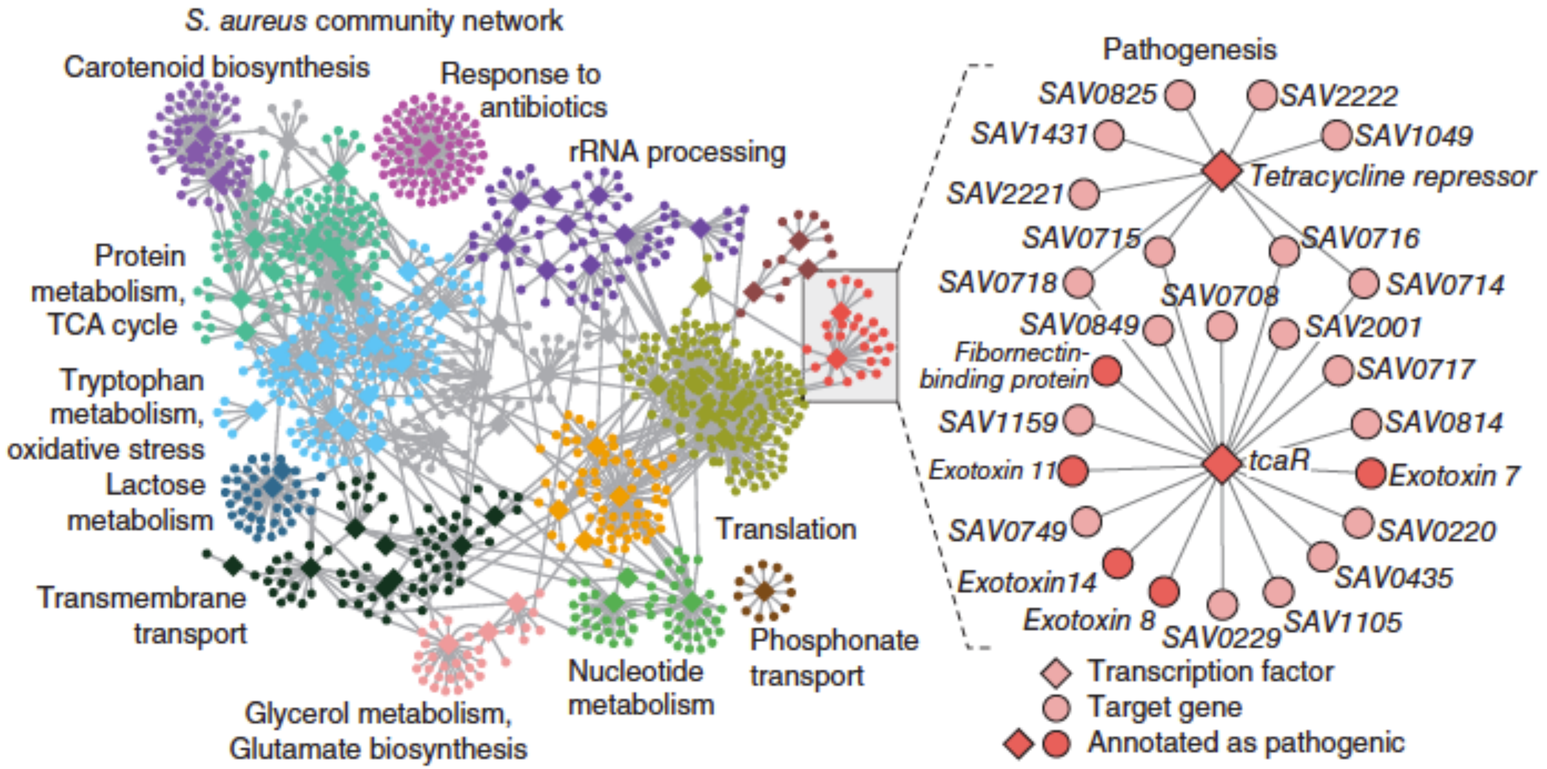
Ontologies & Functional Enrichment Analysis



Ontologies for Structuring Data

- Much data in the biological domain is unstructured
- Of the data that is structured most is bespoke
- There are structured data standards for biology

- The quantity and heterogeneity of data is so great that manual curation into structured data objects is not feasible
- There is an unmet need to retro-fit existing data into formal data structures
- There are surprisingly few contemporary systems that do this as standard for emerging data

- Emergent properties of biological systems and their underlying mechanisms are the result of the integration of multiple biological signatures
- No single type of data can capture this
- Data integration is required that can cope with scale and complexity

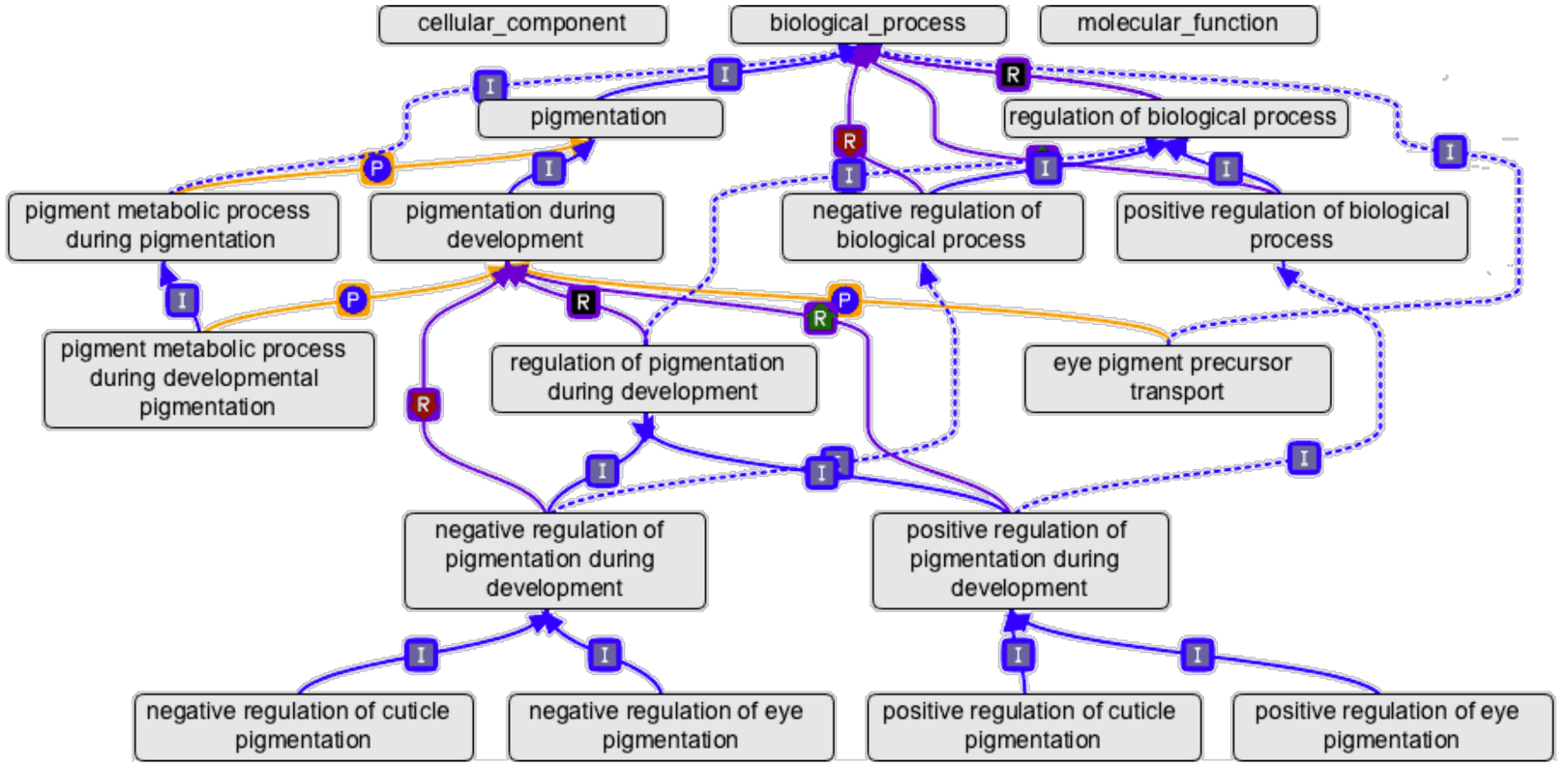
ian.simpson@ed.ac.uk

Structure of an Ontology

- nodes are “Terms”
- edges are “Relations”

low specificity roots

parent

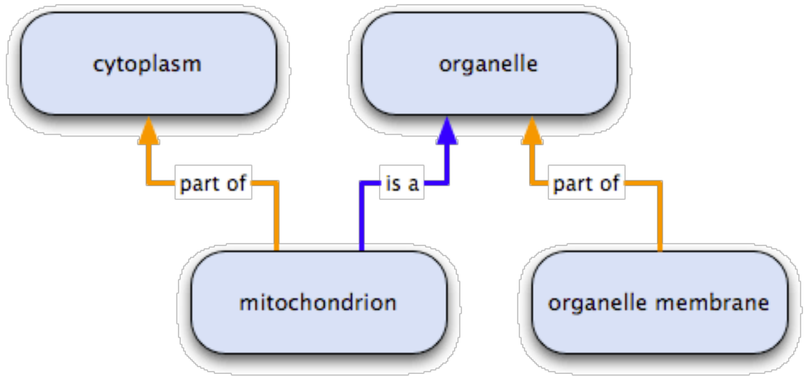


child

high specificity leaves

ian.simpson@ed.ac.uk

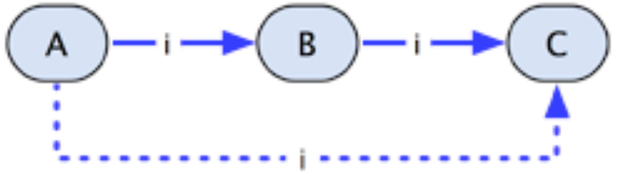
Relationships in Ontologies



parent

child

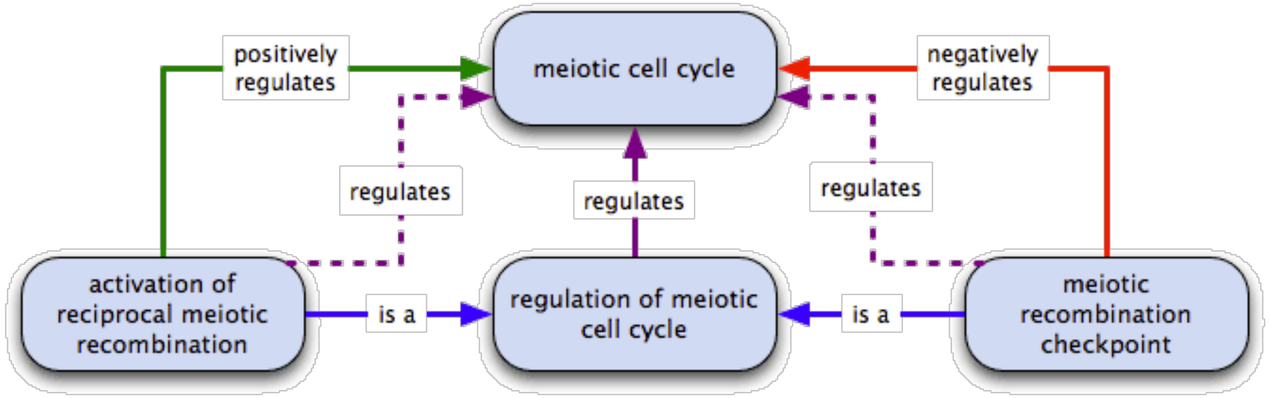
hierarchically structured



reasoning is "*transitive*"

- branching part of
- capable of
- capable of part of
- causally downstream of
- causally downstream of or within
- causally related to
- causally upstream of
- causally upstream of or within
- cell expresses
- child nucleus of
- child nucleus of in hermaphrodite
- child nucleus of in male
- coincident with
- colocalizes_with
- commensually interacts with
- composed primarily of
- concretizes
- conduit for

hundreds of relation types

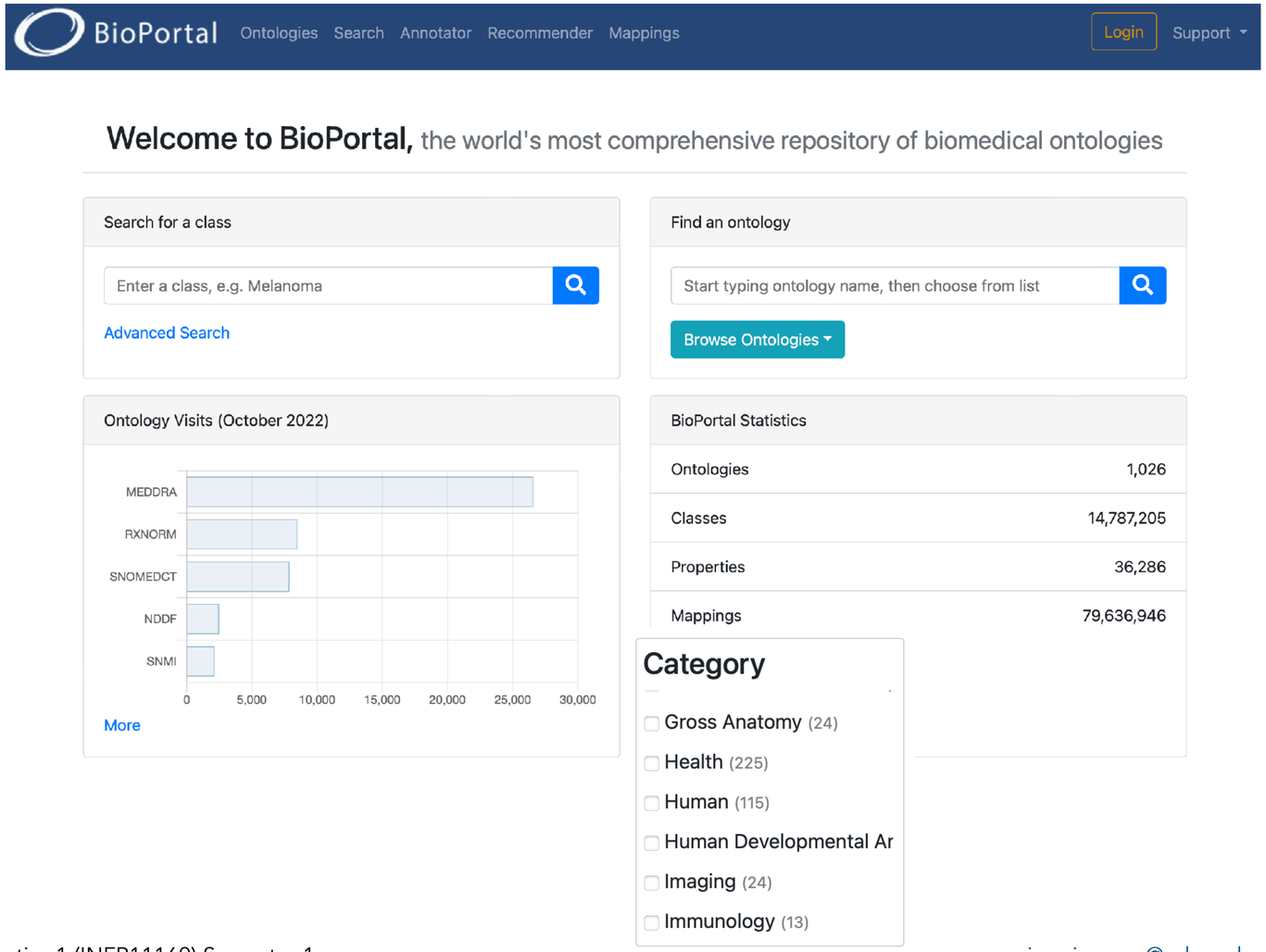


generic AND specific relations exist

ian.simpson@ed.ac.uk

Tour of Bio-Ontologies

http://bioportal.bioontology.org/

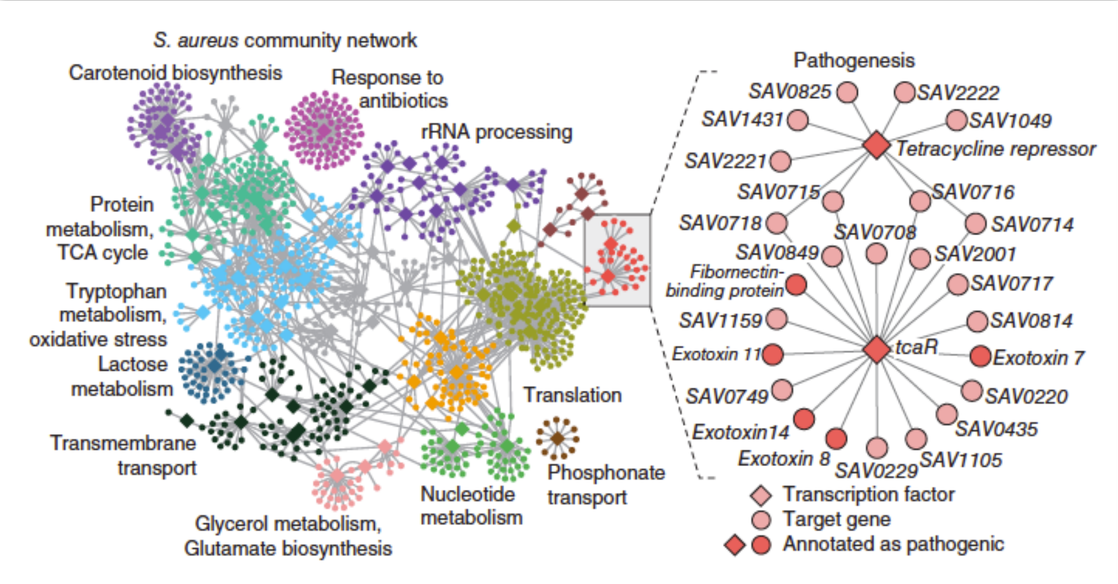


- 1076 ontologies (October 2023)
- OBO, OWL and UMLS format
- links out to original data and direct ontology download
- also available via REST
- among the most popular
  - Disease (DO)
  - Phenotype (HPO)
  - MeSH
  - OMIM
  - Taxonomical
  - Gene function (GO)
  - Anatomical

ian.simpson@ed.ac.uk

The Gene Ontology (GO)

http://geneontology.org/



Gene Ontologies

- (CC) cellular compartment
- (BP) biological process
- (MF) molecular function

Accession: GO:1901632
Name: regulation of synaptic vesicle membrane organisation
Ontology: biological_process
Synonyms
regulation of synaptic vesicle membrane organisation
regulation of synaptic vesicle membrane organisation and biogenesis
regulation of SLMV biogenesis
Definition: Any process that modulates the frequency, rate or extent of synaptic vesicle membrane organisation. Source: GOC:TermGenie, PubMed:22426000

ian.simpson@ed.ac.uk

Gene Ontology - Release Statistics

Ontology

Property	Value
Valid terms	43329 (Δ = -6)
Obsoleted terms	4024 (Δ = 16)
Merged terms	2438 (Δ = 8)
Biological process terms	28050
Molecular function terms	11241
Cellular component terms	4038

Annotations

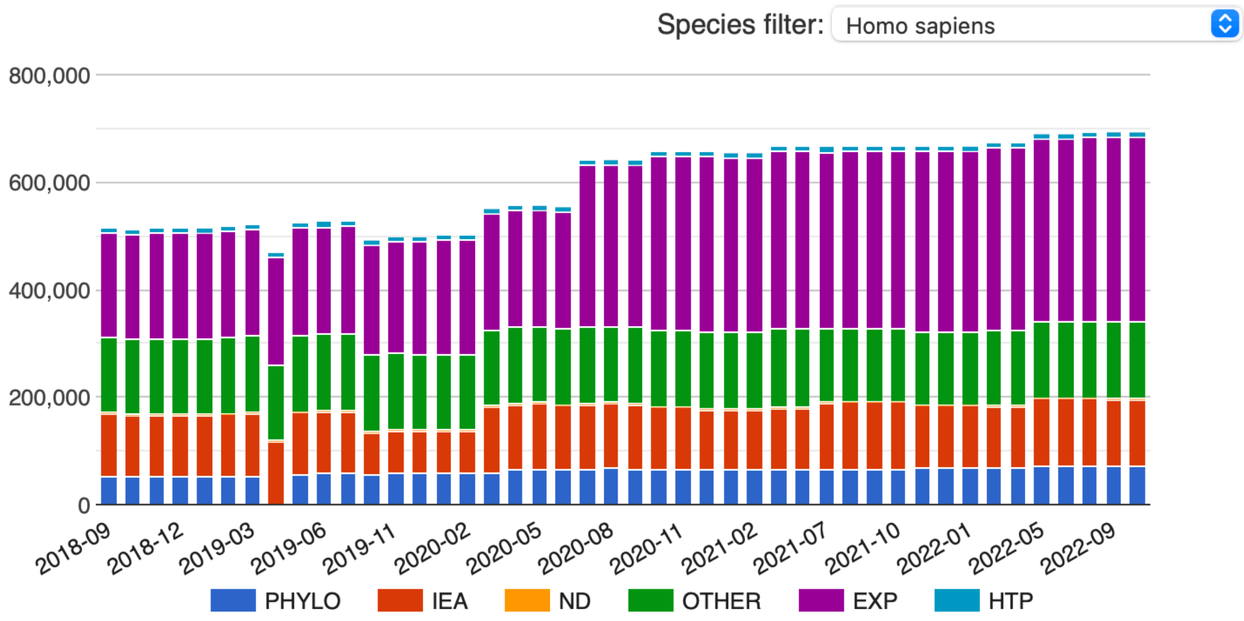
Property	Value
Number of annotations	7,694,564
Annotations for biological process	2,876,072
Annotations for molecular function	2,437,493
Annotations for cellular component	2,380,999
Annotations for evidence PHYLO	3,993,997
Annotations for evidence IEA	1,583,208
Annotations for evidence OTHER	870,221
Annotations for evidence EXP	935,834
Annotations for evidence ND	252,265
Annotations for evidence HTP	59,039
Number of annotated scientific publications	172,648

Gene products and species

Property	Value
Annotated gene products	1,503,740
Annotated species	5,257
Annotated species with over 1,000 annotations	185

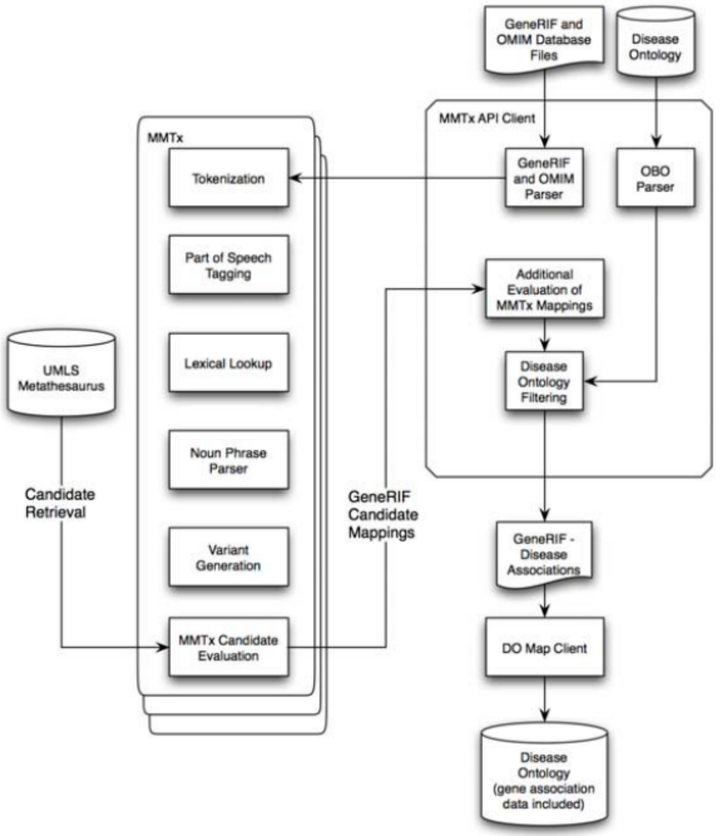
IC	Inferred by curator
IDA	Inferred from direct assay
IEA	Inferred from electronic annotation
IEP	Inferred from expression pattern
IGI	Inferred from genetic interaction
IMP	Inferred from mutant phenotype
IPI	Inferred from physical interaction
ISS	Inferred from sequence or structural similarity
NAS	Non-traceable author statement
ND	No biological data
TAS	Traceable author statement

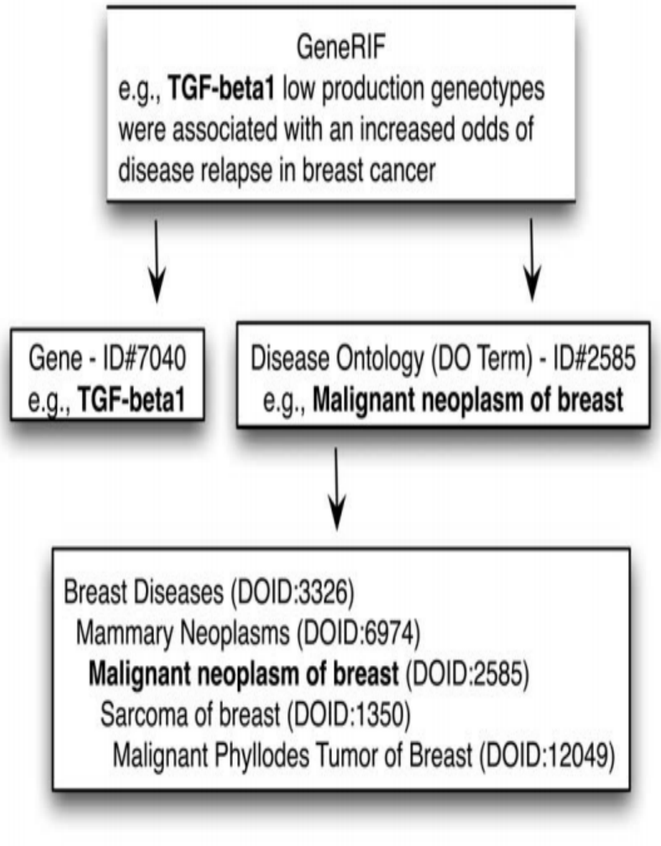
Number of annotations by evidence



ian.simpson@ed.ac.uk

The Human Disease Ontology (DOID)





18354 terms

Organised by disease category

- disease of anatomical entity
- disease of behaviour
- biological process
- environmental origin
- infectious agent and syndromes

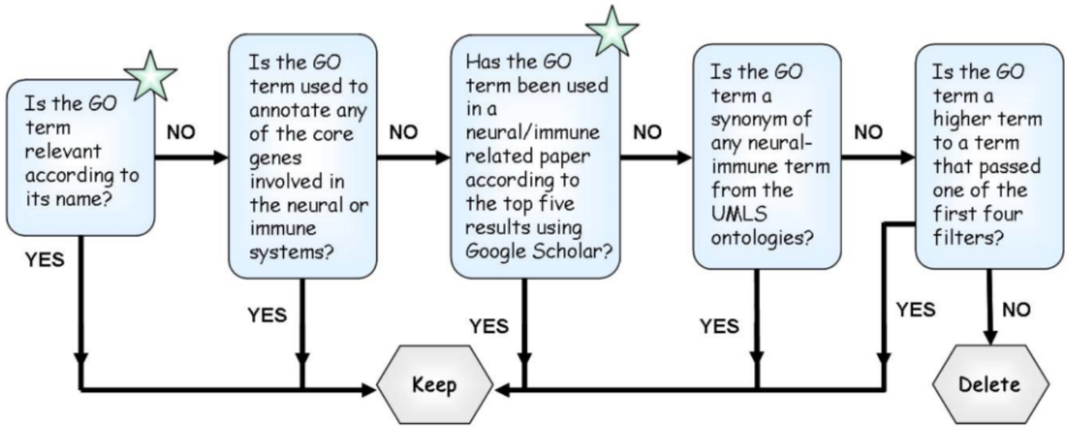
- Build evidence-based mappings phenotype and disease
- Define and validate mappings to other disease relevant
- vocabularies including
  - UMLS
  - MeSH
  - ICD
  - NCI thesaurus
  - OMIM
  - SNOMED

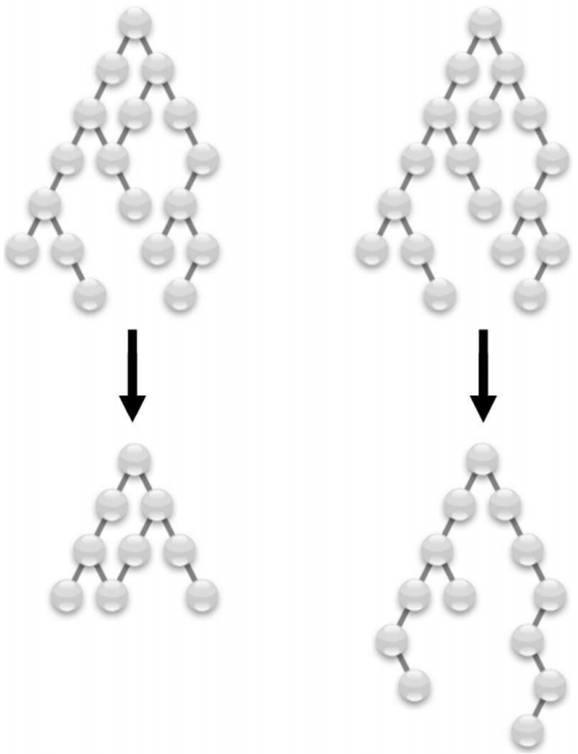
ian.simpson@ed.ac.uk

Clipped Ontologies and Slims

NiGO - the neural/immune Gene Ontology

- subset of GO directed for neurological and immunological systems
- improves statistical scores given to relevant terms
- retrieves functionally relevant terms that did not pass statistical cutoffs with full GO or the slim subset.





