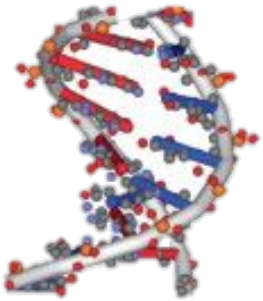
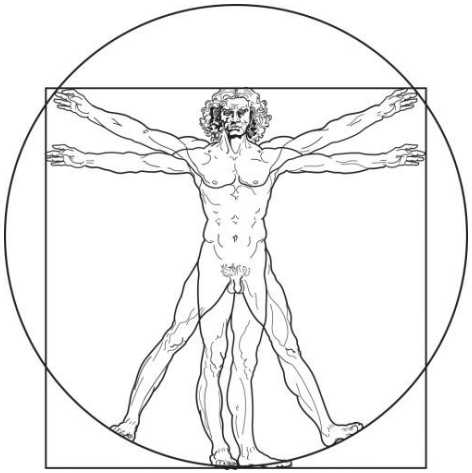
DisGeNET discovery platform 5.0
Illuminating the study of human diseases

Janet Piñero
PRBB Computational Seminars
June 22, 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

From genotype to phenotype





genotype ⟷ phenotype

The availability of ***comprehensive***, ***traceable***, ***high quality*** data on **genotype-phenotype** relations is key to understand the molecular mechanisms underlying human diseases

From genotype to phenotype: data volume

From genotype to phenotype: data silos

#277900

OMIM
Online Mendelian Inheritance in Man

WILSON DISEASE

Alternative titles; symbols

WND; WD

HEPATOLENTICULAR DEGENERATION

Phenotype-Gene Relationships

Location	Phenotype	Phenotype MIM number	Inheritance (in progress)	Phenotype mapping key	Gene/Locus
13q14.3	Wilson disease	277900	AR	3	ATP7B

orphanet

Rare diseases

- Search
- Search by sign
- Classifications
- Genes
- Disability
- Encyclopaedia for patients
- Encyclopaedia for professionals
- Emergency guidelines

Hepatolenticular Degeneration

Basics | Chemical–Gene Interactions | Chemicals | Genes

ctd™

1–50 of 11,475 results.

First | Previous | 1 2 3 4 5 6 7 8 | Next | La

	Gene	Disease	Direct Evidence
1.	CP	Hepatolenticular Degeneration	M
2.	ATP7B	Hepatolenticular Degeneration	M
3.	PRNP	Hepatolenticular Degeneration	M

Genes related to Wilson Disease

Genes related to Wilson Disease (1 elite genes):

★ - Elite gene

id	Symbol ★	Description
1	ATP7B ★	ATPase, Cu++ transporting, beta polypeptide
2	CP	ceruloplasmin (ferroxidase)
3	ATP7A	ATPase, Cu++ transporting, alpha polypeptide
4	COMMD1	copper metabolism (Murr1) domain containing 1
5	ARSA	arylsulfatase A
6	HFE	hemochromatosis
7	SLC31A1	solute carrier family 31 (copper transporter), member 1

MalaCards
HUMAN DISEASE DATABASE

GWAS Catalog

The NHGRI-EBI Catalog of published genome-wide association studies

Search the catalog

Examples: breast cancer, rs7329174, Yang, 2q37.1, HBS1L, 6:16000000-25000000

TATATCT ACCTCAC ClinVar	Gene(s)	Condition(s)
ATP7B, 1-BP DEL, 2511A	ATP7B	Wilson disease
ATP7B, 3-BP DEL, 3892GTC	ATP7B	Wilson disease
ATP7B, 15-BP DEL, NT-441	ATP7B	Wilson disease

From genotype to phenotype: standards

GENE

- ✓ **Lipocalin 2**
- ✓ 24p3
- ✓ 25 KDa Alpha-2-Microglobulin-Related Subunit Of MMP-9
- ✓ HNL
- ✓ lipocalin 2 (oncogene 24p3)
- ✓ Lipocalin-2
- ✓ Migration-Stimulating Factor Inhibitor
- ✓ MSFI
- ✓ neutrophil gelatinase-associated lipocalin
- ✓ NGAL
- ✓ oncogene 24p3
- ✓ P25
- ✓ Siderocalin

DISEASE

- ✓ **Wilson's disease**
- ✓ Cerebral Pseudosclerosis
- ✓ Copper Storage Disease
- ✓ Hepatic Form of Wilson Disease
- ✓ Hepato-Neurologic Wilson Disease
- ✓ Hepatocerebral Degeneration
- ✓ Hepatolenticular degeneration
- ✓ Kinnier-Wilson Disease
- ✓ Neurohepatic Degeneration
- ✓ Progressive Lenticular Degeneration
- ✓ Pseudosclerosis
- ✓ WD
- ✓ Westphal-Strumpell Syndrome
- ✓ Wilson Disease
- ✓ Wilson Disease, Hepatic Form

From genotype to phenotype: standards

GENE

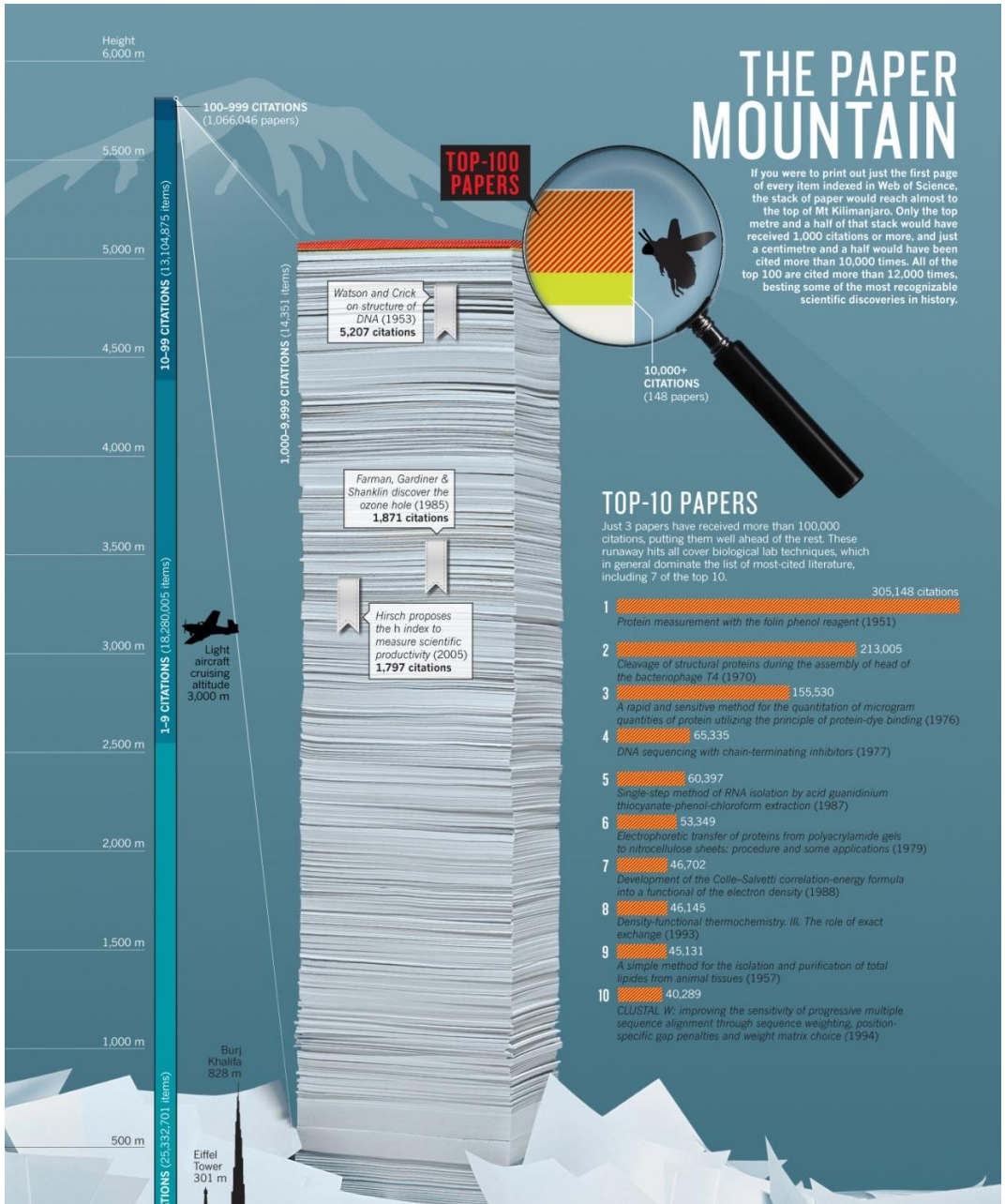
Official Symbol : LCN2
Official Full Name: lipocalin 2
Entrez Gene identifier: 3934
HGNC:6526
Ensembl: ENSG00000148346
UniProtKB Accession: P80188
UniProtKB Entry Name: NGAL_HUMAN
OMIM: 600181

DISEASE

OMIM identifier: 277900
MeSH identifier: D006527
SNOMED CT code : 88518009
NCI Thesaurus code: C84756
ICD 9 code: -
ICD 10 code: E83.01
Disease Ontology identifier: DOID:893
ORPHA number: 905
MEDDRA code: 10019819
UMLS Concept identifier: C0019202

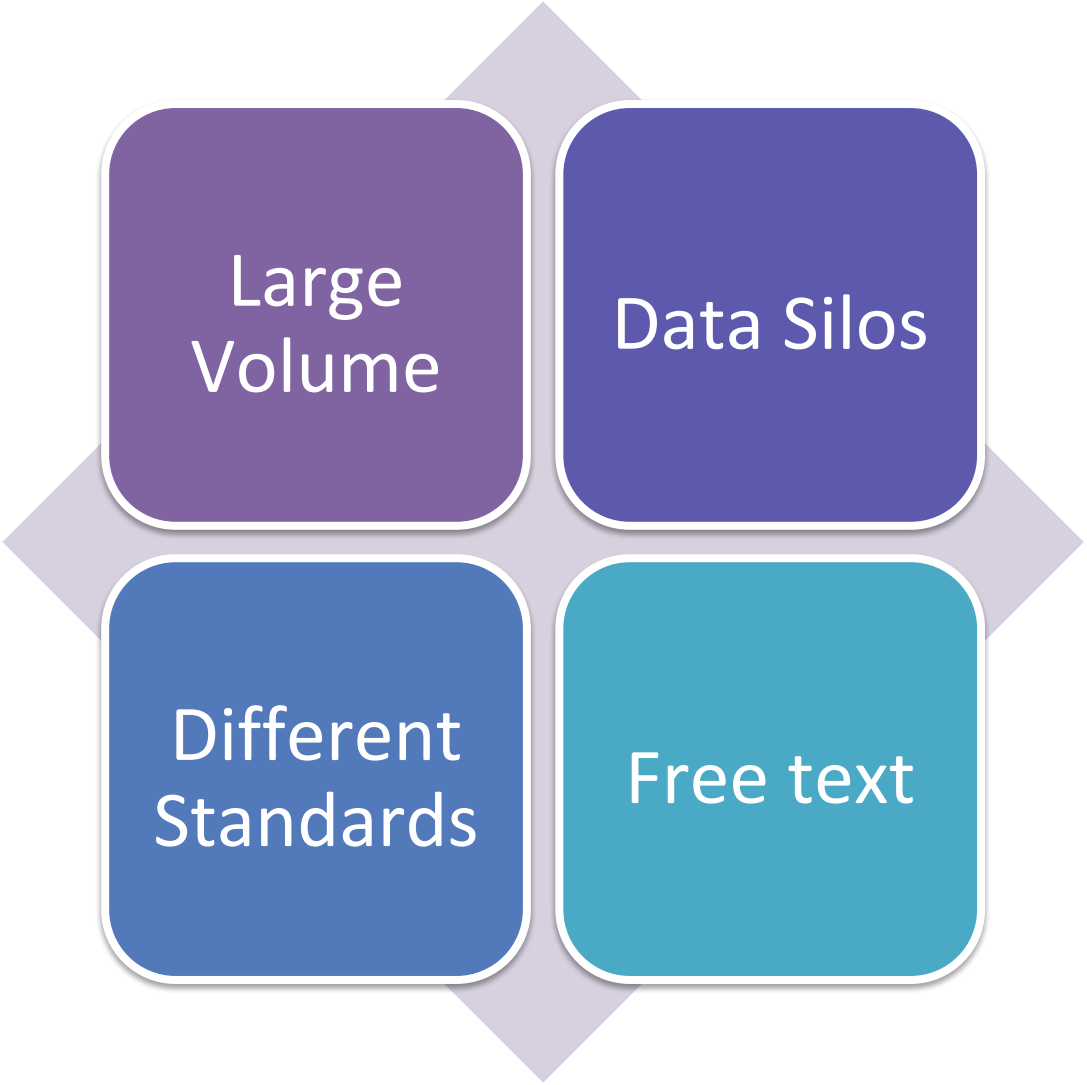
From genotype to phenotype: Free text

- ✓ 25,000 peer-reviewed journals
- ✓ 2.5 10^6 articles published per year
- ✓ 2 papers/minute in life sciences
- ✓ ~ half of the items comprise works that have been cited only once, if at all!!



