DisGeNET discovery platform 5.0
Illuminating the study of human diseases

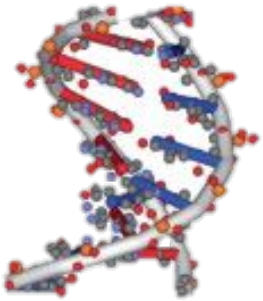
Laura I. Furlong

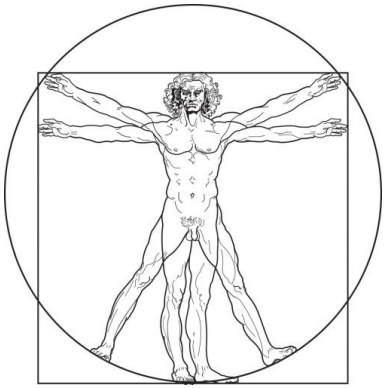
Translational Bioinformatics
Wellcome Genome Campus
UK
12-13 June 2017

GRIB RESEARCH PROGRAMME ON BIOMEDICAL INFORMATICS

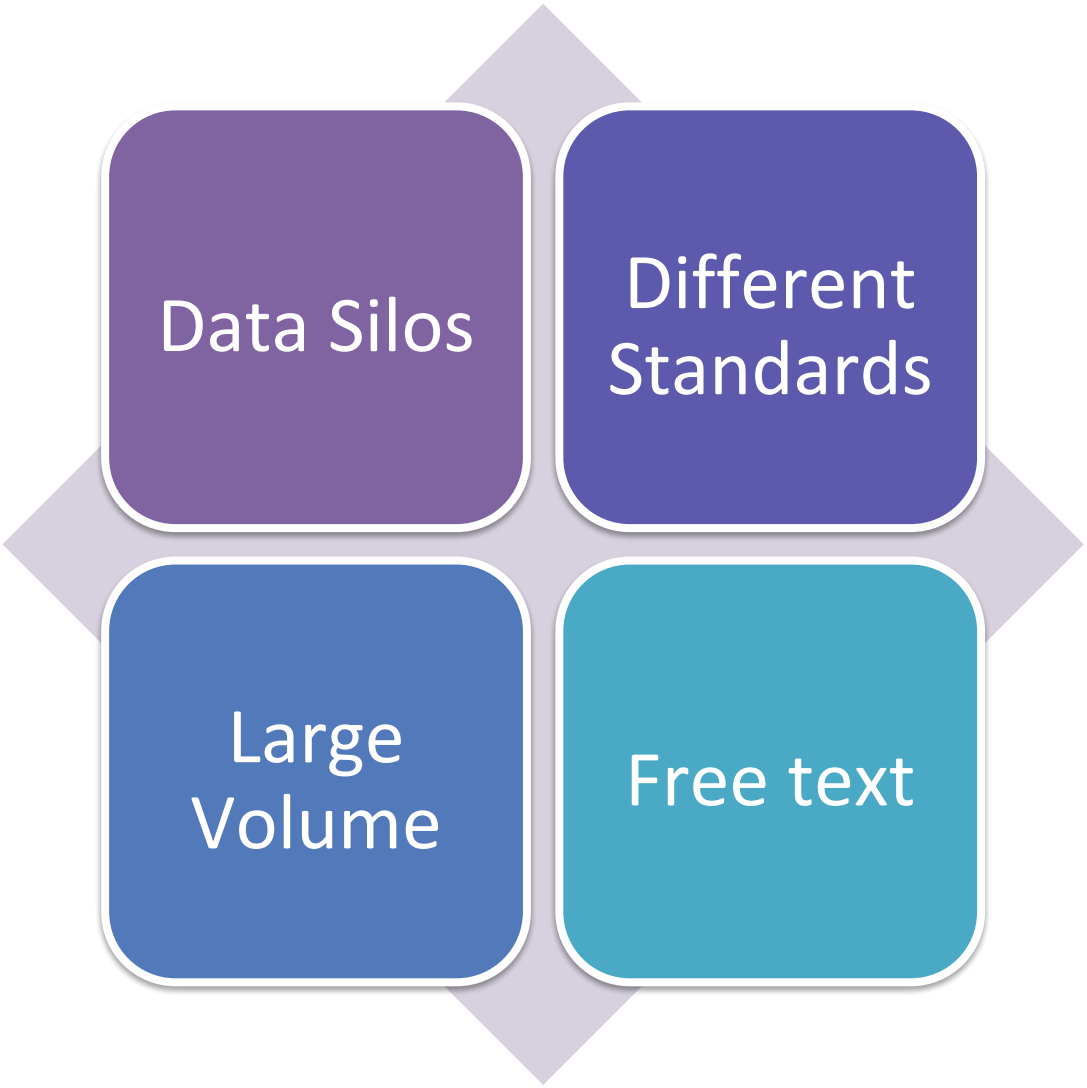
upf. Universitat Pompeu Fabra Barcelona

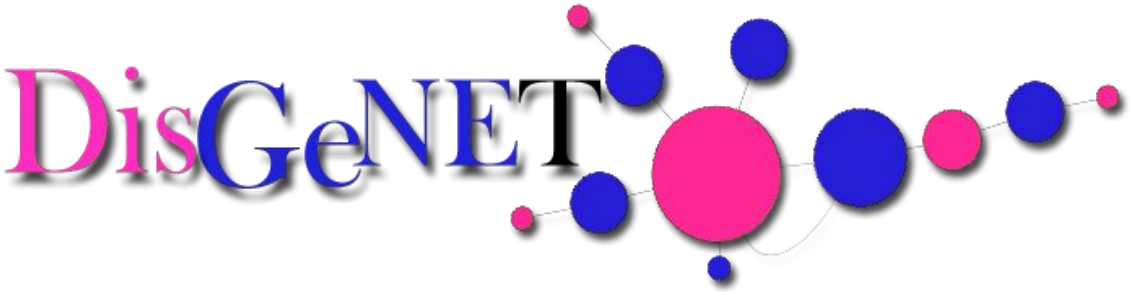
IMIM Parc de Salut MAR Barcelona
Institut Hospital del Mar d'Investigacions Mèdiques





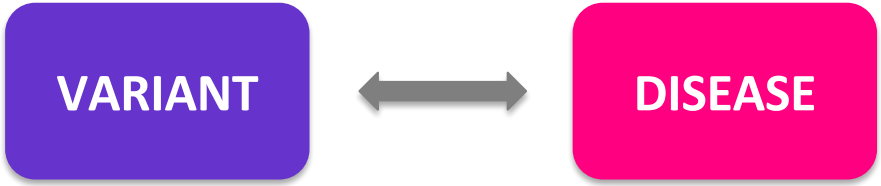
genotype ⟷ **phenotype**


Data Silos
Different Standards
Large Volume
Free text

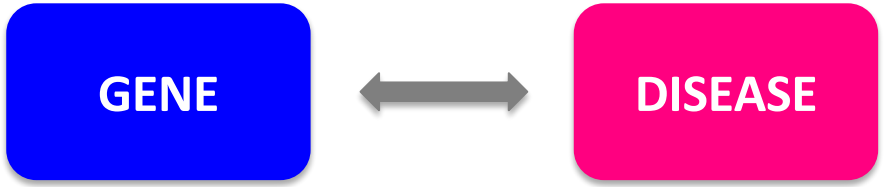


- One of the most comprehensive catalog of genes and variants associated to human diseases
- Developed by integration of different public resources, including information extracted from the literature by text mining
- Publicly available at: http://www.disgenet.org

