`psygenet2r`: Case study on GWAS on bipolar disorder

Alba Gutierrez-Sacristan Carles Hernandez-Ferrer
Juan R. Gonzalez Laura I. Furlong

September 7, 2016

Contents

1 Introduction

Psychiatric disorders have a great impact on morbidity and mortality [**?**, **?**]. According to the World Health Organization (WHO), one of every four people will suffer mental or neurological disorders at some point in their lives[**?**]. It has been suggested that most psychiatric disorders display a strong genetic component [**?**, **?**, **?**]. During the last years there has been a growing research in psychiatric disorders' genetics [**?**], and therefore the number of publications that focus on psychiatric disorders have increased steadily (Figure **??**).

Figure 1: Number of publications for psychiatric disorders in PubMed. It has been obtained querying `'psychiatric disorder[Title/Abstract]'` from 1955 to 2016.

However, there is still limited understanding on the cellular and molecular mechanisms leading to psychiatric diseases, which has limited the application of this wealth of data in the clinical practice. This situation also applies to psychiatric comorbidities. Some of the factors that explain the current situation is the heterogeneity of the information about psychiatric disorders and its fragmentation into knowledge silos, and 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 [**?**] has been developed to fill this gap. `psygenet2r` has been developed to facilitate statistical analysis of PsyGeNET data, allowing its integration with other packages available in R to develop data analysis workflows.

`psygenet2r` package allows to retrieve the genes associated to psychiatric diseases, or explore the association between a disease of interest and PsyGeNET diseases based on shared genes. In addition, `psygenet2r` allows the annotation of genes with psychiatric diseases based on expert-curated information. This functionality can be of interest to interpret the results of GWAS or Whole Exome Sequencing studies, in which a list of gene variants is obtained and there is a need to prioritize them based on their functional and clinical relevance. In this context, it would be of interest to know if there is information on their implication in psychiatric diseases. In this Case study we will describe how we can analyze the genes identified in a GWAS study in the context of psychiatric diseases using `psygenet2r`. For this purpose, we will use as an example the data obtained from a GWAS study on bipolar disorder published by [**?**]. In this study, the authors analyzed the brain expression of 58 genes, previously identified in a GWAS of bipolar disorder [**?**], and correlated this information with structural MRI studies to identify brain regions that are abnormal in bipolar disorder. We will use this list of 58 genes from the bipolar disorder study to show the functionality of `psygenet2r` package.

1.1 Objective

The goal of the study is to analyze a set of genes discovered by GWAS in the context of PsyGeNET. More specifically, we want to answer the following questions:

1. Are the genes associated to psychiatric disorders according to PysGeNET?
2. What is the level of evidence of these associations?
3. What is the function of the proteins encoded by these genes related to bipolar disorder?
4. Is bipolar disorder associated to other psychiatric disorders?

2 Implementation

2.1 `psygenet2r` package

PsyGeNET, a knowledge resource for the exploratory analysis of psychiatric diseases and their genes, contains information on eight psychiatric disorders: depression, bipolar disorder, schizophrenia, alcohol, cocaine and cannabis use disorders, substance-induced depressive disorder and psychoses. PsyGeNET database has been developed by automatic extraction of information from the literature using the text mining tool BeFree (http://ibi.imim.es/befree/), followed by curation by experts in the domain. The current version of PsyGeNET (version 2.0) contains 3,771 associations between 1,549 genes and 117 psychiatric disease concepts. `psygenet2r` package contains functions to query and analyze PsyGeNET data, and to integrate with other information, as exemplified in this case study.

2.2 Installation

`psygenet2r` package is provided through Bioconductor [**?**]. To install `psygenet2r` the user must type in the two following commands in R session:

```
> source( "http://bioconductor.org/biocLite.R" )
> biocLite( "psyGeNET2R" )

> library( psygenet2r )
```

3 Questions that can be answered using `psygenet2r`

The first step that has to be done before doing any analysis is saving the genes in an R vector. For this case-study the 58 genes obtained from McCarthy et al. [**?**] are saved into a vector called `genesOfInterest`.

Genes can be identified using the NCBI gene identifier or the Official Gene Symbol from HUGO.

```
> genesOfInterest <- c("ADCY2", "AKAP13", "ANK3", "ANKS1A",
+ "ATP6V1G3", "ATXN1", "C11orf80", "C15orf53", "CACNA1C",
+ "CACNA1D", "CACNB3", "CROT", "DLG2", "DNAJB4", "DUSP22",
+ "FAM155A", "FLJ16124", "FSTL5", "GATA5", "GNA14", "GPR81",
+ "HHAT", "IFI44", "ITIH3", "KDM5B", "KIF1A", "LOC150197",
+ "MAD1L1", "MAPK10", "MCM9", "MSI2", "NFIX", "NGF", "NPAS3",
+ "ODZ4", "PAPOLG", "PAX1", "PBRM1", "PTPRE", "PTPRT",
+ "RASIP1", "RIMBP2", "RXRG", "SGCG", "SH3PXD2A", "SIPA1L2",
+ "SNX8", "SPERT", "STK39", "SYNE1", "THSD7A", "TNR",
+ "TRANK1", "TRIM9", "UBE2E3", "UBR1", "ZMIZ1", "ZNF274")
```

3.1 How many of these genes are in PsyGeNET?

In order to know how many of the genes of interest are present in PsyGeNET, `psygenetGeneList` function is used. This function requires as input the genes' vector and the selected database. For this analysis `"ALL"` database are selected.

```
> m1 <- psygenetGene(
+     gene     = genesOfInterest,
+     database = "ALL",
+     verbose  = FALSE,
+     warnings = FALSE
+ )
> m1

Object of class 'DataGeNET.Psy'
 . Type:         gene
 . Database:     ALL
 . Term:         ADCY2 ... SYNE1
 . N. Results:   48
 . U. Diseases:  15
 . U. Genes:     16
```

The output is a `DataGeNET.Psy` object. It contains all the information about the different diseases associated with the genes of interest retrieved from PsyGeNET. By looking at the `DataGeNET.Psy` object, it can be observed that, according to PsyGeNET and by querying in `ALL` databases, 16 of the initial genes are found in PsyGeNET. These genes appear associated with 15 different disorders, involving a total of 48 gene-disease associations (GDAs).

3.2 Which diseases are associated to these genes according to PsyGeNET database?

In order to visualize the 48 GDAs between the 16 genes found in PsyGeNET and the 15 different disorders, `psygenet2r` provides several options. One of them is the GDA network, which can be obtained by applying the `plot` function to

the `DataGeNET.Psy` object (`m1`), obtained from `psygenetGene` function (section 3.1). In the GDA network, green nodes represent diseases and orange nodes represent genes.

```
> plot( m1 )
```

Figure 2: Gene-Disease Association Network

As shown in Figure **??**, most of the genes (10) are associated to bipolar disorder (umls:C0005586), in agreement with McCarthy et al. [**?**], but some of these genes are also associated to other disorders related to alcohol UD, depression and schizophrenia.

In PsyGeNET it is important to keep track of both "positive" and the "negative" findings, and let the user make their own judgements based on the available evidence. We can visualize the association type (Association, No Association,Both) using heat-maps.

In the example shown in the next figure (Figure **??**), the gene-disease associations related to alcohol UD and depresison are all of type "Association". On the other hand, for bipolar disorder and schizophrenia there are some GDAs that have both types of associations. Finally, in the schizophrenia category it can be seen that there is one GDA with "no association" type.

The barplot can be obtained using the `geneAttrPlot` function and by setting the `type` argument to `"index"`.

```
> geneAttrPlot( m1, type = "index" )
```

Figure 3: Association type barplot according to psychiatric category

3.3 What are the functions of the proteins encoded by these genes?

`psygenet2r` package can be used to analyze gene attributes such as the panther class. The *PANTHER Protein Class Ontology* includes commonly used classes of protein functions. The Panther class can be analyzed using the function `pantherGraphic`.

`pantherGraphic` function requires as input the list of genes (`genesOfInterest` vector) and the database (for instance `ALL`). The output of `pantherGraphic` function is a bar-plot with the different Panther classes in the Y-axis and the percentage of genes in the X-axis, grouped by PsyGeNET psychiatric disorders.

```
> pantherGraphic( genesOfInterest, "ALL")
```

Figure 4: Panther class analysis of the genes of interest.

The bar-plot shown in Figure **??** is obtained from the gene-list of interest. All the genes in the list that are associated with the psychiatric disease class of alcohol use disorders are signaling molecules. On the other hand, it can be observed that those genes associated with bipolar disorder, depression and schizophrenia are in a variety of Panther classes.

3.4 What is the level of evidence for each GDA?

In PsyGeNET, each GDA is ranked with the PsyGeNET evidence index, that reflects the association type for each GDA. We can use `psygenet2r` package to visualize the level of evidence in a heatmap.

In order to obtain the heatmap, the `plot` function can be applied to the `DataGeNET.Psy` object (`m1`). The argument type must to be set to `"heatmapGenes"`.

As a result a heat-map is obtained with genes in the X axis and disorders in the Y axis. Heatmap cells will be coloured in green, yellow or red according to the evidence index value. Green color represents those GDAs where all the evidences reviewed by the experts support the existence of an association between the gene and the disease (Association, EI = 1); it will be yellow when there is contradictory evidence for the GDA (some publications support the association while others publications do not support it, $1 > \text{EI} > 0$), and it will be red when all the evidences reviewed by the experts report that there is no association between the gene and the disease (Association, EI = 0).

```
> plot( m1, type="heatmapGenes")
```

Figure 5: Gene-Disease Association Heatmap

Figure **??** shows that from the total 48 GDAs, only 4 of them have an evidence index different from 1. For example, all the publications supporting C15orf53-schizophrenia association report that there is no association; for the other 3 GDAs (C15orf53-bipolar disorder, DLG2-schizophrenia and SYNE1-bipolar disorder)contradictory evidences have been found.

3.5 For the disorder of interest, how many publications support each gene-disease associations?

In addition to the PsyGeNET evidence index, we can also inspect the number of publications that support a GDA. `psygenet2r` allows the visualization of this information in a bar-plot, using the `plot` function with the `disorder` argument to indicate the disease of interest, and the `type` argument set to `"barplot"`. Figure **??** shows an example, with the genes in the X-axis and the number of PMIDs in the y-axis.

```
> plot( m1, name="bipolar disorder", type="barplot")
```

Figure 6: Publications that report each gene association with bipolar disorder

The results show that the CACNAC1 gene is associated with bipolar disorder in more than 50 publications, followed by ANK3 gene with around 40 publications. The rest of genes have been associated with bipolar disorder in less than 10 publications.

3.6 What are the sentences that report the association between genes and the disease of interest?

`psygenet2r` package also allows the extraction of the sentences and the pmids for each one of the GDAs for a particular disease. Two functions are required, `psygenetGeneSentences` and `extractSentences`. So, first, `psygenetGeneSentences` function is applied to `genesOfInterest` with `ALL` databases. A `DataGeNET.Psy` object is obtained.

```
> m2 <- psygenetGeneSentences(
+     geneList = genesOfInterest,
+     database = "ALL"
+ )
> m2

Object of class 'DataGeNET.Psy'
 . Type:         gene
 . Database:     ALL
 . Term:         ADCY2 ... SYNE1
 . N. Results:   103
```

```
 . U. Diseases:  15
 . U. Genes:     16
```

Then, the `extractSentences` function is applied to the previous `DataGeNET.Psy` object and `disorder` argument is set to the disorder of interest, in this case, `"bipolar disorder"`. Notice that if the disorder name is used, it must be written as it appears in PsyGeNET, otherwise results will not be found. The result is a data frame that contains the gene symbol, gene identifier, disease code, disease name, original db, the pmid, the annotation type and the sentence. As an example the first pmids are shown.

```
> sentences <- extractSentences( m2,
+   disorder = "bipolar disorder" )
> head(sentences$PUBMED_ID)

[1] 24618891 24655771 25304227 24016415 25711502 24809399
```

3.7 Is bipolar disorder significantly associated with other diseases?

An interesting question is to know which diseases are similar to a target disease based on shared genes. Since PsyGeNET database contains information on genes associated to psychiatric diseases we can use it to estimate disease similarity. The Jaccard Index is a statistic used for comparing the similarity of two sets. In our case, these sets are the genes associated to each one of the target diseases. In the `psygenet2r` package, the Jaccard Index is calculated by using the function `jacacardEstimation`.

The function `jacacardEstimation` allows us to calculate the Jaccard Index using both PsyGeNET's data (like disease names or CUIs) and external information as vectors of genes. Moreover, this function calculates p-value for the index to assess its statistical significance.

The strategy to calculate the Jaccard Index and its p-value as follows:

(A) Calculate the Jaccard Index between the pair of diseases. Let's call it *rJI*.

(B) Randomly select a set of genes from DisGeNET for each one of the input diseases (or set of genes) and compute their Jaccard Index (iJI).

(C) Calculate the p-value by dividing the count of the *iJI* higher than the real *rJI* by the number of attempts we performed the step B plus one (`nboot + 1`).

Let's calculate the Jaccard Index for the genes of interest and bipolar disorder:

```
> xx <- jaccardEstimation( genesOfInterest, "bipolar disorder",
+   database = "ALL", nboot = 500 )
> xx
```

```
Object of class 'JaccardIndexPsy'
 . #Boot:               500
 . Type:                gene-list - dise
 . #Results:            1
```

The result shows the number of iterations used to calculate the p-value and the type of input, in this case a list of genes and a disease. The Jaccard Index and the p-value can be extracted using the function `extract`:

```
> extract( xx )
```

```
   Disease1          Disease2 NGenes1 NGenes2 JaccardIndex pval
1 gene-list bipolar disorder      58     502   0.01818182    0
```

Now we have seen the Jaccard Index and its p-value, between our genes of interest and bipolar disorder (JI: 0.018; $pval$: 0). The function `jaccardEstimation` also allows to calculate the Jaccard Index of our input, the list of genes, and all the diseases in PsyGeNET.

```
> xx <- jaccardEstimation( genesOfInterest,
+   database = "ALL", nboot = 500 )
```

The result from the Jaccard Index estimation can be visualized using the function `plot`. The result is a bar-plot where the p-value of each comparison between our genes and PsyGeNET's disease is shown. A `cutOff` argument can be added in order to visualize only those diseases with an statistically significant p-value:

```
> plot( xx )
```

Figure 7: Bar-plot where the Jaccard Index of each comparison between the list of genes of interest and PsyGeNET's diseases is shown.

The similarity between diseases according to shared genes can be visualized using a barplot. `psygenet2r` package allows to visualize how many of our genes of interest are associated to each psychiatric disorder present in PsyGeNET and how many of them are exclusively associated to a particular psychiatric disorder.

In order see the similarity between our genes of interest and PsyGeNET's psychiatric disorder, the `type` argument of `geneAttrPlot` function is set to `category`. This will allow to see how many of the 16 genes, that appear in PsyGeNET, are associated with each psychiatric disorder.

```
> geneAttrPlot( m1, type = "category" )
```

Figure 8: Barplot: Genes associated to each of the psychiatric disorders

Figure **??** clearly shows that none of our genes of interest are associated to the three disorders. There are 2 exclusively associated to bipolar disorder, 2 in the case of depression and 3 in schizophrenia. The rest of the genes appears in more

than one psychiatric disorder.

In order see which are the genes that are associated only to one psychiatric category, the `type` argument of `geneAttrPlot` function can be set to `gene`. This will allow to see how many of the CUIs and categories are associated to each one of the 16 genes.

```
> geneAttrPlot( m1, type = "gene" )
```

Figure 9: Barplot: CUIs and psychiatric categories associated to each gene

Figure **??** clearly shows that these genes are: ADCY2, ATXN1, CACNA1C, HHAT, MAD1L1, NFIX and SH3PXD2A. Yellow bar for these genes show that they are only associated with one psychiatric disorder.

References

[1] Murray Christopher J.L., Lopez Alan D. **Measuring the global burden of disease**. N.Engl. J. Med. 2013. doi:10.1056/NEJMra1201534

[2] Whiteford Harvey A., Degenhardt Louisa, Rehm Jürgen, Baxter Amanda J., Ferrari Alize J., Erskine Holly E., et al. **Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010**. Lancet 2013.

[3] World Health Organization. **The world health report 2001 - Mental Health: New Understanding, New Hope**. ISBN 92-4-156201-3

[4] Issler Orna, and Chen Alon. **Determining the role of microRNAs in psychiatric disorders**. Nature Reviews Neuroscience 2015. doi:10.1038/nrn3879

[5] Levinson Douglas F., Mostafavi Sara, Milaneschi Yuri, Rivera Margarita, Ripke Stephan, Wray Naomi R., Sullivan Patrick F. **Genetic studies of major depressive disorder: why are there no genome-wide association study findings and what can we do about it?** Biol. Psychiatry 2014. doi:http://dx.doi.org/10.1016/j.biopsych.2014.07.029

[6] Schizophrenia Working Group of the Psychiatric Genomics Consortium. **Biological insights from 108 schizophrenia-associated genetic loci**. Nature 2014. doi:10.1038/nature13595

[7] Sullivan Patrick F., Daly Mark J., O'Donovan Michael. **Genetic architectures of psychiatric disordlers: the emerging picture and its implications**. Nat. Rev. Genet. 2012 doi:10.1038/nrg3240.

[8] Alba Gutiérrez-Sacristán, Solène Grosdidier, Olga Valverde, Marta Torrens, Àlex Bravo, Janet Piñero, Ferran Sanz, Laura I. Furlong. **PsyGeNET: a knowledge platform on psychiatric disorders and their genes**. Bioinformatics 2015. doi:10.1093/bioinformatics/btv301

[9] McCarthy Michael J., Liang Sherri, Spadoni Andrea D., Kelsoe John R., Simmons Alan N. **Whole brain expression of bipolar disorder associated genes: structural and genetic analyses**. PLoS One 2014. doi:10.1371/journal.pone.0100204

[10] PGCBD Consortium. **Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4**. Nat Genet 2011. doi:10.1038/ng.943.

[11] Bioconductor `http://www.bioconductor.org`