Van Noorden, et al. 2014 doi:10.1038/514550a

From genotype to phenotype

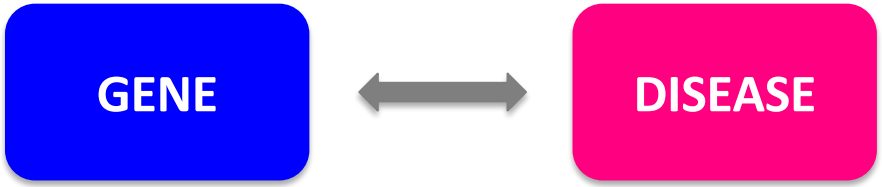


The DisGeNET platform

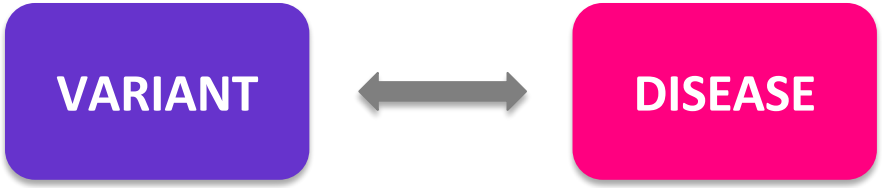
- A catalog of genes and variants associated to human diseases, abnormal phenotypes, and traits.
- Developed by integration of different public resources with data obtained from text mining the scientific literature
- Freely available at: http://www.disgenet.org

The DisGeNET platform

Gene-Disease Association (GDA)

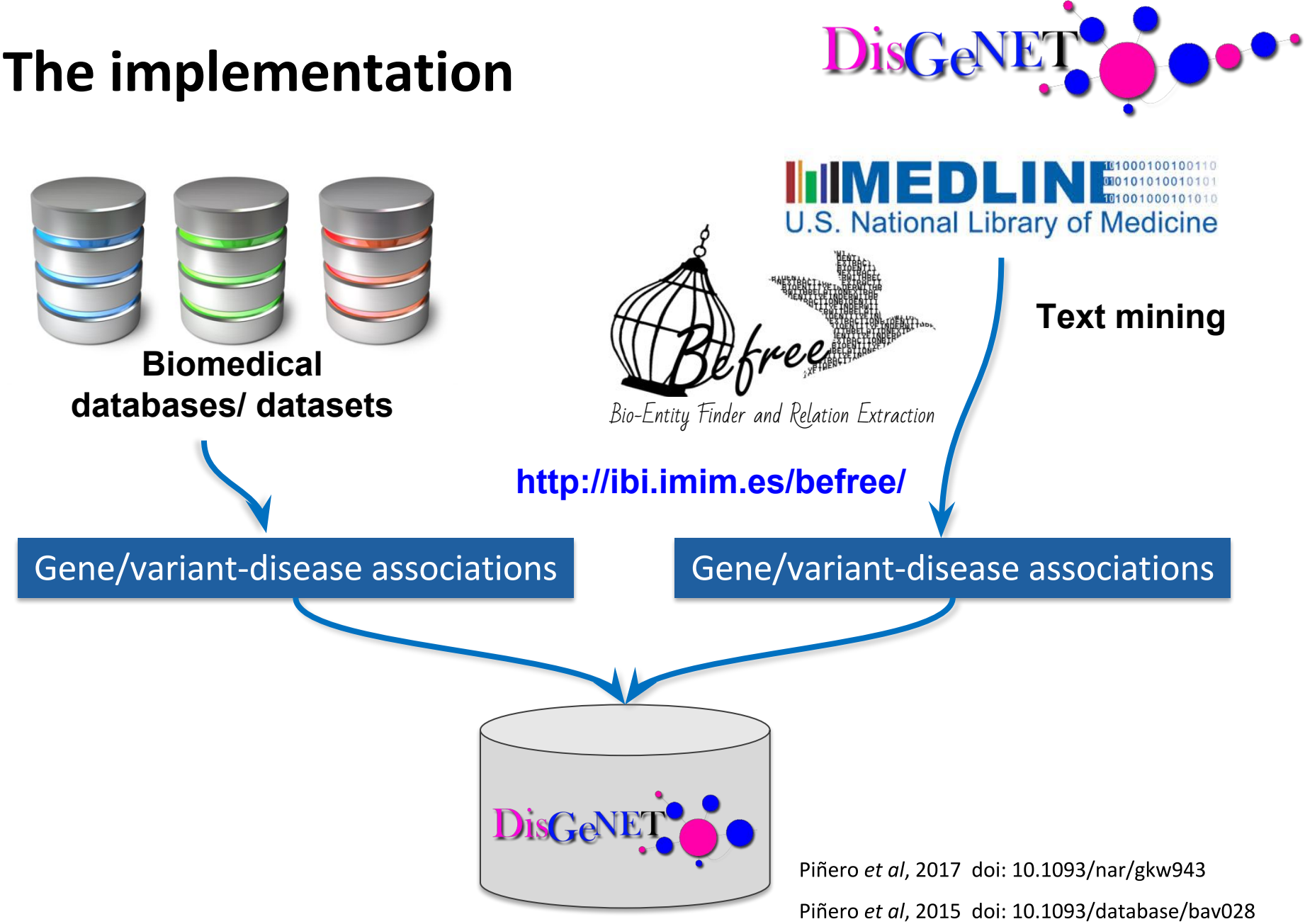


Variant-Disease Association (VDA)



- ✓ Provenance
- ✓ Standards
- ✓ Tools

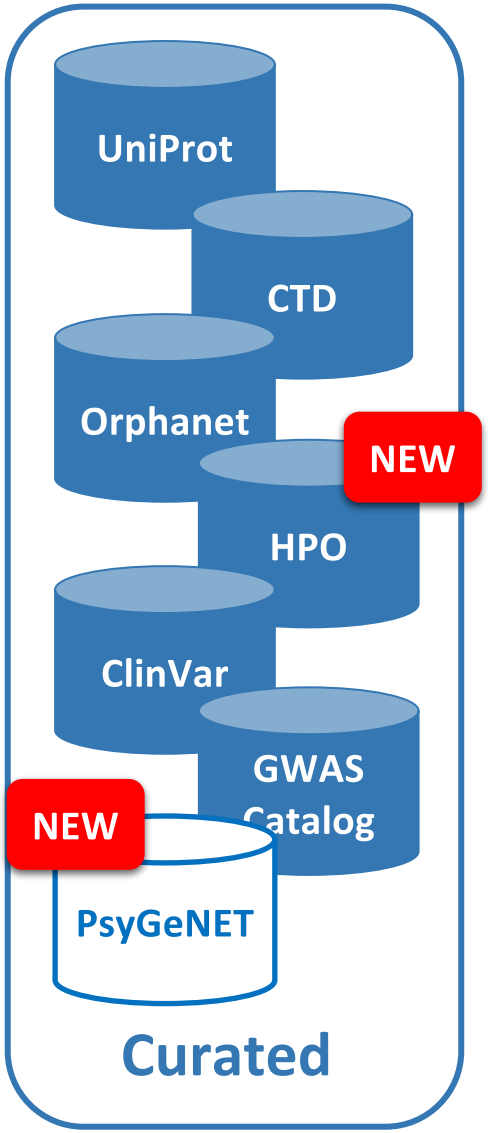
The implementation



Piñero *et al*, 2017 doi: 10.1093/nar/gkw943

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

The data sources



UniProt is a resource of protein sequence and functional information.

The **Comparative Toxicogenomics Database** focuses on how environmental exposures affect human health.

Orphanet is the portal for rare diseases and orphan drugs

The **GWAS Catalog** is a curated collection of published GWA studies assaying >100,000 SNPs with p-values < 1.0×10^{-5}

ClinVar aggregates information about genomic variation and its relationship to human health

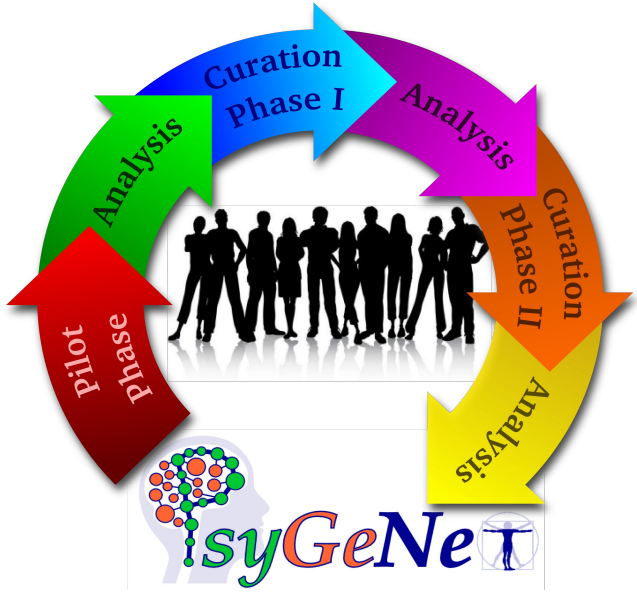
The **Human Phenotype Ontology** provides a standardized vocabulary of phenotypic abnormalities in human diseases

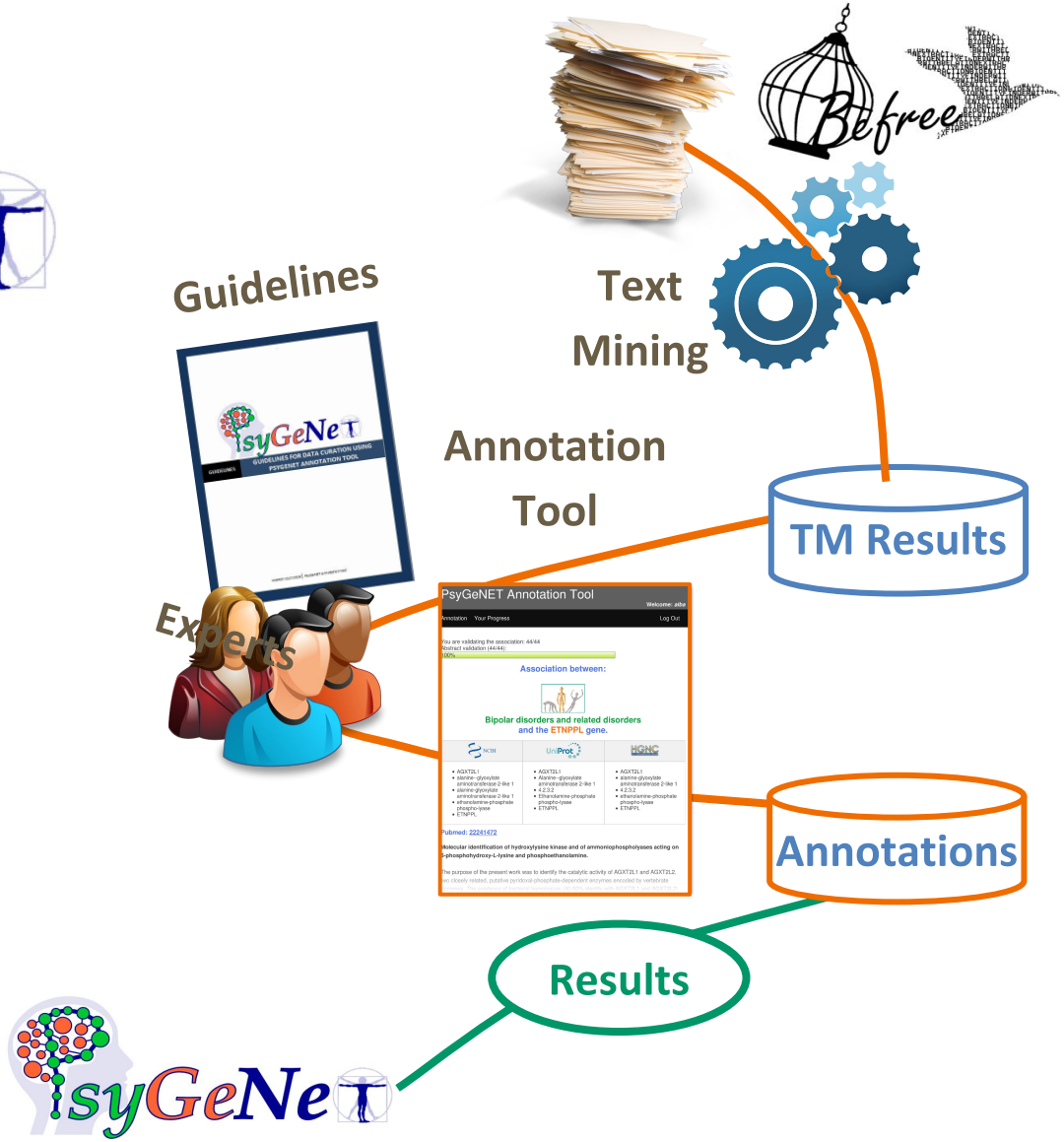
The data sources

Psychiatric disorders **Ge**ne association **NET**work

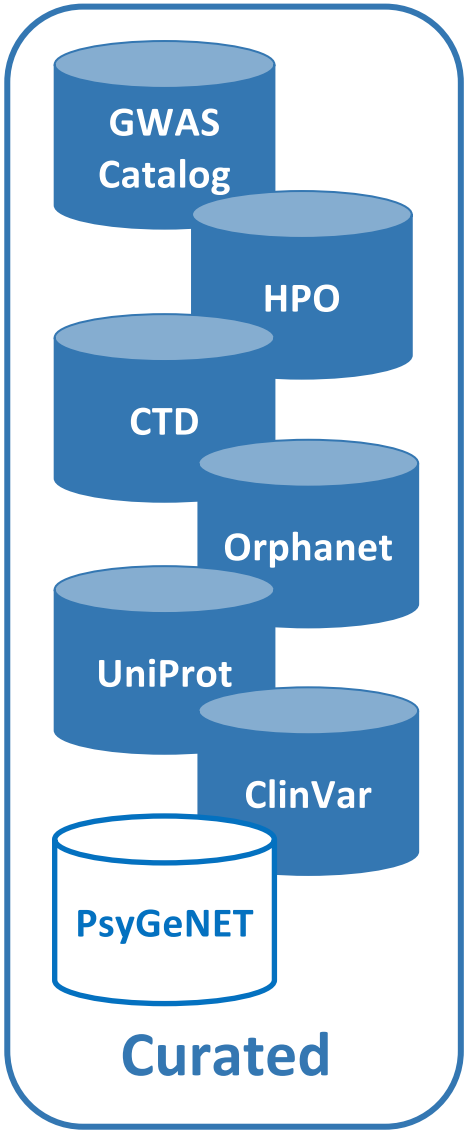
- ✓ Alcohol use disorders
- ✓ Bipolar disorders
- ✓ Depressive disorders
- ✓ Schizophrenia
- ✓ Cocaine use disorders
- ✓ Substance induced depressive disorder
- ✓ Cannabis use disorders
- ✓ Substance induced psychosis

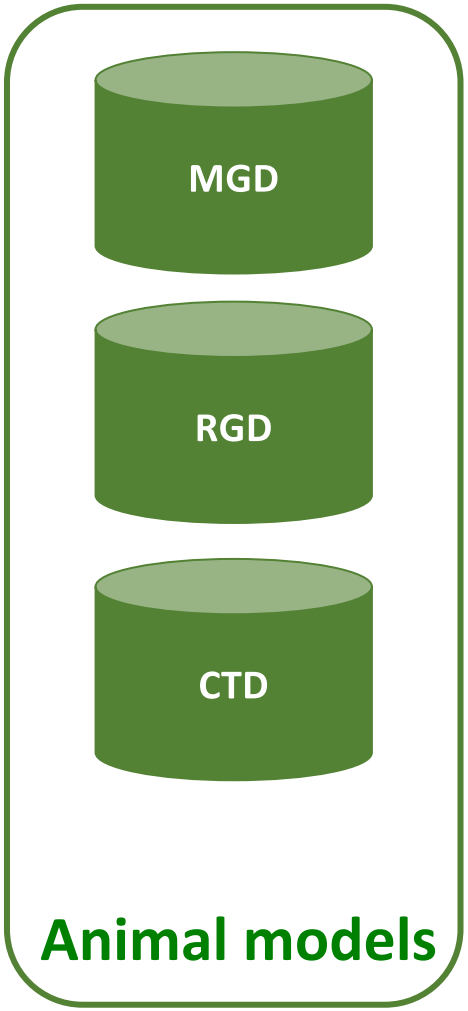
The data sources





The data sources

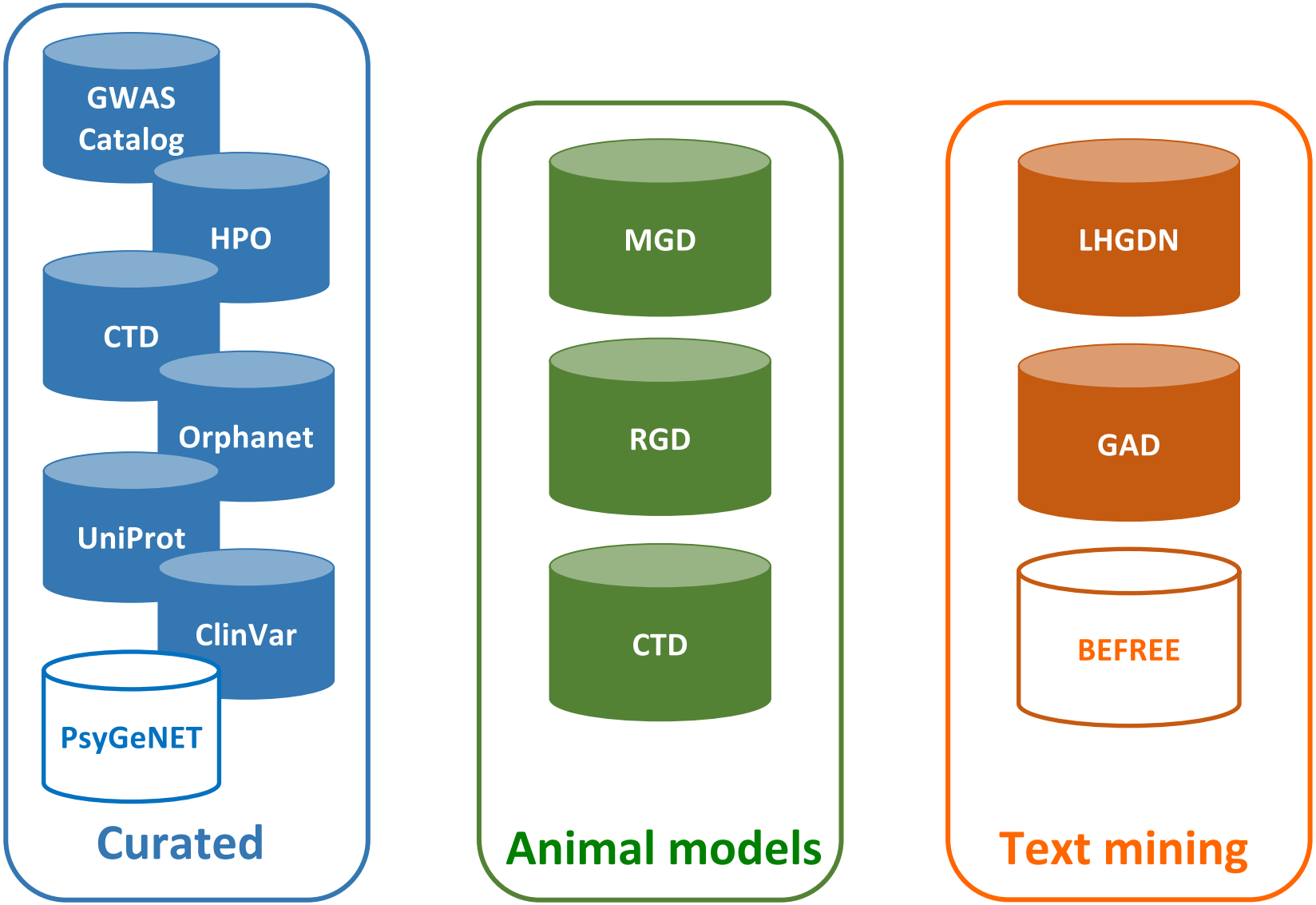




The **Rat Genome Database** is the premier site for genetic, genomic, phenotype, and disease data generated from rat research.

Mouse Genome Informatics is the international resource for the laboratory mouse, providing integrated genetic, genomic, and biological data to facilitate the study of human health and disease.

The data sources



The data sources

Befree Bio-Entity Finder and RElation Extraction

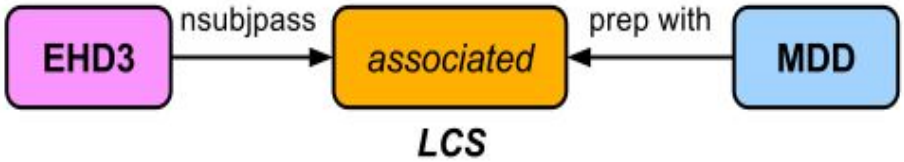
GENE CANDIDATE

DISEASE CANDIDATE

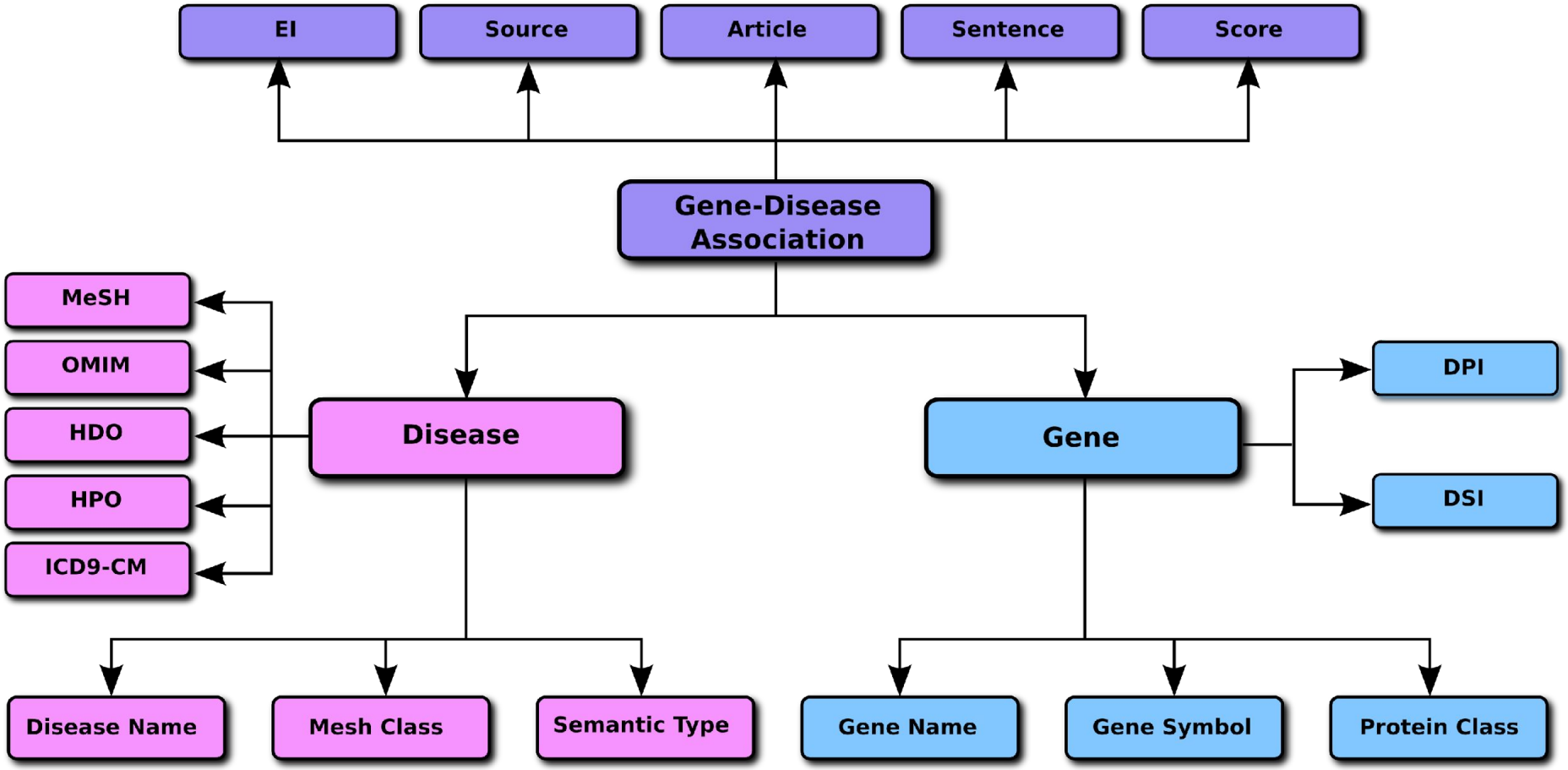
Of the 16 genes tested, **EHD3** and **FREM3** were associated with **MDD** in the Chinese population.

Gene ID: 30845
EH-domain containing 3

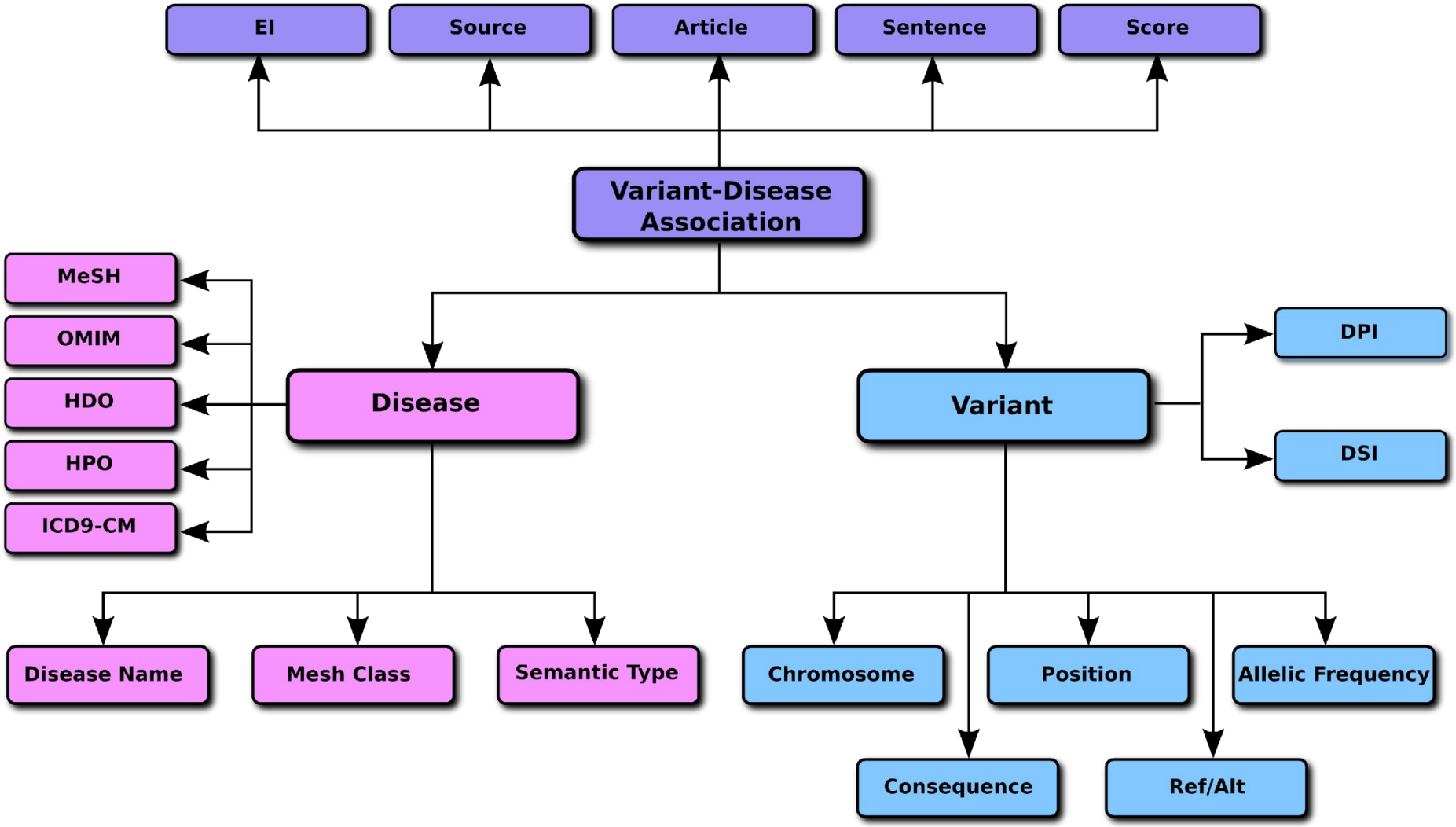
Disease ID: C1269683
Major depressive disorder



Data homogenization and standardization



Data homogenization and standardization

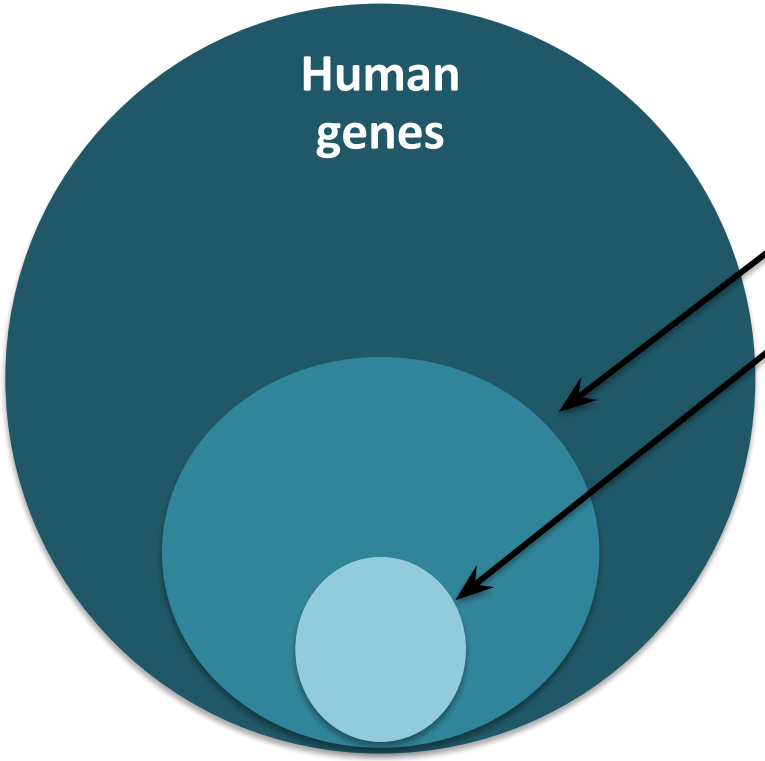


Statistics (v. 5.0)

Gene-Disease Associations (GDAs)

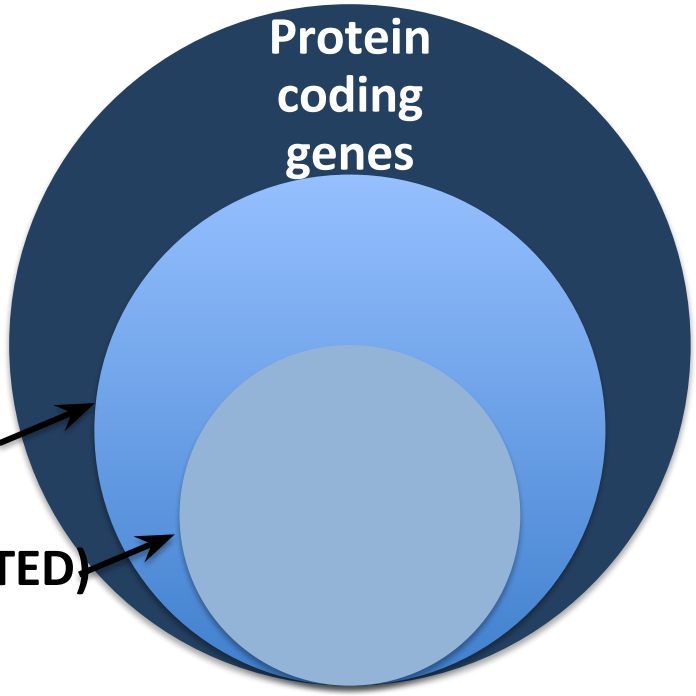
Source	Genes	Diseases*	Associations
Curated	8,948	13,074	130,821
Animal Models	2,300	1,943	6,455
Text mining	16,119	12,604	448,732
All	**17,074**	**20,370**	**561,119**

Statistics (v. 5.0)



Disease genes 28 % (ALL)

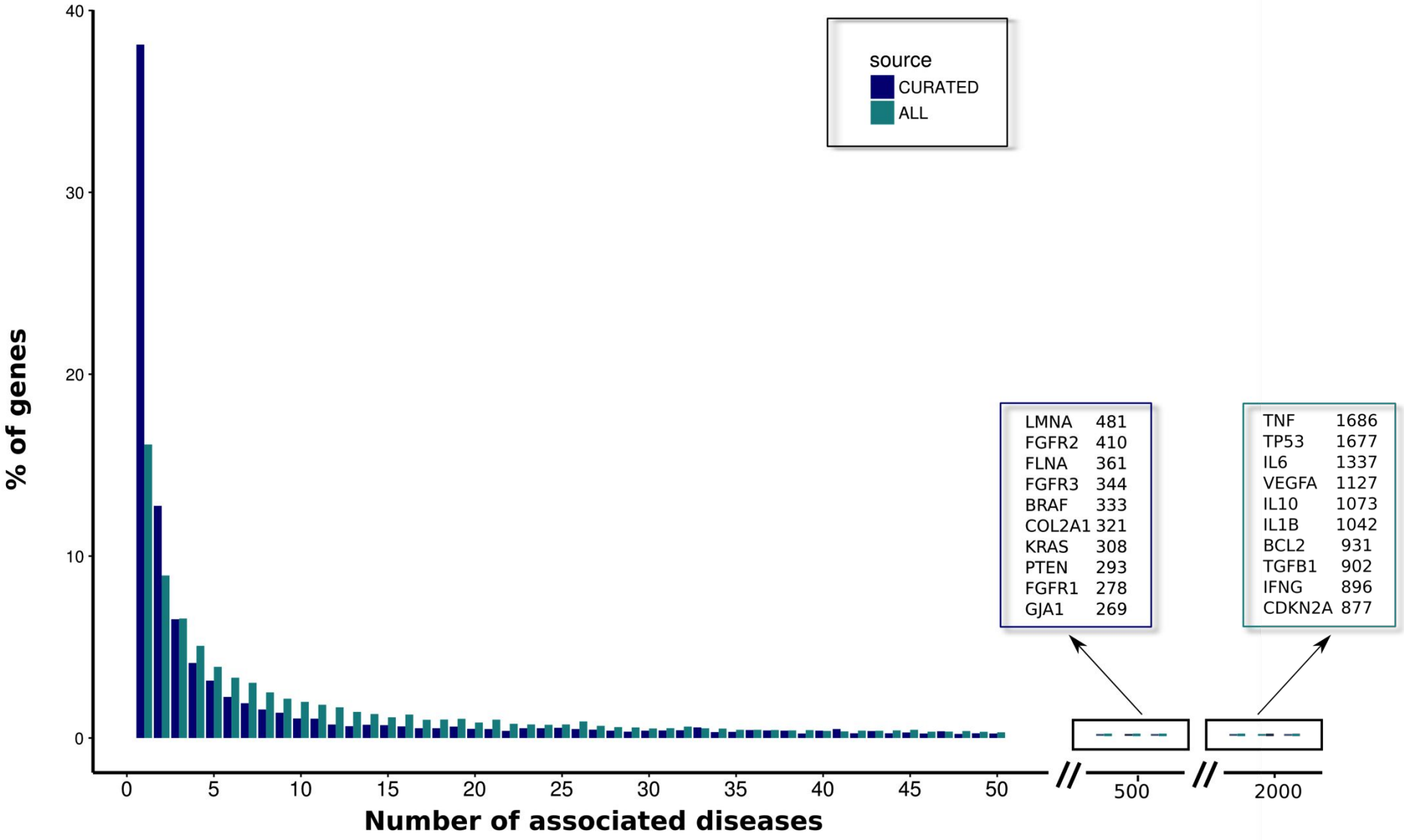
Disease genes 15 % (CURATED)



Protein coding disease genes 71 % (ALL)

Protein coding disease genes 37 % (CURATED)

Statistics (v. 5.0)



Statistics (v. 5.0)

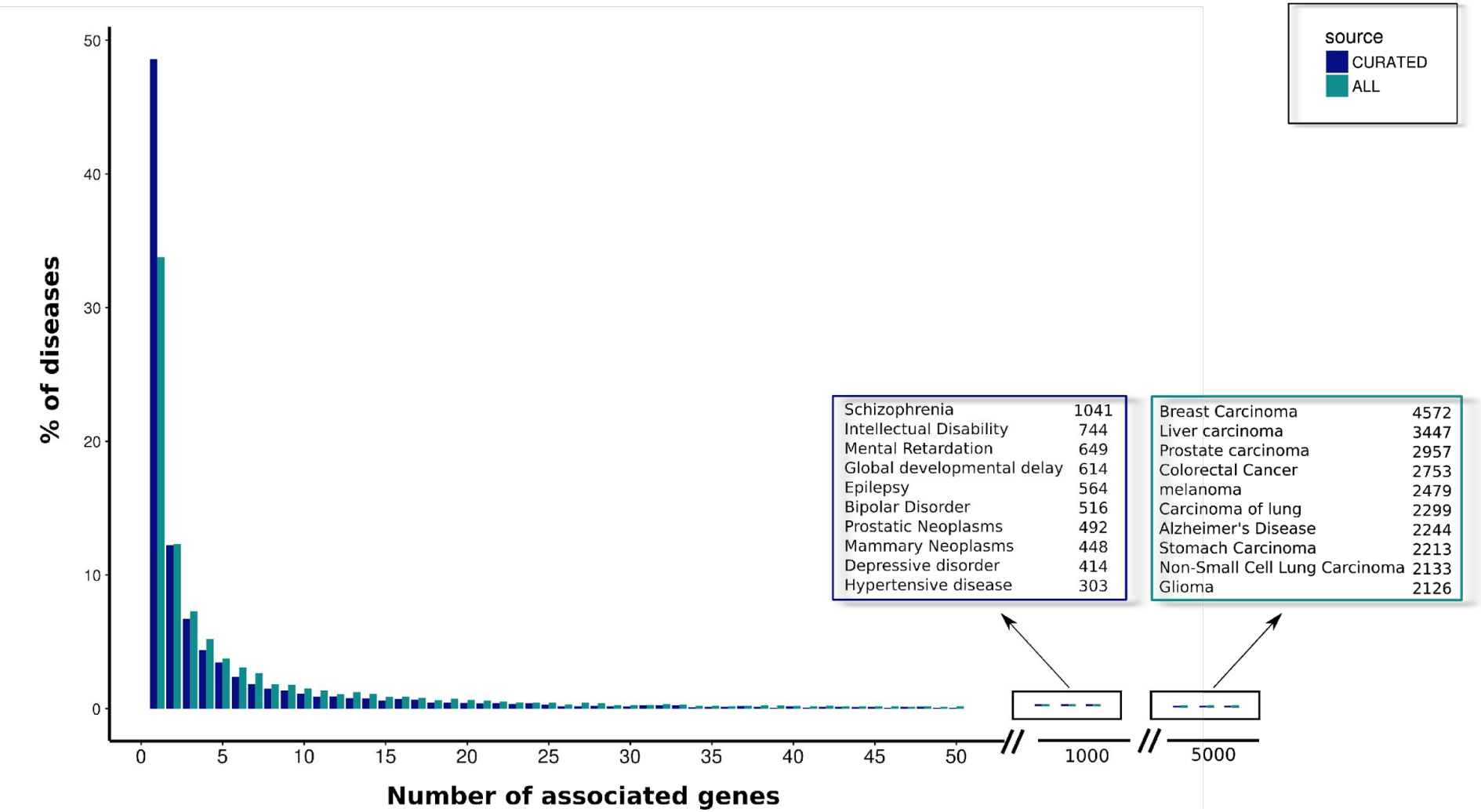
Gene-Disease Associations (GDAs)