Variant-Disease Association (VDA)

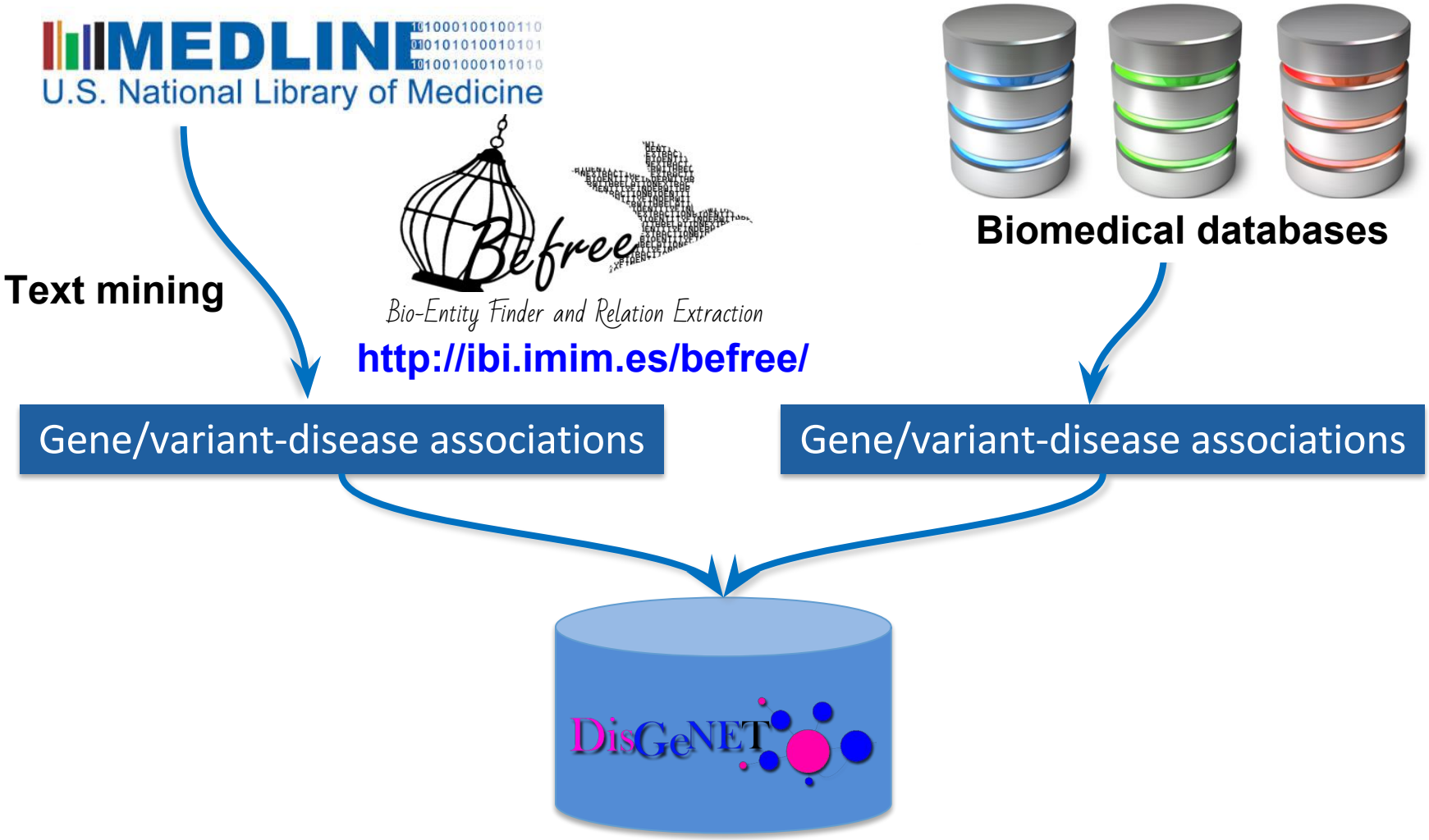


Gene-Disease Association (GDA)



