DisGeNET: a discovery platform for the exploration of human diseases and their genes

Janet Piñero, Núria Queralt-Rosinach, Àlex Bravo, Jordi Deu-Pons, Ferran Sanz, Laura I. Furlong
Hospital del Mar Medical Research Institute (IMIM), Pompeu Fabra University (UPF)

Abstract

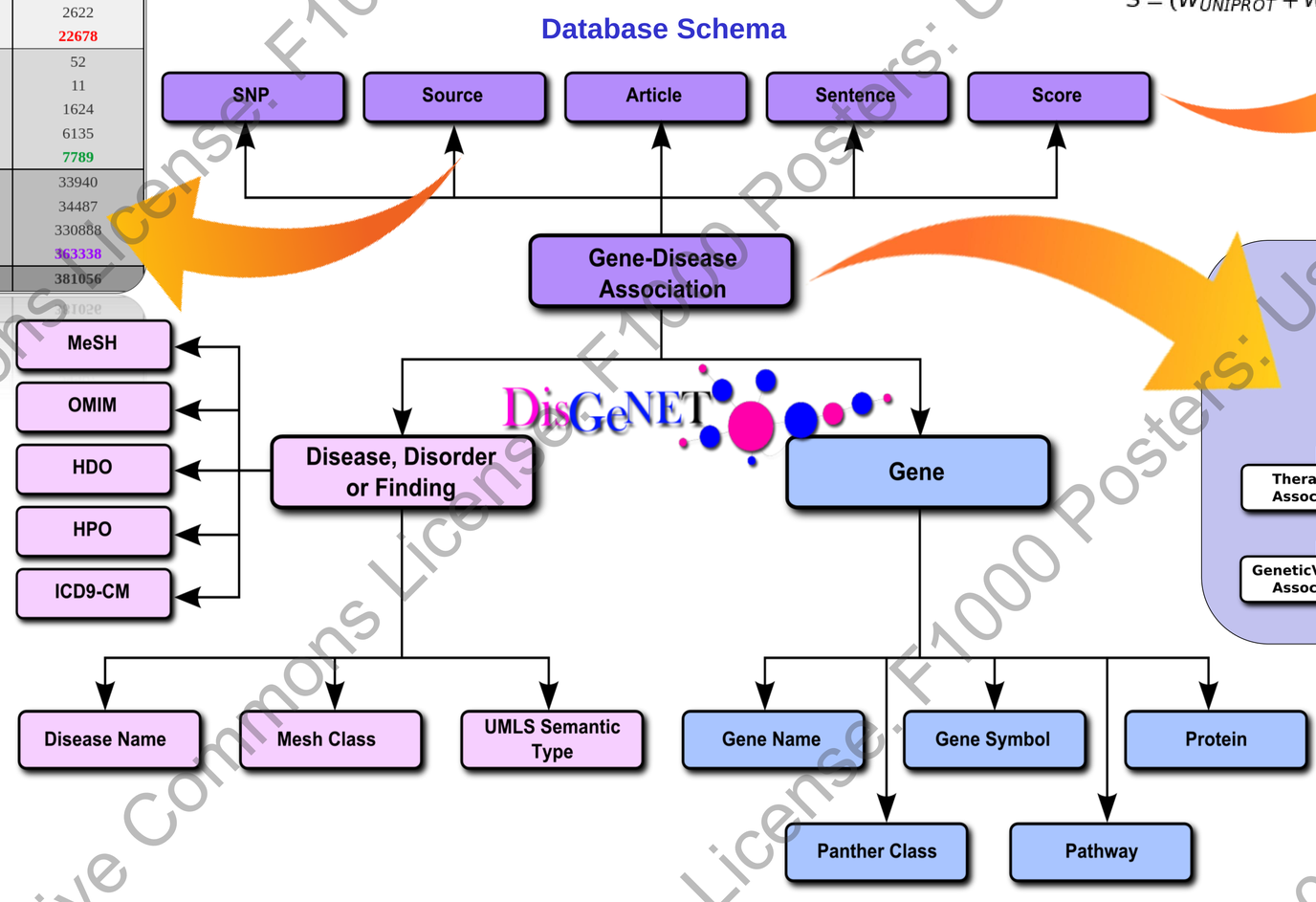
Researchers of the genetic determinants of human disease currently face two main hurdles: the large volume of information that connects genomic elements to disease phenotypes, and the fragmentation of this information across resources that employ different vocabularies and standards. Integrative platforms are therefore essential to gather and homogeneously annotate clinically relevant information on the genetic causes of diseases. In keeping with this spirit, we have developed DisGeNET (**www.disgenet.org**), a discovery platform that integrates human gene-disease relationships from several public sources, covering all disease areas. DisGeNET is one of the largest repositories of gene-disease relationships currently available to researchers, containing more than 300,000 associations between 13,172 diseases and 16,666 genes. Besides compiling information from several expert curated data sources, DisGeNET contains a unique repository of gene-disease associations obtained by text mining biomedical publications using the BeFree system (**http://ibi.imim.es/befree/**), which exploits syntactic and semantic information to find relations between biomedical entities. DisGeNET allows prioritization of gene-disease associations based on data provenance by using the DisGeNET score. The user can explore the information by using standard disease and protein classifications (e.g. MeSH and Panther), and inspect specific sentences describing a gene-disease association. The information in DisGeNET can be accessed in several ways that include a user friendly search and browse web interface, a Cytoscape plugin for network analysis and data visualization, and a SPARQL endpoint that enables to browse DisGeNET as linked data in the Semantic Web. DisGeNET data is available for download, either as text files, or as SQLite database. List of genes or diseases provided by the user can be annotated with DisGeNET information using the web interface or the plugin. In addition, by using the platform, customized queries in R, Perl, Python and bash scripts can be automatically generated and saved by the user, allowing to reproduce their analysis or incorporate them in their own programs. This makes DisGeNET a tool of choice to a broad variety of users, from the ones with basic informatics skills, such as clinician and bench biologists, to the hard-core bioinformaticians.