Geifman N, Monsonego A, Rubin E. The Neural/Immune Gene Ontology: clipping the Gene Ontology for neurological and immunological systems. BMC Bioinformatics. 2010 Sep 12;11:458. doi: 10.1186/1471-2105-11-458

ian.simpson@ed.ac.uk

Over Representation Analysis (ORA) - The Fisher Exact Test Statistic

- consideration of foreground and background lists is crucial
  - genome ?
  - array content ?
  - mappable elements ?

contingency table	foreground	background	Row Total
genes with Term in list	a 10	b 50	60
genes with Term not in list	c 190	d 13950	14140
Column Total	200	14000	n 14200

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ \ b!\ \ c!\ \ d!\ \ n!}$$

probability given by the hypergeometric distribution

in order to calculate significance we need to

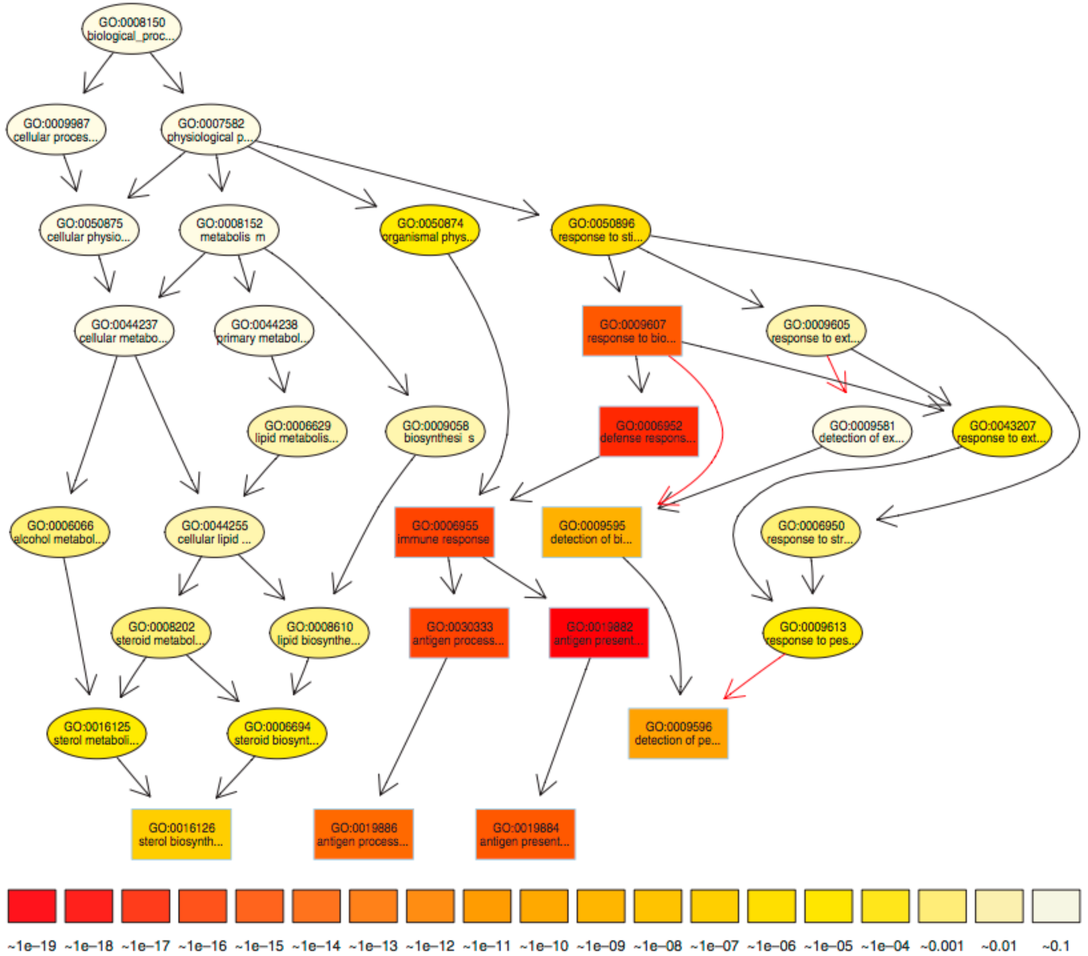
- one-tailed test
  sum all probabilities as extreme or more extreme
- two-tailed test
  sum as above but also all that are as extreme or more extreme in both directions

```
fisher.test(rbind(c(10,50),c(190,13950)),alternative = 'two.sided')
```

p-value = 1.003e-08 odds ratio = 14.68

ian.simpson@ed.ac.uk

Terms in Ontologies are not Statistically Independent



ian.simpson@ed.ac.uk

ORA Using Topology Aware Analyses

Algorithm 1 elim

```
markedGenes ← Ø:   nodeSig ← Ø
get the DAG levels list   dagLevels
for i from max(dagLevels) to 1
  for u in nodes(dagLevels, i)
    genes[u] ← genes[u]\markedGenes[u]
    nodeSig[u] ← FisherTest(genes[u], sigGenes)
    if nodeSig[u] ≤ thresheold then
      for x in upperInducedGraph(u)
        markedGenes[x] ← markedGenes[x] ∪ genes[u]
  end
end
return   nodeSig
```

Problem

- In classical analysis terms are inherited by parents from children
- conditional dependence between nodes invalidates test
- end-up with a "chain" of probability

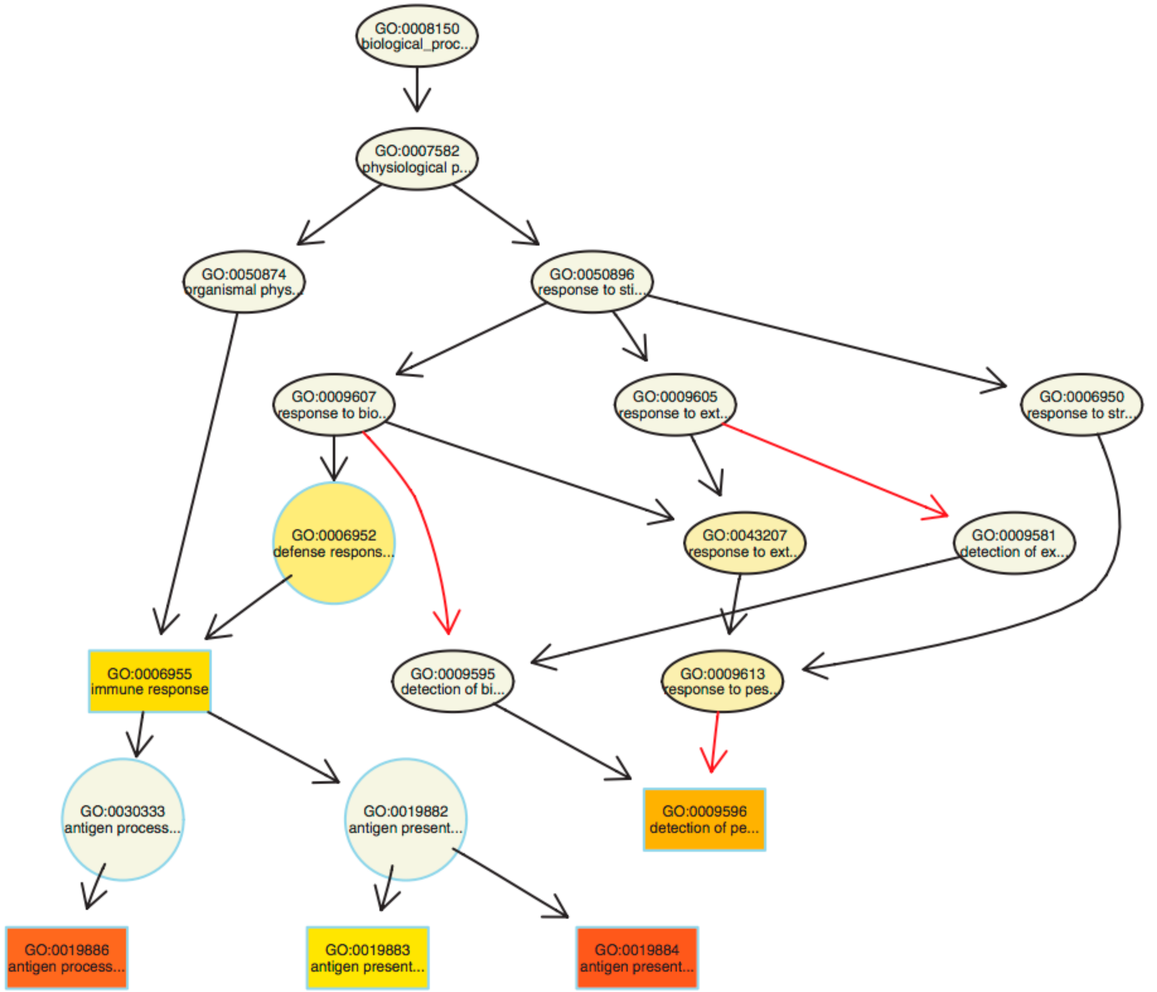
Solution

- de-correlate the GO-graph by removing genes associated with significant nodes from parent term - *elimination*

	GO ID	Term	Observed	Expected	Annotated	*p*-values				
						classic	elim	weight.log	weight.ratio	all.M
1	GO:0019882	antigen presentation	22	2.287	41	1.6e−17	0.2821	1.6e−17	1.6e−17	1.8e−13
2	GO:0006952	defense response	107	47.143	845	8.3e−17	0.0065	1.4e−09	1.1e−06	5.4e−09
3	GO:0030333	antigen processing	20	2.12	38	7.8e−16	1.0000	7.8e−16	7.8e−16	4.7e−12
4	GO:0006955	immune response	98	43.293	776	2.7e−15	5.9e−06	3.0e−05	0.024	3.3e−07
5	GO:0019884	antigen presentation, exogenou...	14	1.004	18	5.9e−15	5.9e−15	2.2e−10	0.054	1.4e−10
6	GO:0009607	response to biotic stimulus	112	53.949	967	9.5e−15	0.6873	1.0e−05	0.404	1.3e−05
7	GO:0019886	antigen processing, exogenous ...	14	1.116	20	6.8e−14	6.8e−14	1.5e−11	0.054	2.5e−10
8	GO:0009596	detection of pest, pathogen or...	9	0.725	13	2.9e−09	2.9e−09	2.9e−09	2.9e−09	2.9e−09
9	GO:0009595	detection of biotic stimulus	9	0.893	16	3.9e−08	1.0000	1.0e−05	0.107	0.00046
10	GO:0016126	sterol biosynthesis	9	1.395	25	4.5e−06	0.0015	4.5e−06	4.5e−06	1.9e−05