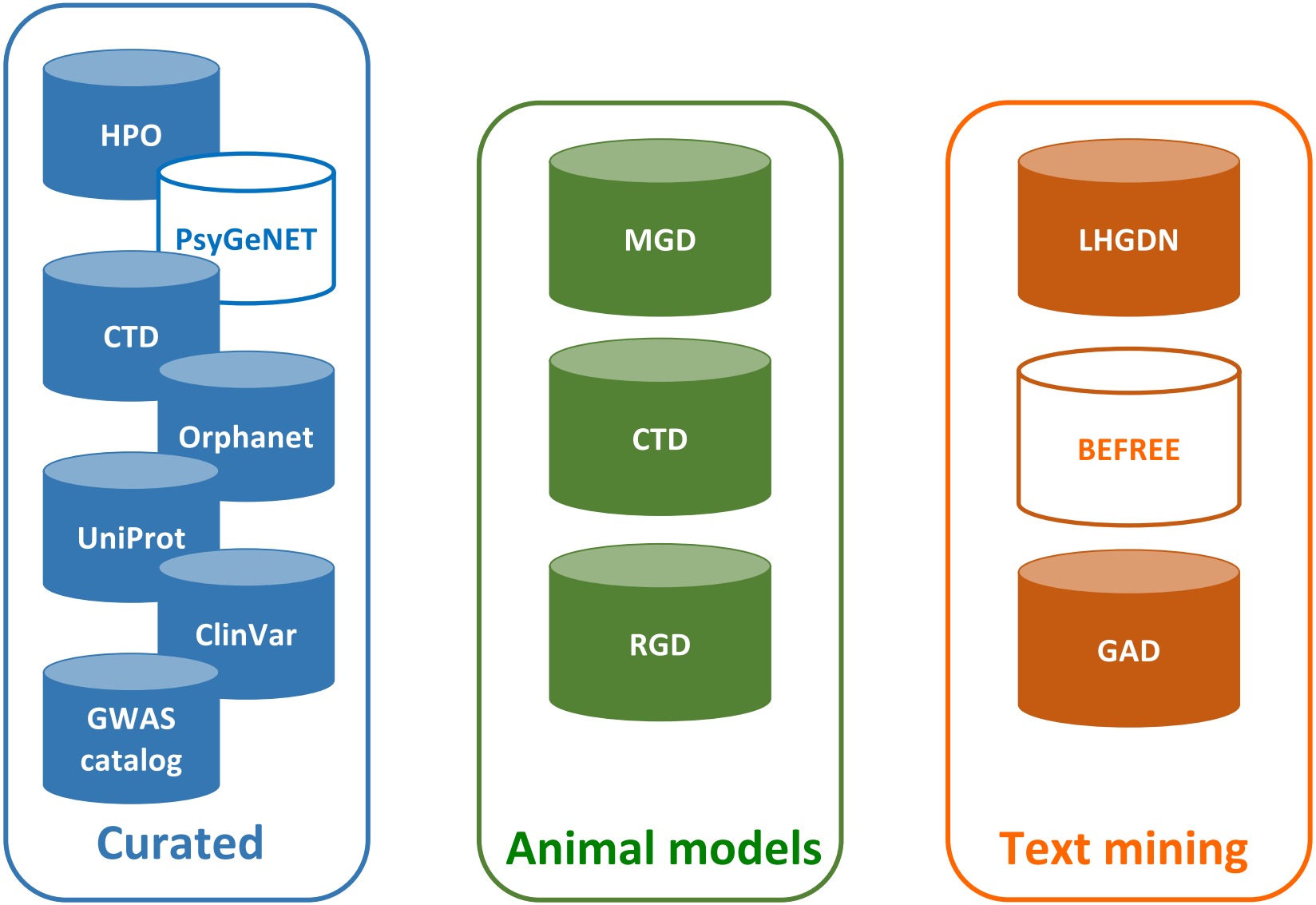
- ✓ Provenance
- ✓ Standards
- ✓ Tools

DisGeNET: the implementation

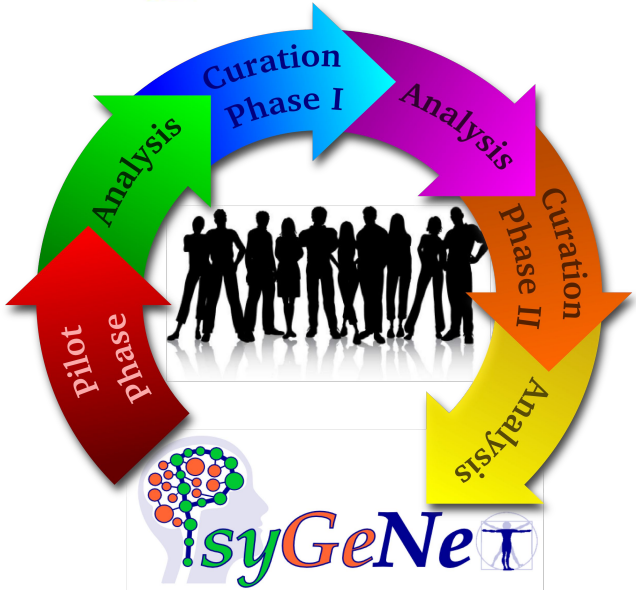


Piñero *et al*, 2015 doi: 10.1093/database/bav028

DisGeNET: data sources



http://www.psygenet.org/



GDAs and VDAs identified with

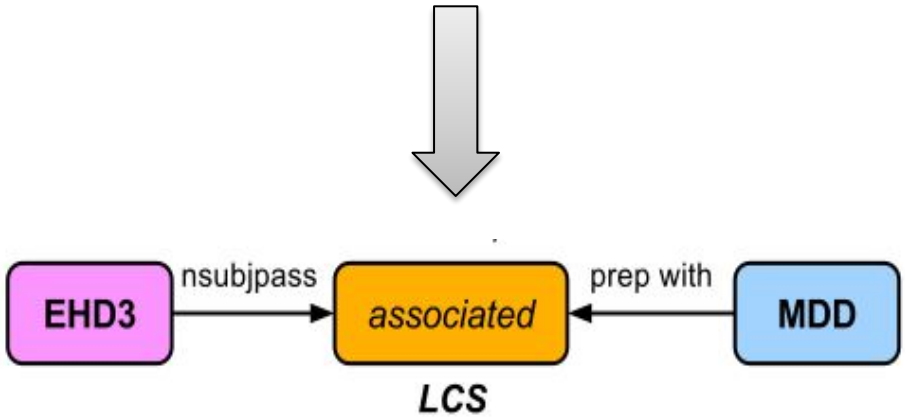
GENE CANDIDATE

DISEASE CANDIDATE

Of the 16 genes tested, **EHD3** and **FREM3** were associated with **MDD** in the Chinese population.