Database Features

Database statistics

Source	Genes	Diseases	Associations
CTD human	6983	4860	21925
UNIPROT	1839	2376	2622
CURATED	**7108**	**5466**	**22678**
CTD mouse	51	40	52
CTD rat	11	10	11
MGD	1197	1059	1624
RGD	1392	737	6135
PREDICTED	**2272**	**1758**	**7789**
GAD	9045	1737	33940
LHGDN	6136	1846	34487
BeFree	13402	10557	330888
LITERATURE	**15821**	**10926**	**363338**
ALL	**16666**	**13172**	**381056**

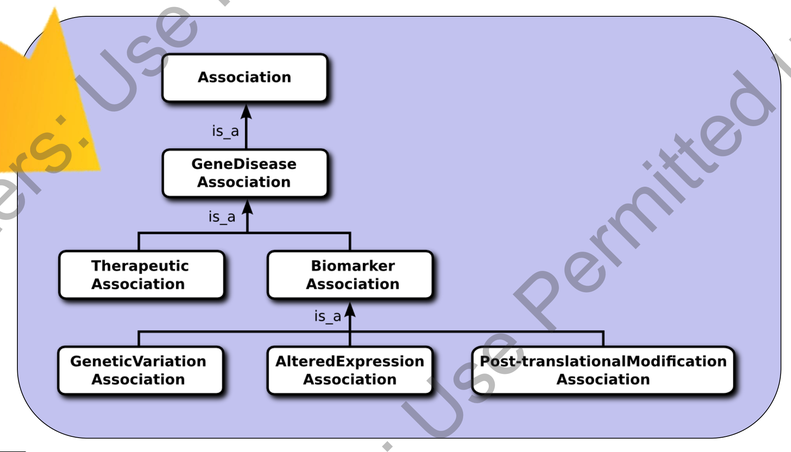
Database Schema



DisGeNET score

$$S = (W_{UNIPROT} + W_{CTD\ human}) + (W_{Rat} + W_{Mouse}) + (W_{GAD} + W_{LHGDN} + W_{BeFree})$$

DisGeNET association type ontology



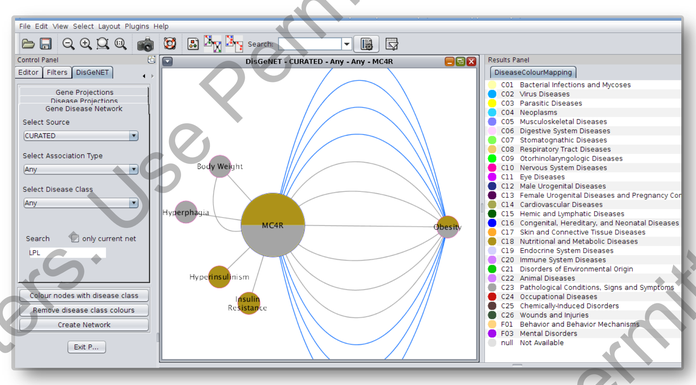
DisGeNET Discovery Platform

DisGeNET Cytoscape Plugin

The DisGeNET plugin for Cytoscape allows to visualize, query and analyze a network representation of DisGeNET data.

