Text mining and expert curation to develop a database on psychiatric diseases and their genes

Alba Gutiérrez-Sacristán
GRIB
IMIM - UPF

Àlex Bravo
GRIB
IMIM - UPF

Marta Portero
GReNeC
IMIM - UPF

Olga Valverde
GReNeC
IMIM - UPF

Antonio Armario
Universitat Autònoma
de Barcelona

M.Carmen Blanco-Gandía
Universitat de Valencia

Adriana Farré
Parc de Salut Mar
UAB

Lierni Fernández-Ibarrondo
Program in Cancer Research
IMIM

Francina Fonseca
Parc de Salut Mar
UAB

Jesús Giraldo
Universitat Autònoma
de Barcelona

Angela Leis
GRIB
IMIM - UPF

Anna Mané
Parc de Salut Mar
UAB

Miguel A. Mayer
GRIB
IMIM - UPF

Sandra Montagud-Romero
Universitat de Valencia

Roser Nadal
Institut de Neurociències
UAB

Jordi Ortiz
School of Medicine
UAB

Francisco Javier Pavón
Instituto de Investigación
Biomédica de Málaga

Ezequiel Perez
Parc de Salut Mar
UAB

Marta Rodríguez-Arias
Universitat de Valencia

Antonia Serrano
Instituto de Investigación
Biomédica de Málaga

Marta Torrens
Parc de Salut Mar
UAB

Vincent Warnault
GReNeC
IMIM - UPF

Ferran Sanz
GRIB
IMIM - UPF

Laura I. Furlong
GRIB
IMIM - UPF

Abstract

During the last years there has been a growing research in the genetics of psychiatric diseases. However, there is still a limited understanding of the cellular and molecular mechanisms leading to these diseases, which has hampered the application of this wealth of knowledge into the clinical practice to improve diagnosis and treatment of psychiatric patients. PsyGeNET (`http://www.psygenet.org/`) has been developed to improve the understanding of psychiatric diseases, by facilitating the access to the vast amount of their genetic information in a structured manner, providing a set of analysis and visualization tools. In this communication we describe the protocol we put in place for the sustainable update of this knowledge resource. It includes the recruitment of a team of experts to perform the curation of the data previously extracted by text mining. Annotation guidelines and a web-based annotation tool were developed to support curators tasks. A curation workflow was designed including a pilot phase, and two rounds of curation and analysis phases. We report the results of the application of this workflow to the task of curation of gene-disease associations

for PsyGeNET, including the analysis of the inter-annotator agreement, and suggest that this model is a suitable approach for the sustainable development and update of knowledge resources.

1 Introduction

Psychiatric disorders have a great impact on morbidity and mortality (Murray and Lopez, 2013; Whiteford et al., 2013). According to the World Health Organization (WHO), one of every four people will suffer mental or neurological disorders (Kessler et al., 2005; Baldacchino et al., 2009). It has been suggested that most psychiatric disorders display a strong genetic component (Sullivan et al., 2012). During the last years there has been a growing research in the genetics of psychiatric disorders, and its findings have been reported on hundreds of thousands of publications. This literature constitutes a rich and diverse source of information essential for any psychiatric research line. However, the huge amount and continuous growth of the number of publications refrain scientists to efficiently explore such large volume of data.

PsyGeNET (Psychiatric disorders Gene association NETwork) (Gutiérrez-Sacristán et al., 2015) has been developed to establish a curated resource on psychiatric diseases and their associated genes. PsyGeNET integrates knowledge extracted from the scientific literature by text-mining which has been curated by experts in psychiatry and neurosciences.

In this communication we describe the process put in place for the update of the PsyGeNET database. This involved i) the recruitment of a team of experts to curate the information extracted by text-mining; ii) the extraction of information of gene-disease associations (GDAs) from the literature using the text mining system BeFree (Bravo et al., 2015), iii) the development of a curation workflow (Figure 1), iv) the development of a web-based annotation tool in order to facilitate the curation task and v) the definition of detailed guidelines to assist the curation task.

In particular, we present the results of the Pilot phase and Curation I phase of the workflow, including the analysis of the inter-annotator agreement, and suggest that this protocol is a suitable approach for the sustainable development and update of knowledge resources.

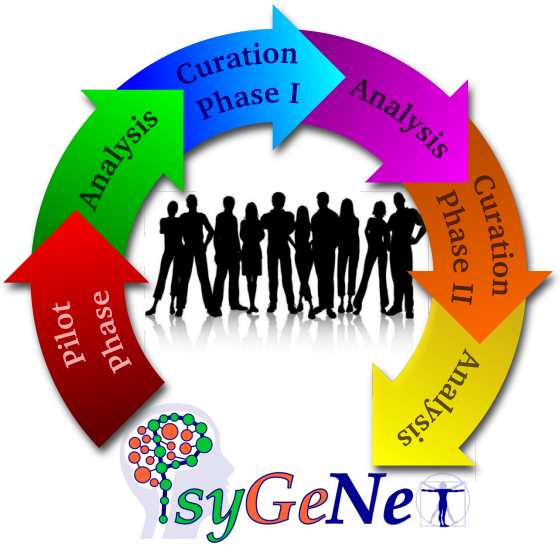

Figure 1: PsyGeNET curation workflow

2 Methods

2.1 Curation team

A team of 22 experts from different domains (such as psychiatry, neuroscience, medicine, psychology and biology) was recruited from the Spanish Network of Addiction and other collaborators of the coordination team (Research Group on Integrative Biomedical Informatics (GRIB)) to participate in the curation process. The incentives for participation were to be part of the PsyGeNET team and to be co-authors in the publication(s) originated from the project. The curators were trained during an initial session where the PsyGeNET annotation guidelines were presented and then during the Pilot phase. Communication with the coordination team through e-mail was established to resolve questions during the curation process. In addition, on-line and f2f meeting were organized after key points of the curation process (analysis phases) to share experiences among all curators and solve curation issues.

2.2 Defining the Psychiatric Diseases in terms of UMLS concepts

In PsyGeNET, the psychiatric diseases are identified by UMLS Metathesaurus concepts. Three experts reviewed the terminology included in more than 2,000 UMLS concepts related to the psychiatric disorders of interest, and assigned them to the following psychiatric disease categories (DCs): 1) Depressive disorders, 2) Bipolar disorders and related disorders, 3) Substance/drug induced depres-

sive disorder, 4) Schizophrenia spectrum and other psychotic disorders, 5) Drug-induced psychosis, 6) Alcohol use disorders, 7) Cannabis use disorders and 8) Cocaine use disorders. This information was used both for text mining of gene-disease associations by BeFree (see below) and for identification of disease classes during the curation.

2.3 Text mining of gene-disease associations

BeFree, a text-mining tool that exploits morphosyntactic information from the text to identify relationships between biomedical concepts, was used to identify associations between genes and the psychiatric diseases of interest from a corpus of $\sim$1M of MEDLINE abstracts focused on human genetic diseases. 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). The identified disorders were grouped according to the eight psychiatric disease categories (described in section 2.2). As a result, BeFree identified 6,349 associations between genes and DCs (gene-disease category association or GDCA) supported by 4,065 publications. A subset of the associations was initially evaluated by our group to identify the most frequent text mining errors. For instance, the word depression is often used in other context in addition to psychiatry. This initial evaluation was performed to identify this kind of errors and improve the text mining system before the identification of GDCAs. We then applied a number of filters to reduce the size of the curation task and make it feasible with the resources at hand. For instance, we removed associations already present in curated resources (DisGeNET (Piñero et al., 2015) and the previous release of PsyGeNET), kept only those associations published recently (after year 2,000) in journals with Impact Factor greater than 1, and we did not take into account reviews. After this process we obtained 2,507 GDCAs, which were submitted to expert curation.