Gene ID: 30845
EH-domain containing 3

Disease ID: C1269683
Major depressive disorder

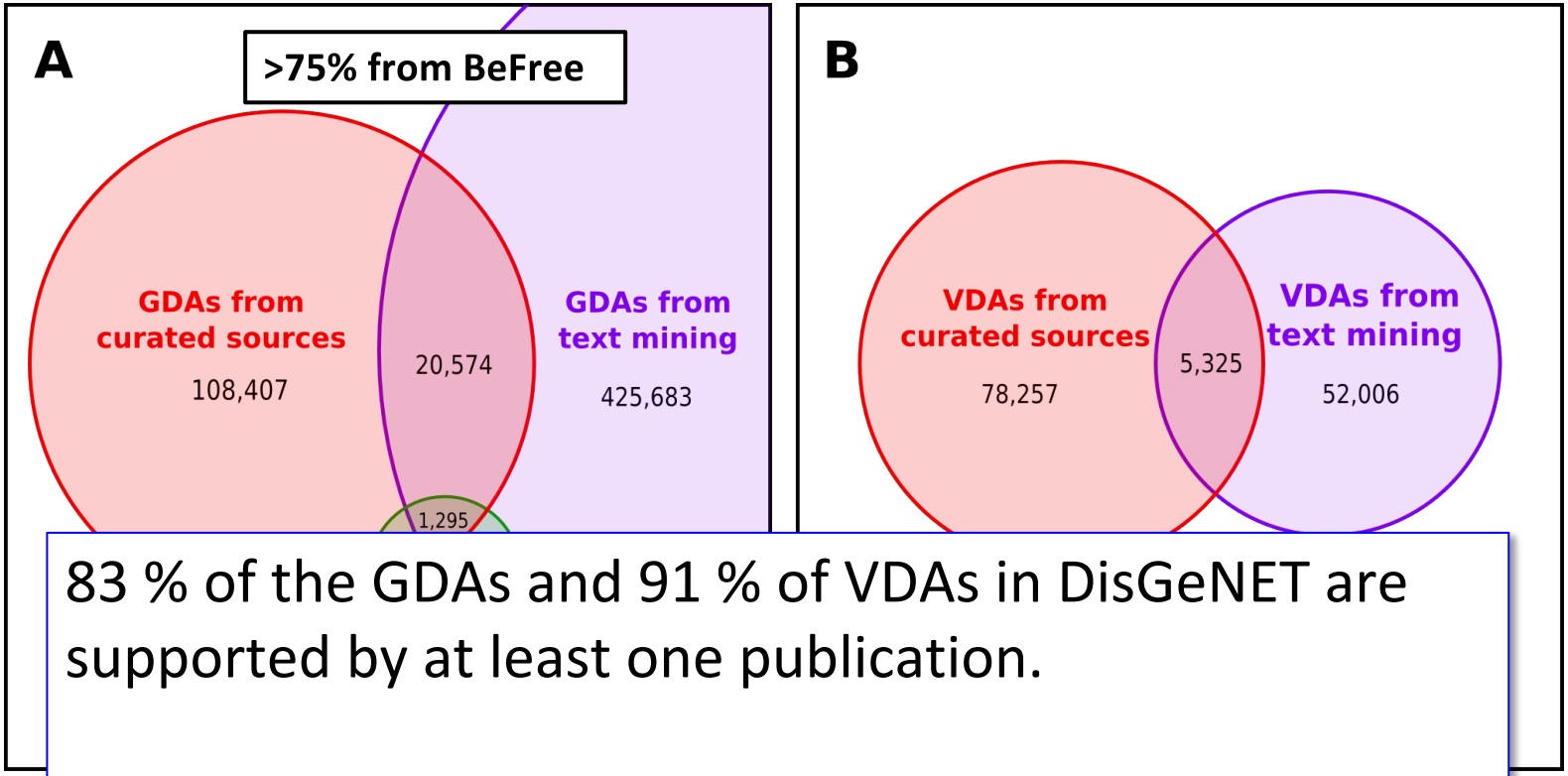


Association types classified according to the DisGeNET ontology

DisGeNET: statistics

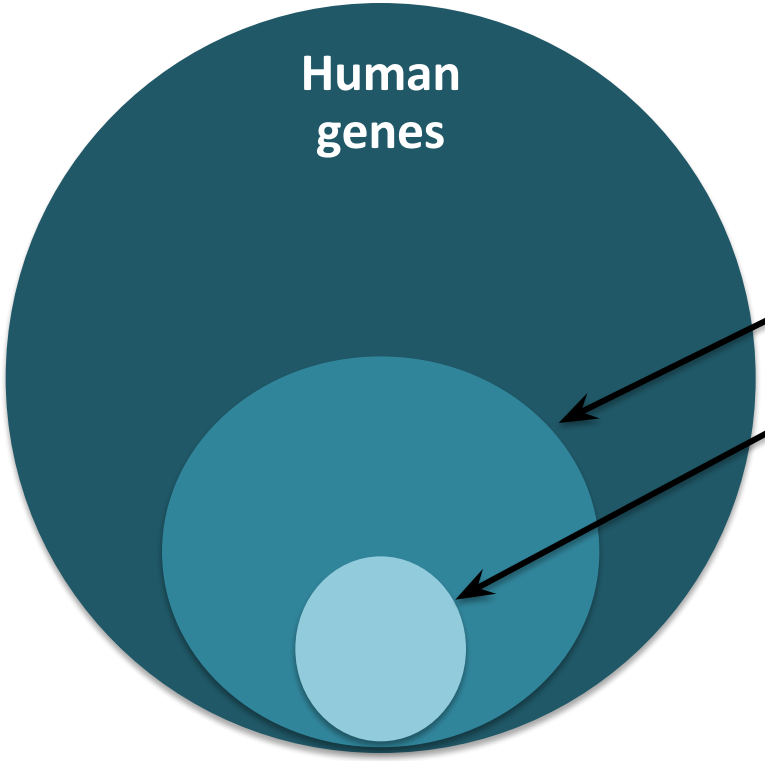
Gene-Disease Associations (GDAs)

Source	Genes	Diseases	Associations
Curated	8,948	13,074	130,821
Animal Models	2,300	1,943	6,455
Text mining	16,119	12,604	448,732
All	**17,074**	**20,370**	**561,119**



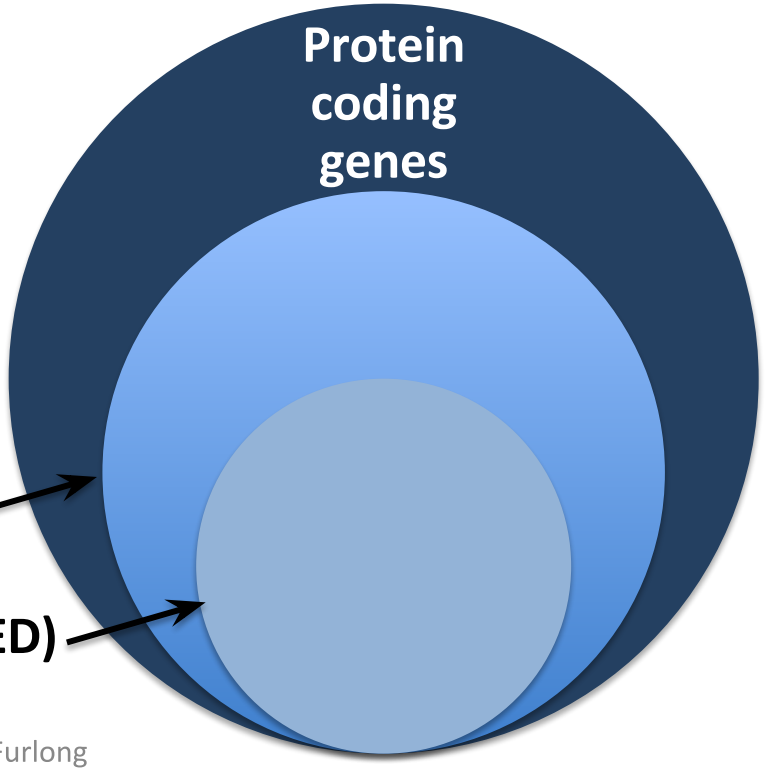
83 % of the GDAs and 91 % of VDAs in DisGeNET are supported by at least one publication.

DisGeNET contains a corpus of 422,027 publications with information on GDAs and VDAs.



Disease genes 28 % (ALL)

Disease genes 15 % (CURATED)



Protein coding disease genes 71 % (ALL)

Protein coding disease genes 37 % (CURATED)

Analysis of genomic information suggests that there are 3000 “druggable” proteins.

Only 10% of these potential targets have an FDA approved drug.

Hopkins et al . The druggable genome. Nat. Rev. Drug Discov. 2002; 1:727–730

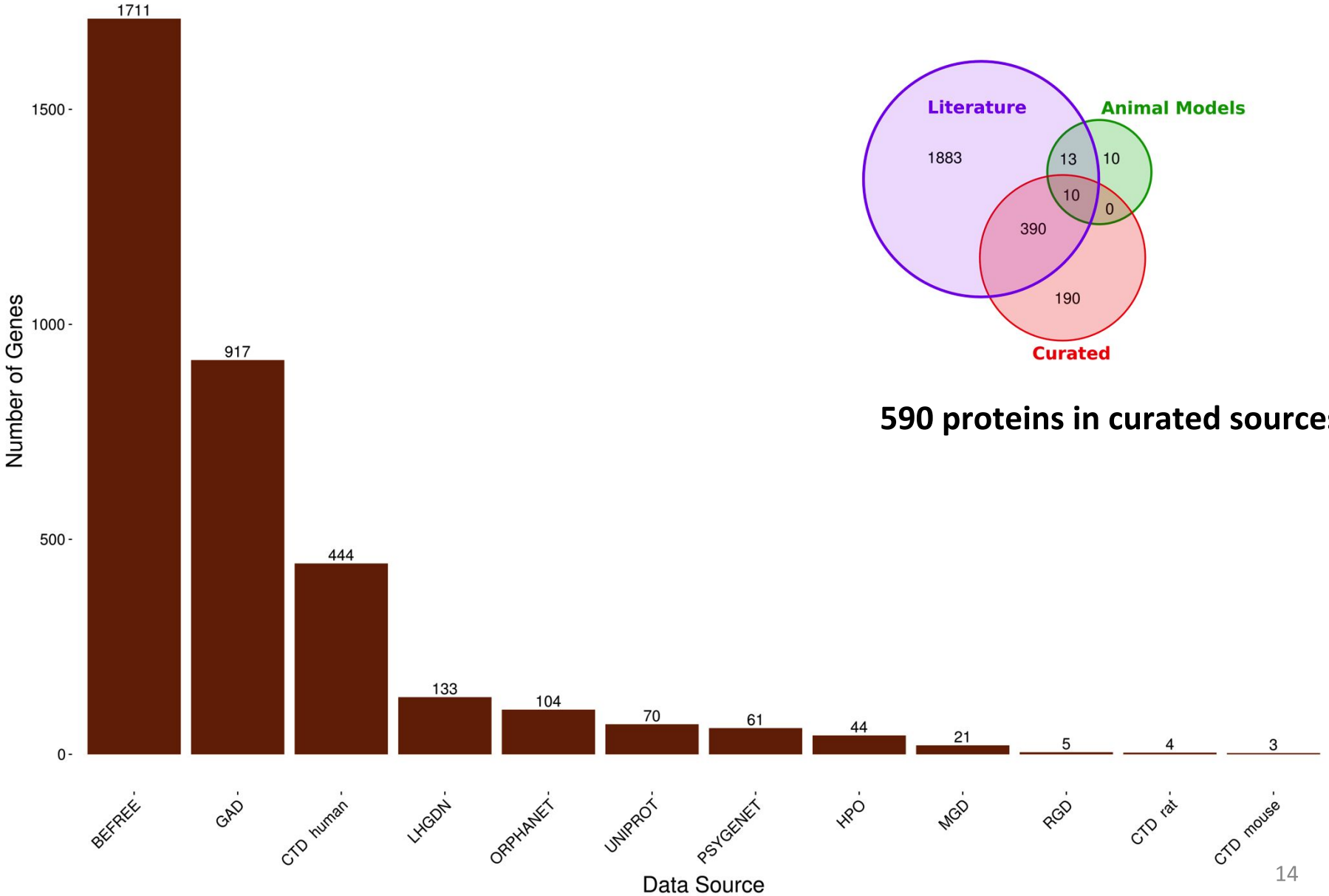
NIH Illuminating the Druggable Genome (IDG) program:

To shed light on **poorly characterized proteins** that can potentially be modulated using small molecules or biologics

Tdark proteins

PHAROS https://commonfund.nih.gov/idg/index

DisGeNET annotates 37 % of Pharos Tdark proteins



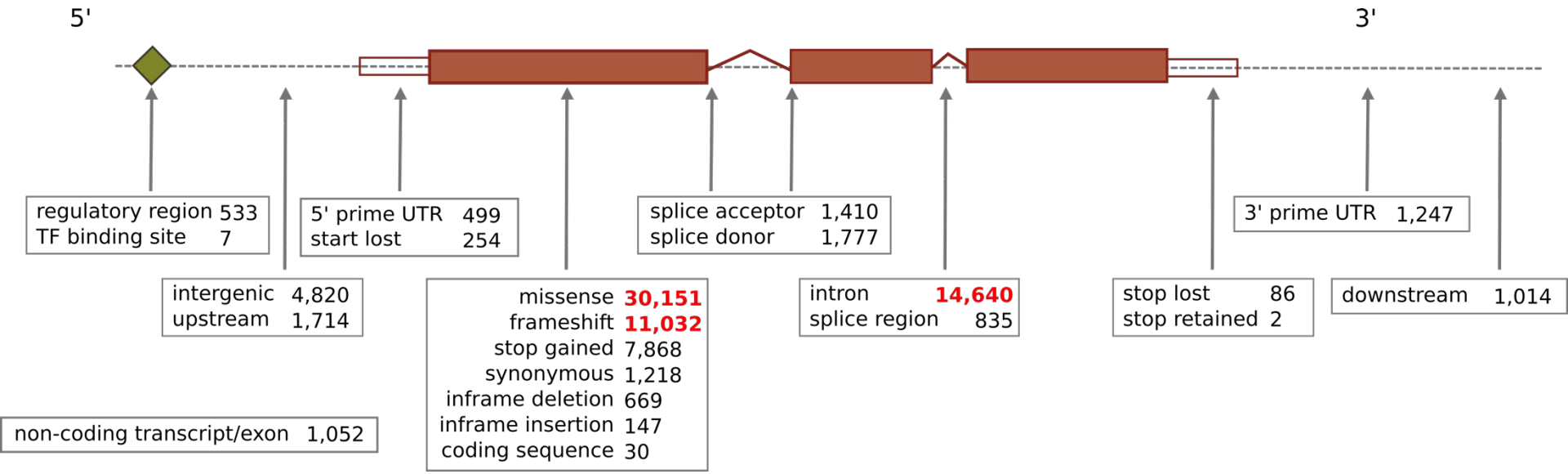
590 proteins in curated sources

DisGeNET: statistics

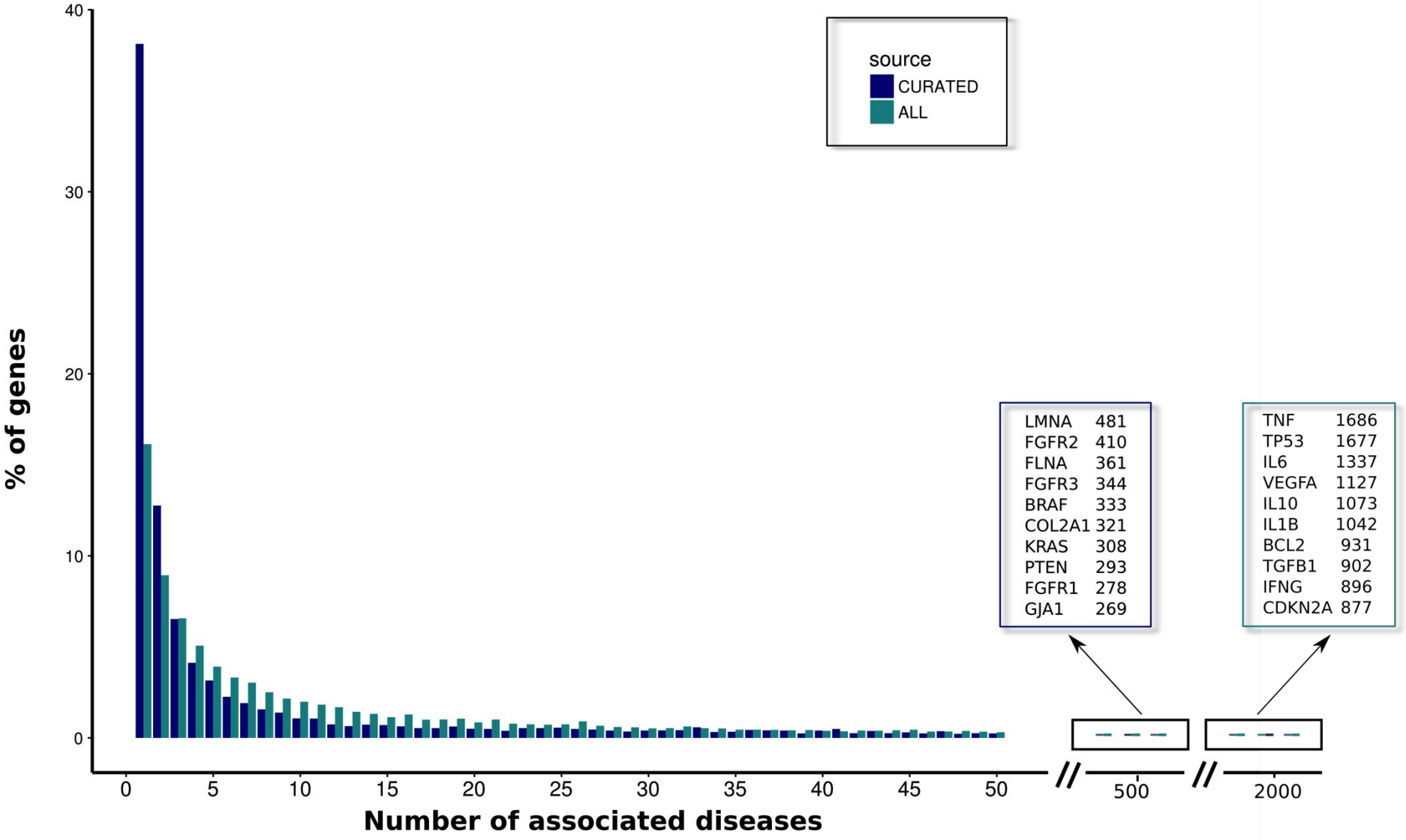
Variant-Disease Associations (VDAs)