- Several types of queries restricted to (i) data source, (ii) the association type, (iii) the disease class or (iv) specific gene(s)/disease(s).
- Disease-disease and gene-gene networks view of the data
- Multicolouring of nodes (genes/diseases) according to standard disease classification for expedient visualization
- Annotation of your own networks with DisGeNET data
- Integration with other Cytoscape plugins for further visualization and analysis

What are the diseases associated to *melanocortin 4 receptor* (MC4R)?

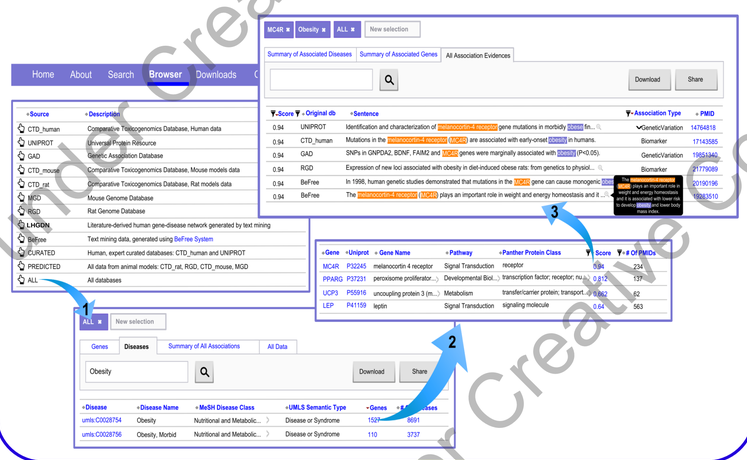


DisGeNET Web Interface

The DisGeNET Web Interface is designed to make it easier for the user to search, visualize, filter and share the data.

- Free text search of genes and diseases
- Prioritization of data by the DisgeNET score, the number of publications and the data source
- Downloads of the results of analysis as data files
- Uploads of data for annotation with DisGeNET data
- Automatic Bash, Perl, Python and R scripts generation to perform and reproduce analysis

What are the genes associated to *Obesity*?



DisGeNET in the Semantic Web

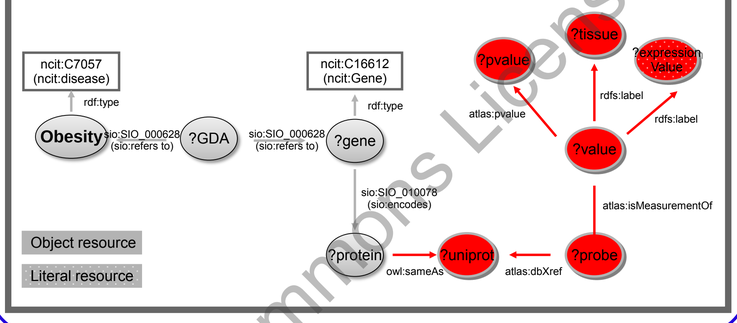
DisGeNET Linked Dataset introduces a harmonized and semantically enriched description of the gene-disease association concept into the Semantic Web by means of the DisGeNET association type ontology.

- Access the data from our SPARQL endpoint: http://rdf.disgenet.org/sparql/
- Access the data from our Faceted Browser: http://rdf.disgenet.org/fct/
- Federated queries to interrogate DisGeNET data in combination with other resources in the Linking Open Data cloud
- Downloads of dump files (serialized in RDF/Turtle): gda.ttl, disease.ttl, gene.ttl, diseaseClass.ttl, pathways.ttl, pantherClass.ttl and umlsSTY.ttl

What is the pattern of tissue expression of the genes associated to *Obesity*?

SPARQL query graph

Linking DisGeNET to Gene Expression Atlas



Funding:
The research leading to these results has received support from Instituto de Salud Carlos III-Fondo Europeo de Desarrollo Regional (CP10/00524 and PI13/00082), the Innovative Medicines Initiative Joint Undertaking under grants agreements n° [115002] (eTOX) and n° [115191] (Open PHACTS)], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. À.B. and L.I.F received support from Instituto de Salud Carlos III Fondo Europeo de Desarrollo Regional (CP10/00524). The Research Programme on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB).