**DETAILED DESCRIPTION OF DISGENET DATA PROCESSING PIPELINE**

The DisGeNET data processing pipeline starts by downloading gene-disease, or variant-gene-disease information files from the each source database.

The data from **Comparative Toxicogenomics Database** (CTD) (1) are retrieved from http://ctdbase.org/reports/CTD\_genes\_diseases.tsv.gz (Release February, 2016). We keep associations from *Homo sapiens*, *Rattus norvergicus*, or *Mus musculus*. We include in DisGeNET only the GDAs marked as “direct”, that is, not inferred (mediated by chemicals). We also keep the publication supporting each GDA.

The data from **UniProt** (2) are obtained from <http://www.UniProt.org/docs/humsavar> (Release 2016\_01 of 20-Jan-2016). From this file, we keep the GDAs with variants annotated as “disease”, and their corresponding variants. We also programmatically retrieve the record for each disease-associated protein, and parse the file to obtain the publication annotated to each variant.

We download the file ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab\_delimited/variant\_summary.txt.gz (Downloaded on 2016-01-27) from **ClinVar** (3). We keep genes, diseases, and variants labelled as “pathogenic”, “likely pathogenic”, “risk factor”, and “affects”.

**Orphanet** (4) GDAs are retrieved from the file <http://www.orphadata.org/data/xml/en_product6.xml> (date 2016-01-01).

The gene, disease, variant, and publication information from **the GWAS Catalog** (5) is obtained from <http://www.ebi.ac.uk/gwas/api/search/downloads/> (Downloaded on 2016-01-27). We keep the data corresponding to SNP-trait associations with p-values < 1.0 x 10-6, and SNPs that are not intergenic.

We obtain GDAs from the **Rat Genome Database** (RGD) (6) from the file <ftp://ftp.rgd.mcw.edu/pub/data_release/annotated_rgd_objects_by_ontology/with_terms/rattus_terms_rdo> (generated on 2016/01/15). We keep genes, diseases, and publications for GDAs not labelled with the evidence codes “Inferred from electronic annotation”, “Inferred from sequence or structural similarity”, and “Non-traceable author statement”, and GDAs not marked as “resistance”, “induced” or “no association”.

The **Mouse Genome Database** (MGD) (7) provides human and mouse GDAs in the file <ftp://ftp.informatics.jax.org/pub/reports/MGI_OMIM.rpt> (Downloaded on 2016-01-27). Only mouse GDAs associations are kept.

The **Genetic Association Database** (GAD) (8) has been retired as of 2014. We keep a copy of the dump data files. We filter out the associations marked as “Normal Variation”, “Negative”, and the ones that have no supporting publication. We also remove intergenic variants.

The **Literature-Derived Human Gene-Disease Network** (LHGDN) is a text mining derived dataset obtained by an approach that uses machine learning algorithms to extract semantic gene-disease relations from the Entrez Gene Reference Into Function (9, 10) dataset (release March, 2009).

**BeFree** GDAs were extracted from MEDLINE abstracts using the BeFree system. BeFree is composed of a Biomedical Named Entity Recognition module to detect diseases and genes (11), and a relation extraction module based on morphosyntactic information (12). The document set used to extract the gene-disease associations was defined by the following PubMed query:

("Psychiatry and Psychology Category"[Mesh] AND "genetics"[Subheading]) OR ("Diseases Category"[Mesh] AND "genetics"[Subheading]) AND (hasabstract[text] AND ("1980"[PDAT] : "2016"[PDAT]) AND "humans"[MeSH Terms] AND English[lang])

BeFree also detects and normalizes the mentions of variants in the sentences where the GDAs were found. In the last step, sentences containing phrases indicative of negative results are removed. We also use the BeFree system to obtain representative sentences for curated GDAs from the publications reported by their original sources.

Once the information is filtered, disease and gene identifier mappings are performed. For genes, HGNC symbols are converted to NCBI Entrez Gene identifiers using a dictionary that cross-references HGNC, UniProt, and NCBI Gene information. To map mouse and rat genes to their human orthologs, we use files provided by RGD (<ftp://rgd.mcw.edu/pub/data_release/RGD_ORTHOLOGS.txt>) and MGD (<ftp://ftp.informatics.jax.org/pub/reports/HOM_MouseHumanSequence.rpt>). For diseases, MeSH and OMIM identifiers are mapped to Concept Unique Identifiers (CUIs) using the Unified Medical Language System (UMLS) Metathesaurus (2015AB release). We use the disease cross-references provided in <http://www.orphadata.org/cgi-bin/inc/product1.inc.php> to map Orpha Numbers (Orphanet disease identifiers) to UMLS CUIs. For GAD, and the GWAS Catalog, disease names are mapped to UMLS CUIs using a strict match search in the UMLS Metathesaurus. We filter the CUIs by semantic type, to keep only the traits that are diseases, or disease-related (we keep CUIs with semantic types that include Congenital Abnormality, Acquired Abnormality, Finding, Injury or Poisoning, Pathologic Function, Disease or Syndrome, Mental or Behavioral Dysfunction, Sign or Symptom, Anatomical Abnormality, and Neoplastic Process).