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**

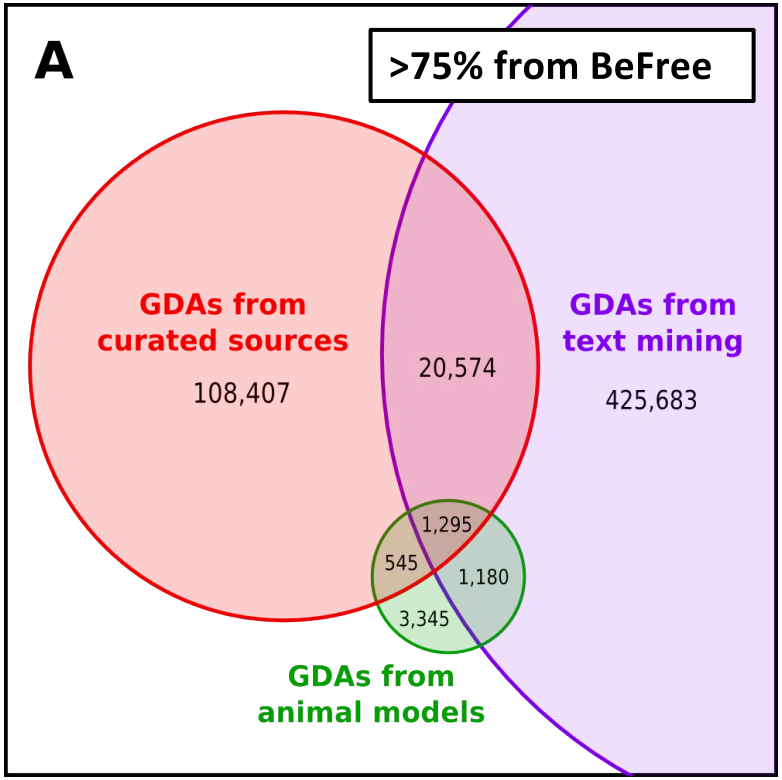
Statistics (v. 5.0)

- Abnormal phenotypes, signs and symptoms, traits 5,200
  - Inflammation
  - Seizures
  - Overweight
  - Hemoglobin measurement NEW
- Diseases 15,684
  - Breast carcinoma
  - Diabetes Mellitus Type II
- Disease groups 798
  - Cardiovascular Diseases
  - Autoimmune Diseases
  - Neurodegenerative Diseases

Statistics (v. 5.0)



Statistics (v. 5.0)



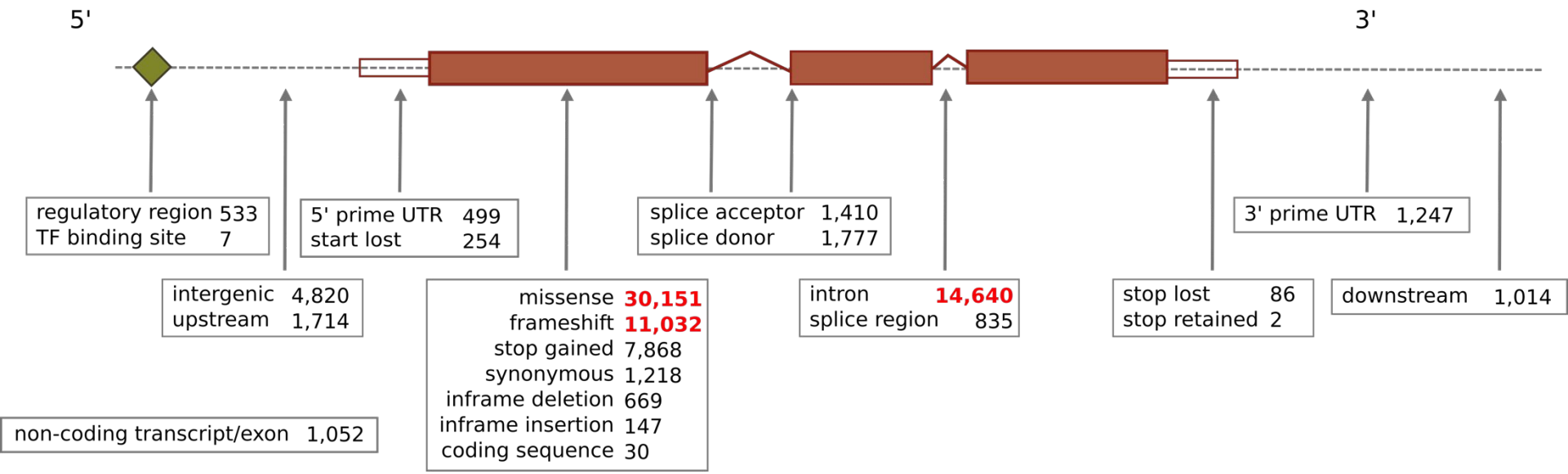
Statistics (v. 5.0)

Variant-Disease Associations (VDAs)

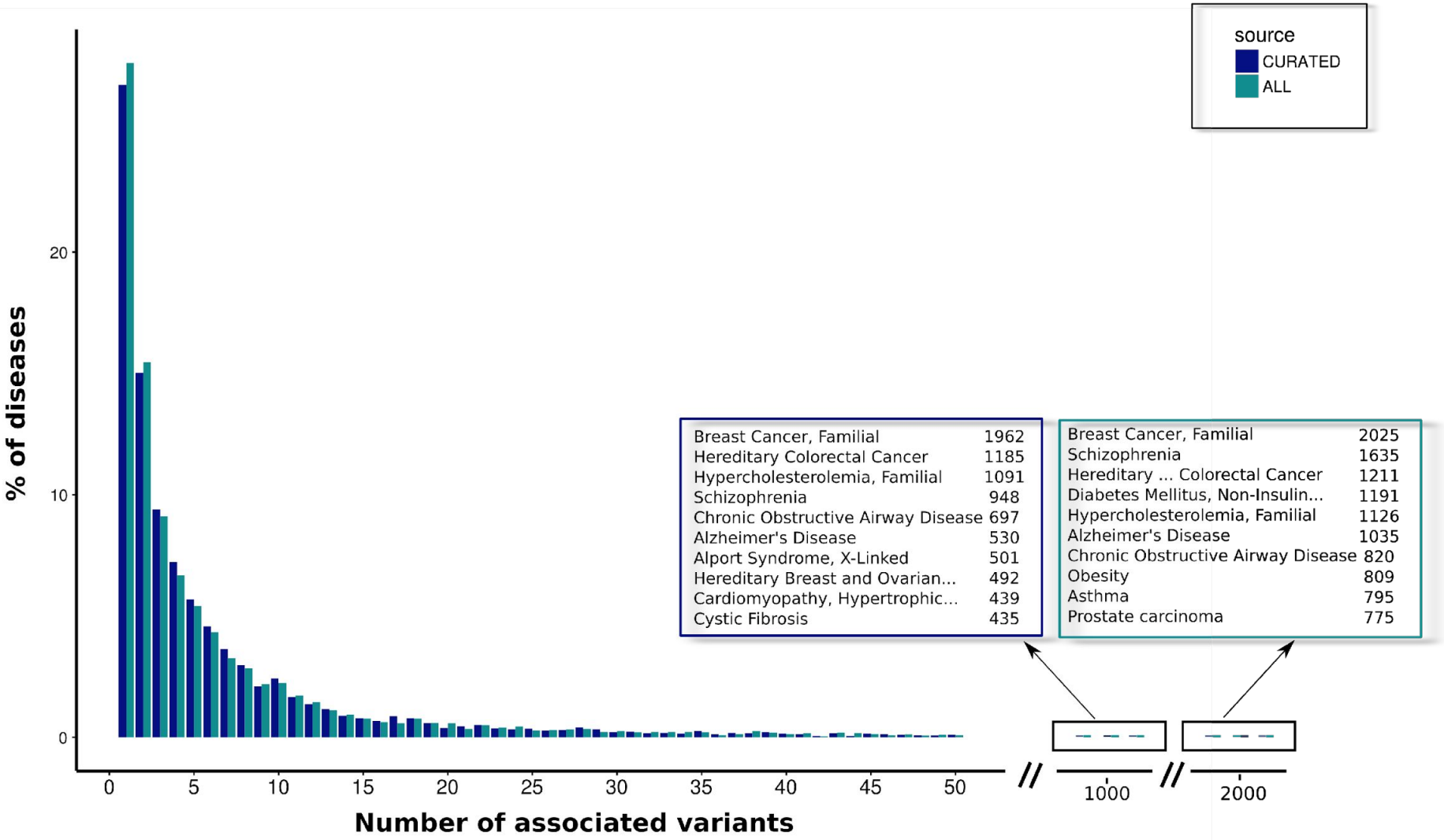
Source	Variants	Diseases	Associations
Curated	66,903	6,388	83,582
Text mining	24,455	4,432	57,331
All	**83,002**	**9,169**	**135,588**

Statistics (v. 5.0)

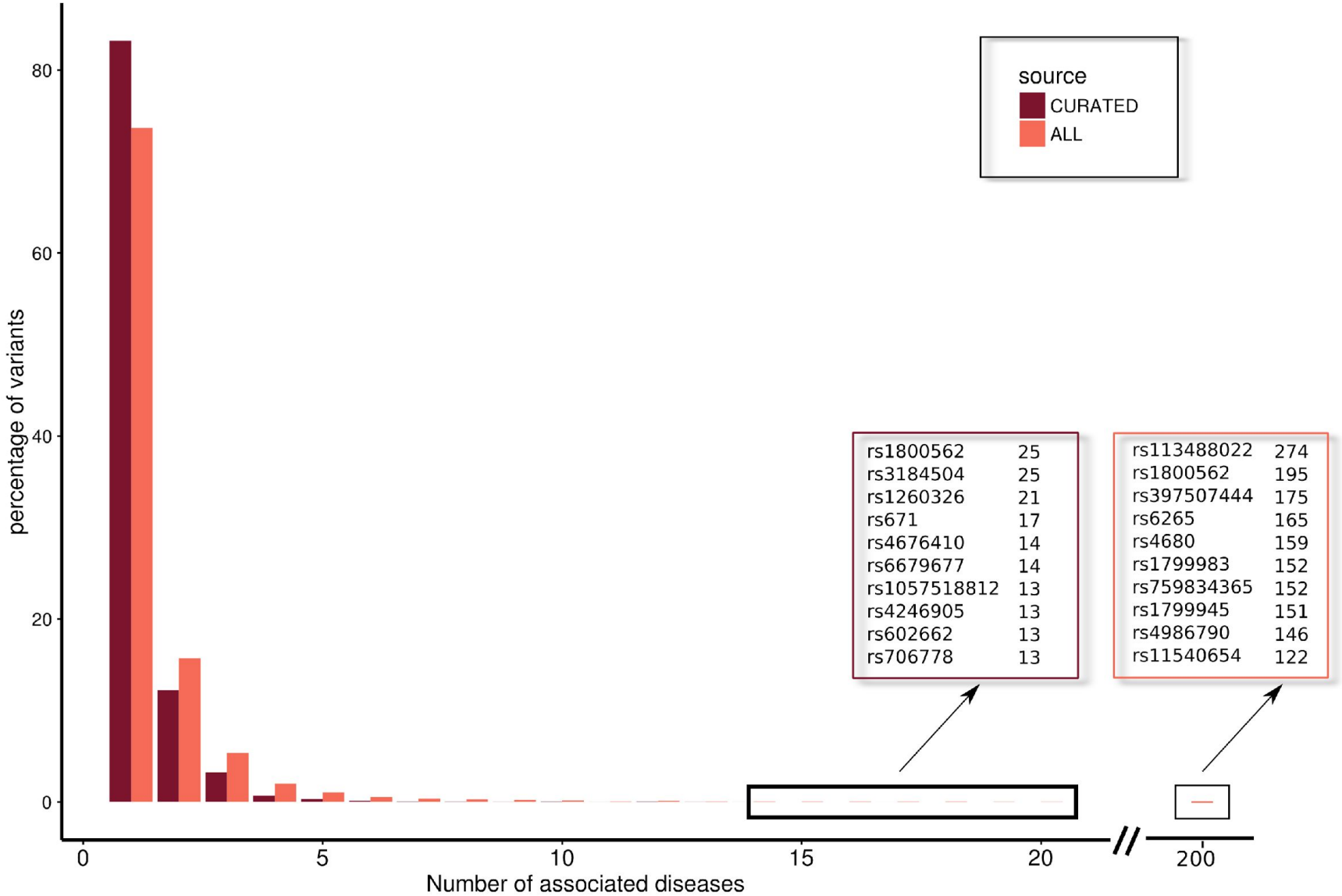
Distribution of disease- associated variants according to the variant consequence type (VEP)