ian.simpson@ed.ac.uk

ORA Using Topology Aware Analyses

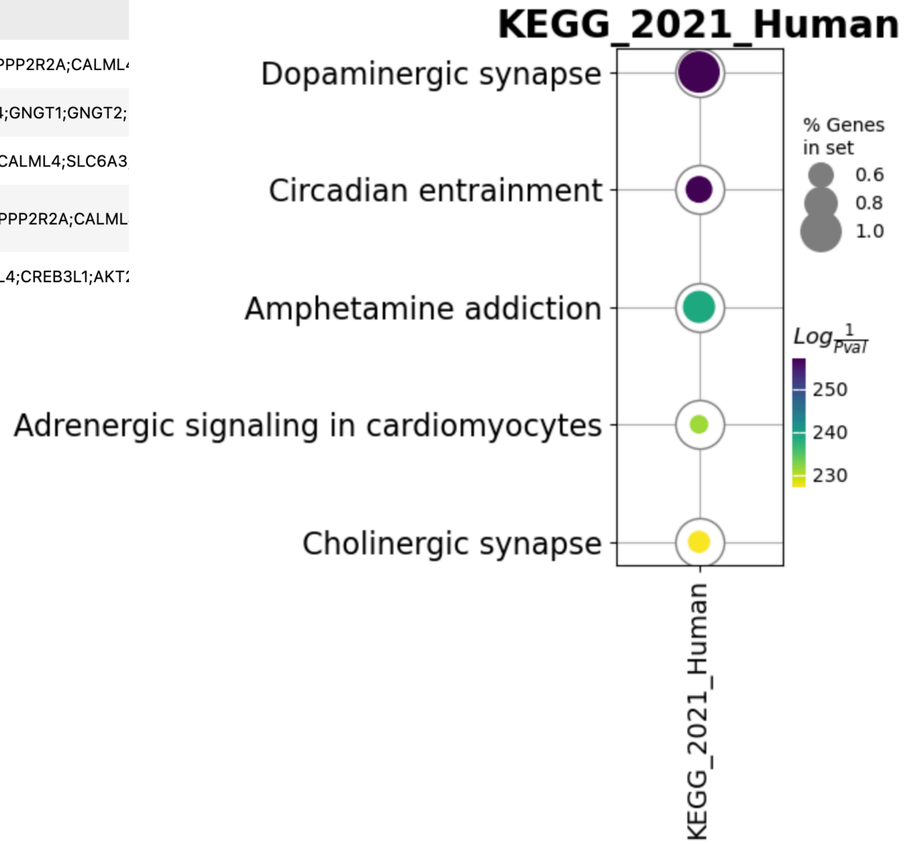


ian.simpson@ed.ac.uk

Example 1 - Over-Representation Analysis

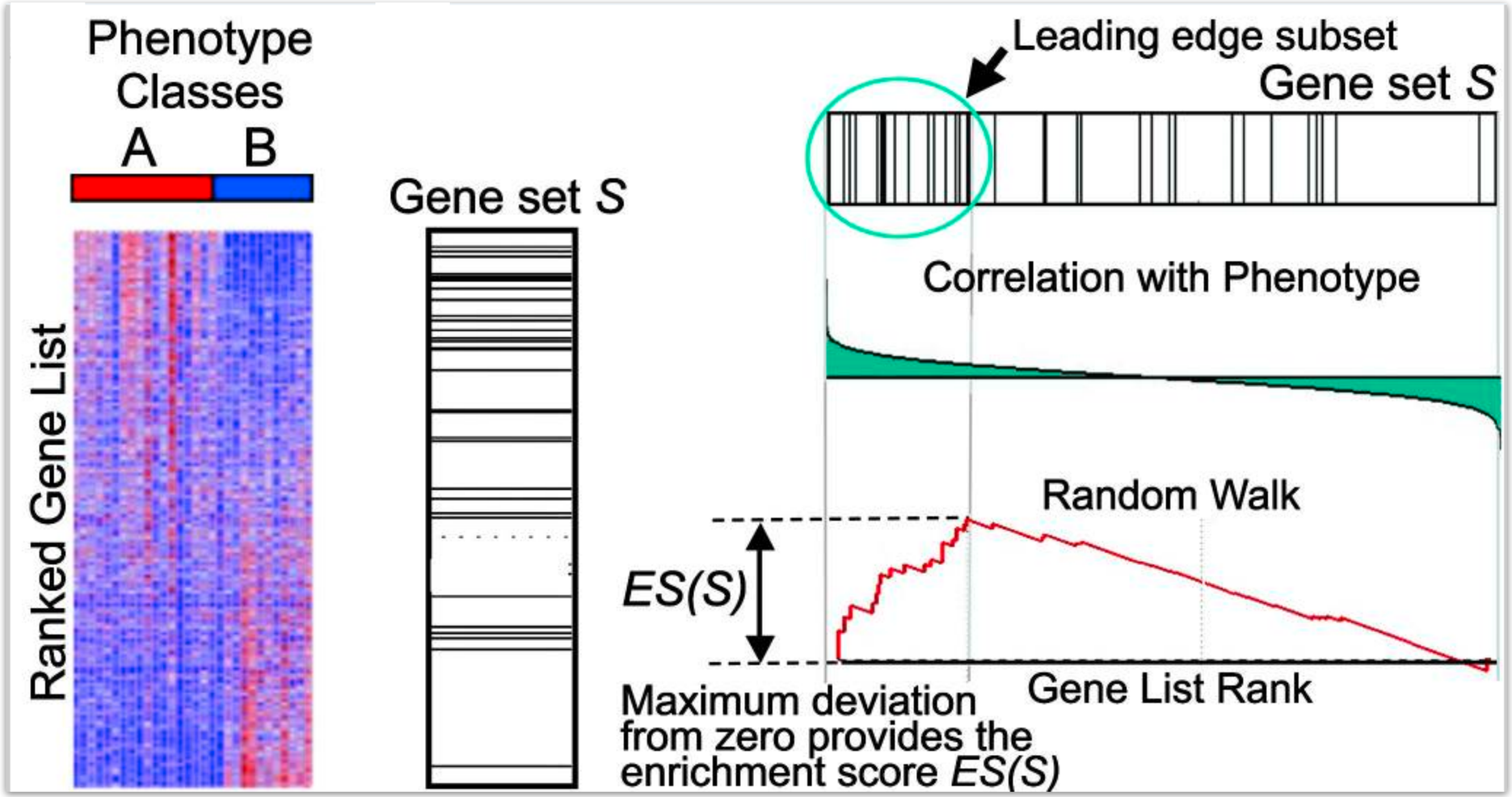
Taking a “toy” example using the dop_genes.txt file from last week and running a straight ORA against KEGG

	Gene_set	Term	Overlap	P-value	Adjusted P-value	Old P-value	Old Adjusted P-value	Odds Ratio	Combined Score	
0	KEGG_2021_Human	Dopaminergic synapse	132/132	0.000000e+00	0.000000e+00	0	0	2.622576e+06	inf	ATF2;CALML5;CALML6;CALML3;PPP2R2A;CALML4
1	KEGG_2021_Human	Circadian entrainment	61/97	2.715484e-114	2.634020e-112	0	0	4.732989e+02	1.237656e+05	CALML5;CALML6;CALML3;CALML4;GNGT1;GNGT2;
2	KEGG_2021_Human	Amphetamine addiction	53/69	2.598422e-106	1.680313e-104	0	0	8.324019e+02	2.023728e+05	ATF2;CALML5;CALML6;CALML3;CALML4;SLC6A3
3	KEGG_2021_Human	Adrenergic signaling in cardiomyocytes	63/150	7.726235e-103	3.747224e-101	0	0	2.075967e+02	4.881048e+04	ATF2;CALML5;CALML6;CALML3;PPP2R2A;CALML
4	KEGG_2021_Human	Cholinergic synapse	58/113	7.593595e-101	2.946315e-99	0	0	2.823474e+02	6.509062e+04	GNGT1;GNGT2;CREB3L3;CREB3L4;CREB3L1;AKT3



ian.simpson@ed.ac.uk

Gene Set Enrichment Analysis (GSEA)



Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50

ian.simpson@ed.ac.uk

Gene Set Enrichment Analysis (GSEA)

Input data

- Expression data set D with N genes and k samples
- Ranking procedure to produce Gene List L against a phenotype or profile of interest C.
- Independently derived Gene Set S of NH genes (e.g., a pathway, a cytogenetic band, or a GO category).

Enrichment Score (ES)

- Rank order the N genes in D to form L = {g1,...,gN} according to the correlation of their expression profiles with C.
- Evaluate the fraction of genes in S ("hits") weighted by their correlation and the fraction of genes not in S ("misses") present up to a given position i in L.
- The ES is the maximum deviation from zero of $P_{hit} - P_{miss}$.

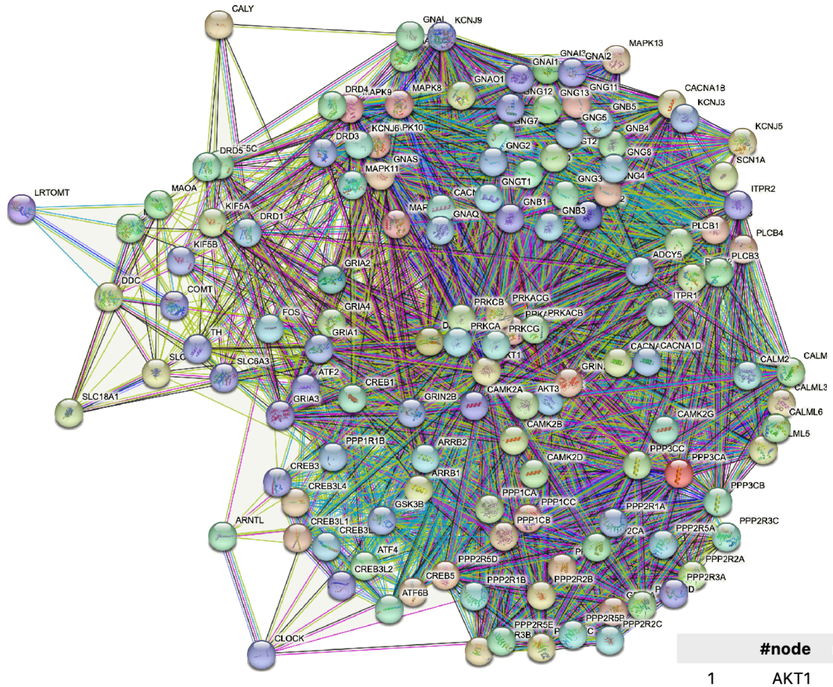
Estimating significance

- Randomly assign the original phenotype labels to samples, reorder genes, and re-compute ES(S).
- Repeat for 1,000 permutations, and create a histogram of the corresponding enrichment scores ES_{Null}.
- Estimate nominal P value for S from ES_{Null} by using the positive or negative portion of the distribution corresponding to the sign of the observed ES(S).

ian.simpson@ed.ac.uk

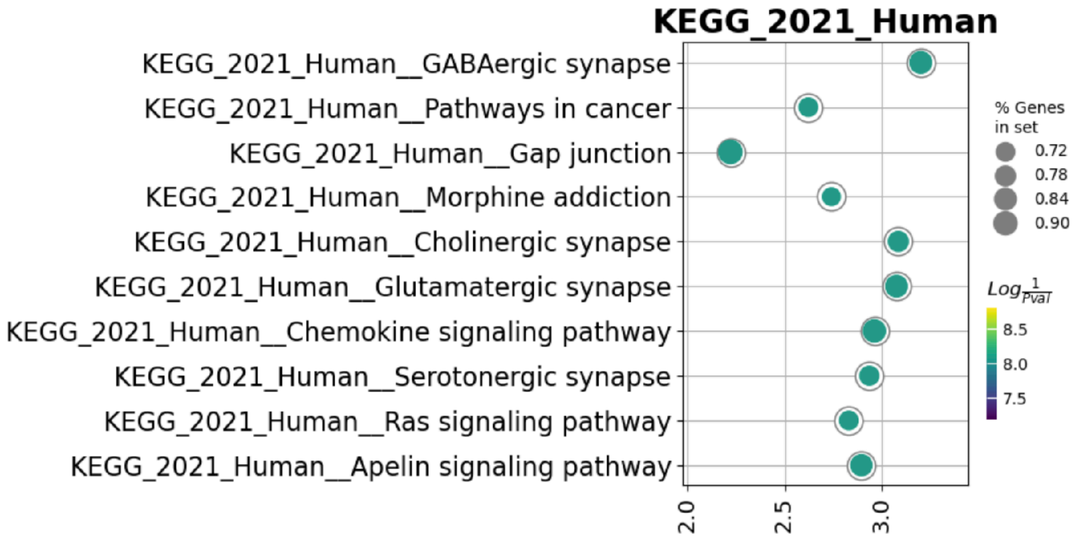
Example 2 - Gene Set Enrichment Analysis

Taking a “toy” example using the dop_genes.txt file from last week and using protein-protein interaction network node-degree to rank the genes - then we do a prerank GSEA analysis



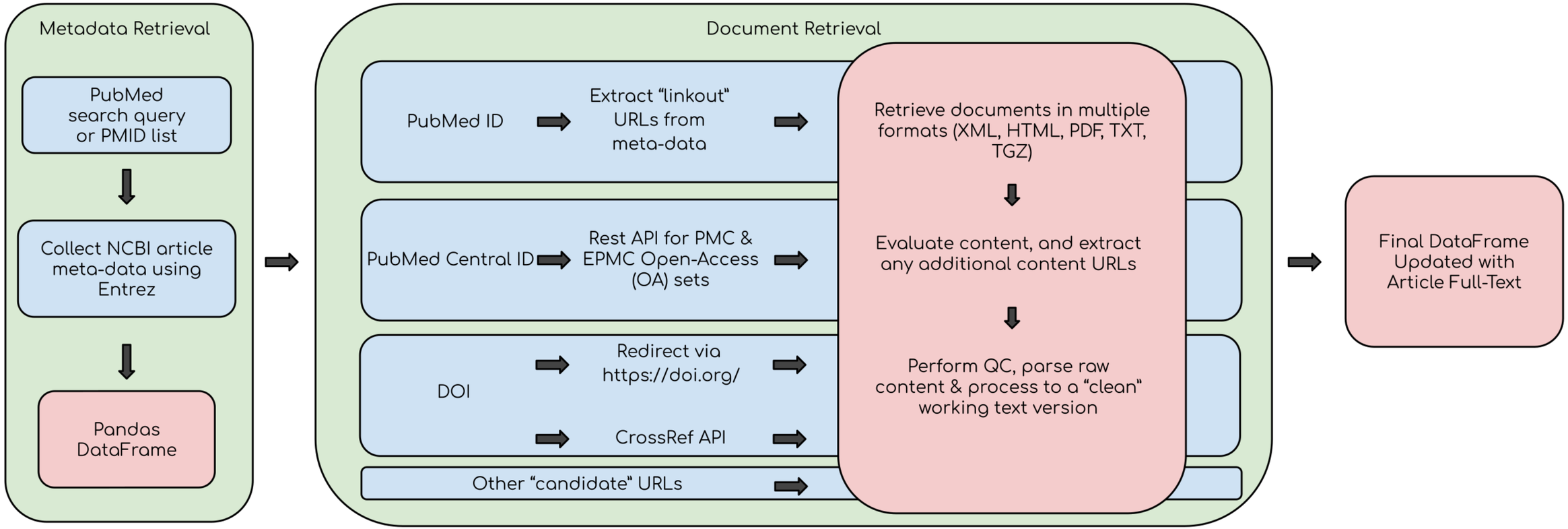
	#node	node_degree
1	AKT1	82
12	CACNA1C	75
13	CACNA1D	68
35	DRD2	66
120	PRKACB	64
123	PRKCB	63
121	PRKACG	63
122	PRKCA	62
119	PRKACA	62
72	ITPR1	61

	Name	Term	ES	NES	NOM p-val	FDR q-val	FWER p-val	Tag %	Gene %	Lead_genes
0	prerank	KEGG_2021_Human__GABAergic synapse	0.843099	3.197717	0.0	0.0	0.0	28/33	24.62%	CACNA1C;CACNA1D;PRKACB;PRKCB;PRKACG;PRKCA;PRKA...
1	prerank	KEGG_2021_Human__Cholinergic synapse	0.77439	3.083102	0.0	0.0	0.0	45/58	40.00%	AKT1;CACNA1C;CACNA1D;PRKACB;PRKCB;PRKACG;PRKCA...
2	prerank	KEGG_2021_Human__Glutamatergic synapse	0.77133	3.073702	0.0	0.0	0.0	42/50	40.00%	CACNA1C;CACNA1D;PRKACB;PRKCB;PRKACG;PRKCA;PRKA...
3	prerank	KEGG_2021_Human__Chemokine signaling pathway	0.769659	2.96008	0.0	0.0	0.0	32/37	35.38%	AKT1;PRKACB;PRKCB;PRKACG;PRKACA;GNB1;GNB3;GNG3...
4	prerank	KEGG_2021_Human__Serotonergic synapse	0.741352	2.9343	0.0	0.0	0.0	37/50	35.38%	CACNA1C;CACNA1D;PRKACB;PRKCB;PRKACG;PRKCA;PRKA...



ian.simpson@ed.ac.uk

Example 3 - Mining for Biomedical Concepts (*Cadmus & MetaMap)*



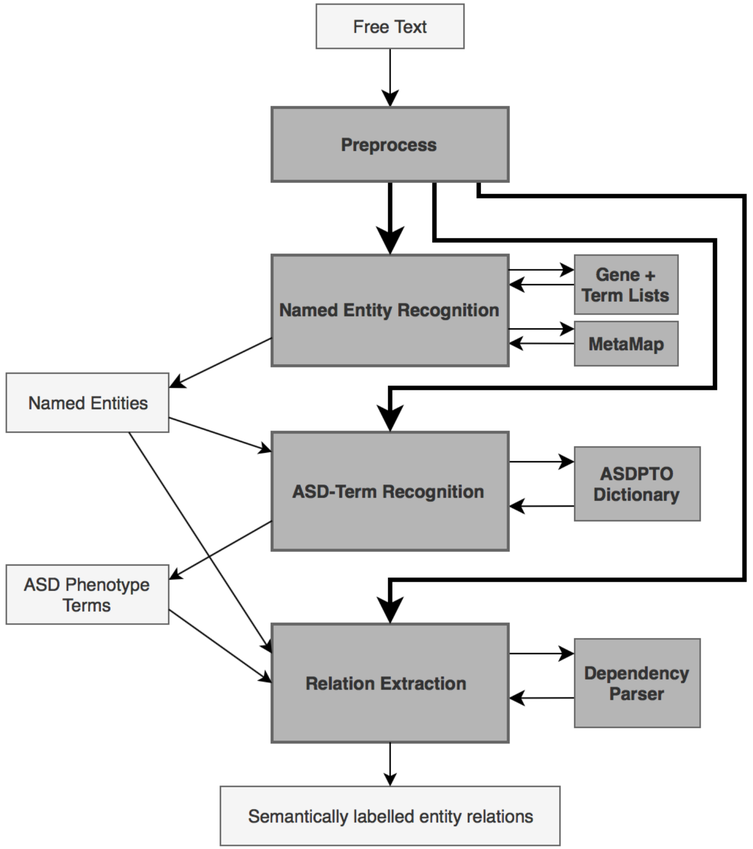
Literature corpora for Autism Spectrum Disorder
OA and UoE licensed for Data Science
Full-Text Retrieval HTML, XML, & PDF
68,329 target papers 04/02/22 (**54,591** - 86% retrieved)

https://github.com/biomedicalinformaticsgroup/cadmus

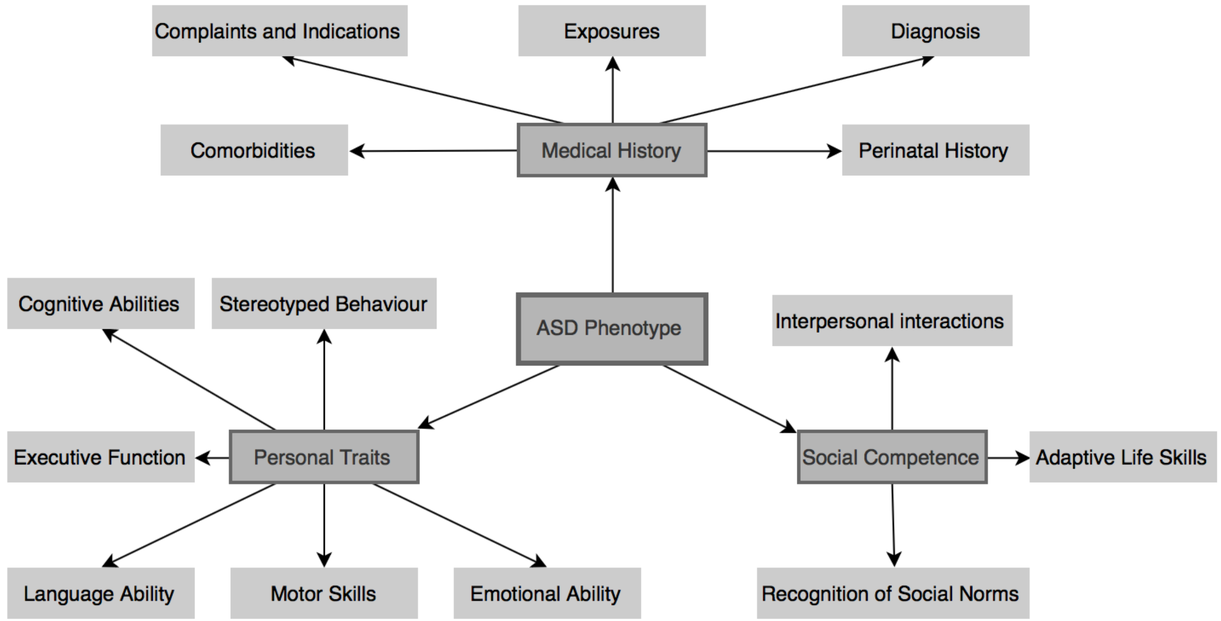
ian.simpson@ed.ac.uk

Developing a Phenotype Knowledgeable for ASDs

Customised ASD NER pipelines



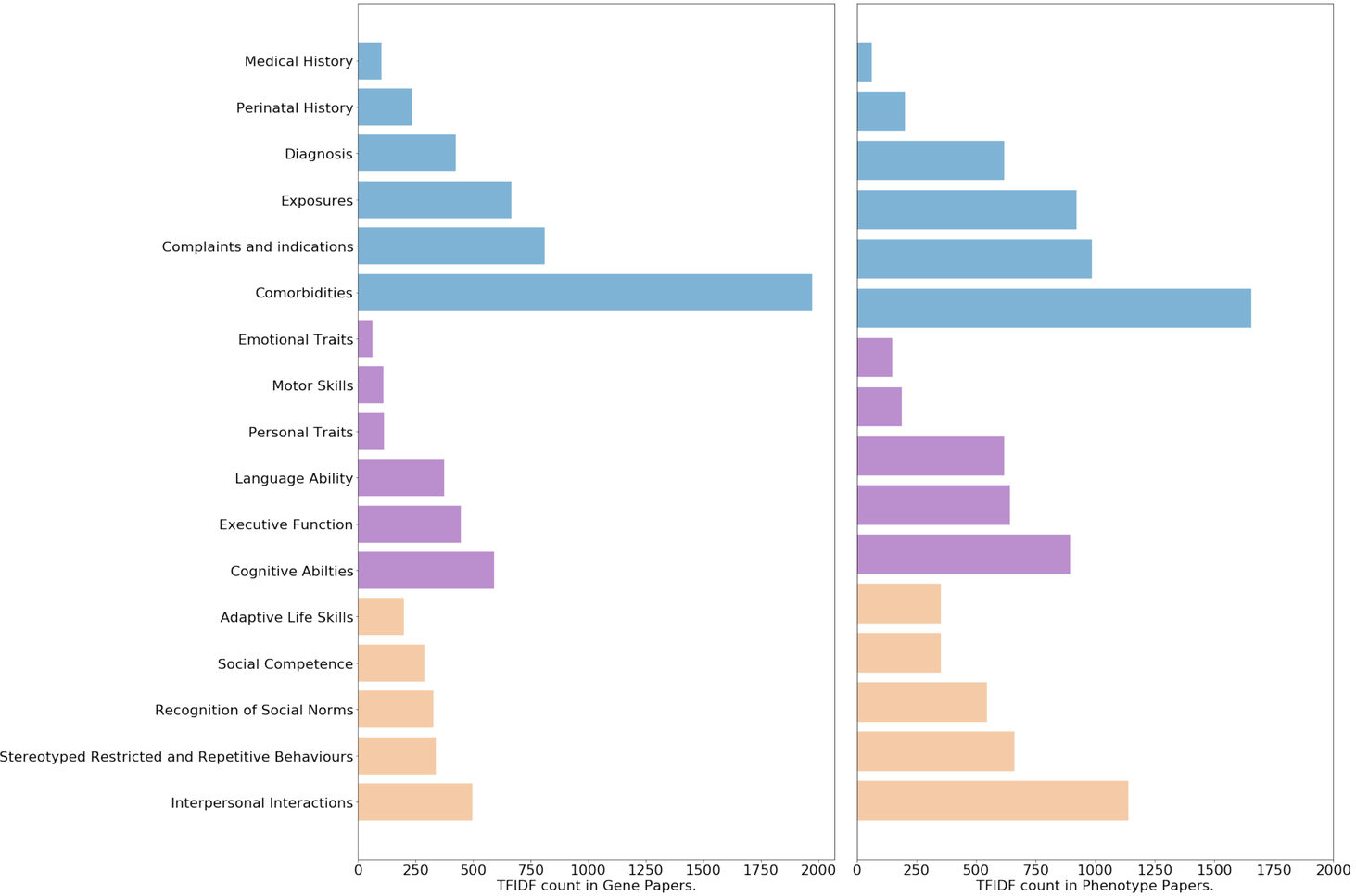
ASDPT Ontology



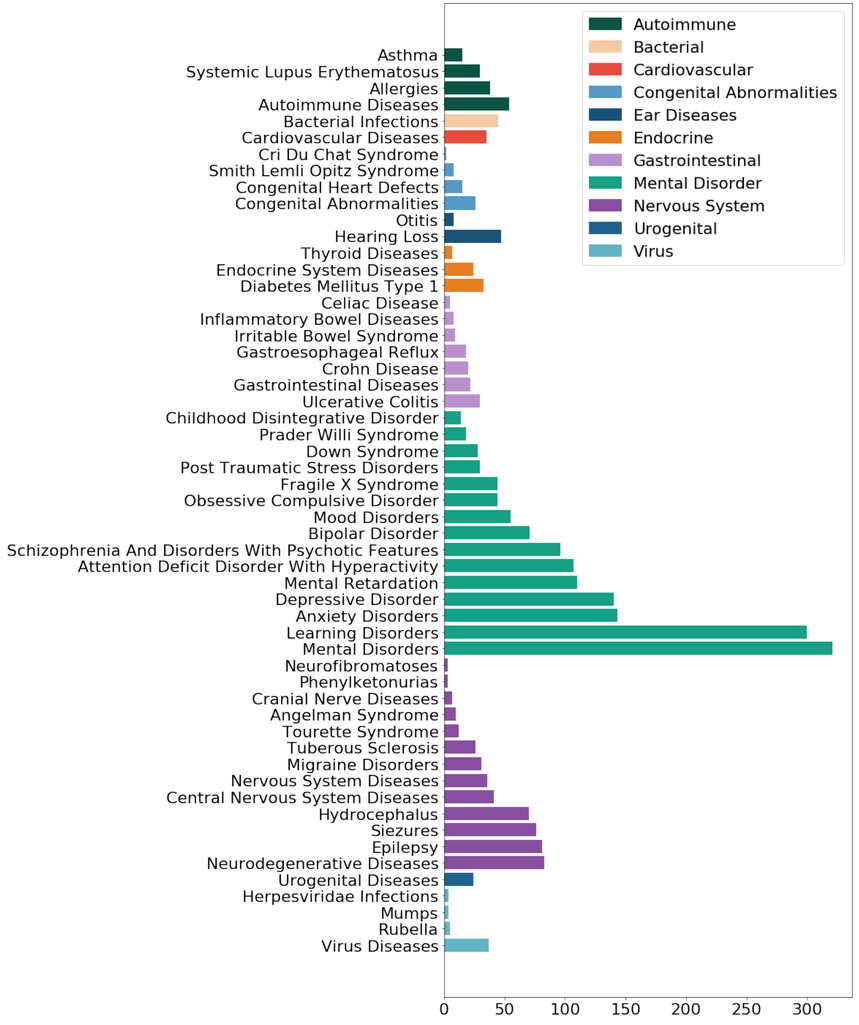
ian.simpson@ed.ac.uk

ASD Ontology Representation & Co-morbidities Using NCBI-MetaMap

ASDPT representation

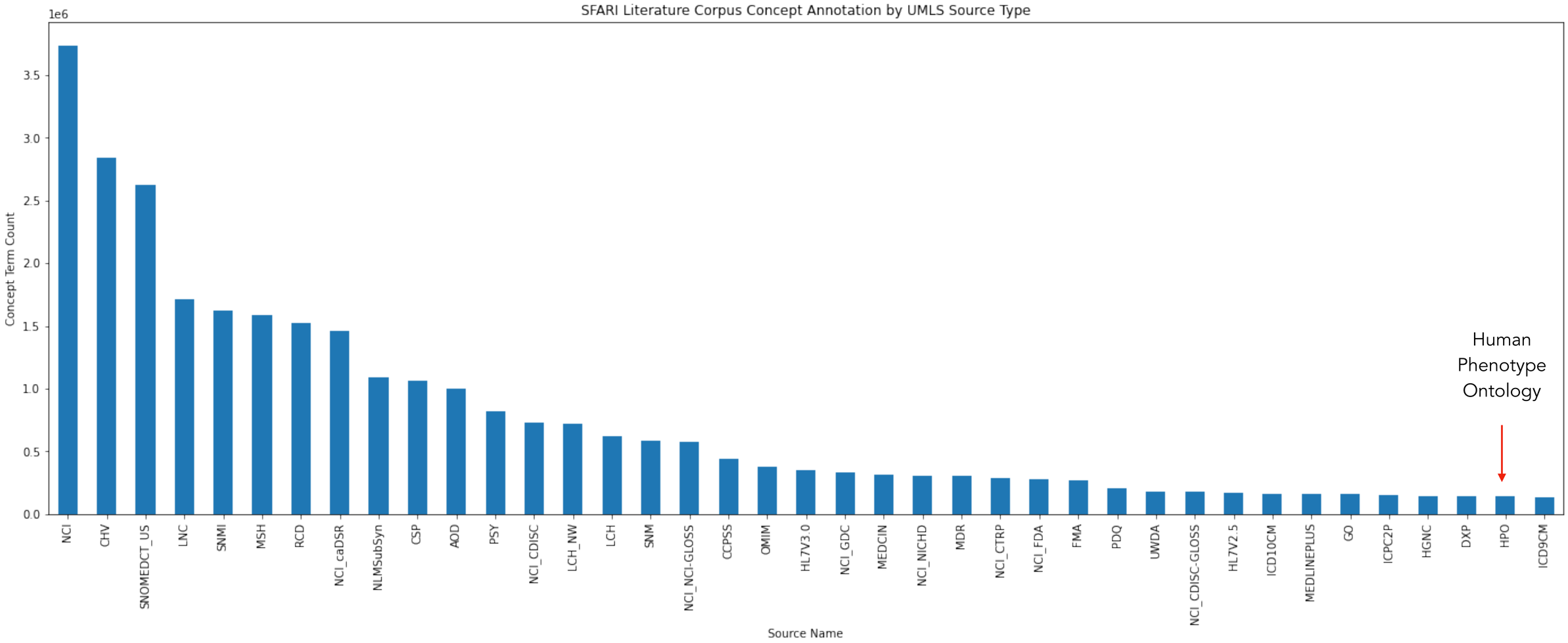


Co-morbidities



ian.simpson@ed.ac.uk

Biomedical Concept Extraction with Full-Text SFARI (Autism) Publication Corpus



*with Natalia Volfovsky (SFARI, NY)

ian.simpson@ed.ac.uk

Example 4 - Literature Driven Phenotypic Gene Models

Database, 2022, 1–10
DOI: https://doi.org/10.1093/database/baac038
Original article