2.4 Annotation Guidelines

The PsyGeNET annotation guidelines were developed with the purpose of guiding the manual curation process. The guidelines included the definition of a gene-disease association, how it should be classified according to the level of evidence, what information should be considered for the annotation and provided real examples of the association types. Finally, it also included a tutorial on how to use the PsyGeNET annotation tool. The goal of the curation was to validate the association of a gene to a particular disease. We consider that a gene is associated to a disease if the gene or the product of the gene plays a role in the disease pathogenesis, or is a marker for the disease. The PsyGeNET annotation tool was used to help in this curation task. For each gene-disease association identified by text mining, the annotation tool displayed the evidence that supports the association, more concretely the abstracts and the sentences in which the gene-disease association is stated. Then, by inspecting the evidence, the curator had to determine the type of association (Association, No Association, False, Error and Not Clear). The types of association are described as follows: i) **Association**: the publication clearly states that there is an association between the gene and the disease - it can be a causative association (e.g. a mutation in the gene causes the disease), or a marker association (e.g. a SNP in the gene identified in a GWAS study); ii) **No Association**: the publications clearly states that there is no association between the gene and the disease (e.g. a publication that reports a negative finding on the association between the gene and the disease), iii) **False Association**: The gene and the disease are found co-occurring in a sentence, but there is no clear evidence from the publication that the gene plays a role or is a marker of the disease and iv) **Error**: when there is a text mining error in the correct identification of the gene and/or the disease.

Table 1 shows some examples of the association types considered in PsyGeNET. In the example for False Association, the study is on children that do not meet the criteria for the disease (FASD) therefore the association between the gene and the disease has to be classified as false. In the example of Error, note that in this abstract OCT is not a gene but an acronym of optical coherence tomography (OCT). The document describing the guidelines is available on the PsyGeNET web page (`http://www.psygenet.org/Psytool_manual_v5.0.pdf`). Here we provide the general instructions for the curation of the gene-disease associations in PsyGeNET:

1. The curation has to be performed at abstract

Association Type	PMID	Sentence
Association	267012	The **D-amino acid oxidase activator gene (G72)** has been found associated with several psychiatric disorders such as **schizophrenia**, major depression, and bipolar disorder.
No Association	17692928	There was no association between **TPH-2** gene variants and **MD** in the same population that had shown a strong association with TPH-1.
False	25225167	Two children referred for suspicion of **FASD** (neither of which were exposed to alcohol or met the criteria for **FASD**) had a pathogenic microstructural chromosomal rearrangement (**del16p11.2** of 542 KB and dup1q44 of 915 KB).
Error	21174530	**OCT** demonstrated loss of foveal **depression** with distortion of the foveal architecture in the macula in all patients

Table 1: Examples of Association types. Disease and genes that have to be evaluated are highlighted in the sentence in green and orange, respectively.

level. For those cases in which abstract is not clear enough, the full text article should be reviewed.

2. Annotate only relationships between the gene and disease. Other types of relationships should not be annotated.

3. Annotate relationships according to the provided categories: association, no association, error, and false.

2.5 Annotation tool

A user-friendly web-based tool was developed to assist both the definition of the psychiatric disorders of interest and curation of gene-disease associations. The tool was designed to support a multi-user environment by user and password assignment. Figure 2 shows a screenshot of the tool for the curation of GDCAs. The tool shows the GDCA to be evaluated (in this example the association between the ETNPPL gene and Bipolar disorders class), and a publication at a time. The curator has to review the publication and decide if the association of the gene and the disease class holds, and decide on the association type using the drop-down menu. To aid the curators task, the tool displays the terminology for the gene according to standard resources (NCBI Gene, UniProt and HGNC), and highlights the sentences in which BeFree identified an association between the gene and the disease under consideration. If required, the curator can access the full text article using the PubMed hyperlink. The curator is also asked to select a sentence that best represents their validation decision, if available. This was implemented in order to collect example sentences to improve the performance of the BeFree system. In addition, the tool also provides a progress bar indicating the number of validations and associations performed by the expert, and allows to review previous annotations. We refer to a validation to each publication supporting a particular GDCA. Note that each publication can have more than one GDCA.

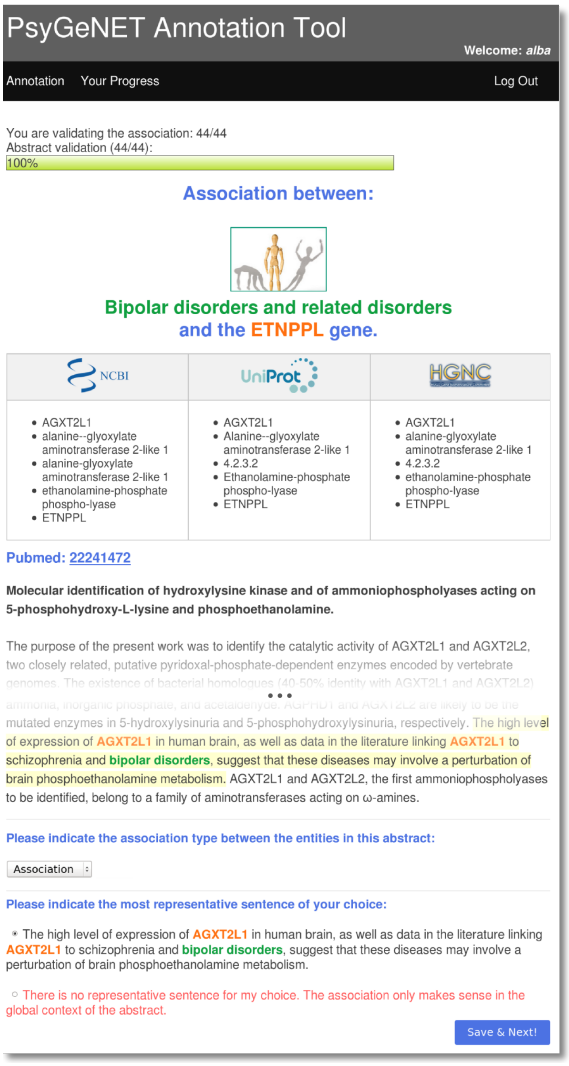

Figure 2: Annotation web-based tool

2.6 Curation workflow

We put in place a curation workflow including a pilot phase and two curation and analysis phases

(see Figure 1). During the pilot phase, the initial training of the curators was carried out including how to use the curation tool. A set of 100 abstracts was validated and analyzed during the pilot phase. After this process both the curation tool and the annotation guidelines were improved and the first curation phase was launched (Curation Phase I), to evaluate 2,507 GDCAs identified by text mining and supported by 4,065 publications. The results of the curation were analyzed to estimate the inter-annotator agreement at the level of abstract. The validations for which an agreement was not found in Curation Phase I are then reviewed by a third expert during Curation Phase II (results not reported here). Four experts are participating in this phase. Only the validations for which agreement of at least 2 experts is found will be included in the database.

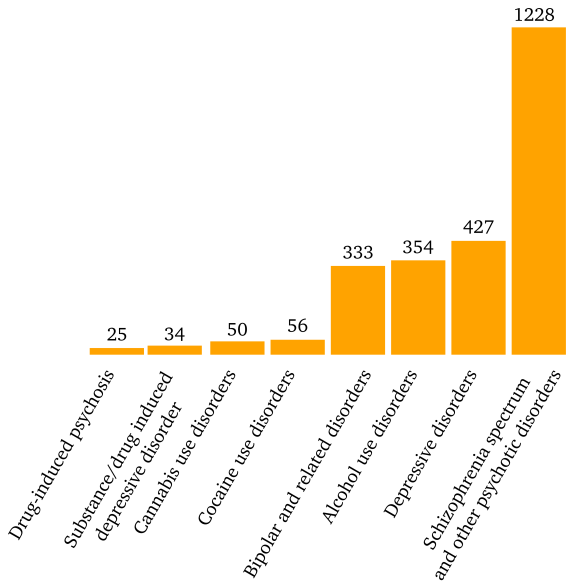

Figure 3: Psychiatric disease categories and the number of associated genes.

3 Results and discussion

Firstly, three experts reviewed the terminology of 2,523 UMLS concepts related to psychiatric disorders of interest. As a result, 1,942 UMLS concepts were assigned to one of the 8 disease categories, being alcohol use disorder, depression and schizophrenia defined by more than 300 concepts (321, 368 and 488, respectively). On the other hand, 581 UMLS concepts were excluded at this stage. Then, BeFree was used to identify gene-disease associations from the literature based on the above disease definition and a subset of the associations focused on the disorders of interest was selected (see methods section 2.3). The 2,507 genes associated to DCs identified by BeFree were submitted to expert curation. These genes were unevenly distributed across the disease categories, being schizophrenia the disease category with more associations followed by depression and alcohol use disorders (see Figure 3).

Of note, most of the GDCAs were supported by only one publication (70.6 %). We included up to the 5 most recent publications for each GDCA for the validation process. This led to 242-284 GDCAs to be validated by each curator, depending on the disease category. Since most of GDCAs are supported by only one publication, the number of publications to be reviewed by the curators ranged between 322 and 491. Before starting the curation of the 2,507 GDCAs, a Pilot curation phase was designed with the purpose of training the curators, testing the PsyGeNET annotation tool, and reviewing the PsyGeNET annotation guidelines. One hundred publications were reviewed during the Pilot phase, distributed in 10 publications per 2 experts. The average agreement between the experts pairs in the Pilot Phase was 60%. The main sources of discrepancies were the handling of speculations, the proper identification of text mining errors, in particular for genes, and the distinction between False and Error Association types. The annotation tool was modified to show the terminology of the genes in order to help the curators to find potential errors in the identification of genes, and by improving the Review function. Then, the proper curation (Curation Phase I in the workflow in Figure 1) was launched and it was completed in 33 days. During Curation Phase I, 2,507 GDCAs supported by 4,065 publications were reviewed by the curators. Each expert was assigned with a set of approx. 275 GDCAs (corresponding to 450 publications) according to their field of expertise (e.g. Major depression vs Schizophrenia). Some curators evaluated associations from all the disease categories, while others focused in a single category. The results of the curation phase I were analyzed to identify agreements and disagreements between the experts. Table 2 shows the number of abstracts validated by each curator team (composed of two experts) and the agreement achieved. The average agreement between all the experts was 68.95%, higher that the one obtained in the Pilot Phase. For one cura-

tor team the agreement was higher (89%) than for the rest of the teams. We can attribute this higher agreement to the fact that there was some communication between the two experts to discuss on the curation criteria during the Curation Phase I.

Teams	Validations	Agreem.	Disagr.	% Agreem.
Team 1	494	325	169	65.79
Team 2	319	194	125	60.89
Team 3	489	342	147	69.94
Team 4	450	402	48	89.33
Team 5	492	308	184	62.60
Team 6	508	341	167	67.12
Team 7	463	317	146	68.46
Team 8	516	363	153	70.35
Team 9	334	221	113	66.17

Table 2: Agreement for each expert pair.

From the validations in which agreement was found (2,813 validations), 1,880 were classified as Association or No Association; 901 were classified as False or Error, and only in 32 of them, the evidence extracted from the publication was not enough to classify them within any of the previous categories, falling in the not clear category (Figure 4). The set of 1,880 validations will be part of the next release of PsyGeNET. Notably, an important fraction of these associations (24.7%) are classified as No association, meaning that there is evidence reporting negative findings on the association between the gene and the disease. This highlights the importance of recording negative findings from the literature in knowledge resources. On the other hand, collecting these information is relevant for the development of corpora for training text mining systems able to identify negative findings regarding gene-disease associations from the literature.

We observe that for 30% of the total GDCAs validated, agreement between curators was not found. A substantial fraction of the disagreements involved the annotation of an association as False by one of the experts (53.28%, see Figure 5). The results of Curation Phase I were discussed with the experts in order to identify the main difficulties during the annotation. The main sources of the discrepancies between curators were the following: i) difficulty in assessing if the studies using animal models captures well the disease pathophysiology, ii) the studies focused on pharmacogenomics or response to drug treatments, iii) studies assessing disease phenotypes (e.g. low mood) in otherwise normal populations, and iv) the assessment of validity of the statistical

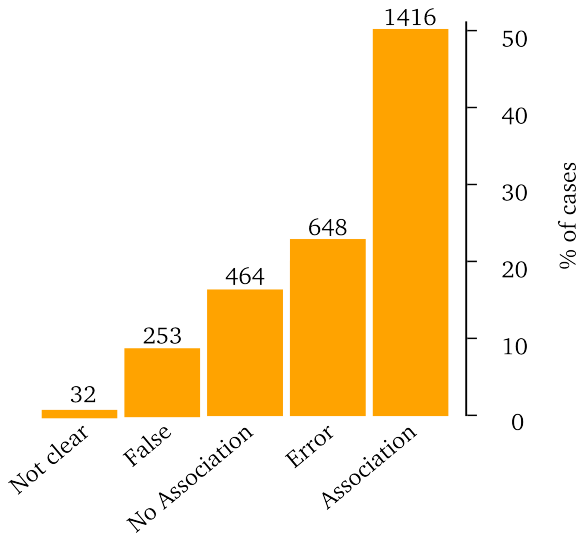

Figure 4: Summary of the agreement results. Each bar in the barplot represents the number of validations annotated as: Association, No association, False, Error and Not clear, respectively.

analysis in some studies (e.g. GWAS studies). In the first case, the decision on the association type will depend on the expertise of the curator on animal model research in psychiatry, that was not the same among the team of experts. In the other three cases the experts expressed difficulties in correctly identifying if an association has to be annotated or not. Overall, although the curation task was very focused to the domain of genetics of psychiatric diseases, the wide variety of studies covered by the publications (GWAs studies, sequencing studies, animal models, etc) require an equivalent diversity of expertise among the experts. We think that this complexity in the task is one of the main reasons for the inter-annotator agreement achieved. Ongoing work includes revisiting the annotation guidelines to further clarify the curation issues raised, in order to improve the agreement in the annotations.

In recent years, many efforts have been made to develop and contribute with novel corpora in the biomedical domain. Nevertheless, the number of corpora annotated with information on gene-disease associations is particularly low (Neves, 2014). For example, the Craven corpus (Craven et al., 1999), contains annotations of gene-disease associations, but there is no information on data quality such as inter-annotator agreement in the original publication. The EU-

ADR corpus (Van Mulligen et al., 2012) includes associations between genes and diseases from 100 MEDLINE abstracts, with an inter-annotator agreement of 86%. Wiegers et al. presented the manual curation of chemical-gene-disease network for the Comparative Toxicogenomics Database (CTD) (Wiegers et al., 2009). For this study 112 articles were distributed between three curators (each one revised less than 60 articles), achieving an inter-annotator agreement of 77%. The CoMAGC corpus (Lee et al., 2013), focused on genes associated to prostate, breast and ovarian cancer, is based on 821 sentences. The authors report an agreement 72%. In another study, agreement over 70% was reported in the development of a sentence-based corpus on prostate cancer-gene associations (Chun et al., 2006). In summary, compared to other corpora annotation initiatives, our inter-annotator agreement results are lower. As described in the paragraphs above, we think that the agreement obtained is due to the complexity of the annotation task. In addition, the large number of experts (for instance, 22 in our case vs 5 in the case of the EU-ADR corpus) and also the large size of our corpus (4,065 publications vs approx. 100 in EU-ADR and CTD corpora) could also explain the lower agreement obtained compared to other curation initiatives.

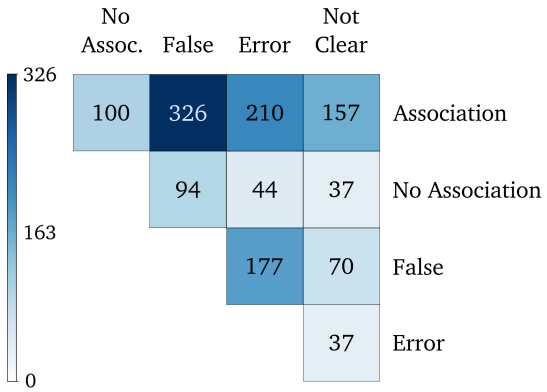

Figure 5: Summary of the disagreement results at the abstract level. Each cell in the heatmap represents the number of abstracts in which disagreement was found for each pair of experts. The darkest the blue, the higher is the disagreement. For example, there were 100 abstracts that one expert annotated as Association while the paired expert annotated as No association.

The Curation Phase II is aimed at reviewing the associations in which no agreement was found among two experts in the first phase of curation. Currently, this involves 1,252 validations, which are being reviewed by a third expert (ongoing work at the time of writing). Finally, the information that will be included in PsyGeNET are the associations in which at least two experts agreed on the annotation.

4 Conclusions

In this communication we report the development of a protocol for the sustainable update of a knowledge resource on the genetics of psychiatric diseases, PysGeNET. We combined state-of-the-art text-mining, data filtering and curation by a community of domain experts for the release of a new version of the database. We designed a protocol that includes curators' training and the iterative improvement of both the tools and annotation guidelines. The proposed approach is allowing to update the database in a timely manner with expert-validated information. Importantly, our curation protocol included the identification of negative findings from the literature. Note that 24.7% of the GDCAs were classified as No association, indicating the importance of properly annotating this information in a knowledge resource. This information will be taken into account for the ranking of the gene-disease association in the next release of PsyGeNET. In addition, the corpus of annotated sentences and abstracts developed during the curation constitutes a valuable resource for the development and evaluation of relation extraction systems. In this era of biomedical big data, we present this approach involving the expert community for the curation of the information as a suitable approach for the development and maintenance of knowledge resources.

5 Fundings

We received support from ISCIII-FEDER (PI13/00082, CP10/00524), IMI-JU under grants agreements n 115002 (eTOX), n 115191 (Open PHACTS)], n 115372 (EMIF) and n 115735 (iPiE), resources of which are composed of nancial contribution from the EU-FP7 (FP7/2007-2013) and EFPIA companies in kind contribution, and the EU H2020 Programme 2014-2020 under grant agreements no. 634143 (MedBioinformatics) and no. 676559 (Elixir-Excelerate). The Research Programme on Biomedical Informatics (GRIB) is a node of the Spanish National Institute

of Bioinformatics (INB).

References

[Baldacchino et al.2009] A Baldacchino, N Groussard-Escaffre, C Clancy, C Lack, K Sieroslavrska, C-L Hodges, L-B Merinder, T Greacen, M Sorsa, H Laijarvi, et al. 2009. Epidemiological issues in comorbidity: lessons learnt from a pan-european isadora project. *Mental Health and Substance Use: Dual Diagnosis*, 2(2):88–100.

[Bravo et al.2015] Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC bioinformatics*, 16(1):1.

[Chun et al.2006] Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, and Jun'ichi Tsujii. 2006. Automatic recognition of topic-classified relations between prostate cancer and genes using medline abstracts. *BMC bioinformatics*, 7(3):1.

[Craven et al.1999] Mark Craven, Johan Kumlien, et al. 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, volume 1999, pages 77–86.

[Gutiérrez-Sacristán et al.2015] Alba Gutiérrez-Sacristán, Solène Grosdidier, Olga Valverde, Marta Torrens, Àlex Bravo, Janet Piñero, Ferran Sanz, and Laura I Furlong. 2015. Psygenet: a knowledge platform on psychiatric disorders and their genes. *Bioinformatics*, page btv301.

[Kessler et al.2005] Ronald C Kessler, Patricia Berglund, Olga Demler, Robert Jin, Kathleen R Merikangas, and Ellen E Walters. 2005. Lifetime prevalence and age-of-onset distributions of dsm-iv disorders in the national comorbidity survey replication. *Archives of general psychiatry*, 62(6):593–602.

[Lee et al.2013] Hee-Jin Lee, Sang-Hyung Shim, Mi-Ryoung Song, Hyunju Lee, and Jong C Park. 2013. Comagc: a corpus with multi-faceted annotations of gene-cancer relations. *BMC bioinformatics*, 14(1):1.

[Murray and Lopez2013] Christopher JL Murray and Alan D Lopez. 2013. Measuring the global burden of disease. *New England Journal of Medicine*, 369(5):448–457.

[Neves2014] Mariana Neves. 2014. An analysis on the entity annotations in biological corpora. *F1000Research*, 3.

[Piñero et al.2015] Janet Piñero, Núria Queralt-Rosinach, Àlex Bravo, Jordi Deu-Pons, Anna Bauer-Mehren, Martin Baron, Ferran Sanz, and Laura I Furlong. 2015. Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015:bav028.

[Sullivan et al.2012] Patrick F Sullivan, Mark J Daly, and Michael O'Donovan. 2012. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews Genetics*, 13(8):537–551.

[Van Mulligen et al.2012] Erik M Van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A Kors, and Laura I Furlong. 2012. The eu-adr corpus: annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, 45(5):879–884.

[Whiteford et al.2013] Harvey A Whiteford, Louisa Degenhardt, Jürgen Rehm, Amanda J Baxter, Alize J Ferrari, Holly E Erskine, Fiona J Charlson, Rosana E Norman, Abraham D Flaxman, Nicole Johns, et al. 2013. Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010. *The Lancet*, 382(9904):1575–1586.

[Wiegers et al.2009] Thomas C Wiegers, Allan P Davis, K Bretonnel Cohen, Lynette Hirschman, and Carolyn J Mattingly. 2009. Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (ctd). *BMC bioinformatics*, 10(1):326.