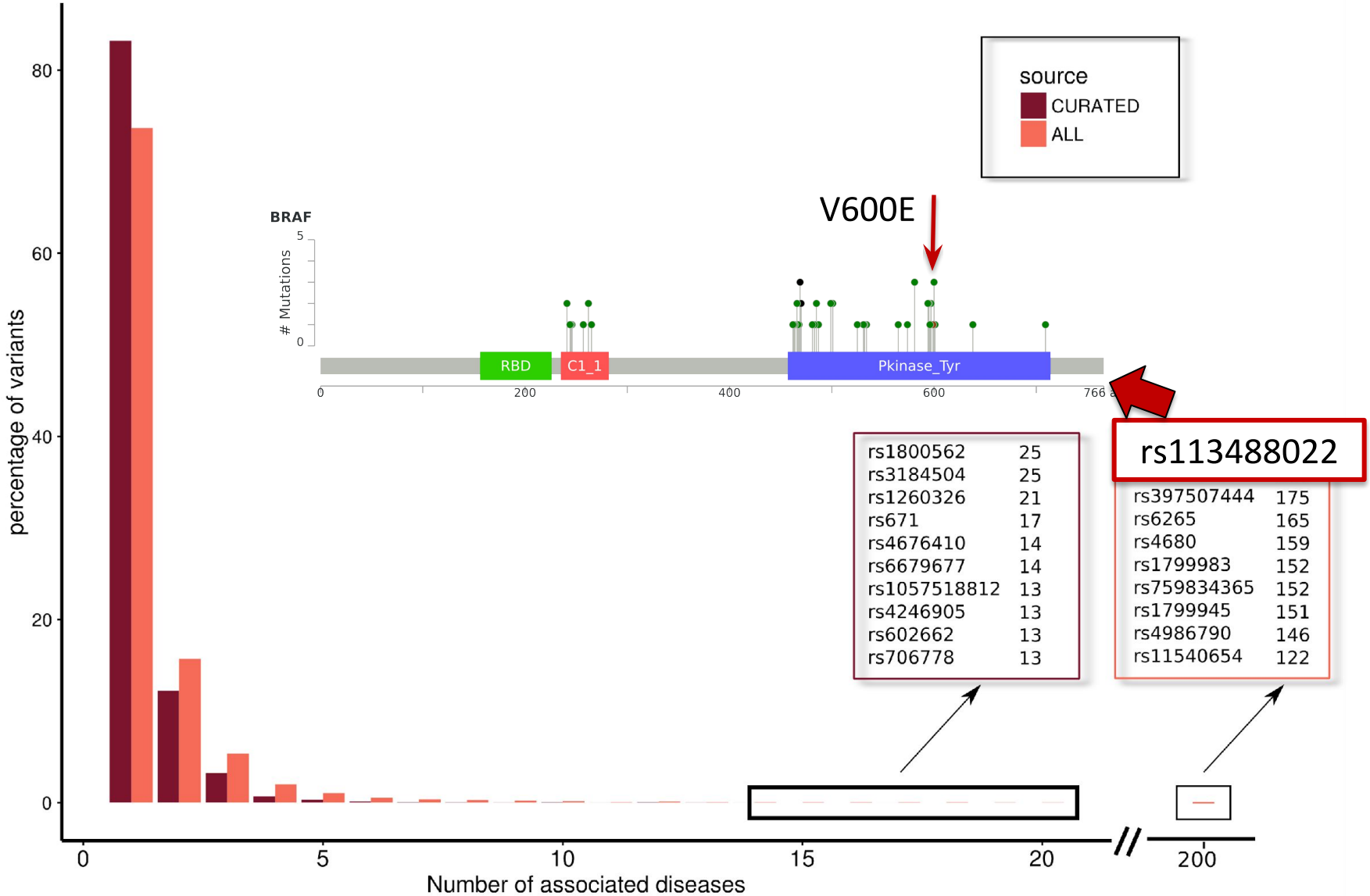
Statistics (v. 5.0)



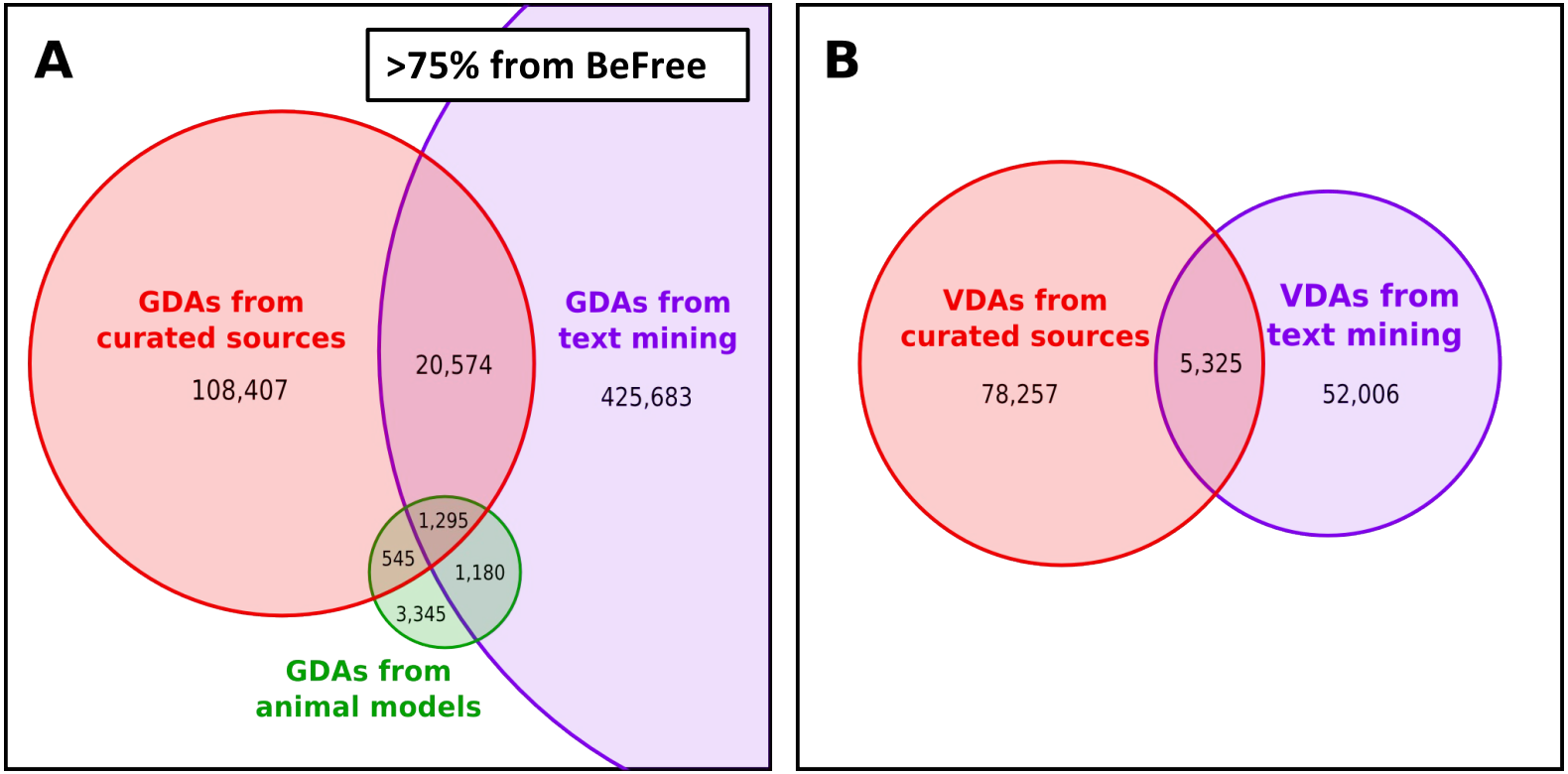
Statistics (v. 5.0)



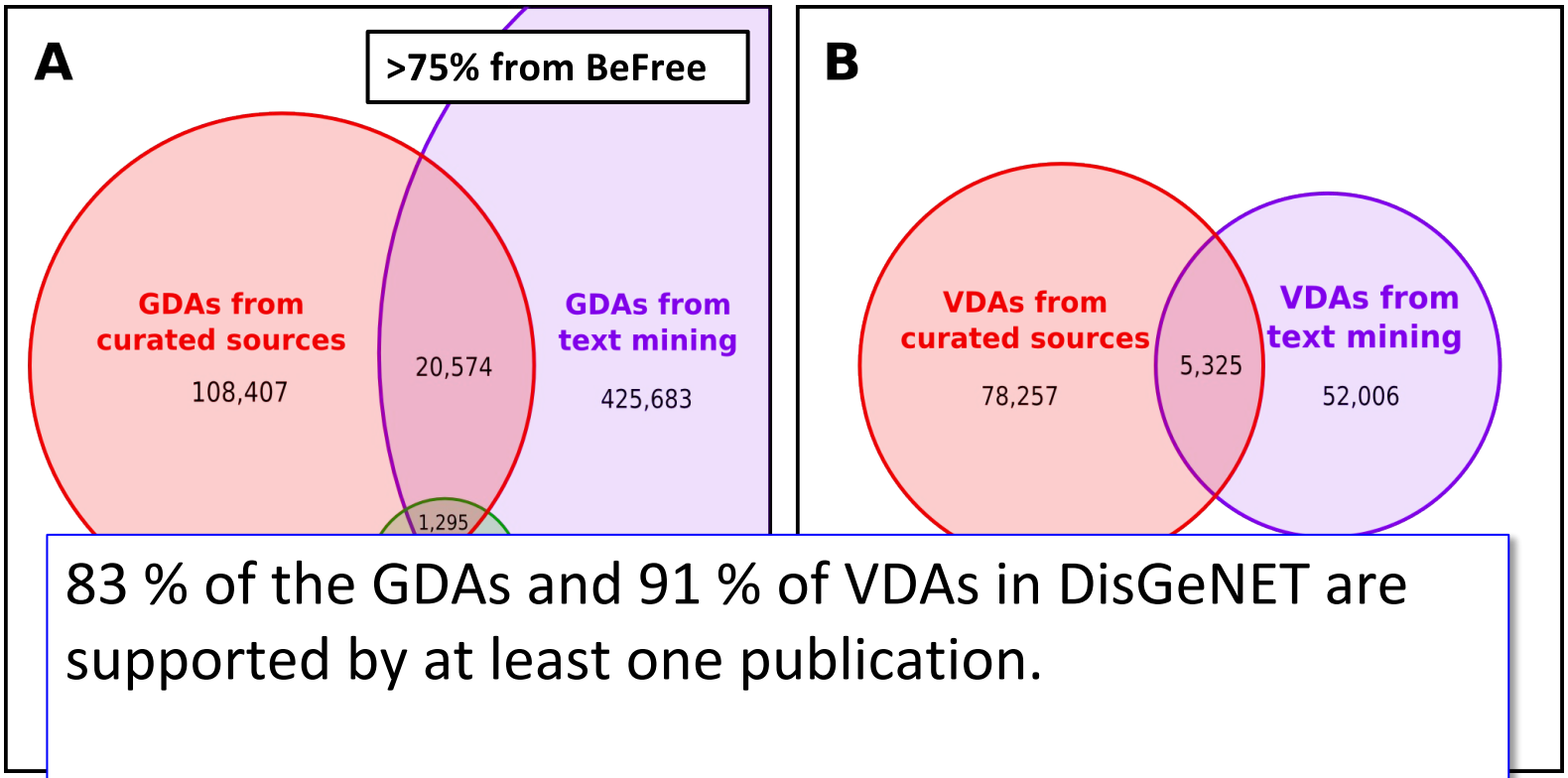
Statistics (v. 5.0)



Statistics (v. 5.0)