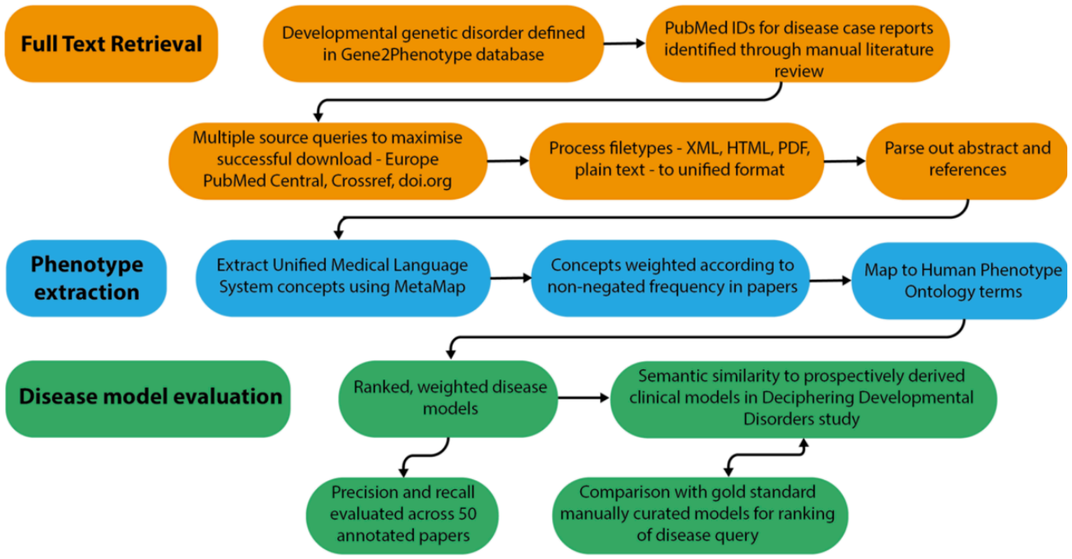
Creation and evaluation of full-text literature-derived, feature-weighted disease models of genetically determined developmental disorders

T.M. Yates[1,2], A. Lain[3], J. Campbell[1,4], D.R. FitzPatrick[1,2,4] and T.I. Simpson[3,4,*]

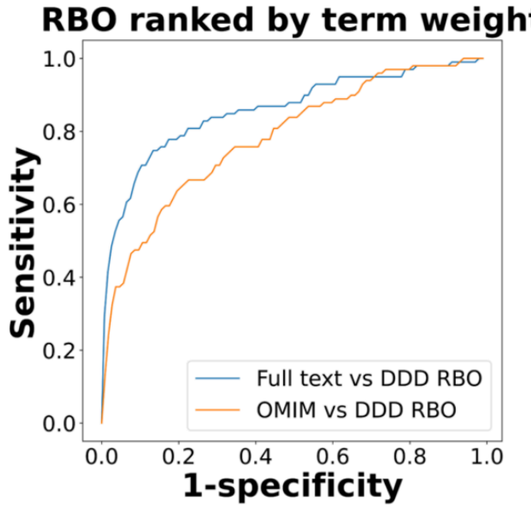
[1]MRC Human Genetics Unit, Western General Hospital, Institute of Genetics and Cancer, The University of Edinburgh, Crewe Road South, Edinburgh EH4 2XU, UK
[2]Transforming Genetic Medicine Initiative, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[3]Institute for Adaptive and Neural Computation, Informatics Forum, The University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK
[4]Simons Initiative for the Developing Brain, The University of Edinburgh, Hugh Robson Building, George Square, Edinburgh EH8 9XF, UK

*Corresponding author: Tel: +44 (0)131 6515637; Email: ian.simpson@ed.ac.uk

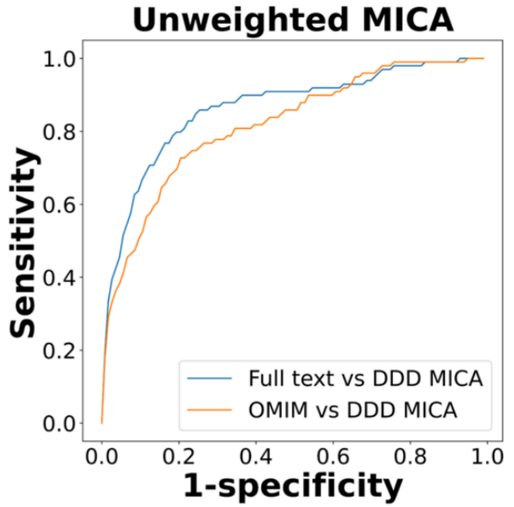
Citation details: Yates, T., Lain, A., Campbell, J. *et al.* Creation and evaluation of full-text literature-derived, feature-weighted disease models of genetically determined developmental disorders. *Database* (2022) Vol. 2022: article ID baac038; DOI: https://doi.org/10.1093/database/baac038



A



B

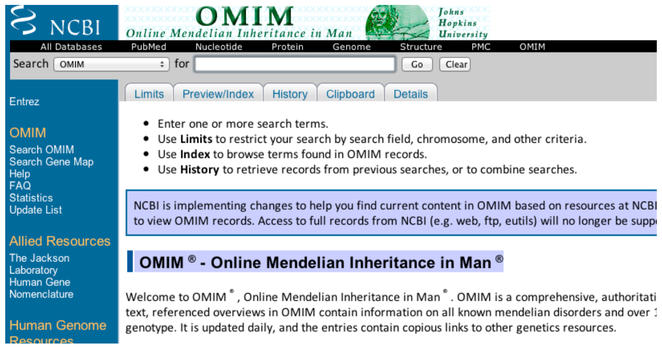


C

Comparator	Similarity metric	AUC
Full text vs DDD	RBO	0.850
OMIM vs DDD	RBO	0.774
Full text vs DDD	MICA	0.853
OMIM vs DDD	MICA	0.808

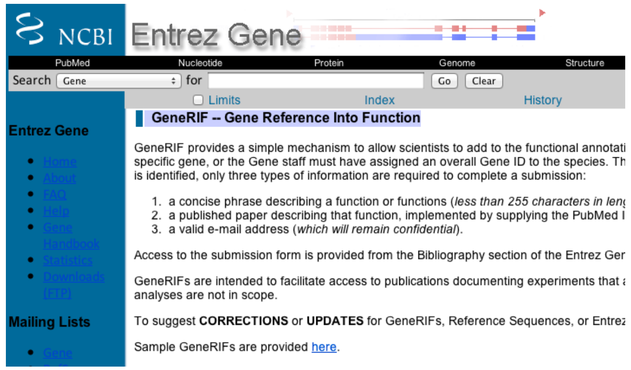
Biomedical Functional Data Resources

Online Mendelian Inheritance in Man



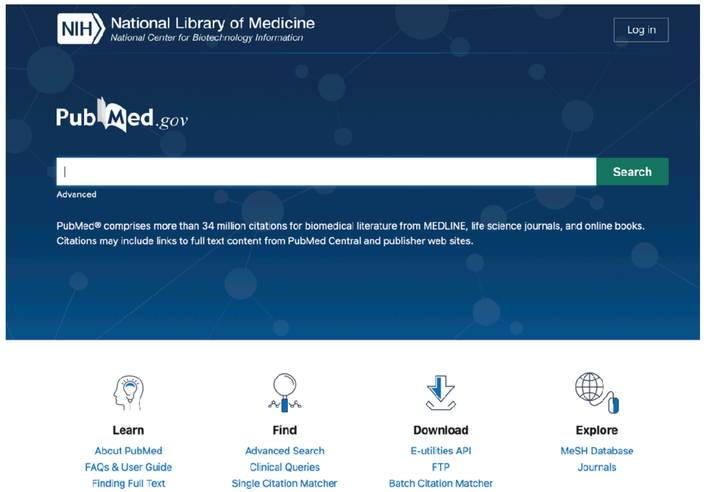
disease database (phenotypic)
gene -> disease association (mutation, linkage)

GeneRIF - gene reference into function

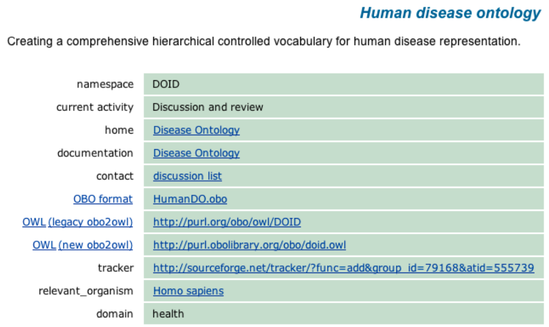


gene functional references in text

PubMed



Human Disease Ontology

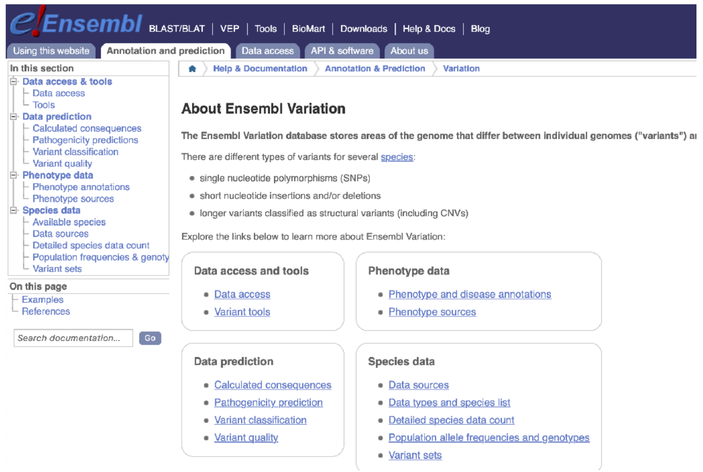


relationships between classes
of diseases (hierarchical tree)

UMLS - Unified Medical Language System

thesaurus of biomedical terms and concepts

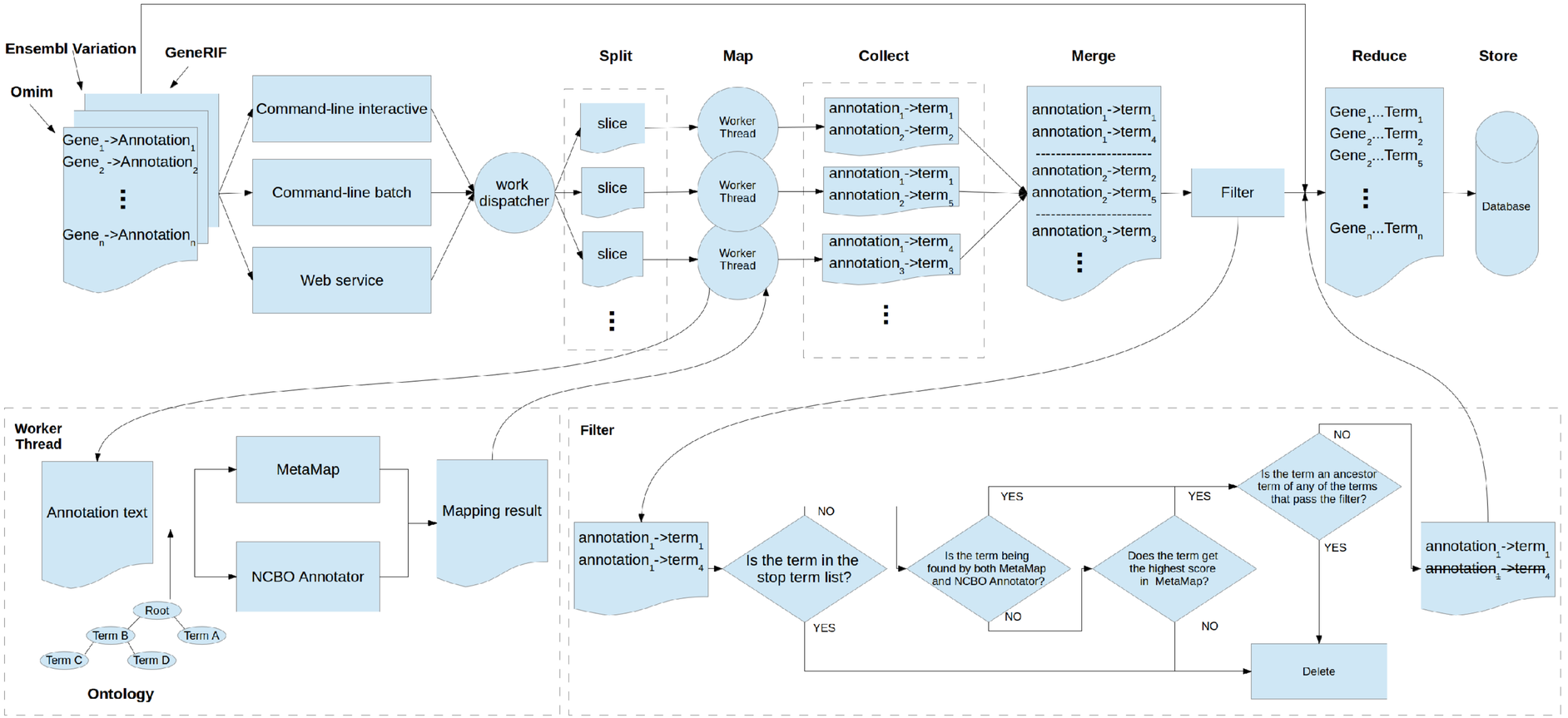
Ensembl Variation



link between genome sequence variation and phenotype

ian.simpson@ed.ac.uk

Example 5 - OntoSuite Miner - Resolving Annotation Inconsistencies



ian.simpson@ed.ac.uk

Annotating Using the Human disease ontology (DOID)

HDO

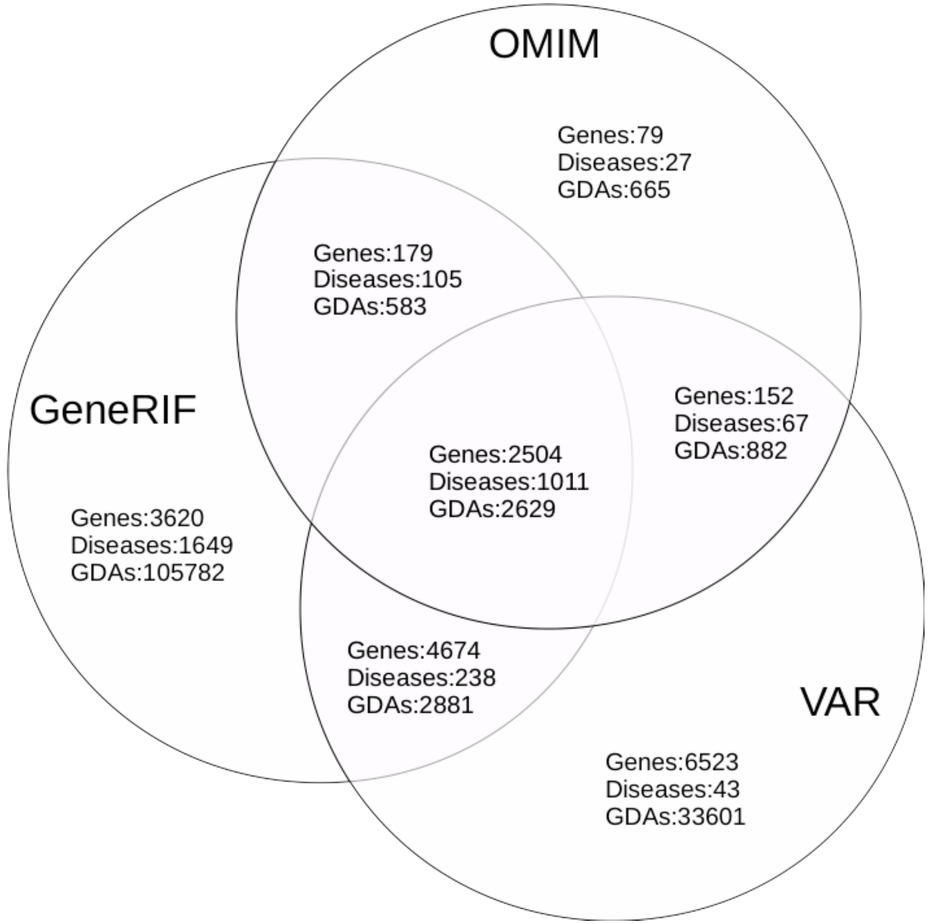
- 18055 terms
- organised by disease category
- diseases of anatomical entity
- diseases of behaviour
- biological processes
- environmental origin
- infectious agent and syndromes

Primary data sources

- Online Menelian Inheritance In Man (OMIM)
- PubMed GeneRIF
- Ensembl Variation

Low agreement between sources

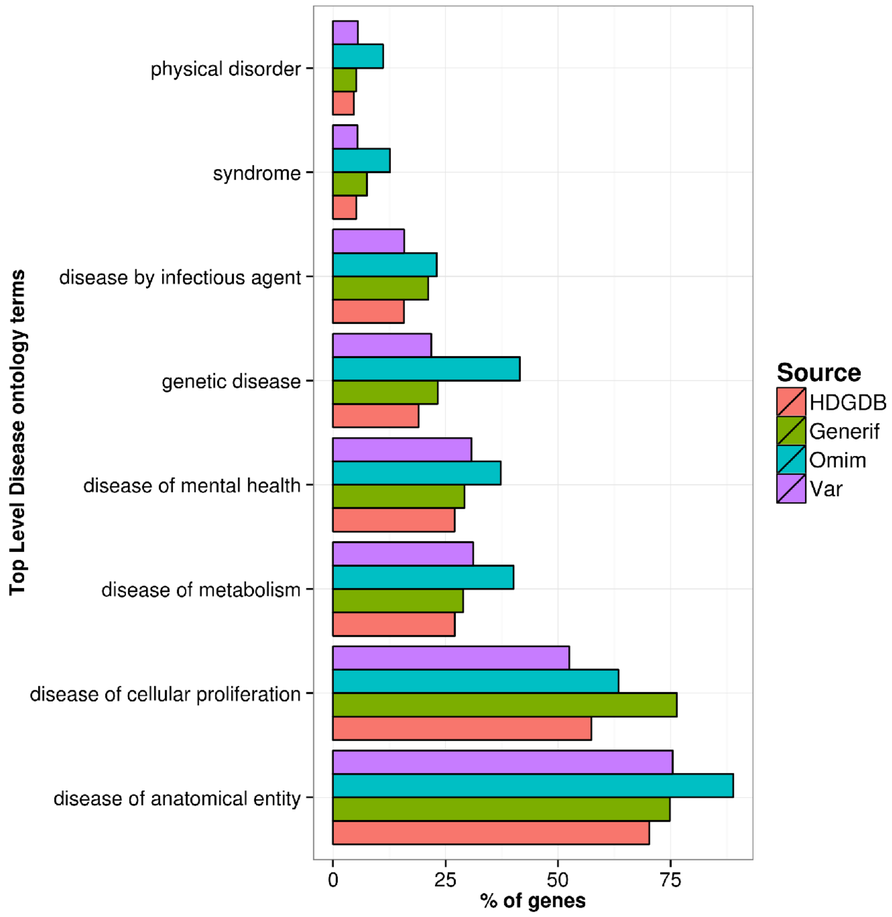
- 32.2% (1101) diseases
- 14% (2504) of genes
- 1.8% (2881) of GDAs



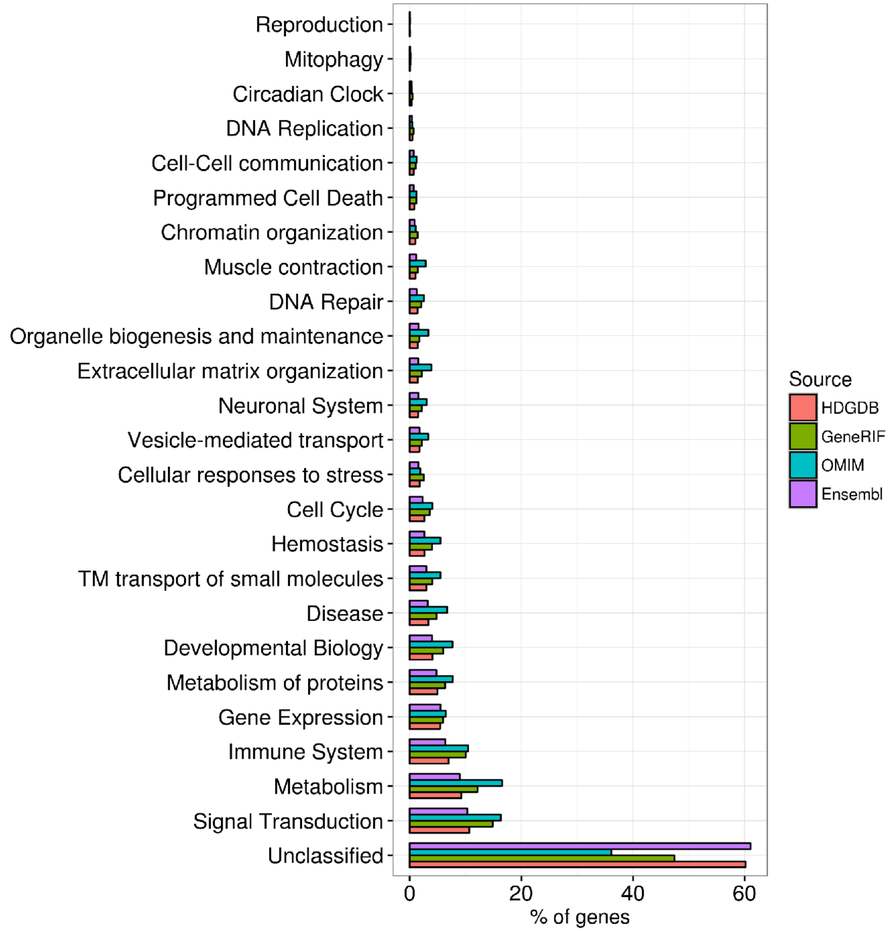
ian.simpson@ed.ac.uk

Overview of Human Disease Gene Database (HDGDB)

Top Level Disease



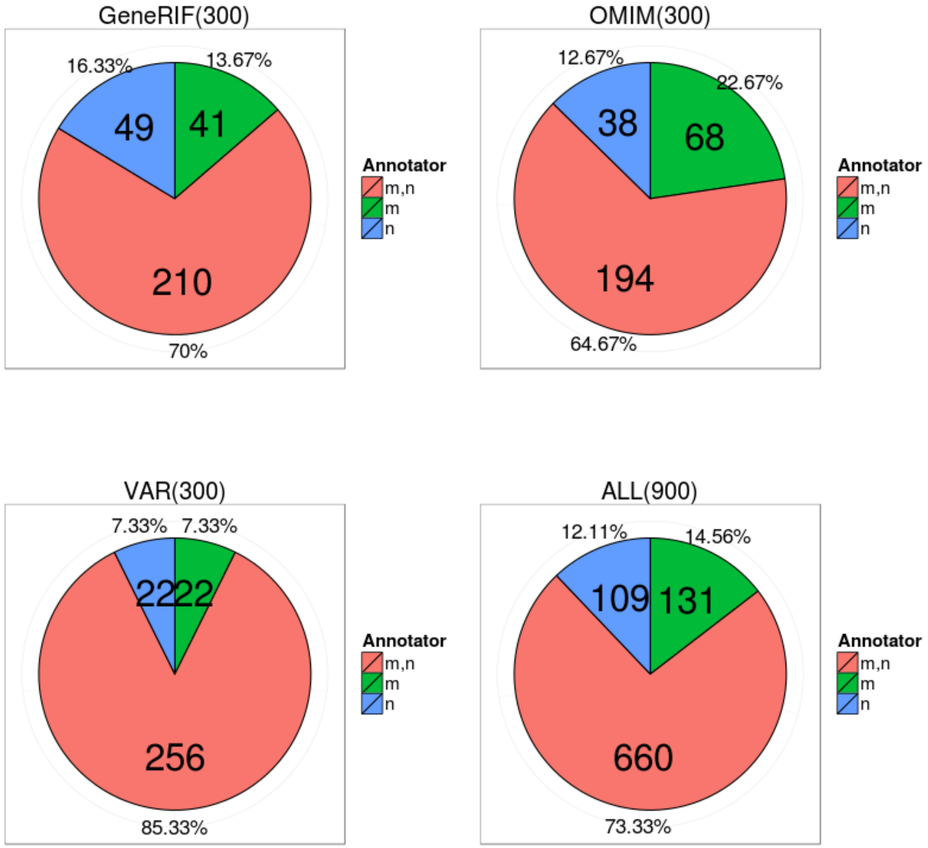
Top Level Pathways



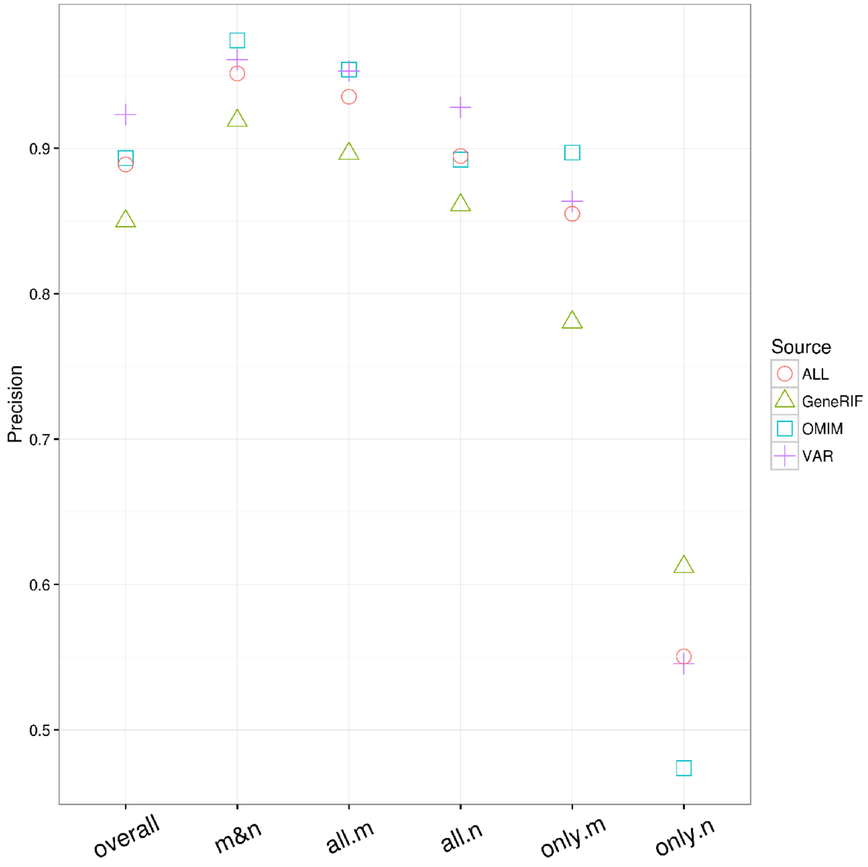
ian.simpson@ed.ac.uk

Validation of the Human Disease Gene DataBase (HDGDB)

Annotator Performance by Source



Precision by Annotator

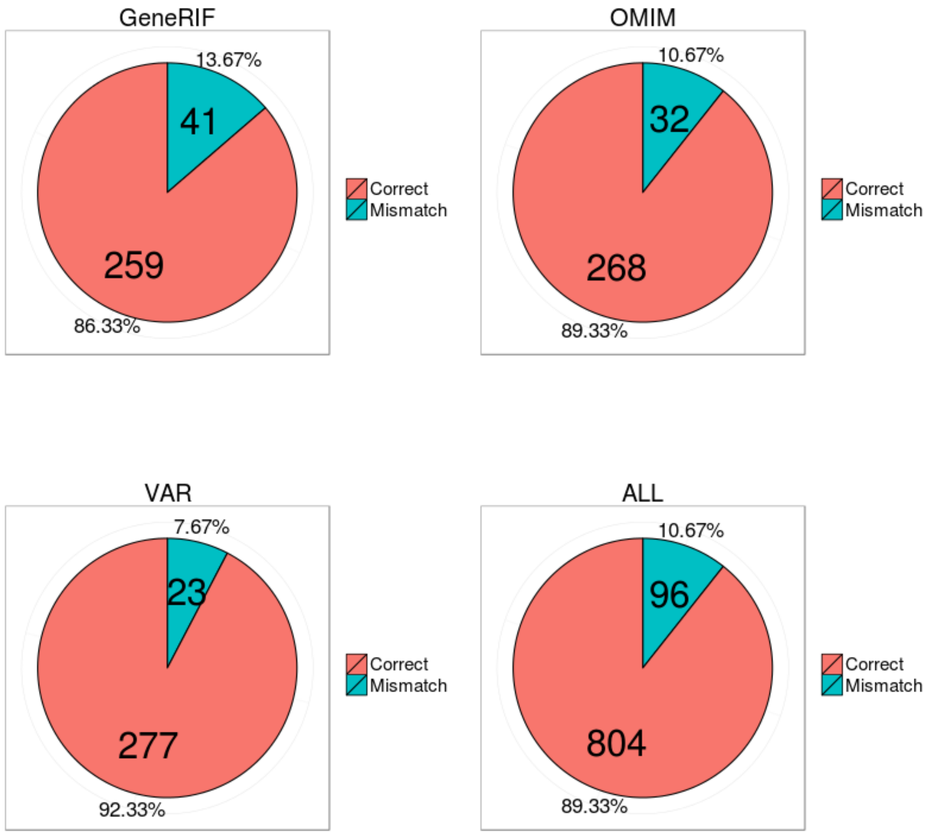


900 manually inspected mappings

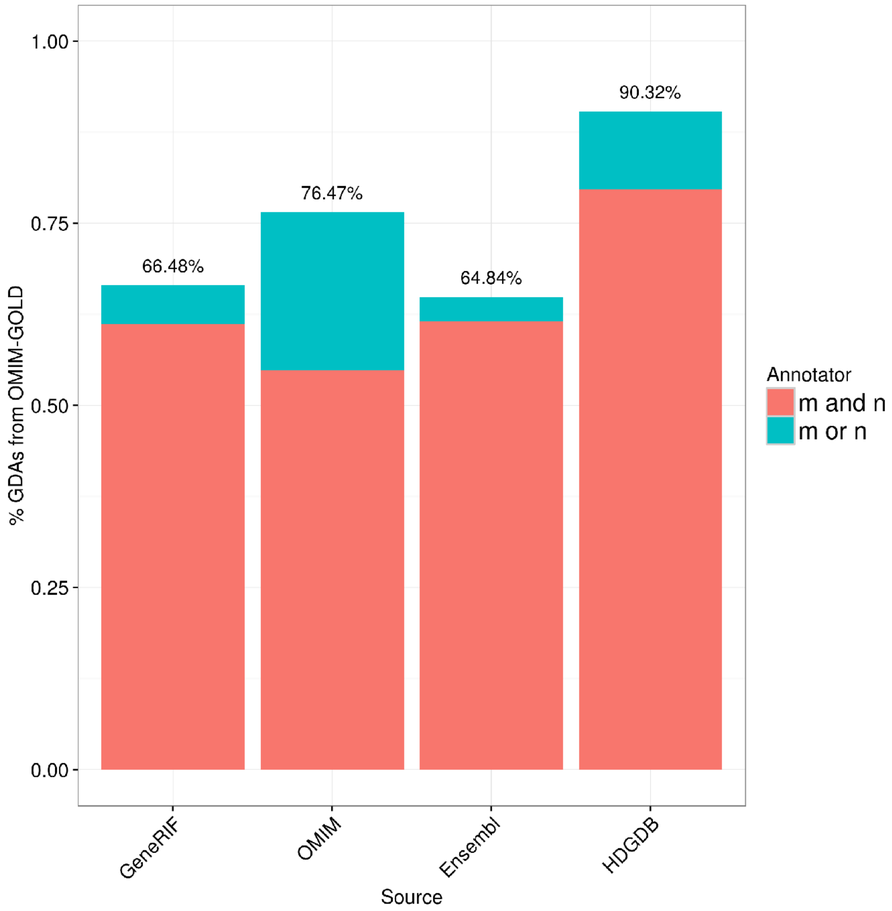
ian.simpson@ed.ac.uk

Validation of the Human Disease Gene DataBase (HDGDB)

