Bioinformatics, 31(18), 2015, 3075–3077
doi: 10.1093/bioinformatics/btv301
Advance Access Publication Date: 11 May 2015
Applications Note

OXFORD

Databases and ontologies

PsyGeNET: a knowledge platform on psychiatric disorders and their genes

Alba Gutiérrez-Sacristán[1,†], Solène Grosdidier[1,†], Olga Valverde[2], Marta Torrens[3], Àlex Bravo[1], Janet Piñero[1], Ferran Sanz[1] and Laura I. Furlong[1,*]

[1]Research Group on Integrative Biomedical Informatics, Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM), Department of Experimental and Health Sciences (DCEXS), Universitat Pompeu Fabra (UPF), [2]Neurobiology of Behaviour Research Group (GReNeC), IMIM, DCEXS, UPF and [3]Institute of Neuropsychiatry and Addiction, Parc de Salut Mar, Universitat Autònoma de Barcelona, Barcelona 08003, Spain

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: John Hancock

Received on February 13, 2015; revised on May 4, 2015; accepted on May 5, 2015

Abstract

Summary: PsyGeNET (Psychiatric disorders and Genes association NETwork) is a knowledge platform for the exploratory analysis of psychiatric diseases and their associated genes. PsyGeNET is composed of a database and a web interface supporting data search, visualization, filtering and sharing. PsyGeNET integrates information from DisGeNET and data extracted from the literature by text mining, which has been curated by domain experts. It currently contains 2642 associations between 1271 genes and 37 psychiatric disease concepts. In its first release, PsyGeNET is focused on three psychiatric disorders: major depression, alcohol and cocaine use disorders. PsyGeNET represents a comprehensive, open access resource for the analysis of the molecular mechanisms underpinning psychiatric disorders and their comorbidities.
Availability and implementation: The PysGeNET platform is freely available at http://www.psygenet.org/. The PsyGeNET database is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/).
Contact: lfurlong@imim.es
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Psychiatric disorders have a great impact on morbidity and mortality, accounting for 1:3 of disability worldwide (Murray and Lopez, 2013; Whiteford *et al.*, 2013) and generating most of healthcare expenditure (Collins *et al.*, 2011). Psychiatric diagnoses are syndromes rather than diseases with operative diagnostic criteria based on symptoms and signs (e.g. the Diagnostic and Statistical Manual of Mental Disorders, and the International Classification of Diseases Diagnostic Criteria) and lacking reliable biological markers. Co-occurrence of two or more psychiatric diagnoses ('psychiatric comorbidity') has been reported to be very frequent. This co-occurrence raises two major clinical questions: whether there are underlying common environmental, genetic and neurobiological ethiopathological factors and what is their impact on clinical care. For instance, there is a high prevalence of comorbidity between major depression (MD) and substance use disorders (Pettinati *et al.*, 2013) which might be related to poor therapeutic outcome for both disorders (WHO.Global Status Report on Alcohol and Health, 2014). In the past years, there has been a great progress on the genetics of psychiatric disorders (Sullivan *et al.*, 2012). However, there

is still limited understanding of the cellular and molecular mechanisms underlying these diseases and their comorbidities. Some of the factors limiting progress in this area are (i) the heterogeneity and fragmentation of the available data on psychiatric disorders into silos and (ii) the lack of resources that collect these wealth of data, integrate them and supply the information in an intuitive, open access manner to the community. PsyGeNET (**Psy**chiatric disorders and **G**enes association **NET**work) has been developed with these issues in mind. PsyGeNET integrates curated information on psychiatric disorders and their genes, offering exploratory tools for their query and analysis. In its current version, we focused on three psychiatric disorders: MD, alcohol use disorder (AUD) and cocaine use disorder (CUD). The selection of the diseases has been made on the basis of the following factors: (i) the significant burden of each of the diseases for the society, (ii) the large body of literature reporting results on the genetics of these diseases, in particular for MD and (iii) the comorbidity between MD and substance use disorders. Here, we describe the development of the database, the functionalities of the platform and the questions that can be addressed with PsyGeNET.

2 PsyGeNET database

PsyGeNET database (version 1.0, release January 25, 2015) contains 2642 gene-disease associations (GDAs) between 1271 genes and 37 psychiatric disease concepts collected from DisGeNET (Piñero *et al.*, 2015) and the literature by text mining followed by expert curation. The psychiatric disorders were defined by experts in the domain of psychiatry and neurobiology using 59 (MD), 13 (AUD) and 11 (CUD) UMLS concepts (the list of concepts is available at http://www.psygenet.org/).

2.1 Curated data

We included expert-curated, direct associations between human genes and the selected disease concepts from the Comparative Toxicogenomics Database™ [CTD (Davis *et al.*, 2013)]. On the other hand, a dataset of GDAs identified by text mining and curated by domain experts was developed (PsyCUR). To this end, Medline abstracts published between 1980 and 2013 in journals with a SCI impact factor greater than one were processed by BeFree (Bravo *et al.*, 2015) to select sentences with GDAs. The diseases were identified using the UMLS concepts that define each disorder, whereas an in-house developed gene dictionary was used to identify the genes, as described in Bravo *et al.* (2015). A web-based annotation tool was developed to assist the curation process. The curators were presented with a putative GDA and additional information such as the supporting publications with links to the PubMed abstracts and to DisGeNET was provided. The curator decided to annotate a particular GDA as valid based on her background knowledge and the publications that support the GDA identified by text mining. Only valid GDAs were stored as PsyCUR.

2.2 Animal model data

We obtained genes from animal models of the diseases of interest from the Mouse Genome Database [MGD (Blake *et al.*, 2014)] and the Rat Genome Database [RGD (Laulederkind *et al.*, 2013)]. Mouse and rat genes were mapped to their human orthologs using the mapping files provided by MGD (ftp://ftp.informatics.jax.org/pub/reports/HOM_MouseHumanSequence.rpt) and RGD (ftp://rgd.mcw.edu/pub/data_release/RGD_ORTHOLOGS.txt), respectively. These mapping files are derived from NCBI HomoloGene. We only included GDAs from animal models when a human ortholog of the mouse or rat gene could be found and were supported by publications. In the case of RGD, we did not include the associations labeled as 'resistance', 'induced' or 'no association' nor the ones annotated with the following evidence codes 'Inferred from electronic annotation', 'Inferred from sequence or structural similarity' and 'Non-traceable author statement'.

2.3 Other data sources

We extracted GDAs from the Genetic Association Database (GAD), an archive of human genetic association studies of complex diseases (Becker *et al.*, 2004). It includes a summary extracted from published articles in peer-reviewed journals on candidate gene and Genome Wide Association Studies (GWAS). We only included GDAs from GAD that were not labeled as 'negative' or 'normal variation'.

2.4 Data classification

We organized the data according to the source databases in the following classes: CURATED (containing GDAs from PsyCUR and CTD), MODELS (containing GDAs from RGD and MGD) and GAD (containing data from GAD).

2.5 Data attributes

Diseases are annotated to UMLS concepts and classified according to the MeSH disease hierarchy, whereas genes are annotated to NCBI Gene and UniProt and classified according to the Panther Protein Ontology. Each GDA is annotated with the supporting evidence (the publications reporting the GDA, a representative sentence describing the association and the original source of provenance) and the PsyGeNET score. The PsyGeNET score ranks each GDA taking into account the number of sources that report each association, the level of curation of each source and the number of publications supporting the association. The score ranges between 0 and 1 and is computed as follows:

$$S = S_{\text{PSYCUR}} + S_{\text{CTD}} + S_{\text{MODELS}} + S_{\text{LITERATURE}}$$

where

$$S_{\text{PSYCUR}} = \begin{cases} 0.5 \text{ if the association is reported by PSYCUR} \\ 0 \text{ otherwise} \end{cases}$$

$$S_{\text{CTD}} = \begin{cases} 0.2 \text{ if the association is reported by CTD} \\ 0 \text{ otherwise} \end{cases}$$

$$S_{\text{MODELS}} = \begin{cases} 0.1 \text{ if the association is reported by MGD or RGD} \\ 0 \text{ otherwise} \end{cases}$$

$$S_{\text{LITERATURE}} = \text{ratio}_{\text{GAD}} + \text{ratio}_{\text{PSYCUR}}$$

where

$$\text{ratio}_{\text{source}} = \max\left(\frac{n_{\text{gda}_{\text{source}}}}{N_{\text{LITERATURE}_{\text{source}}}} \times 10, 0.1\right)$$

and

source is GAD or PSYCUR

$n_{\text{gda}_{\text{source}}}$ is the number of publications supporting a GDA in the data source

$N_{\text{LITERATURE}_{\text{source}}}$ is the total number of publications in the data source

2.6 Comparison to other sources

We performed a comparison between PsyGeNET, CTD and OMIM in terms of the coverage of genes. We used as query terms the list of

Table 1. Genes associated to psychiatric disorders in PsyGeNET and other public resources

	CTD™	OMIM®	PsyGeNET	PsyCUR
MD	90	6	845	448
AUD	36	6	534	170
CUD	109	0	128	13

PsyGeNET contains information on the genetics of psychiatric disorders curated from the literature that is not available in other databases. CTD™, Comparative Toxicogenomics Database; OMIM®, Online Mendelian Inheritance of Man. MD: Major Depression, AUD: Alcohol Use Disorder, CUD: Cocaine Use Disorder

UMLS concepts that define each psychiatric disorder: 59 concepts for MD, 13 concepts for AUD and 11 concepts for CUD UMLS (the list of concepts is available at http://www.psygenet.org/). We used DisGeNET to query OMIM (downloaded on February 2015) and CTD (version 13949M, downloaded on February 2015), as it allows using UMLS concepts to search for diseases. We used PsyGeNET release 1.1 (January 2015). The results are listed in Table 1, where the PsyCUR column represents the genes only available in PsyGeNET and not present in the other sources.