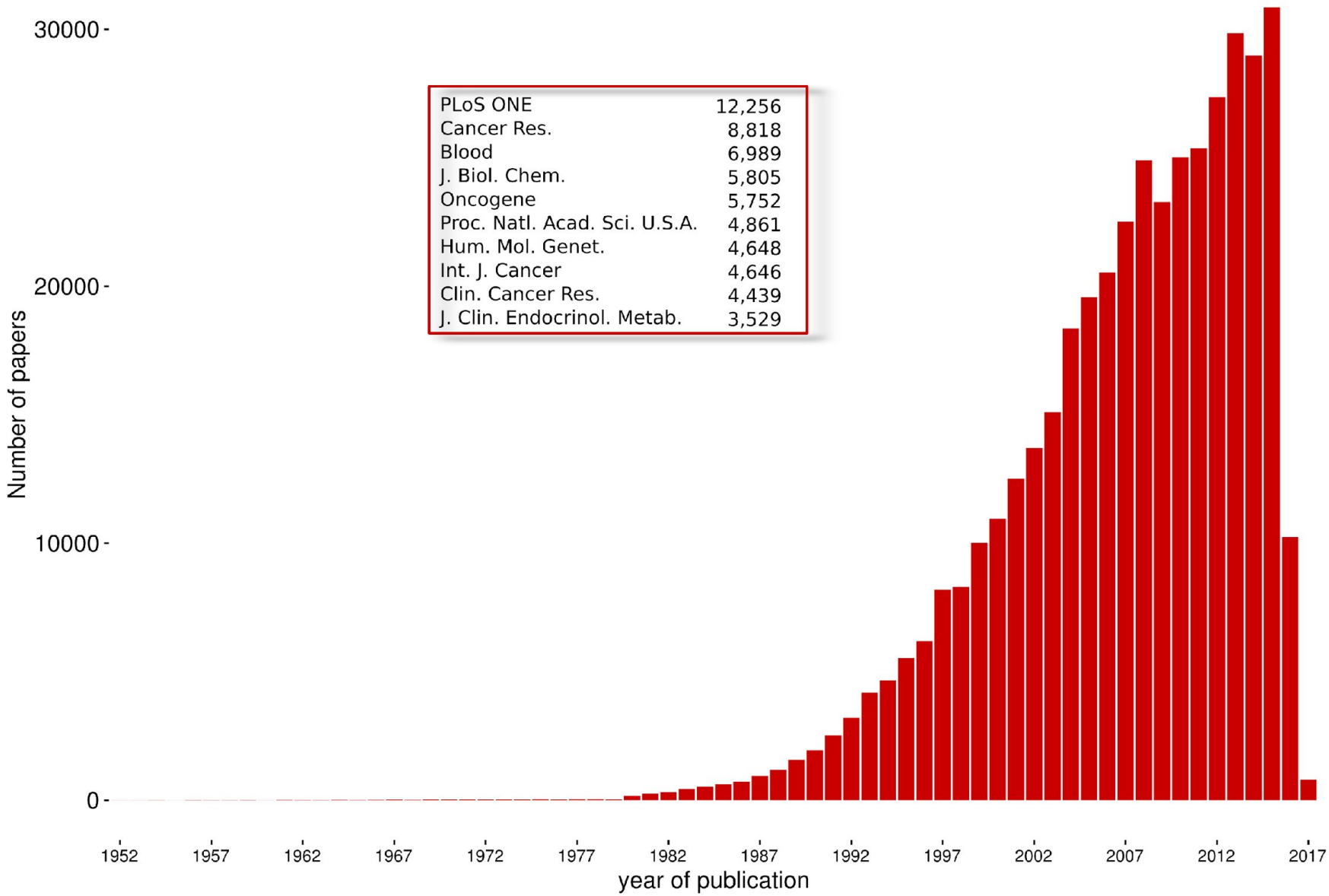
Statistics (v. 5.0)



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

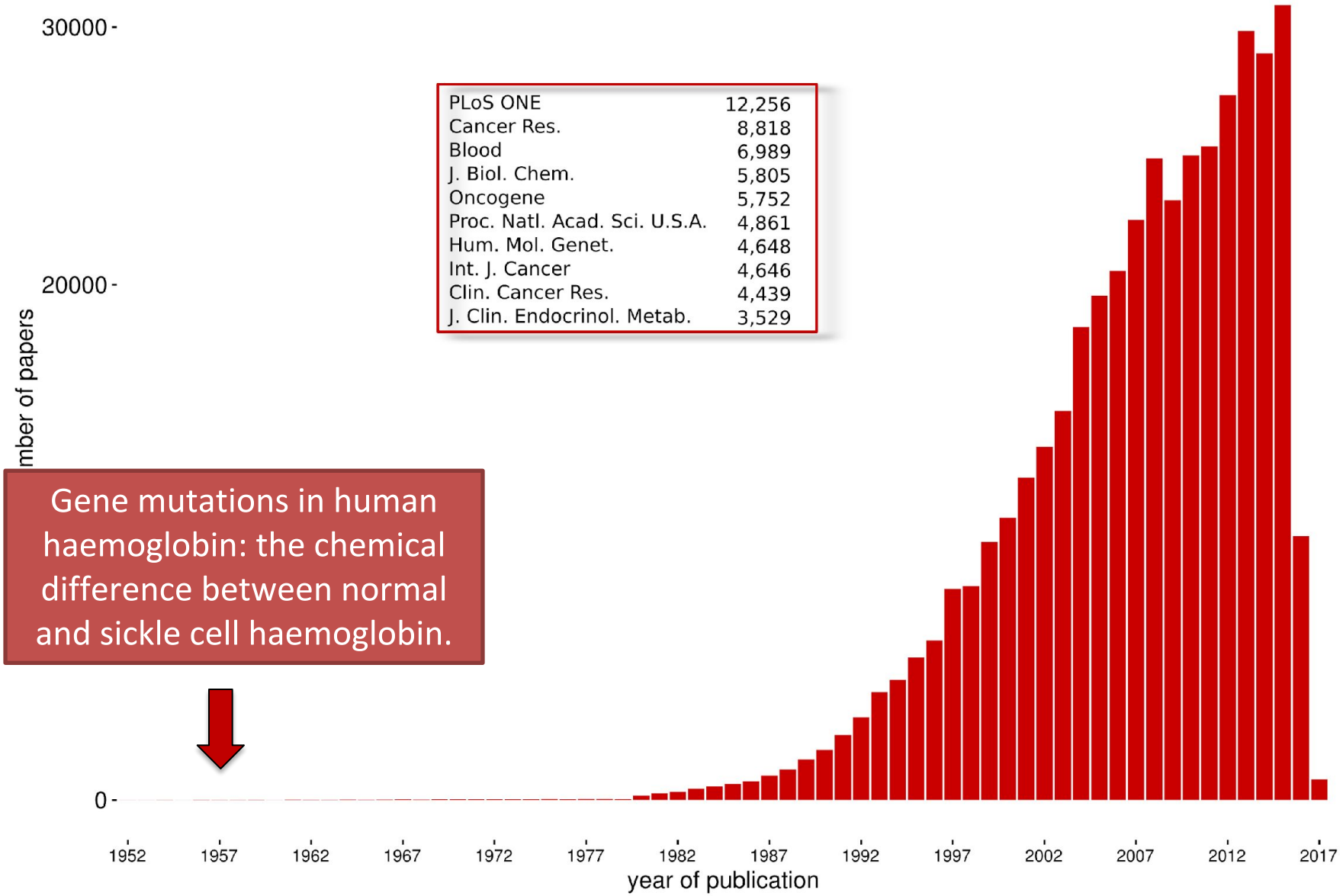
DisGeNET contains a corpus of 422,027 publications with information on GDAs and VDAs.

Statistics (v. 5.0)



PLoS ONE	12,256
Cancer Res.	8,818
Blood	6,989
J. Biol. Chem.	5,805
Oncogene	5,752
Proc. Natl. Acad. Sci. U.S.A.	4,861
Hum. Mol. Genet.	4,648
Int. J. Cancer	4,646
Clin. Cancer Res.	4,439
J. Clin. Endocrinol. Metab.	3,529

Statistics (v. 5.0)



PLoS ONE	12,256
Cancer Res.	8,818
Blood	6,989
J. Biol. Chem.	5,805
Oncogene	5,752
Proc. Natl. Acad. Sci. U.S.A.	4,861
Hum. Mol. Genet.	4,648
Int. J. Cancer	4,646
Clin. Cancer Res.	4,439
J. Clin. Endocrinol. Metab.	3,529

Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin.

Tools for prioritization

DisGeNET

DISEASE

- ✓ UMLS semantic types
- ✓ MeSH disease class
- ✓ Disease Ontology top level class
- ✓ Human Phenotype Ontology top level class
- ✓ Group, disease, phenotype

