



A discovery platform to support translational research on human diseases

ECCB T7 tutorial
September 4 2016

Janet Piñero and Laura I. Furlong



RESEARCH
PROGRAMME
ON BIOMEDICAL
INFORMATICS



Universitat
Pompeu Fabra
Barcelona



Institut Hospital del Mar
d'Investigacions Mèdiques

DisGeNET ECCB 2016 Tutorial

- How can DisGeNET help in your research?
- Overview of the DisGeNET Platform
- Hands-on Tutorial
 - Web interface
 - DisGeNET Cytoscape app
 - DisGeNET RDF and SPARQL endpoint
 - disgenet2r R package

Research questions

- What are the diseases associated to the gene SIRT1?
- What are the genes associated to a Alzheimer's disease?
- What are the genes shared by comorbid diseases?
- What are the genetic variants associated to obesity?
- What are the druggable proteins associated to Schizophrenia?
- Which are the pathways perturbed in Lafora disease?

High throughput genomic technologies are helping to find disease genes and pathogenic variants

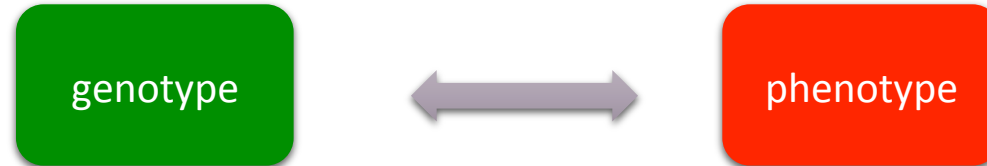
A typical whole exome sequencing experiment produces 30,000–100,000 variants relative to the reference genome

Approximately 10,000 of these variants will have a consequence at the protein function

Only one or few may be causative

Identification of **true pathogenic variants** among
all the variation is still a major challenge

The availability of ***comprehensive, traceable, high quality*** data on **genotype-phenotype** relations is key



DATA SILOS

What is the genetic basis of Wilson Disease?

#277900

WILSON DISEASE



ICD+

Alternative titles; symbols

WND; WD

HEPATOLENTICULAR DEGENERATION

Phenotype-Gene Relationships

Location	Phenotype	Phenotype MIM number	Inheritance (in progress)	Phenotype mapping key	Gene/Locus	Gene/Locus MIM number
13q14.3	Wilson disease	277900	AR	3	ATP7B	606882

Clinical Synopsis

TEXT

A number sign (#) is used with this entry because Wilson disease is caused by homozygous or compound heterozygous mutation in the ATP7B gene ([606882](#)) on chromosome 13q14.

Description

Wilson disease is an autosomal recessive disorder characterized by dramatic build-up of intracellular hepatic copper with subsequent hepatic and neurologic abnormalities.

[De Bie et al. \(2007\)](#) provided a detailed review of the molecular pathogenesis of Wilson disease.

Clinical Features

In Wilson disease, the basal ganglia and liver undergo changes that express themselves in neurologic manifestations and signs of cirrhosis, respectively. A disturbance in copper metabolism is somehow involved in the mechanism. Low ceruloplasmin ([117700](#)) is found in the serum. [Shokeir and Shreffler \(1969\)](#) advanced the hypothesis that ceruloplasmin functions in enzymatic transfer of copper to copper-containing enzymes such as cytochrome oxidase. Supporting the hypothesis was the finding of markedly reduced levels of activity of cytochrome oxidase in Wilson disease and moderate reductions in heterozygotes.

What is the genetic basis of Wilson Disease?

#277900

WILSON DISEASE

OMIM
Online Mendelian Inheritance in Man

Alternative titles: symbols

ATP7B
HEPATO LENTICULAR DEGENERATION

Phenotype-Genes Relationships

Location	Phenotype	Phenotype MIM number	Inheritance (in progress)	Phenotype mapping key	Gene/Locus	Gene/Locus MIM num
13q14.3	Wilson disease	277900	AR	3	ATP7B	606882

TATATCT
ACCTCAC
ClinVar

Variation Location	Gene(s)	Condition(s)
<input type="checkbox"/> ATP7B, 1-BP DEL, 2511A	ATP7B	Wilson's disease
<input type="checkbox"/> ATP7B, 3892GTC	ATP7B	Wilson's disease
<input type="checkbox"/> ATP7B, 15-BP DEL, NT-441	ATP7B	Wilson's disease
<input type="checkbox"/> ATP7B, 1-BP INS, NT2487	ATP7B	Wilson's disease
<input type="checkbox"/> ATP7B, 1-BP DEL, 2337C	ATP7B	Wilson's disease
<input type="checkbox"/> ATP7B, 7-BP DEL, NT2010	ATP7B	Wilson's disease

Comparative Toxicogenomics Database

Home Search Analyze Download Help

Hepatolenticular Degeneration

Bas Chemicals Genes Comps Pathways

1-5 6 7 8 Next Last

1. CP
2. ATP7B
3. PRNP
4. IL6
5. LOX
6. ANXA5
7. TNF
8. APOE

Genes related to Wilson Disease

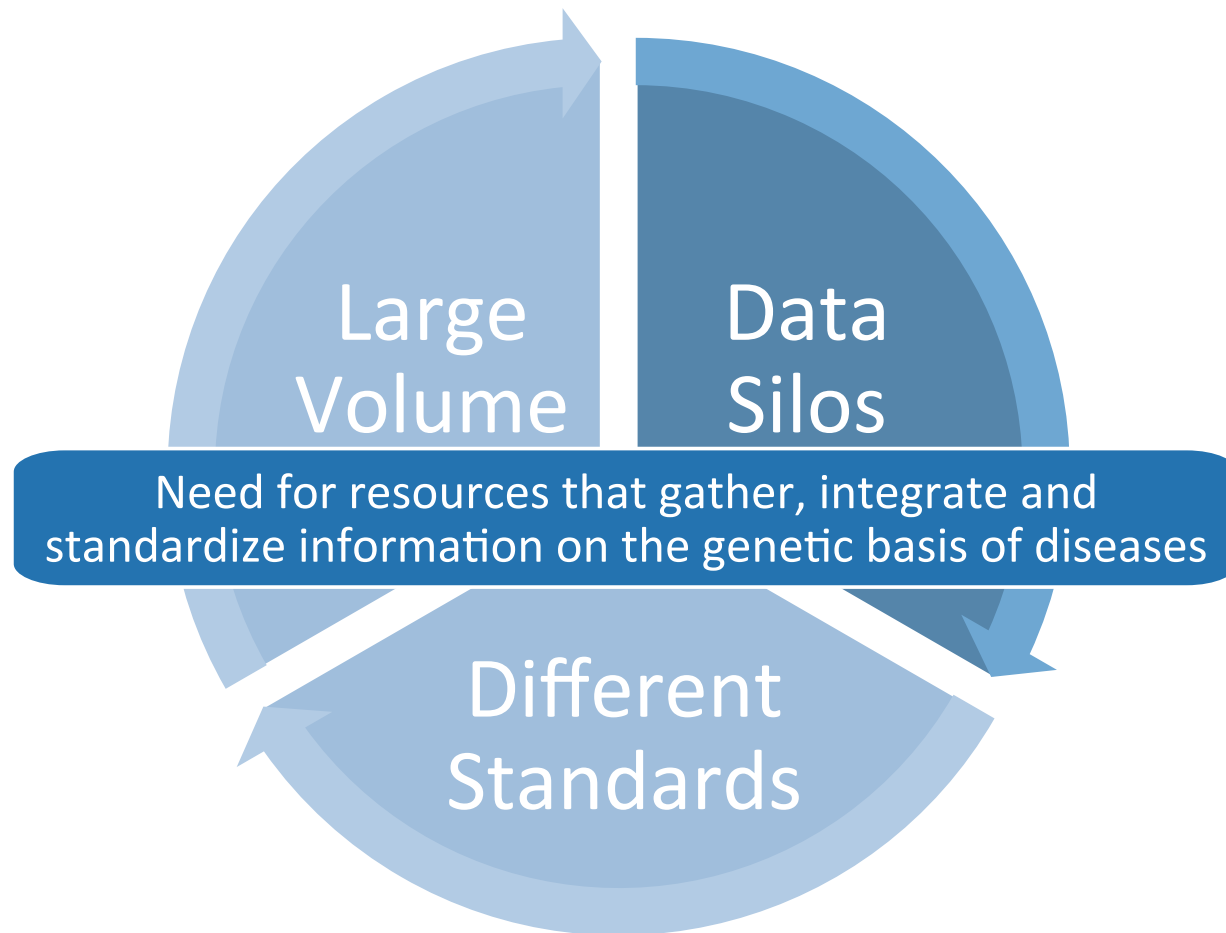
Genes related to Wilson Disease (1 elite genes):

★ - Elite gene

id	Symbol ★	Description
1	ATP7B ★	ATPase, Cu++ transporting, beta polypeptide
2	ATP7A	ATPase, Cu++ transporting, alpha polypeptide
3	COMMD1	Commodin (ferroxidase)
4	ARSA	Arylsulphatase A
5	HFE	HFE (Murr1) domain containing 1
6	SLC31A1	Solute carrier family 31 (copper transporter), member 1