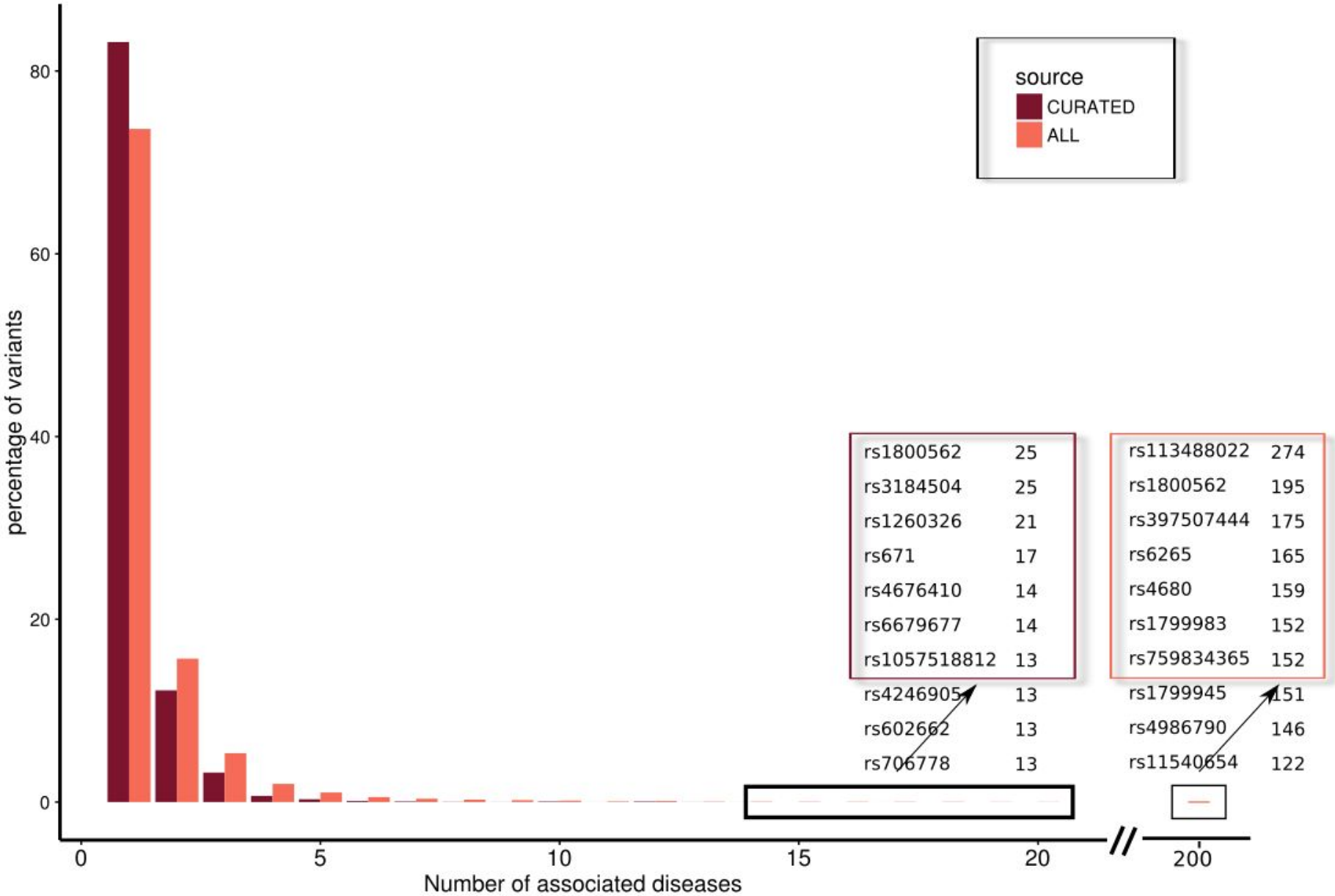
Source	Variants	Diseases	Associations
Curated	66,903	6,388	83,582
Text mining	24,455	4,432	57,331
All	**83,002**	**9,169**	**135,588**

Distribution of disease variants according to variant consequence type



http://www.ensembl.org/info/genome/variation/predicted_data.html


source
CURATED
ALL
% of genes
Number of associated diseases
0
10
20
30
40
0
5
10
15
20
25
30
35
40
45
50
500
2000
LMNA 481
FGFR2 410
FLNA 361
FGFR3 344
BRAF 333
COL2A1 321
KRAS 308
PTEN 293
FGFR1 278
GJA1 269
TNF 1686
TP53 1677
IL6 1337
VEGFA 1127
IL10 1073
IL1B 1042
BCL2 931
TGFB1 902
IFNG 896
CDKN2A 877