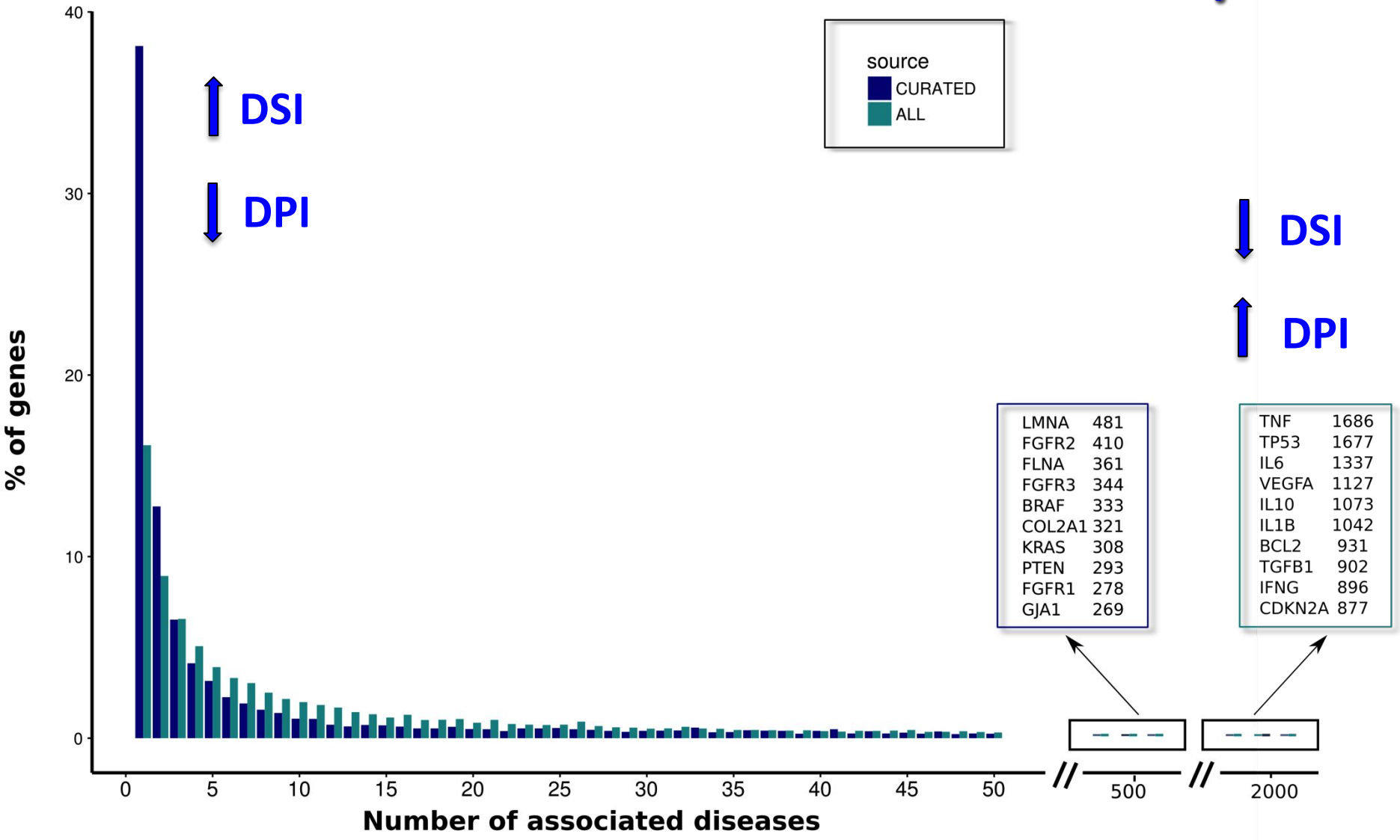
Tools for prioritization

VARIANT

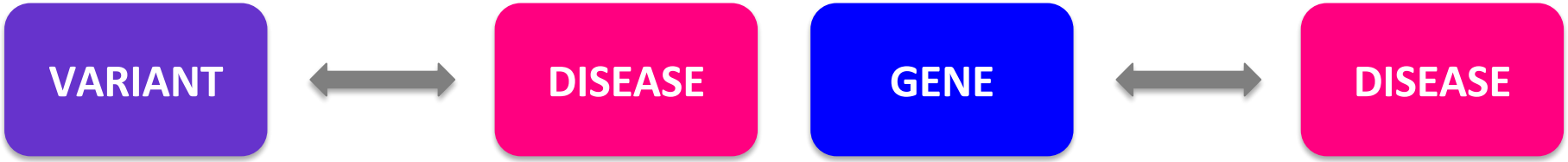
GENE

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

Tools for prioritization



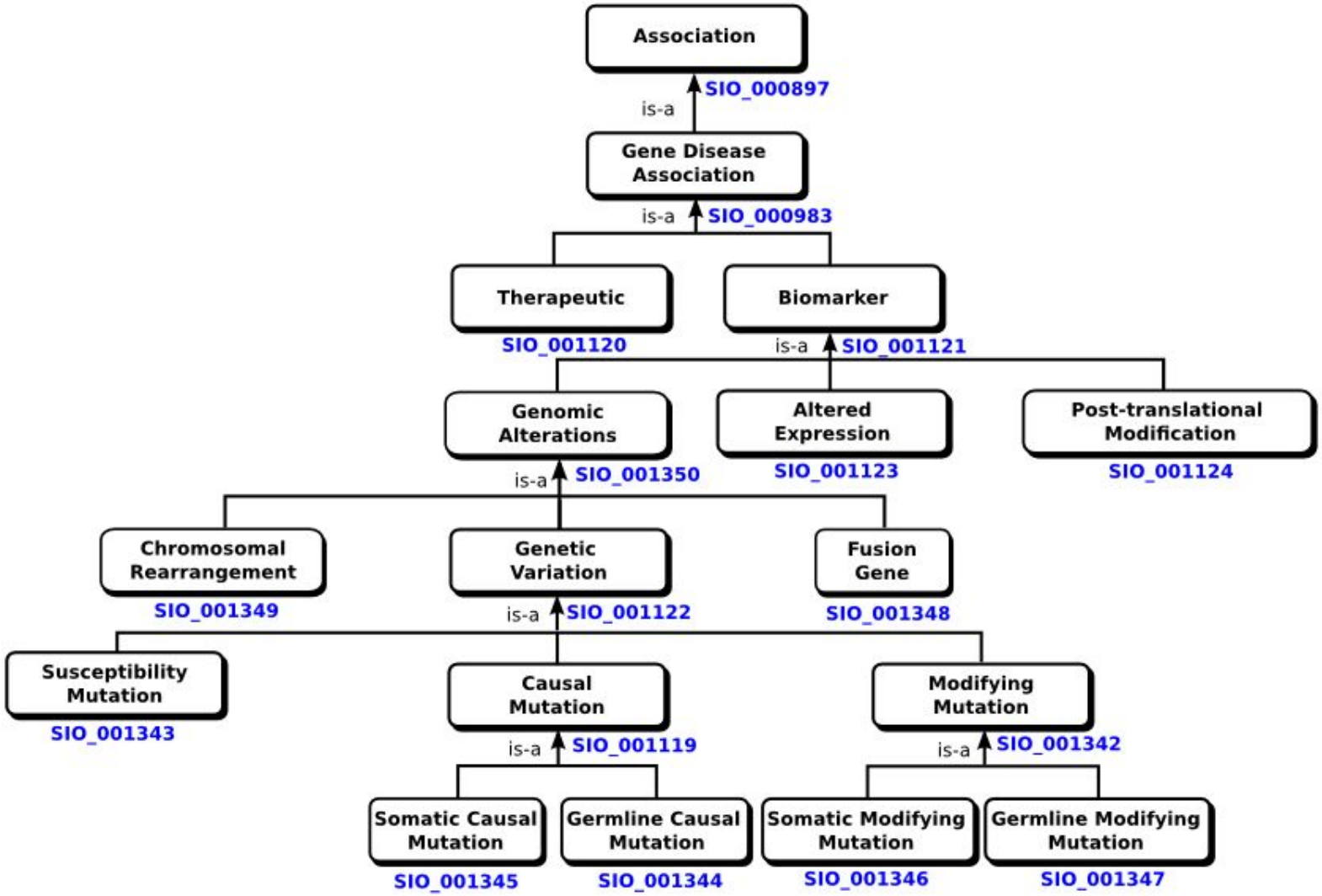
Tools for prioritization



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

Tools for prioritization

DisGeNET association type ontology



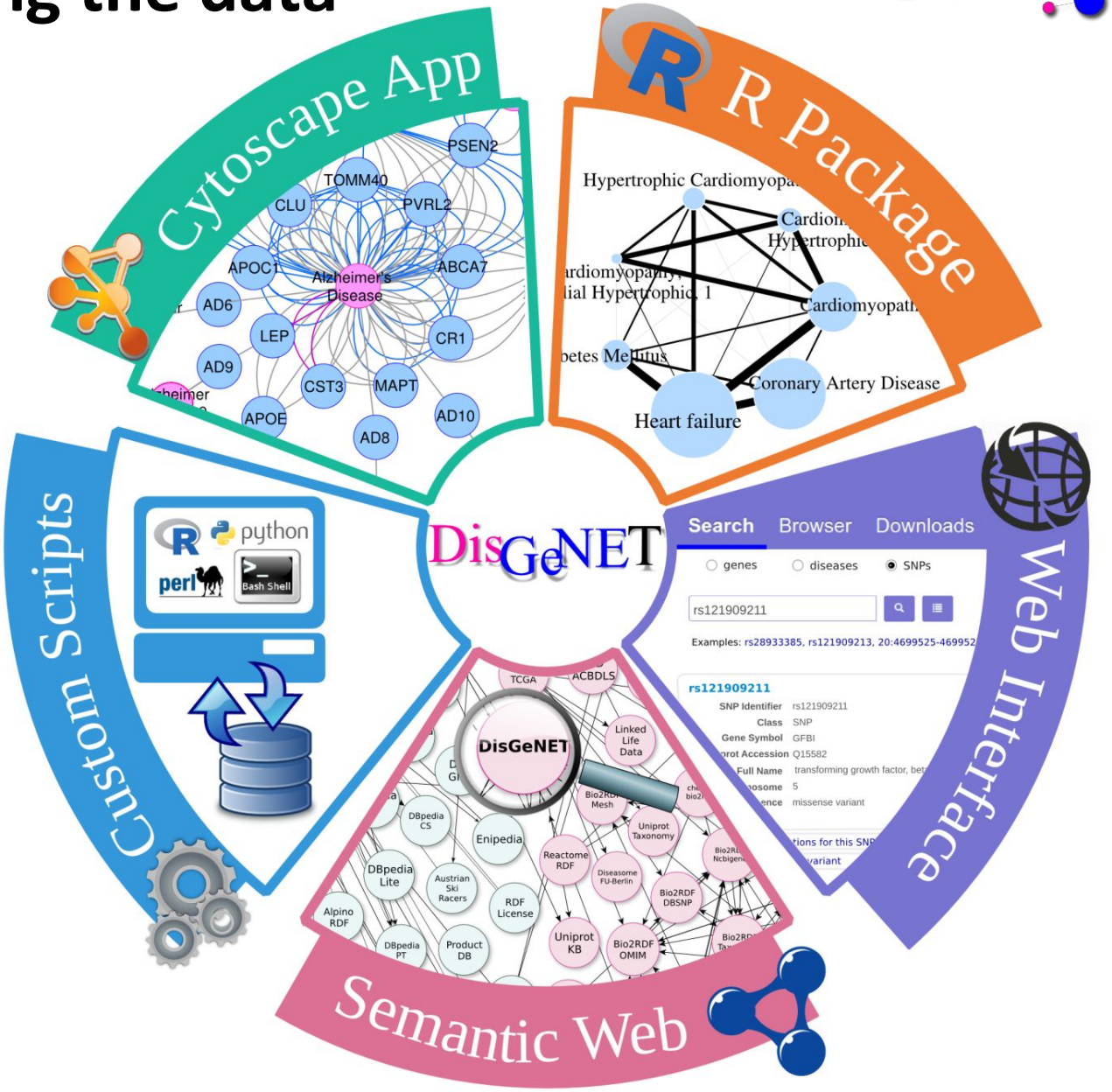
http://sio.semanticscience.org

Tools for prioritization

Top scored genes for Cystic Fibrosis

Gene	Number of diseases	DisGeNET Score	EI	DPI	DSI	Number of PMIDs	Number of SNPs
CFTR	308	1	0,948	0,857	0,422	1596	436
TGFB1	902	0,433	1	0,964	0,314	17	3
DCTN4	83	0,401	1	0,679	0,555	2	0
SCNN1B	69	0,293	1	0,5	0,573	13	0
SCNN1G	65	0,212	1	0,464	0,579	6	1
SCNN1A	64	0,207	0,833	0,536	0,581	9	1
TNFRSF1A	318	0,205	1	0,893	0,419	1	0
CLCA4	18	0,204	1	0,321	0,709	4	0
STX1A	54	0,200	1	0,679	0,598	1	2

Accessing the data



Summary

- 13 sources integrated

- Standardization using community driven ontologies and vocabularies
- Linked Open Data

Data Silos

Different Standards

DisGeNET

Large Volume

Free text

- Variety of prioritization tools
- Flexible data access

- Text mining from scientific publications

Befree

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

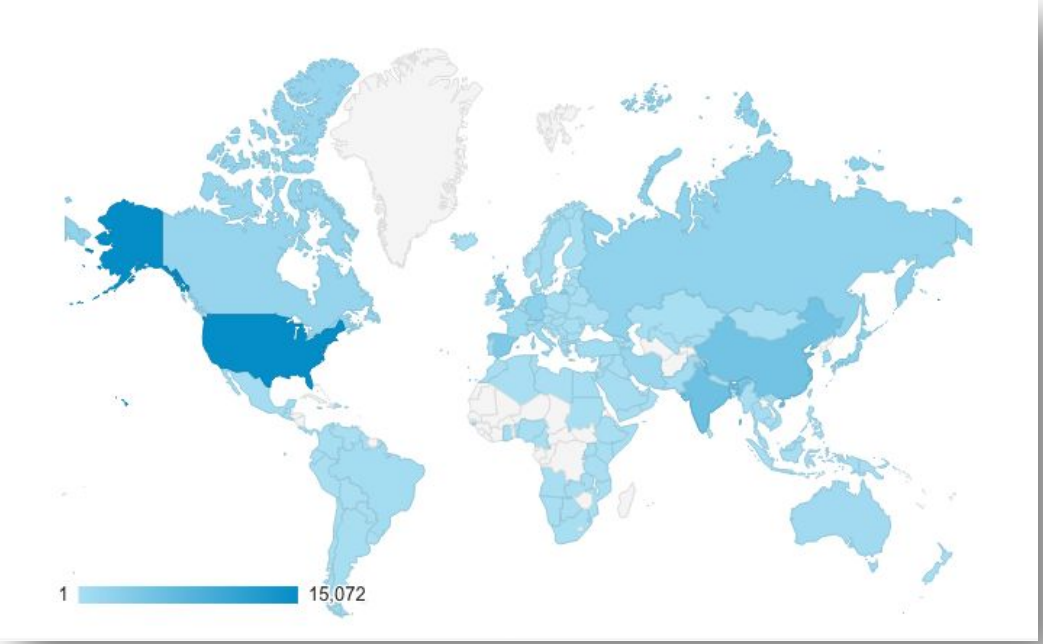
How is DisGeNET used?

- ✓ Investigate the molecular basis of specific diseases
- ✓ Annotate lists of genes/variants produced by different types of omics and sequencing protocols
- ✓ Validate disease genes prediction methods
- ✓ Understand disease mechanisms in the context of protein networks
- ✓ Understand drug action and drug adverse reactions mechanisms
- ✓ Drug repurposing
- ✓ Understand the molecular basis of disease comorbidities
- ✓ Assess the performance of text-mining algorithms
- ✓ As part of other resources

Who is using DisGeNET

> 30,000 users during the last year, from over 150 countries

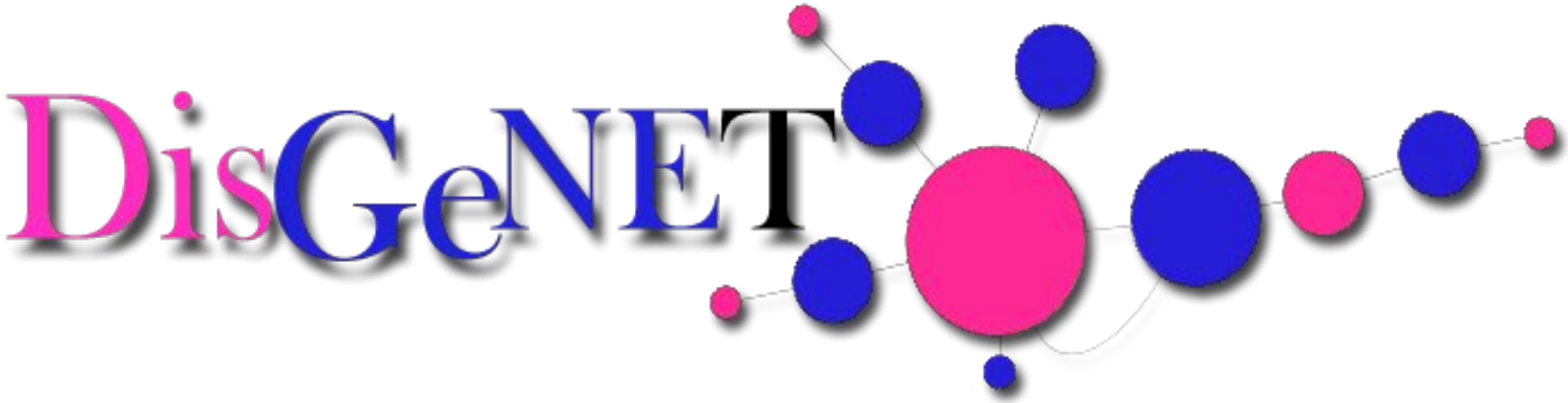
>500 publications cite DisGeNET



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

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

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



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