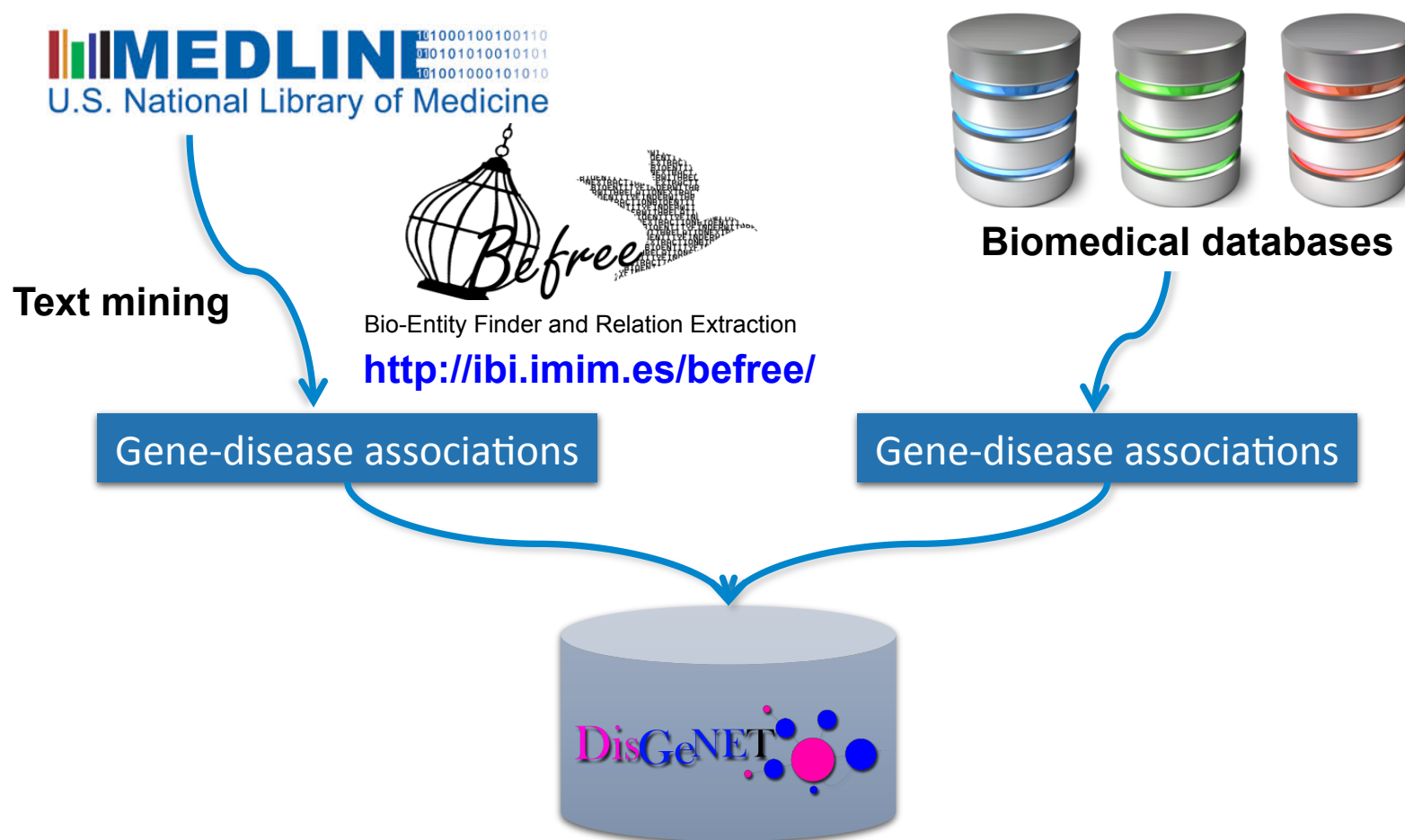
Information on genetic basis of diseases



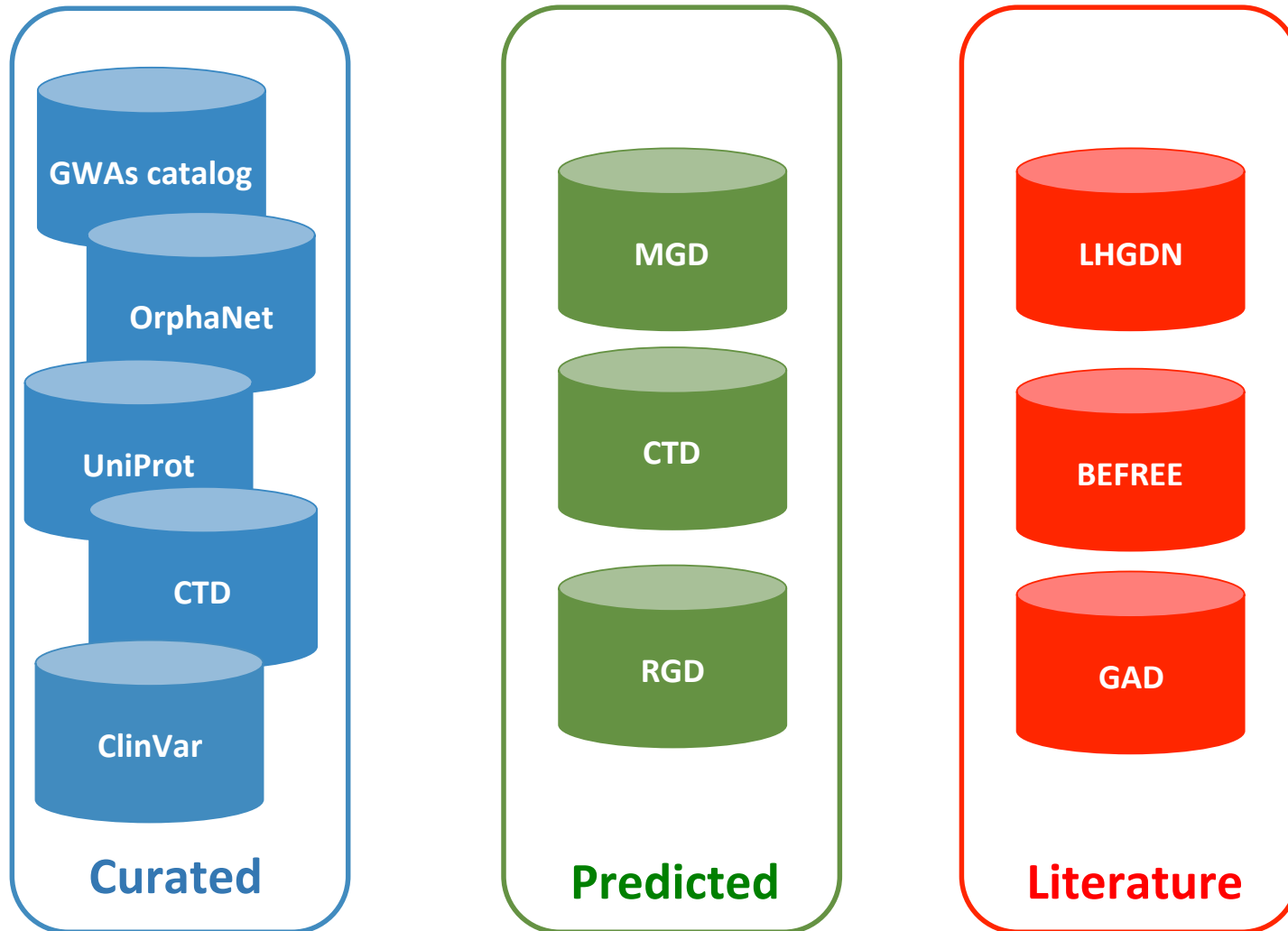


- ✓ Knowledge platform on human diseases and their genes
- ✓ Covers all disease therapeutic areas
- ✓ Integrates information from expert-curated resources and from the literature
- ✓ Focus on gene-disease association (GDA) and its supporting evidence
- ✓ Standardization of the information and provenance

DisGeNET: the implementation



DisGeNET: data sources



DisGeNET v4.0

DisGeNET: statistics (version 4.0)

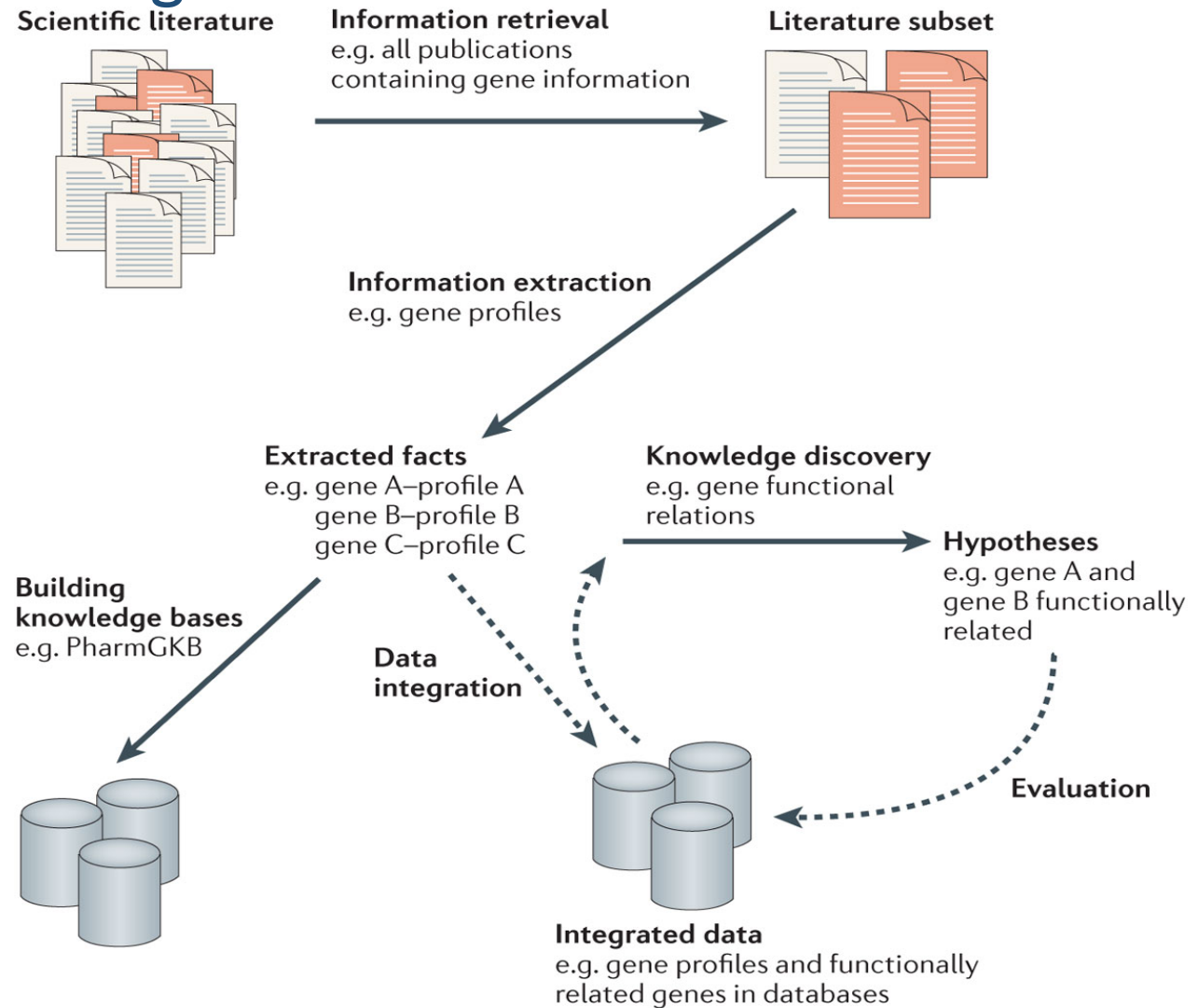
Source	Genes	Diseases	Associations
Curated	7,362	7,607	32,834
Predicted	2,743	2,064	10,264
Literature	16,141	11,447	403,925
All	17,381	15,093	429,036

>94%

Last update: June 2016

What is Text Mining?

Text mining unlocks information by automatically extracting data from free-text resources





<http://ibi.imim.es/tools/befree/>

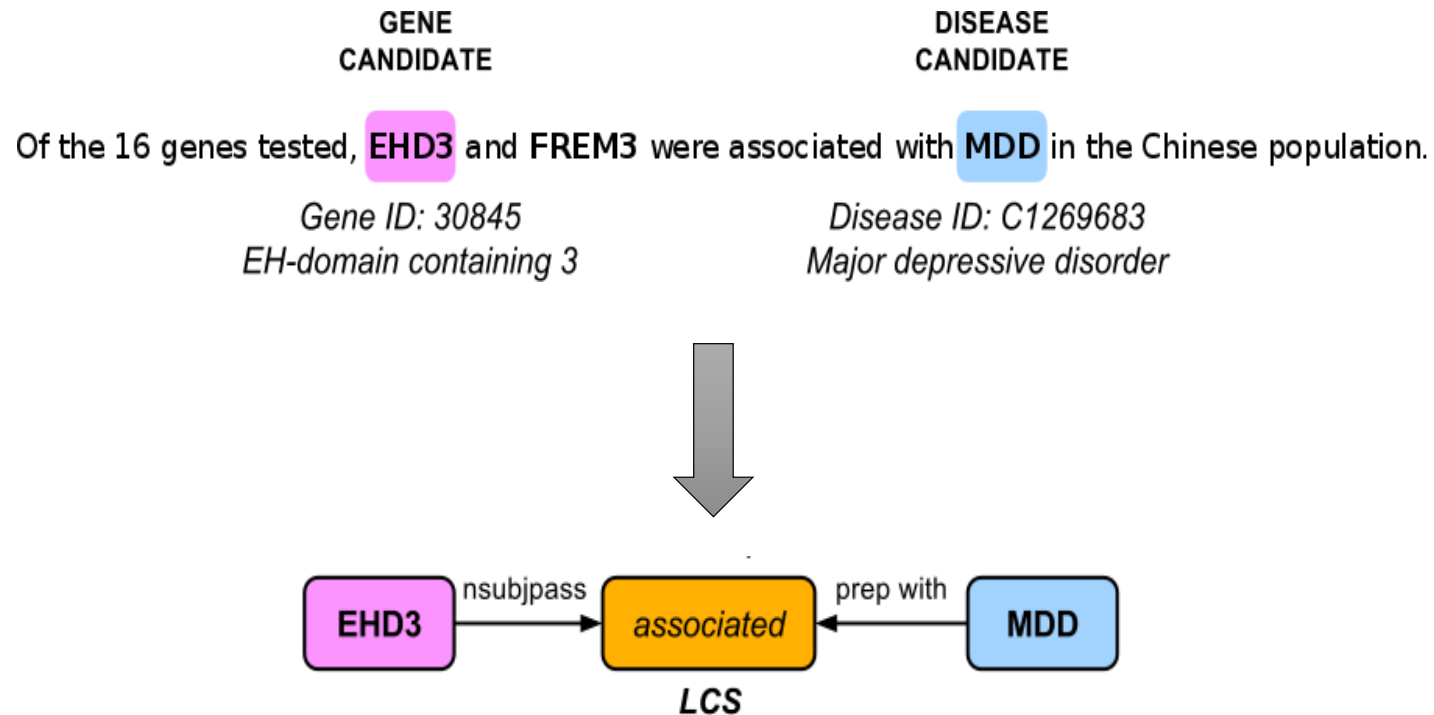
BioNER module

- Entity mention and normalization
- Fuzzy and pattern matching methods + dictionaries
- Disease and genes
- Handles ambiguities between entities

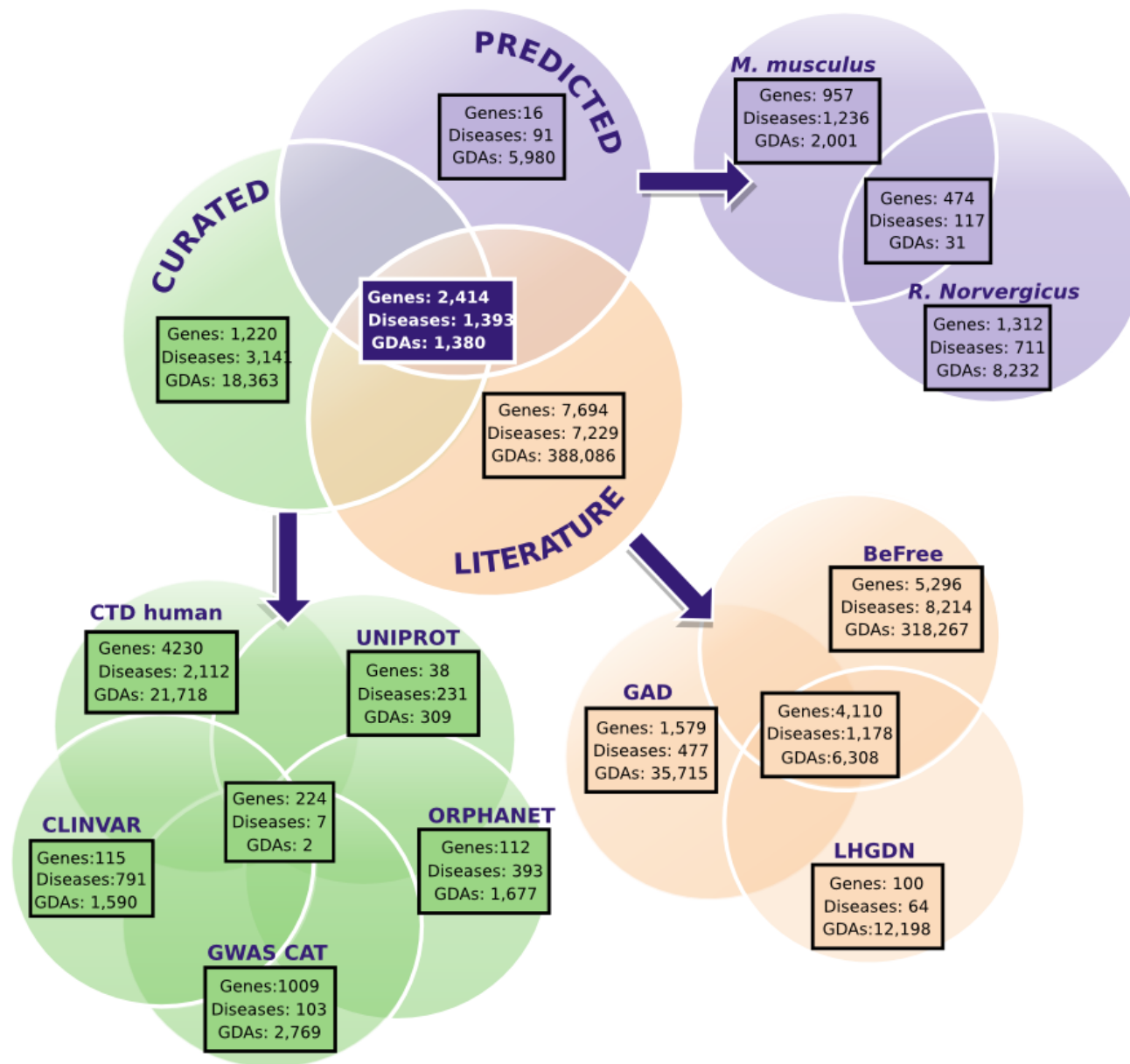
Relation Extraction module

- Based on SVM
- Combines Shallow Linguistic Kernel (K_{SL}) with Dependency Kernel (K_{DEP})
- Exploits shallow and deep syntactic information

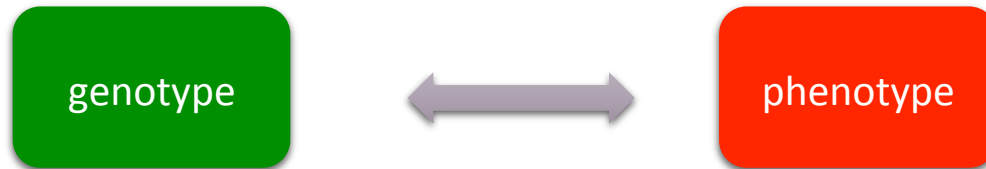
Gene-disease association identification with BeFree



Gene-disease association types according to the DisGeNET ontology



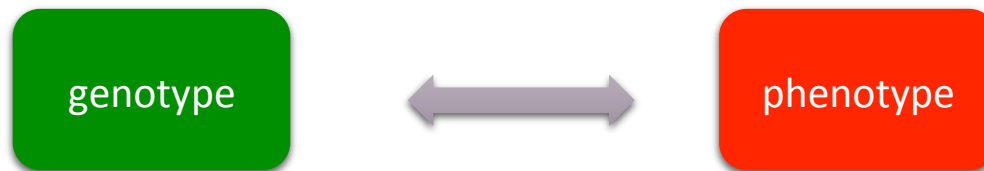
STANDARDS



- Large in scale and growing rapidly (NGS)
- Large studies on genetics of disease available
- HGVS standard for sequence variation nomenclature
- Standards for data exchange
- UniProt, NCBI, Ensembl
- VarioML, Vario

- Phenotype data spans a wide spectrum of possible observations about an individual
- More difficult to capture and to standardize
- Human Phenotype Ontology, Disease Ontology
- Broad phenotype categories used in many studies