3 PsyGeNET web interface

The PsyGeNET platform is composed of a database and a web interface with a set of analysis tools, powered by Onexus (http://www.onexus.org/). The web interface supports searching, visualizing, filtering and sharing of PsyGeNET data. Data files containing the results of the user's search can be downloaded in a variety of formats. Moreover, automatically generated scripts are provided in several programming languages to reproduce the analyses performed by the user and to incorporate them into bioinformatic workflows. Lastly, functionalities are offered to share the results of queries performed with PsyGeNET via e-mail or by embedding the HTML code of exploration results in web pages. The web interface offers two entry points:

Search view: It allows the user to search a specific gene (or disease) in the database using free-text or standard identifiers. As a result, the user retrieves all diseases (or genes) that are associated with the gene (or disease) and the supporting evidence. Queries on multiple genes or diseases are also supported.

Browse view: It allows the user to explore all the information starting by a specific data source (e.g. PsyCUR). The results are shown in the same way as when using the Search view.

The PsyGeNET results can be filtered by the score, data source and attributes of the data. The web interface also provides links to external resources such as NCBI Gene and UniProt for genes, Linked Life Data for diseases and PubMed for publications. The user can inspect the evidence of each GDA by exploring the sentences extracted from the supporting publications in which the gene and disease are highlighted. For more details on the functionalities offered by the web interface, go to the Web Interface Help section (http://www.psygenet.org/).

4 PsyGeNET applications

Examples of questions that can be answered with PsyGeNET are:

1. Which genes are associated to alcoholism in expert-curated databases?
2. What subtypes of depression are available in PsyGeNET?
3. Which genes support the association between MD and AUD?

A step-by-step guide on how to answer these questions is available in the online tutorial (http://www.psygenet.org/).

5 Conclusions

We have developed PsyGeNET as a new resource to foster translational research on psychiatric disorders and their genes, through an open access database and publicly available exploration and analysis tools. A comparison to other databases indicates that PsyGeNET is unique regarding data quality and coverage (Table 1). PsyGeNET can be used to support different case studies on psychiatric diseases and their comorbidities. Because of its comprehensiveness, level of curation and suite of tools, PsyGeNET represents a valuable resource to foster the research on the molecular and biological mechanisms underpinning psychiatric disorders and their comorbidities.

Funding

This work was supported by Instituto de Salud Carlos III-Fondo Europeo de Desarrollo Regional (CP10/00524, PI13/00082 and RD12/028/024 and RD12/0028/0009 Retics-RTA), MINECO (SAF2013-41761-R-Fondo Europeo de Desarrollo Regional), the Innovative Medicines Initiative Joint Undertaking (no. 115372, EMIF and no. 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. The Research Programme on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB).

Conflict of Interest: none declared.

References

Becker,K.G. *et al.* (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.

Blake,J.A. *et al.* (2014) The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res.*, **42**, D810–D817.

Bravo,À. *et al.* (2015) Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, **16**, 55.

Collins,P.Y. *et al.* (2011) Grand challenges in global mental health. *Nature*, **475**, 27–30.

Davis,A.P. *et al.* (2013) The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res.*, **41**, D1104–D1114.

Laulederkind,S.J.F. *et al.* (2013) The Rat Genome Database 2013—data, tools and users. *Brief. Bioinform.*, **14**, 520–526.

Murray,C.J.L. and Lopez,A.D. (2013) Measuring the global burden of disease. *N. Engl. J. Med.*, **369**, 448–457.

Pettinati,H.M. *et al.* (2013) Current status of co-occurring mood and substance use disorders: a new therapeutic target. *Am. J. Psychiatry*, **170**, 23–30.

Piñero,J. *et al.* (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database (Oxford).*, pii: bav028.

Sullivan,P.F. *et al.* (2012) Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat. Rev. Genet.*, **13**, 537–551.

Whiteford,H.A. *et al.* (2013) Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet*, **382**, 1575–1586.

World Health Organization. (2014) *Global status report on alcohol and health*. World Health Organization.