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: a) the heterogeneity and fragmentation of the available data on psychiatric disorders into silos, and b) 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 (Psychiatric disorders and Genes association NETwork) 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: a) the significant burden of each of the diseases for the society, b) the large body of literature reporting results on the genetics of these diseases, in particular for MD, and c) 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 25th, 2015) contains 2,642 gene-disease associations (GDAs) between 1,271 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/).
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.

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 MGD1 and RGD2, 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”.

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 papers in peer-reviewed journals on candidate gene and Genome Wide Association Studies (GWAS). We only included GDAs from GAD that were not labelled as “negative” or “normal variation”.

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).

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

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

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

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

\[ S_{\text{LITERATURE}} = \frac{n_{\text{pubsource}}}{N_{\text{LITERATUREsource}}} \times 10.01 \]

and

\[ n_{\text{pubsource}} \text{ is the number of publications supporting a GDA in the data source} \]

\[ N_{\text{LITERATUREsource}} \text{ is the total number of publications in the data source} \]

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 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.psychenet.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 shown in Table 1, where the PsyCUR column represent the genes only available in PsyGeNET and not present in the other sources.

3 PYGENET 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.

1 ftp://ftp.informatics.jax.org/pub/reports/HOM_MouseHumanSequence.rpt
2 ftp://rgd.mcw.edu/pub/data_release/RGD_ORTHOLOGS.txt
Psychiatric diseases and their comorbidities. Due to its comprehen-

siveness, 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 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 (see Table 1). PsyGeNET can be used to support different case studies on psychiatric disorders and their comorbidities. Due to 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).

REFERENCES


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.


WHO | Global status report on alcohol and health 2014

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

<table>
<thead>
<tr>
<th></th>
<th>CTD*</th>
<th>OMIM®</th>
<th>PsyGeNET</th>
<th>PsyCUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD</td>
<td>90</td>
<td>6</td>
<td>845</td>
<td>448</td>
</tr>
<tr>
<td>AUD</td>
<td>36</td>
<td>6</td>
<td>534</td>
<td>170</td>
</tr>
<tr>
<td>CUD</td>
<td>109</td>
<td>0</td>
<td>128</td>
<td>13</td>
</tr>
</tbody>
</table>

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 (see Table 1). PsyGeNET can be used to support different case studies on psychiatric disorders and their comorbidities. Due to 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.

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/).