We also download the following files to obtain the annotations for genes, variants, and diseases:

Reactome (13): Top level pathway for genes

<http://www.reactome.org/download/current/UniProt2Reactome_All_Levels.txt>

Panther (14): Protein class for genes <ftp://ftp.pantherdb.org/sequence_classifications/10.0/PANTHER_Sequence_Classification_files/PTHR10.0_human>

Human Phenotype Ontology (15): Top level Phenotype Class for diseases <http://compbio.charite.de/jenkins/job/hpo/lastSuccessfulBuild/artifact/hp/hp.obo>

Disease Ontology (16): Top level Disease Ontology Class for diseases

<http://purl.obolibrary.org/obo/doid.obo>

MeSH®: Top level MeSH Disease Class for diseases

<ftp://nlmpubs.nlm.nih.gov/online/mesh/MESH_FILES/asciimesh/d2016.bin>

EXAC (17): Allele Frequency for variants

<ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/ExAC.r0.3.1.sites.vep.vcf.gz>

1000 Genomes (18): Allele Frequency for variants

<ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>

Additionally, we access the batch query server from dbSNP to obtain variant class, and the ENSEMBL Rest API endpoint from to get the most severe consequence type.

Finally, several metrics are computed:

* The DisGeNET score is computed for all GDAs according to:
* The Disease Specificity Index is computed for the genes that are associated to at least one disease according to:

where:

- Nd is the number of diseases (excluding phenotypes and disease groups) associated to the gene

- NT is the total number of diseases (excluding phenotypes and disease groups) in DisGeNET

* The Disease Pleiotropy Index is computed for the genes that are associated to at least one disease with MeSH disease class according to:

where:

- Ndc is the number of the different MeSH disease classes of the diseases associated to the gene

- NTC is the total number of MeSH diseases classes in DisGeNET

References

1. Davis,A.P., Grondin,C.J., Lennon-Hopkins,K., Saraceni-Richards,C., Sciaky,D., King,B.L., Wiegers,T.C. and Mattingly,C.J. (2015) The Comparative Toxicogenomics Database’s 10th year anniversary: update 2015. *Nucleic Acids Res.*, **43**, D914-20.

2. The UniProt Consortium (2014) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204-212.

3. Landrum,M.J., Lee,J.M., Benson,M., Brown,G., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Hoover,J., *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, **44**, D862–D868.

4. Rath,A., Olry,A., Dhombres,F., Brandt,M.M., Urbero,B. and Ayme,S. (2012) Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Hum. Mutat.*, **33**, 803–8.

5. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicek,P., Manolio,T., Hindorff,L., *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.

6. Shimoyama,M., De Pons,J., Hayman,G.T., Laulederkind,S.J.F., Liu,W., Nigam,R., Petri,V., Smith,J.R., Tutaj,M., Wang,S.-J., *et al.* (2015) The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res.*, **43**, D743-50.

7. Eppig,J.T., Blake,J.A., Bult,C.J., Kadin,J.A. and Richardson,J.E. (2015) The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Res.*, **43**, D726-36.

8. Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S.A. (2004) The genetic association database. *Nat. Genet.*, **36**, 431–2.

9. Mitchell,J.A., Aronson,A.R., Mork,J.G., Folk,L.C., Humphrey,S.M. and Ward,J.M. (2003) Gene indexing: characterization and analysis of NLM’s GeneRIFs. *AMIA Annu. Symp. Proc.*

10. Brown,G.R., Hem,V., Katz,K.S., Ovetsky,M., Wallin,C., Ermolaeva,O., Tolstoy,I., Tatusova,T., Pruitt,K.D., Maglott,D.R., *et al.* (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, **43**, D36–D42.

11. Bravo,A., Cases,M., Queralt-Rosinach,N., Sanz,F. and Furlong,L.I. (2014) A knowledge-driven approach to extract disease-related biomarkers from the literature. *Biomed Res. Int.*, **2014**, 253128.

12. Bravo,A., Piñero,J., Queralt-Rosinach,N., Rautschka,M. and Furlong,L.I. (2015) Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, **16**, 55.

13. Fabregat,A., Sidiropoulos,K., Garapati,P., Gillespie,M., Hausmann,K., Haw,R., Jassal,B., Jupe,S., Korninger,F., McKay,S., *et al.* (2015) The Reactome pathway Knowledgebase. *Nucleic Acids Res.*, **44**, D481-487.

14. Mi,H., Poudel,S., Muruganujan,A., Casagrande,J.T. and Thomas,P.D. (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res.*, **44**, D336–D342.

15. Groza,T., Köhler,S., Moldenhauer,D., Vasilevsky,N., Baynam,G., Zemojtel,T., Schriml,L.M., Kibbe,W.A., Schofield,P.N., Beck,T., *et al.* (2015) The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. *Am. J. Hum. Genet.*, **97**, 111–124.

16. Kibbe,W.A., Arze,C., Felix,V., Mitraka,E., Bolton,E., Fu,G., Mungall,C.J., Binder,J.X., Malone,J., Vasant,D., *et al.* (2015) Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.*, **43**, D1071-8.

17. Lek,M., Karczewski,K.J., Minikel,E. V., Samocha,K.E., Banks,E., Fennell,T., O’Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B., *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.

18. Auton,A., Abecasis,G.R., Altshuler,D.M., Durbin,R.M., Abecasis,G.R., Bentley,D.R., Chakravarti,A., Clark,A.G., Donnelly,P., Eichler,E.E., *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.