Standards in DisGeNET

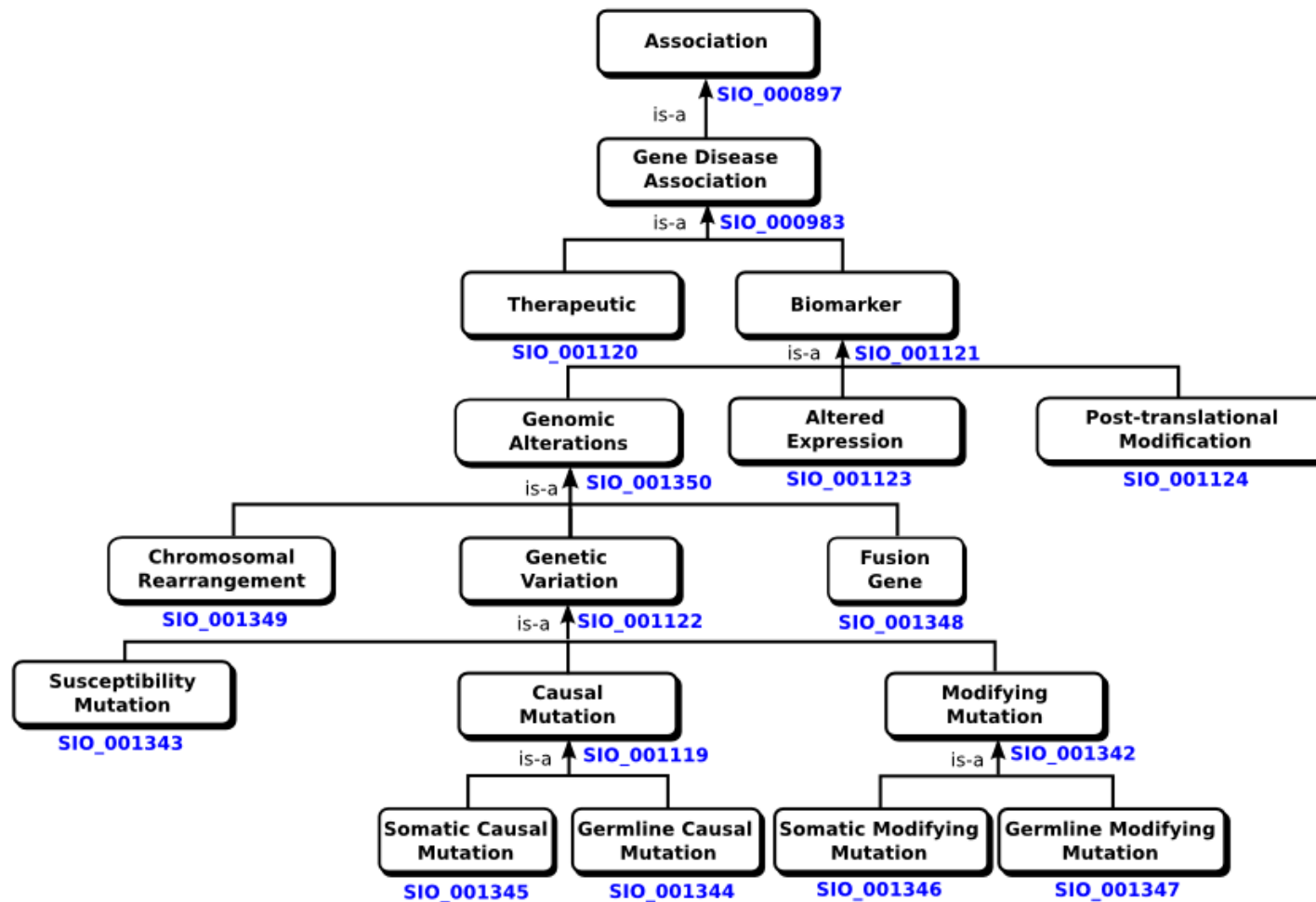


- Gene, protein, SNPs
- Official Gene symbol
- NCBI Gene Id
- Uniprot accession
- dbSNP identifier for variants

- Diseases and phenotypes
- UMLS CUIs
- UMLS semantic types
- Disease Ontology
- Mappings to a variety of phenotype vocabularies and ontologies

DisGeNET association type ontology

DisGeNET association type ontology



<http://sio.semanticscience.org>

Coverage of disease vocabularies and ontologies in DisGeNET

UMLS	MeSH	OMIM	NCIt	DO	ORDO	ICD9CM	EFO	HPO	DECIPH
100	57	40	34	20	14	11	11	8	0.4

Signs, symptoms and diseases in DisGeNET

- Abnormal phenotypes, signs and symptoms

Inflammation

Seizures

Pain

Overweight

- Diseases

Breast carcinoma

Diabetes Mellitus

- Disease class

Cardiovascular Diseases

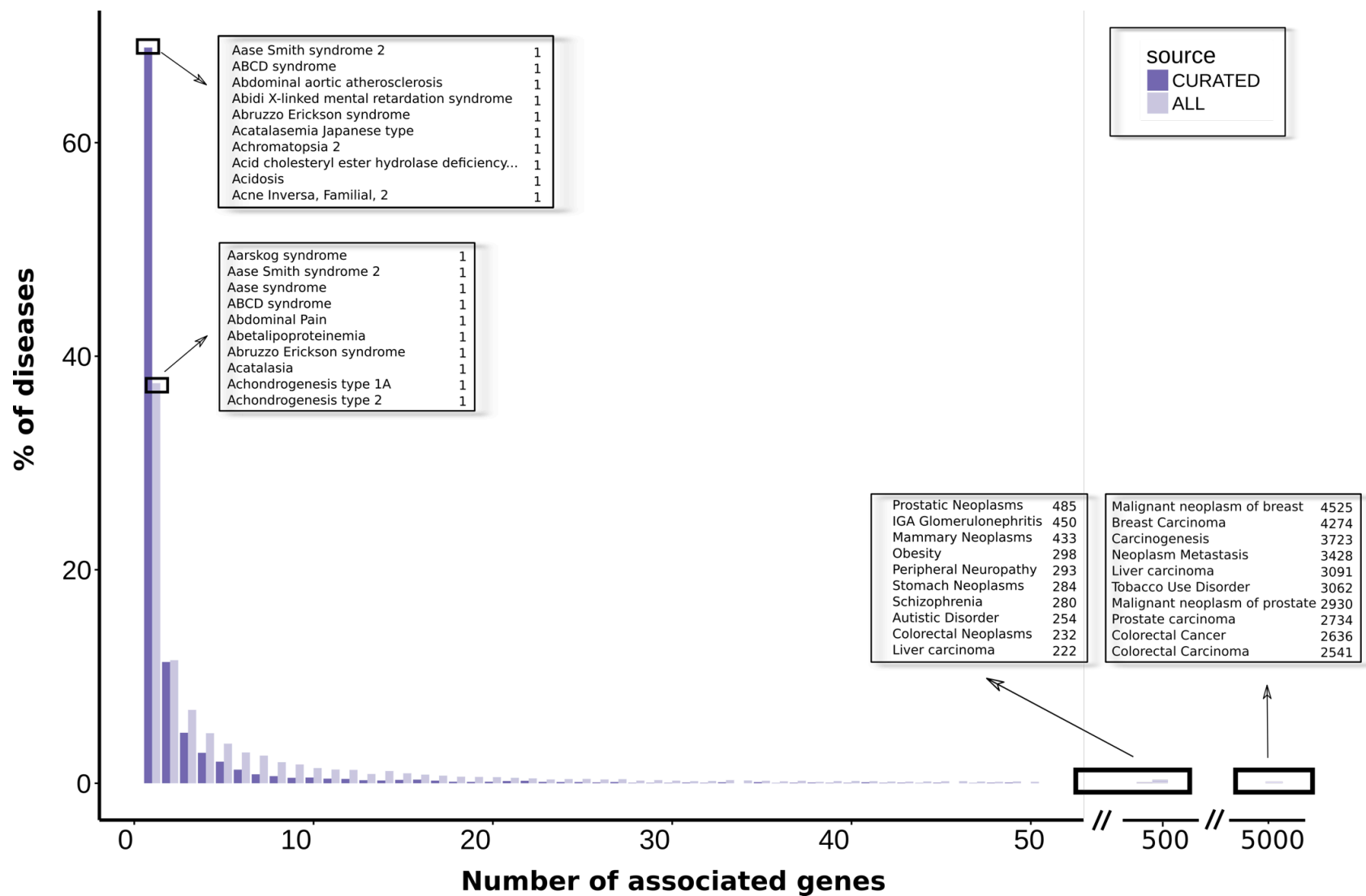
Autoimmune Diseases

Neurodegenerative Diseases

Signs, symptoms and diseases in DisGeNET

	Number of concepts	Number of associated genes	Number of associated SNPs
Disease	13,674	17,005	44,467
Disease class	55	5,739	992
Phenotype	1,364	9,332	2,894

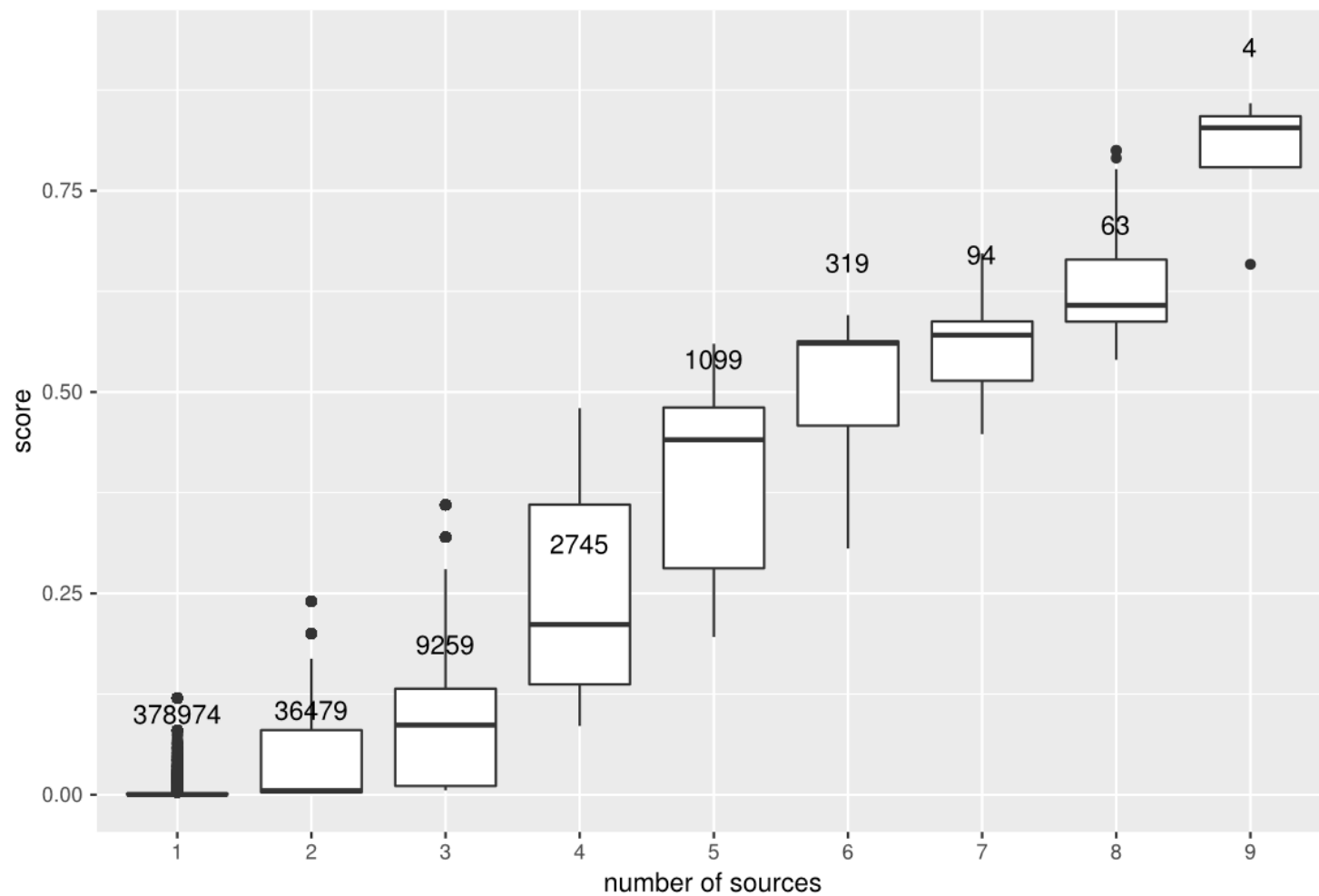
DATA PRIORITIZATION



DisGeNET gene-disease association score

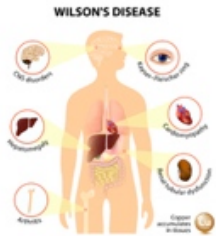
Indicates **popularity** of a **gene-disease association** across all data sources

$$\text{DisGeNET score} = S_{\text{CURATED}} + S_{\text{PREDICTED}} + S_{\text{LITERATURE}}$$



Disease Specificity Index (DSI)

- ✓ Indicates how **specific** is a **gene** with respect to diseases
- ✓ Is inversely proportional to the number of diseases associated to a particular gene (ranges from 0 to 1).
- ✓ A gene associated to a large number of diseases, such as TNF (associated to > 1,500 diseases), is less “specific” for any disease, and has a small DSI value (0.247)
- ✓ A gene associated to only one disease, is more “specific” for that disease and has DSI of 1.



Top scored genes for Wilson disease

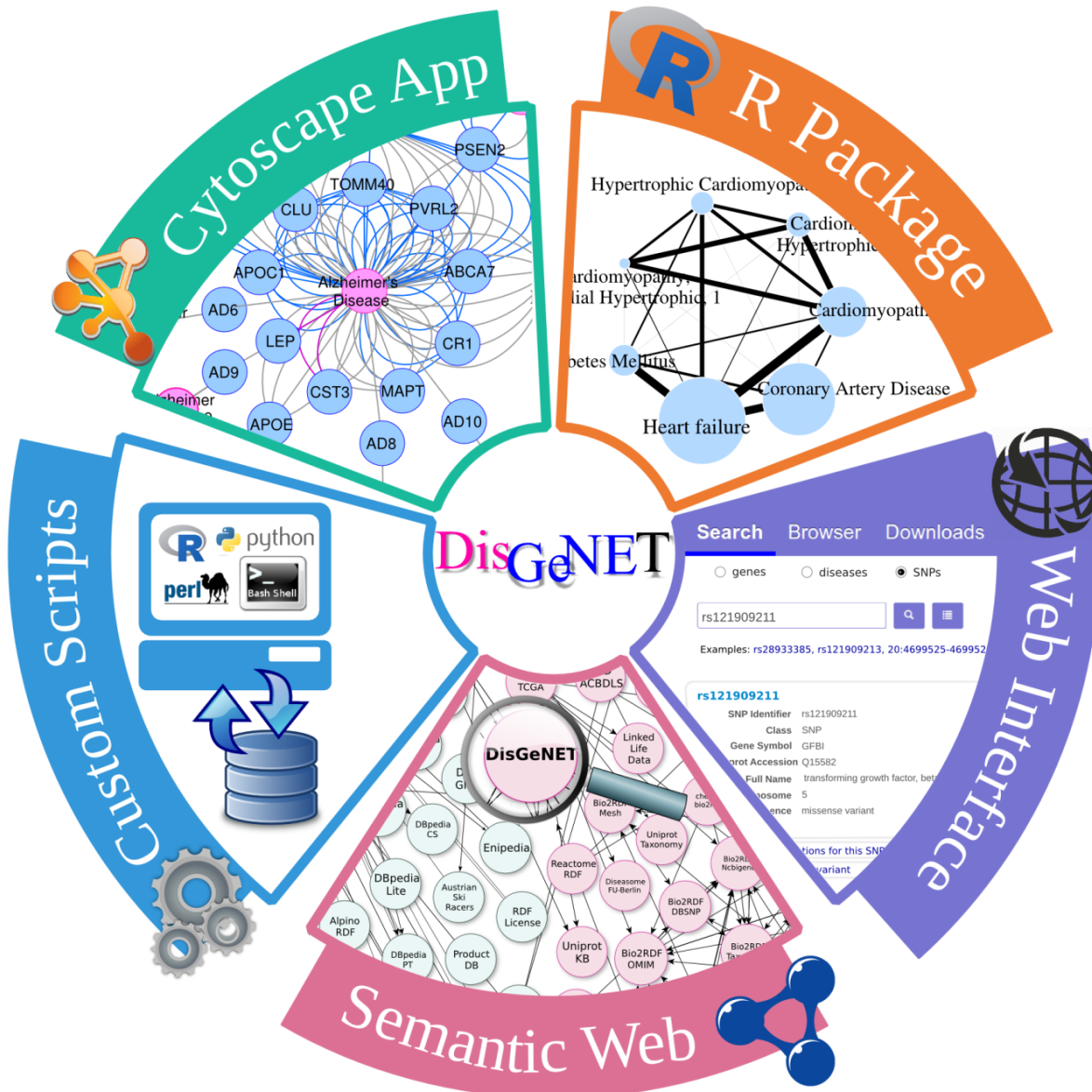
Gene	Number of diseases	DisGeNET score	DSI	Number of PMIDs	Number of SNPs
ATP7B	57	0,819	0,596	234	99
ANXA5	129	0,2	0,505	1	0
PRNP	205	0,128	0,468	4	1
CP	114	0,126	0,532	26	0
LOX	141	0,123	0,498	2	0
LOXL2	48	0,123	0,610	1	0
APOE	729	0,122	0,333	2	0
TNF	1524	0,120	0,247	2	0
IL6	1260	0,120	0,268	2	0
NDUFB7	1	0,120	1	1	0

Top scored genes for Major Depressive Disorder



Gene	Number of diseases	DisGeNET score	DSI	Number of PMIDs	Number of SNPs
SLC6A4	374	0,236	0,411	157	5
TPH2	89	0,211	0,548	26	1
HTR2A	222	0,155	0,463	45	17
PCLO	20	0,130	0,696	12	5
CRHR1	118	0,127	0,531	11	11
CYP2D6	316	0,127	0,4281	11	2
FKBP5	78	0,126	0,563	16	1
SP4	16	0,125	0,739	3	1
GRM7	32	0,123	0,666	5	1
GNAI3	7	0,122	0,812	2	1

FLEXIBLE DATA ACCESS



DisGeNET ECCB 2016 Tutorial

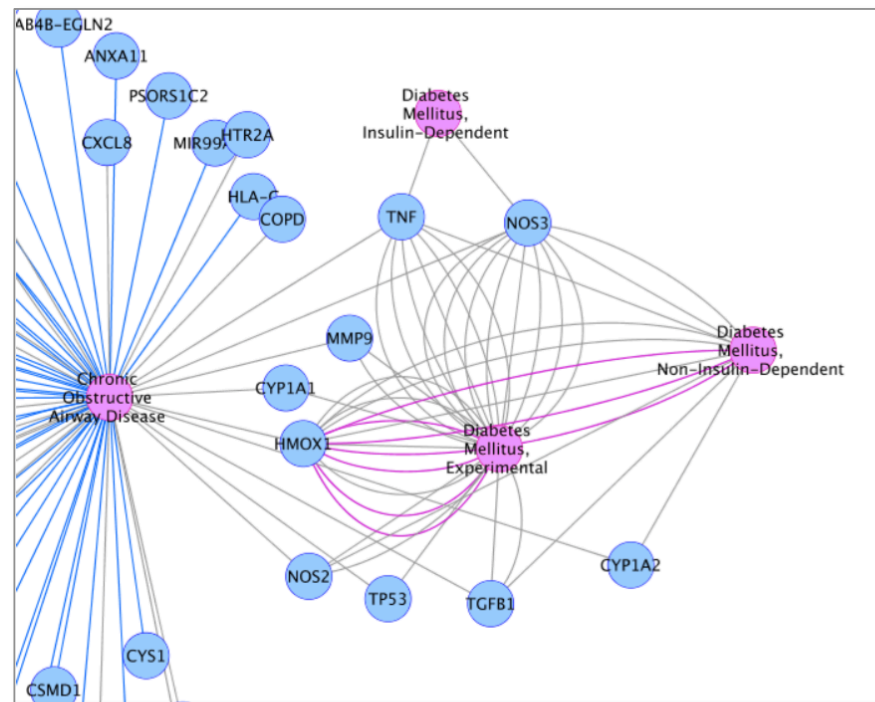
- How can DisGeNET help in your research?
- Overview of the DisGeNET Platform
- Hands-on Tutorial
 - **Web interface**
 - DisGeNET Cytoscape app
 - DisGeNET RDF and SPARQL endpoint
 - disgenet2r R package

DisGeNET ECCB 2016 Tutorial

- How can DisGeNET help in your research?
- Overview of the DisGeNET Platform
- Hands-on Tutorial
 - Web interface
 - **DisGeNET Cytoscape app**
 - DisGeNET RDF and SPARQL endpoint
 - disgenet2r R package

DisGeNET Cytoscape app

- Network representation of gene-disease associations and projections
- Downstream analysis with a variety of network analysis and annotation tools available in Cytoscape



DisGeNET ECCB 2016 Tutorial

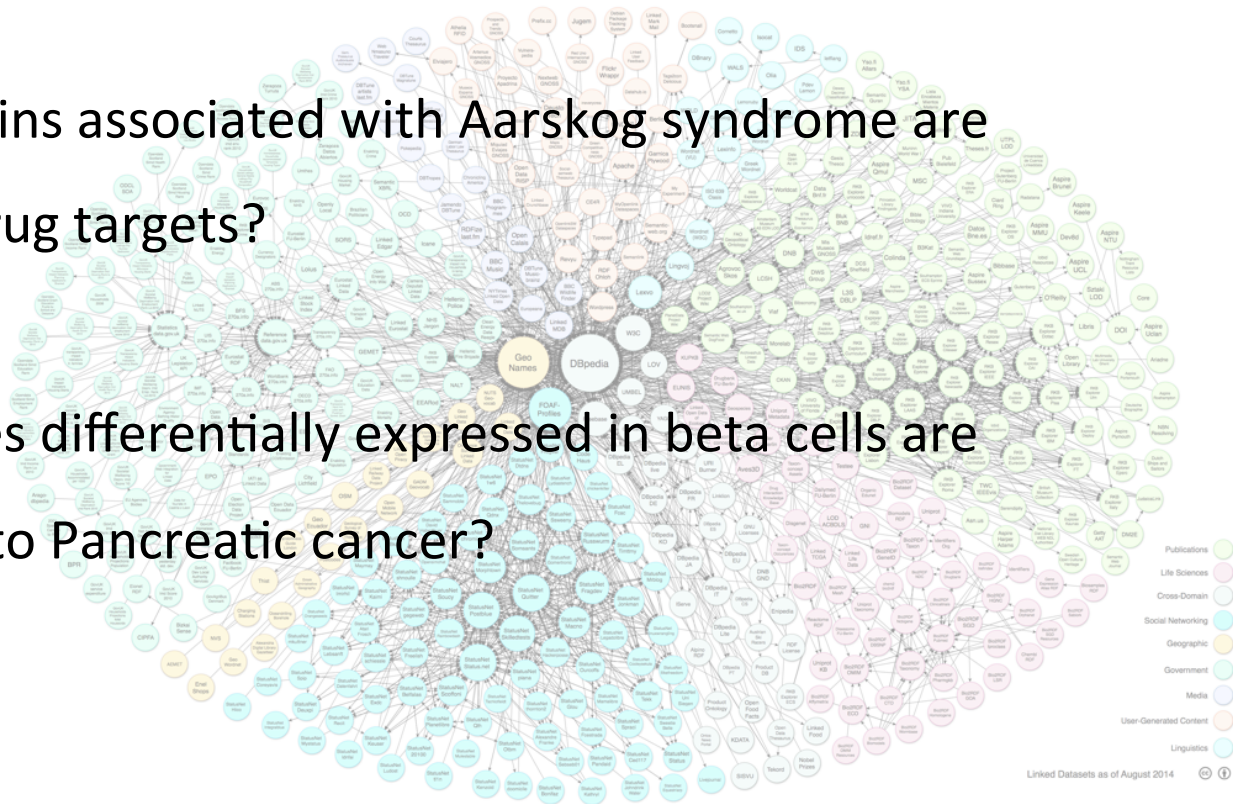
- How can DisGeNET help in your research?
- Overview of the DisGeNET Platform
- Hands-on Tutorial
 - Web interface
 - DisGeNET Cytoscape app
 - **DisGeNET RDF and SPARQL endpoint**
 - disgenet2r R package

DisGeNET as Linked Open Data




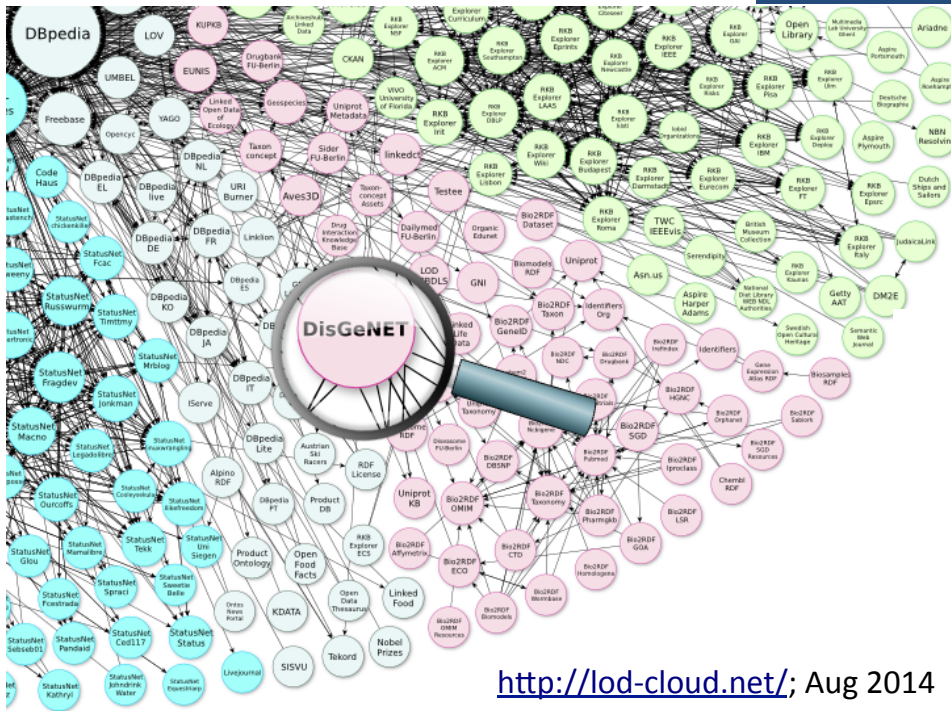
DisGeNET as Linked Open Data

- ✓ What are the perturbed pathways in Lafora disease?
- ✓ What proteins associated with Aarskog syndrome are potential drug targets?
- ✓ Which genes differentially expressed in beta cells are associated to Pancreatic cancer?



DisGeNET as Linked Open Data

- **RDF and nanopublications**
- **URIs:** RDF providers or 
 - **SIO**
 - Use of standards (**11 ontologies** in NCBO)



<http://lod-cloud.net/>; Aug 2014

- **Metadata** description ( HCLS)

- **Interlinking**



- Access

- **Download Data Dump**

- **SPARQL** Endpoint

- **Faceted Browser**



Open PHACTS

Discovery Platform

- Nanopublication Network

- disgenet2R

- **Open license**

- FAIR (ELIXIR and NIH)

- Datahub

- Software

D2RQ

Virtuoso



Semantic Web – Linked Data

Based on W3C standards

RDF: Resource Description Framework

Captures logical structure of the data

Graph representation

SPARQL: RDF query language



Usual Web vs Semantic Web

Website	Dataset
Page/URL	Resource/URI
document, textual	Formal description
HTML: presentation	RDF: semantic
Human readable	Machine readable

SPARQL Query Structure

prefix declarations

PREFIX **foaf:**<http://xmlns.com/foaf/0.1/>

dataset definition

FROM <DATASET GRAPH>

result clause

SELECT /CONSTRUCT/ASK/DESCRIBE ..OUTPUT..

query pattern

WHERE { graph pattern }

query modifiers

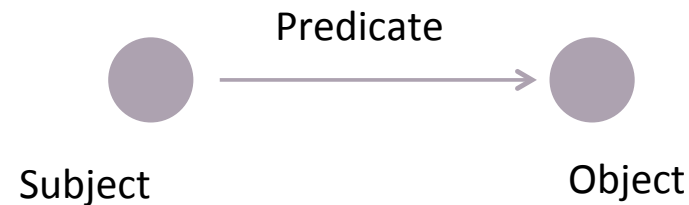
ORDER BY ...

A statement in a publication

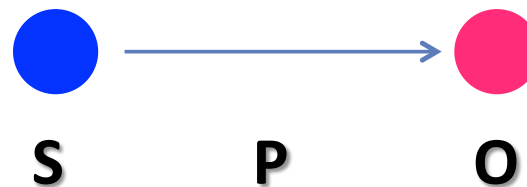


RB is overexpressed in bladder cancer samples as measured by....

In RDF, a statement is a triple



Gene associated **Disease**

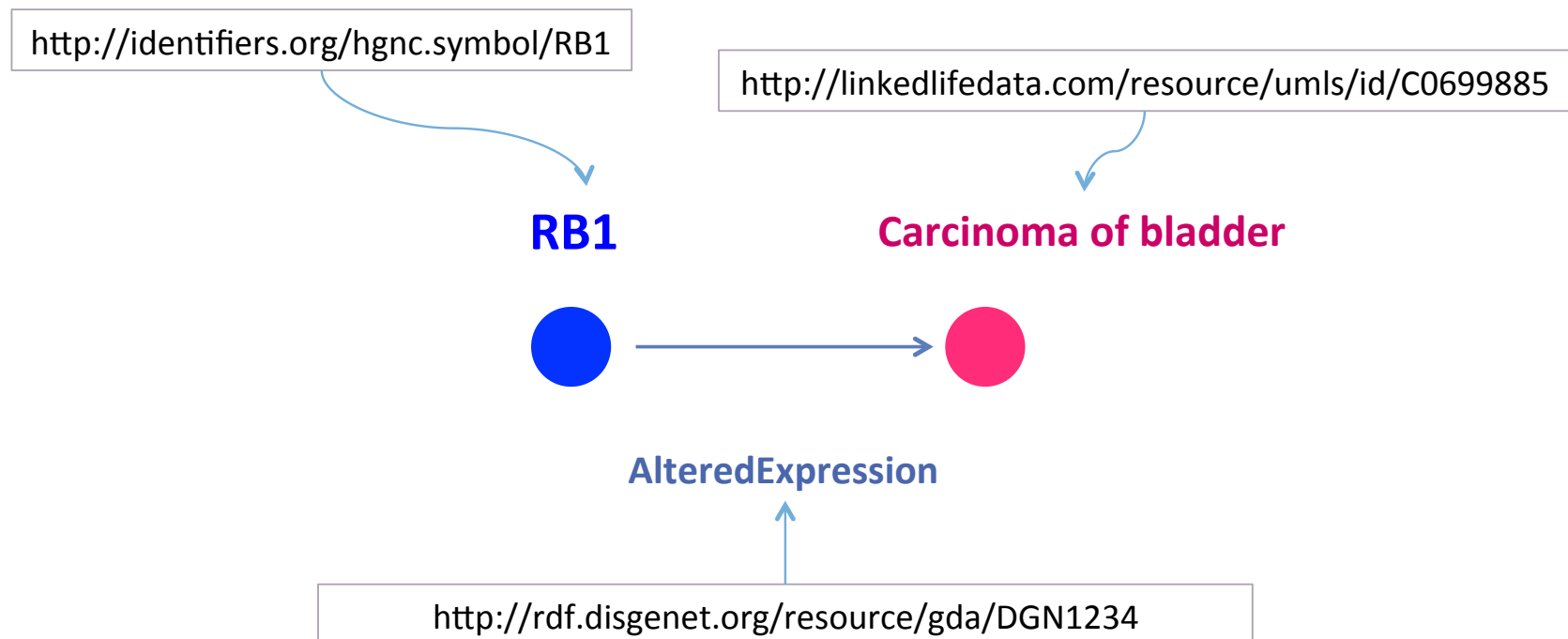


A statement in a publication



RB is overexpressed in bladder cancer samples as measured by...

In RDF, a statement is a triple



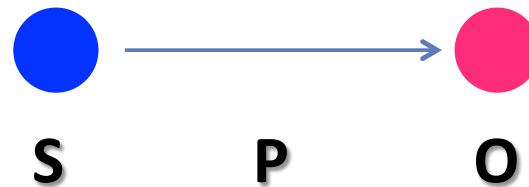
Data Model



- How to describe an **association**?

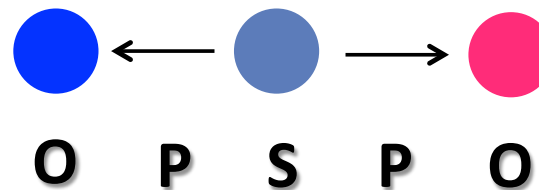
a) As a **property**

Gene associated Disease



b) As a **class**

Gene Association Disease



Data Model



- How to describe an **association**?

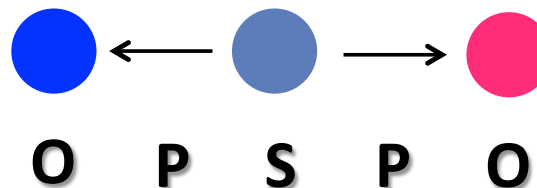
a) As a **property**

Gene associated Disease



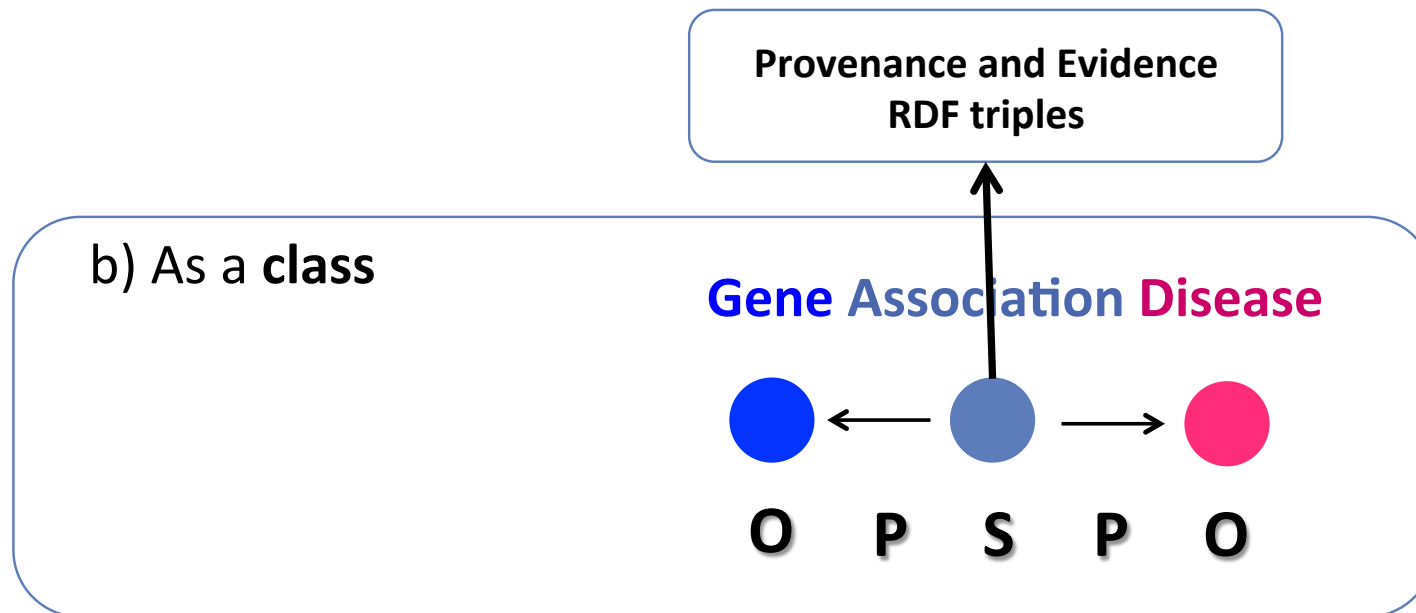
b) As a **class**

Gene Association Disease



Data Model

- How to describe an **association**?



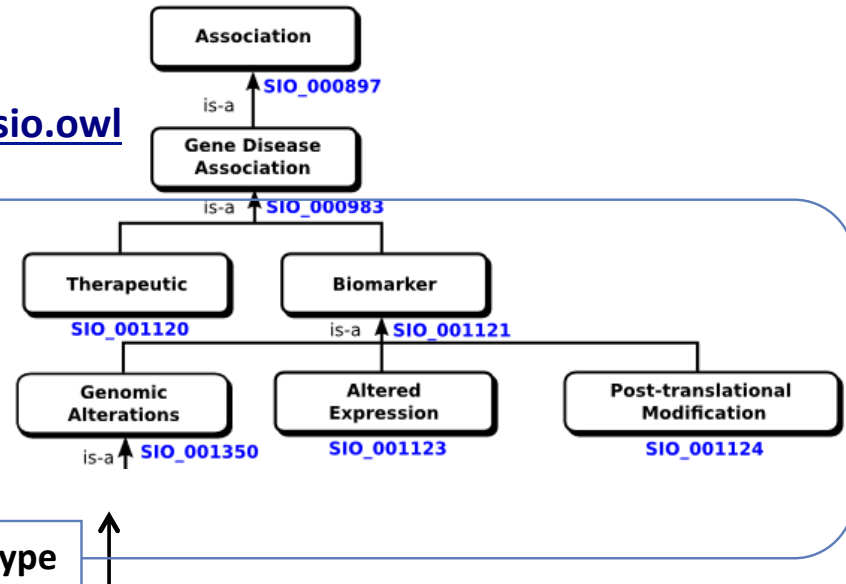
Data Model

- Ontology-based integration

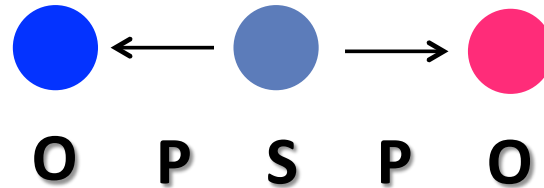
<http://semanticscience.org/ontology/sio.owl>



DisGeNET Association
Type Ontology

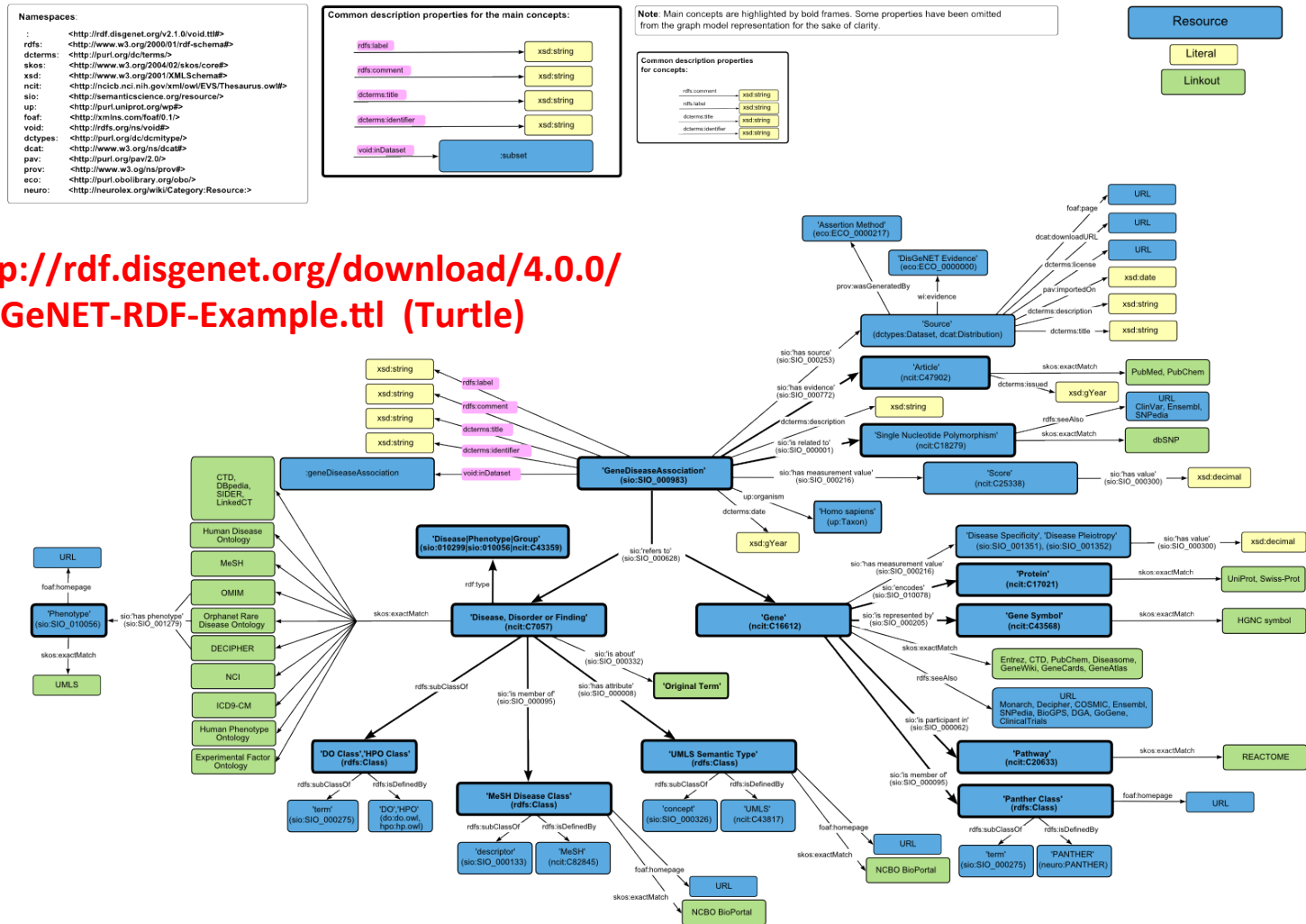


Gene Association Disease

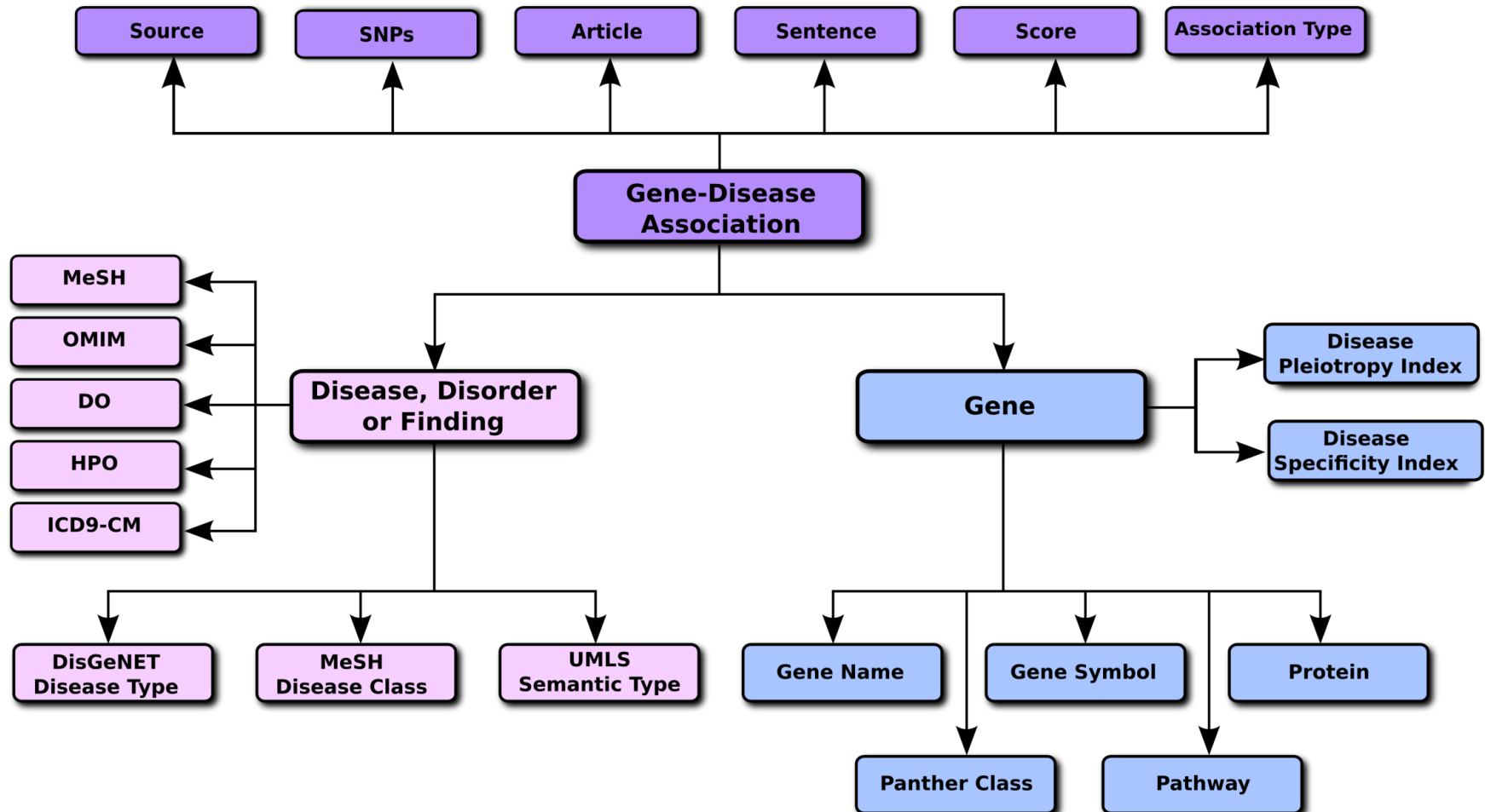


- DisGeNET Standards
 - Shared IDs
 - Standard ontologies

RDF data model

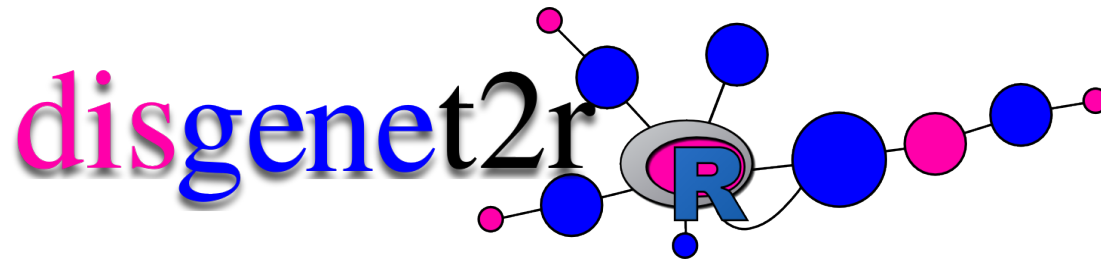


DisGeNET: the data model

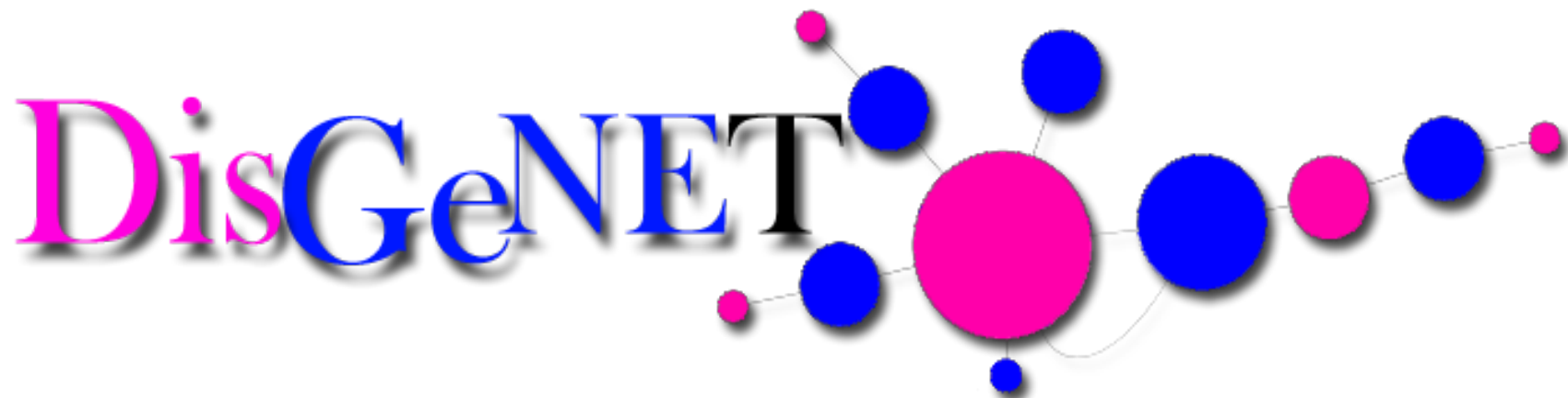


DisGeNET ECCB 2016 Tutorial

- How can DisGeNET help in your research?
- Overview of the DisGeNET Platform
- Hands-on Tutorial
 - Web interface
 - DisGeNET Cytoscape app
 - DisGeNET RDF and SPARQL endpoint
 - **disgenet2r R package**



- R package
- To interrogate DisGeNET data
- To cross DisGeNET data with other resources
- To visualize the results within the powerful R framework
- To engage with the R/Bioconductor community
- Launched within the release of DisGeNET v4.0 (April, 2016)



<http://www.disgenet.org/>
support@disgenet.org
[twitter: @DisGeNET](https://twitter.com/DisGeNET)

IBI Group

<http://ibi.imim.es/>

Alba Gutiérrez-Sacristán

Àlex Bravo

Janet Piñero

Alexia Giannoula

Miguel A. Mayer

Angela Leis

Santiago de la Peña

Emilio Centeno

Laura I. Furlong

Ferran Sanz



RESEARCH
PROGRAMME
ON BIOMEDICAL
INFORMATICS



Universitat
Pompeu Fabra
Barcelona



Institut Hospital del Mar
d'Investigacions Mèdiques

Past Members

Núria Queralt-Rosinach

Montserrat Cases

Solène Grosdidier

Pablo Carbonell

Anna Bauer-Mehren

Michael Rautschka



Unión Europea

Fondo Europeo
de Desarrollo Regional