Errors

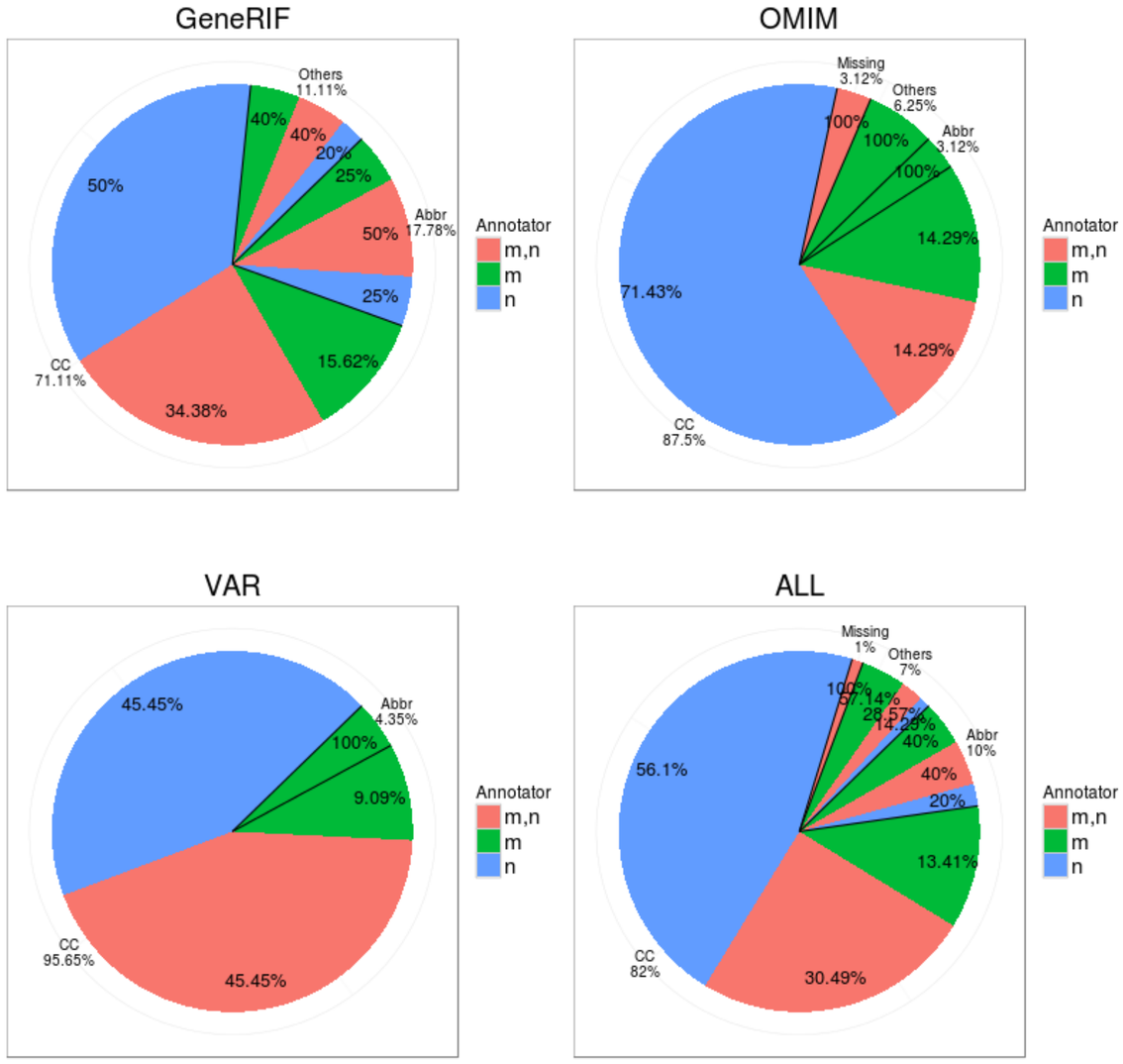


Recovery of Gold Standard Gene:Disease Associations by Source



ian.simpson@ed.ac.uk

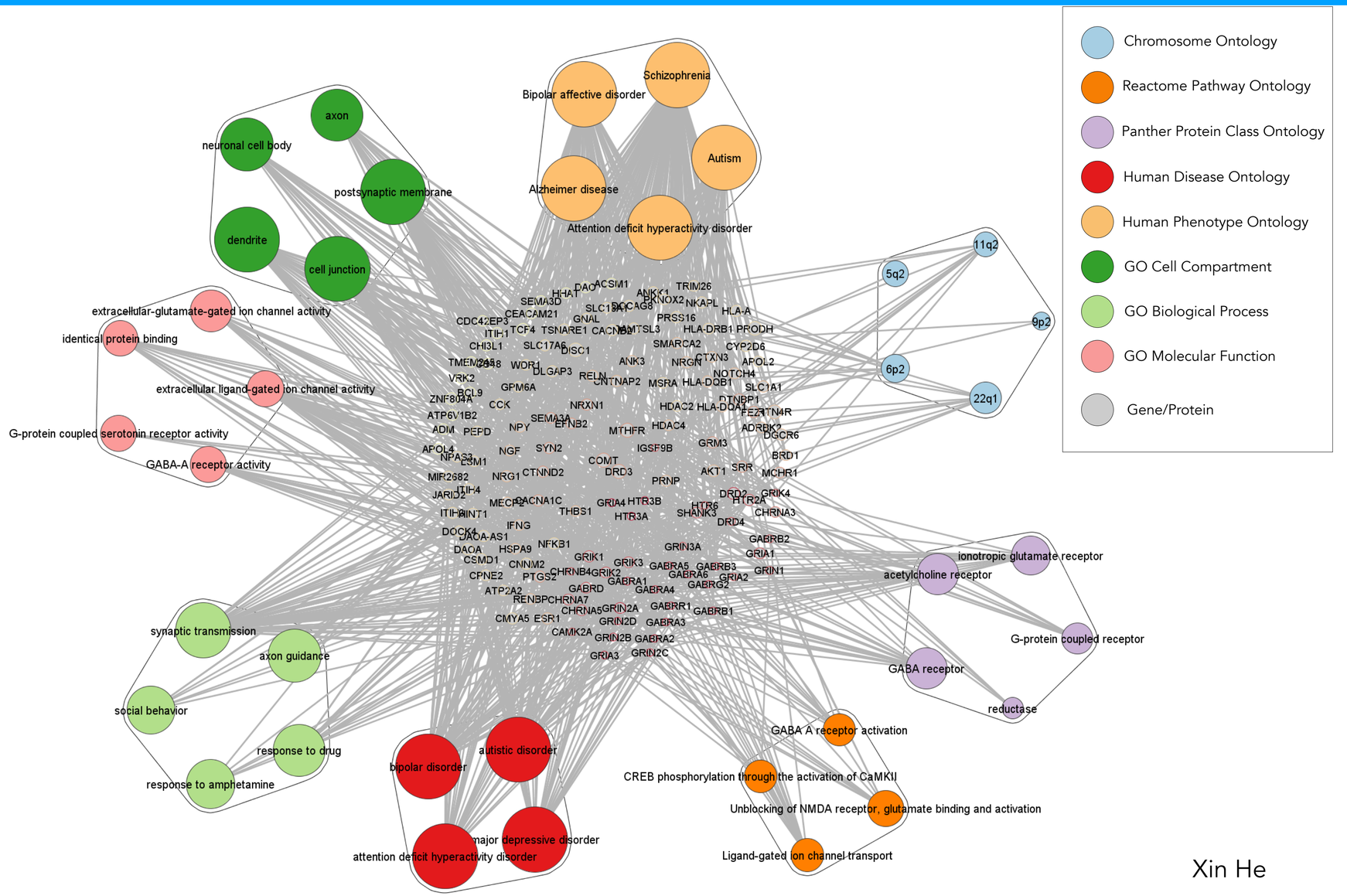
Validation of the Human Disease Gene DataBase (HDGDB)



CC - coorindating conjunctions

ian.simpson@ed.ac.uk

Multi-Ontology Integration - Schizophrenia Disease Environment



Xin He

ian.simpson@ed.ac.uk

References

Bio-Ontologies

Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform. 2008:67-79

Making Ontologies

Ontology Development 101: A Guide to Creating Your First Ontology. Natalya F. Noy and Deborah L. McGuinness, Stanford University.
http://protege.stanford.edu/publicationsontology_development/ontology101.pdf

topGO

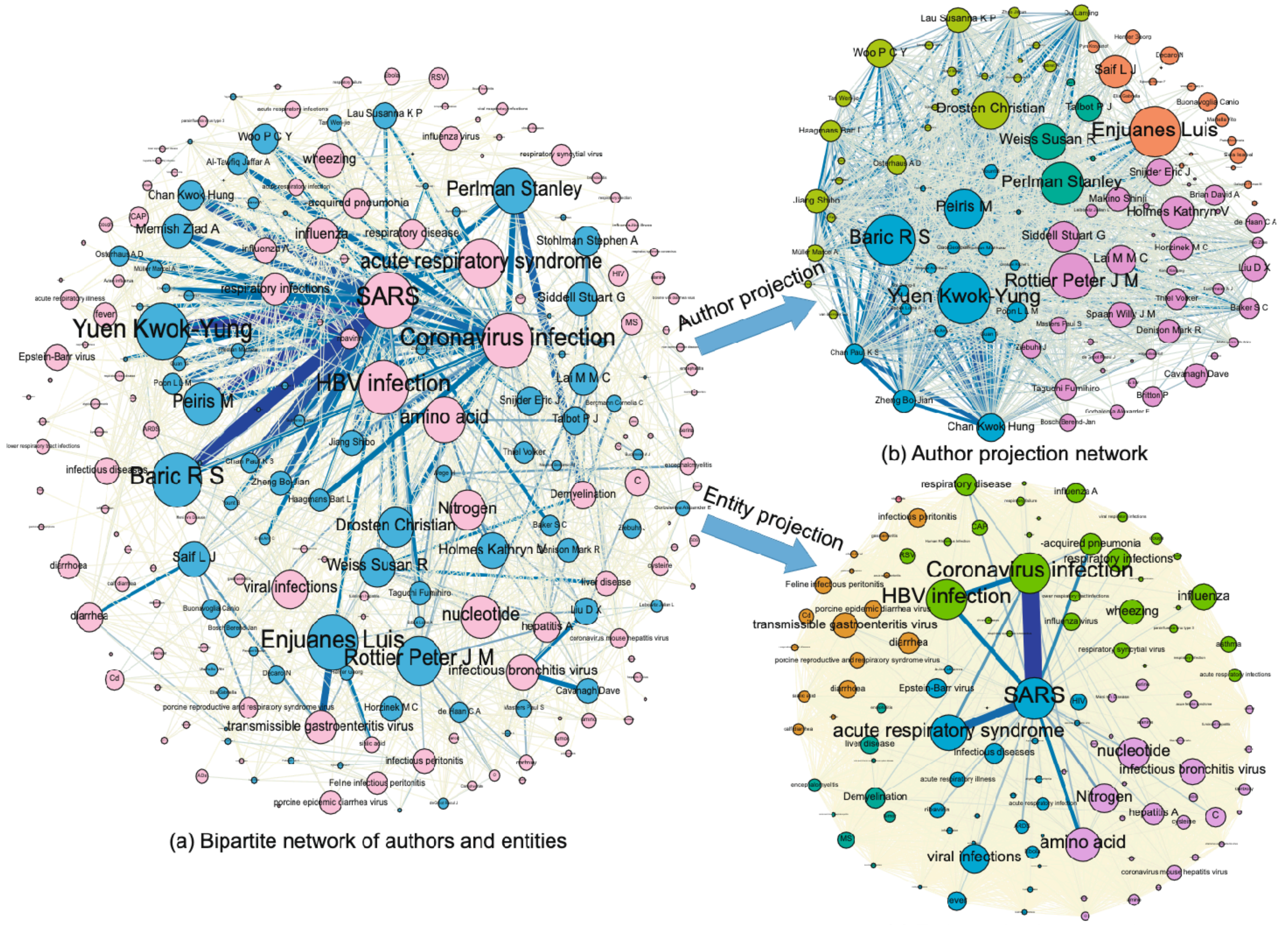
Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. 2006 Jul 1;22(13):1600-7

GSEA

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50

Next Week

Week 8 - Biological Network Analysis



(a) Bipartite network of authors and entities
(b) Author projection network
(c) Entity projection network

<u>Optional</u> Reading - also check the course Resource List

Newman. (2018). Networks / Mark Newman. (Second edition.). Oxford University Press.

Xiaoke Ma, Lin Gao. Biological network analysis: insights into structure and functions.
Briefings in Functional Genomics, Volume 11, Issue 6, November 2012, Pages 434–442
<u>https://doi.org/10.1093/bfgp/els045</u>

Analysis of Biological Networks - Junker, Björn H., and Falk. Schreiber. Hoboken, N.J: Wiley-Interscience, 2008
Chapters 1,2 and 9 (protein-protein interactions, esp. sections 9.3 and 9.4)
This book is available online through the university and is on the Resource List

ian.simpson@ed.ac.uk