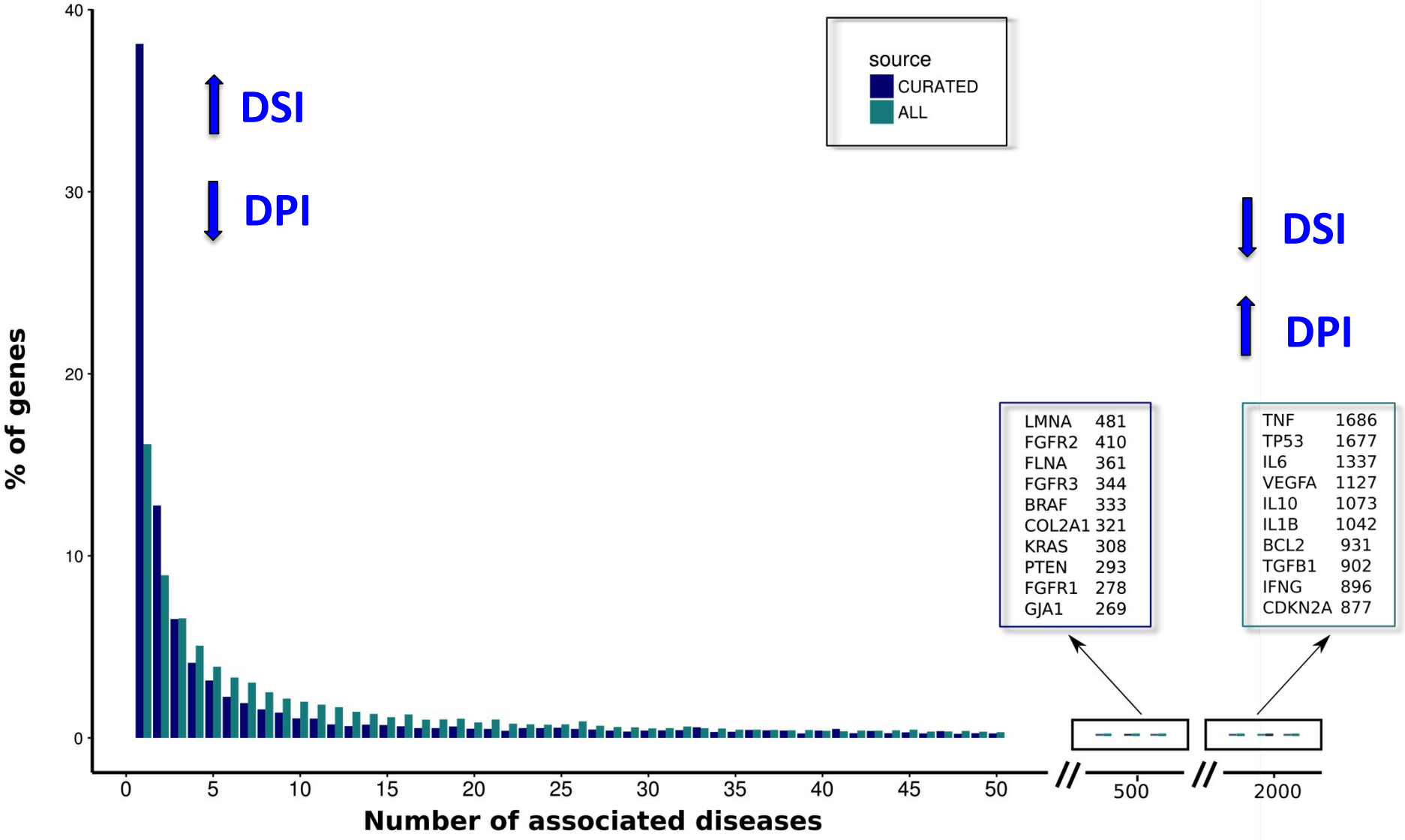

source
CURATED
ALL
percentage of variants
80
60
40
20
0
0
5
10
15
20
200
Number of associated diseases
rs1800562 25
rs3184504 25
rs1260326 21
rs671 17
rs4676410 14
rs6679677 14
rs1057518812 13
rs4246905 13
rs602662 13
rs706778 13
rs113488022 274
rs1800562 195
rs397507444 175
rs6265 165
rs4680 159
rs1799983 152
rs759834365 152
rs1799945 151
rs4986790 146
rs11540654 122

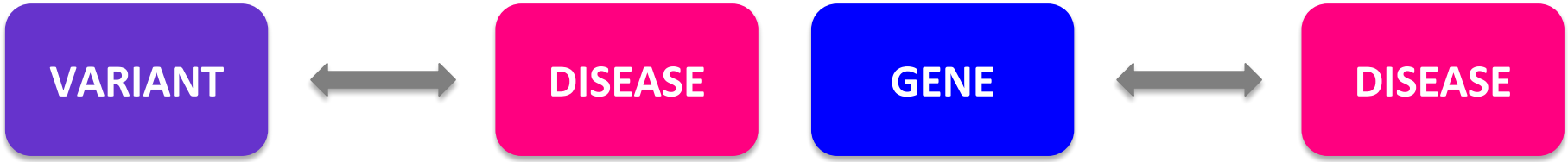
VARIANT

GENE

Tools for prioritization

- ✓ Disease Specificity Index (DSI)
- ✓ Disease Pleiotropy Index (DPI)
- ✓ Panther Protein class
- ✓ Allele frequency, variant consequence type


source
CURATED
ALL
DSI
DPI
DSI
DPI
% of genes
Number of associated diseases
LMNA 481
FGFR2 410
FLNA 361
FGFR3 344
BRAF 333
COL2A1 321
KRAS 308
PTEN 293
FGFR1 278
GJA1 269
TNF 1686
TP53 1677
IL6 1337
VEGFA 1127
IL10 1073
IL1B 1042
BCL2 931
TGFB1 902
IFNG 896
CDKN2A 877



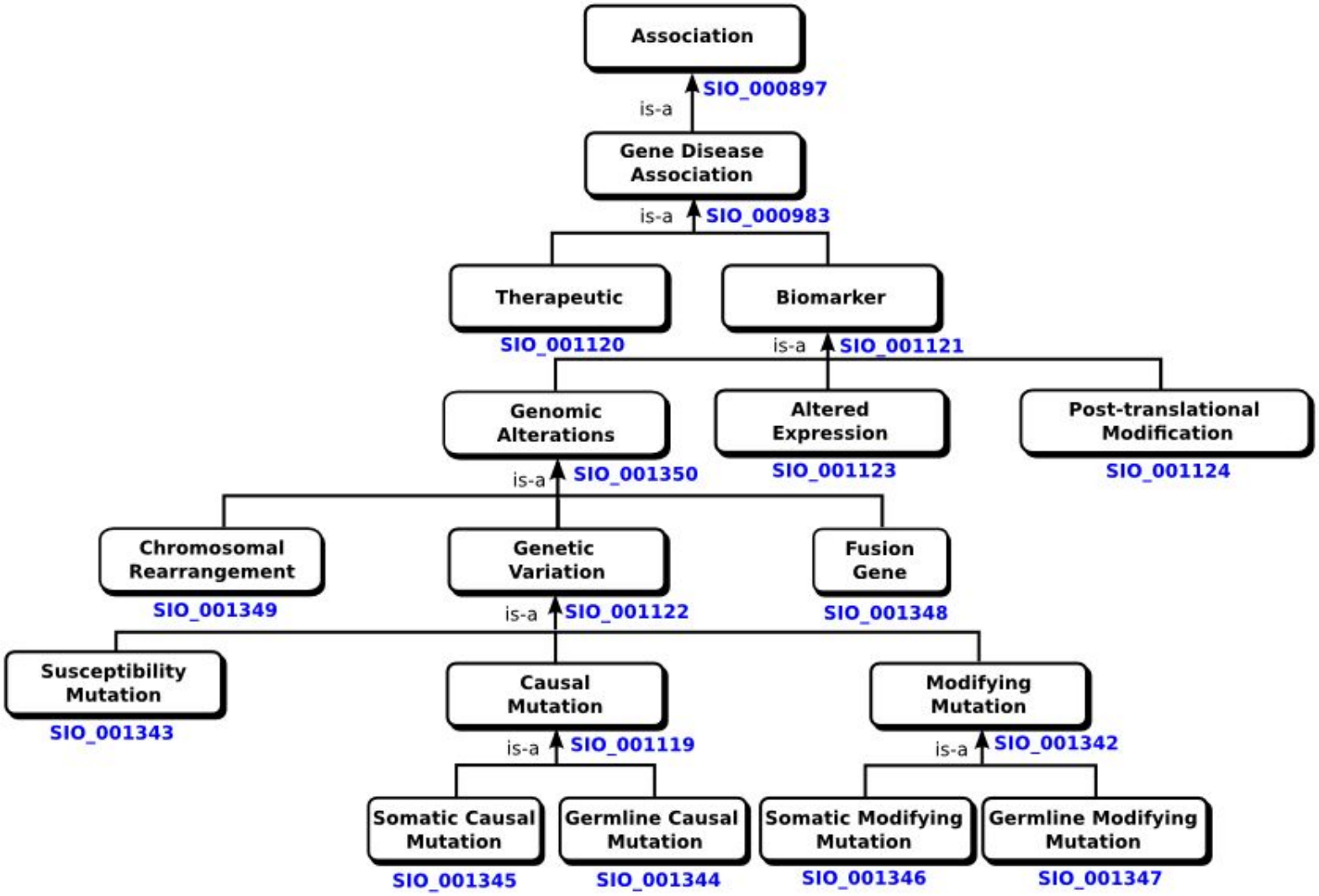
Tools for prioritization

- ✓ **DisGeNET association score:** popularity/novelty
- ✓ **Evidence Index:** controversial field of research
- ✓ **DisGeNET association type:** insight on biology
- ✓ **Number of publications**

