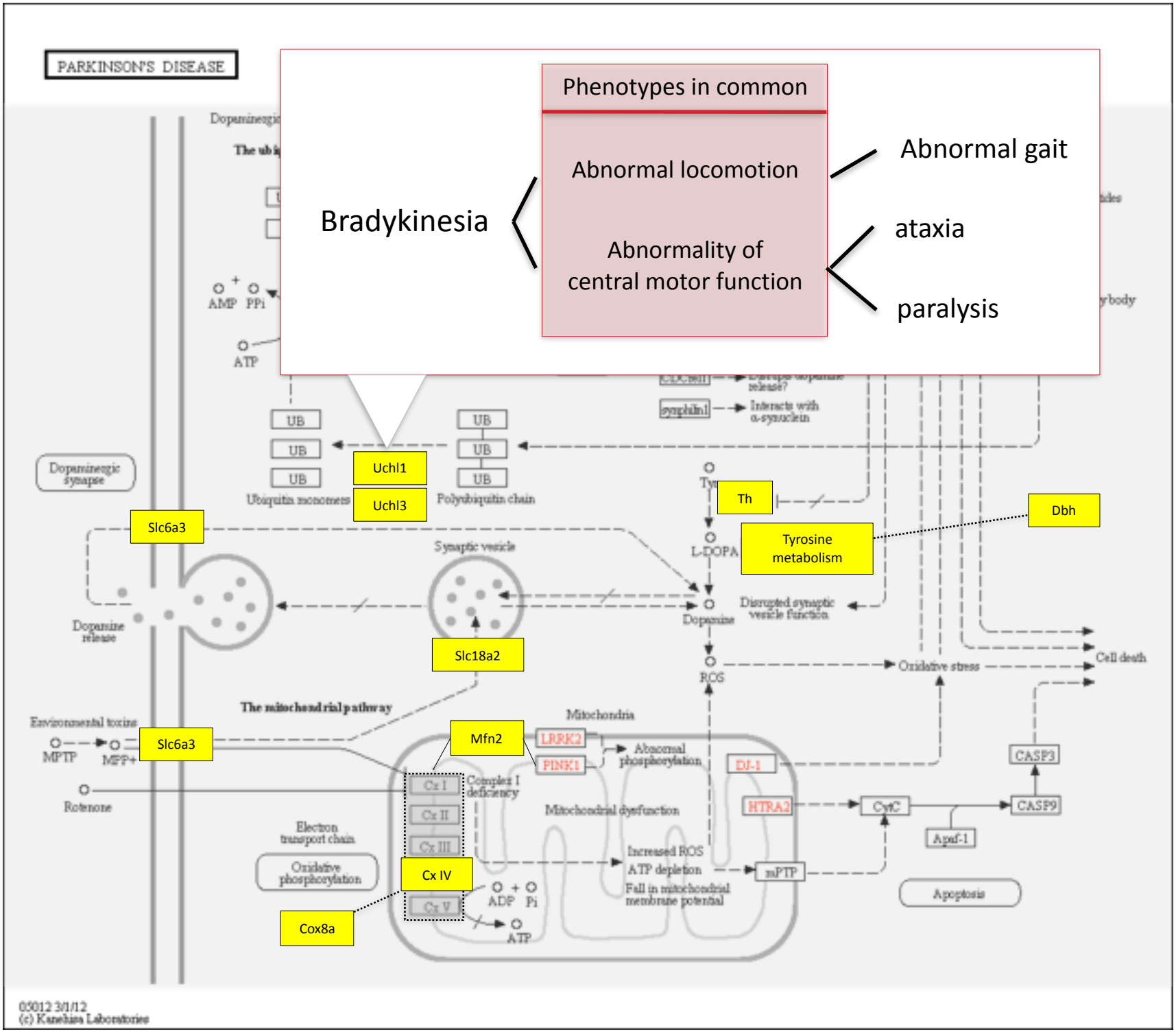
Harry Hochheiser, harryh@pitt.edu

Integrating Biological Pathways



Late-onset Parkinson's Phenotypes
(subset)

- Bradykinesia
- Lewy bodies
- Dysphagia
- Depression



Late-onset Parkinson's Phenotypes

(subset)

- **Bradykinesia**
- Lewy bodies
- Dysphagia
- Depression

What I did not talk about…

Metadata for analysis

- Open provenance model (OPM), PROV ontology, workflows, etc..
  Garijo & Gil 2011

Generalizing

High-quality metadata annotations can increase clarity and utility at all stages in pipeline.

May not be easy, but value is high.

Undoubtedly examples exist in other domains.

Can good practices and clear descriptions help generalize your data pipeline work?

Acknowledgments

FaceBase:

U. Pittsburgh: Mike Becich, Becky Boes, Chuck Borromeo, Lance Kennelty, Annette Krag-Jensen, Tom Maher, Johnson Paul, Linda Schmandt, Shiyi Shen, Bill Shirey, Cristy Spino, Mike Stefanko, Justin Stickel, Mary Marazita

U. Iowa: Jeff Murray

OCDM :James Brinkley; Jose Leonardo Mejino; Landon Detwiler; Ravensara Travillian; Melissa Clarkson; Timothy Cox; Carrie Heike; Michael Cunningham; Linda Shapiro

Support: NIH Grants U01 DE020057, 3U01DE020050-03S1

Monarch:

Pittsburgh: Chuck Borromeo, Jeremy Espino

OHSU: Melissa Haendel, Nicole Vasilevky, Matt Brush

NIH-UDP: Murat Sincan, David Adams, Neal Boerkel, Amanda Links, Bill Gahl

LBNL: Nicole Washington, Suzanna Lewis, Chris Mungall

+ colleagues at Sanger, Charite , Toronto, and JAX

Support: NIH Office of Director: 1R24OD011883, NIH-UDP: HHSN2682013