Top scored genes for Cystic Fibrosis

Gene	Number of diseases	DisGeNET score	DSI	Number of PMIDs	Number of SNPs
CFTR	308	1	0,948	0,422	1596
TGFB1	902	0,433	1	0,314	17
DCTN4	83	0,401	1	0,555	2
SCNN1B	69	0,293	1	0,573	13
SCNN1G	65	0,212	1	0,579	6
SCNN1A	64	0,207	0,833	0,581	9
TNFRSF1A	318	0,205	1	0,419	1
CLCA4	18	0,204	1	0,709	4
STX1A	54	0,200	1	0,598	1

DisGeNET association type ontology

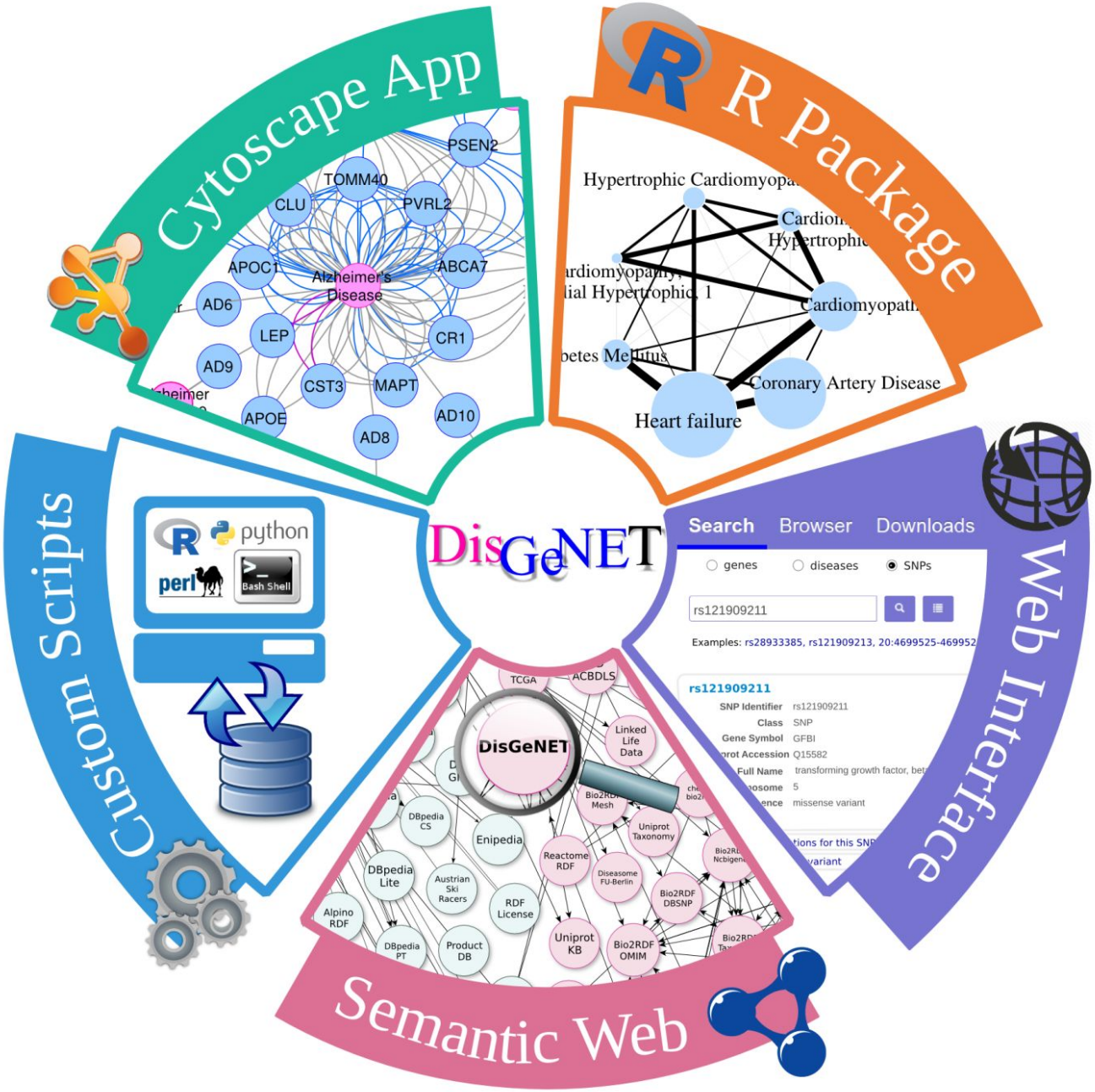


http://sio.semanticscience.org

DISEASE

Tools for prioritization

- ✓ UMLS semantic types
- ✓ DO, HPO class
- ✓ Group, disease, phenotype


Cytoscape App
PSEN2
TOMM40
CLU
PVRL2
APOC1
Alzheimer's Disease
ABCA7
AD6
LEP
CR1
AD9
CST3
MAPT
APOE
AD10
AD8
R Package
Hypertrophic Cardiomyopa
Cardiomyopati
Heart failure
Coronary Artery Disease
Custom Scripts
python
perl
Bash Shell
DisGeNET
Web Interface
Search
Browser
Downloads
genes
diseases
SNPs
rs121909211
Examples: rs28933385, rs121909213, 20:4699525-469952
rs121909211
SNP Identifier rs121909211
Class SNP
Gene Symbol GFBI
Full Name transforming growth factor, bet
missense variant
Semantic Web
DisGeNET
TCGA
ACBDLS
Linked Life Data
DBpedia CS
Enipedia
Uniprot Taxonomy
Reactome RDF
Diseasome FU-Berlin
DBpedia Lite
Austrian Ski Racers
RDF License
Bio2RDF DBSNP
Alpino RDF
DBpedia PT
Product DB
Uniprot KB
Bio2RDF OMIM

http://www.disgenet.org/
support@disgenet.org
twitter: @DisGeNET

IBI Group
http://ibi.imim.es/

Alba Gutiérrez-Sacristán
Àlex Bravo
Janet Piñero
Alexia Giannoula
Miguel A. Mayer
Angela Leis
Emilio Centeno
Francesco Ronzano
Juan Manuel Ramírez
Laura I. Furlong
Ferran Sanz

Past Members
Santiago de la Peña
Núria Queralt-Rosinach
Montserrat Cases
Solène Grosdidier
Pablo Carbonell
Anna Bauer-Mehren
Michael Rautschka