psygenet2r: Case study on GWAS on bipolar disorder

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

August 25, 2016

Contents

1 Introduction

Psychiatric disorders have a great impact on morbidity and mortality [1, 2]. According to the World Health Organization (WHO), one of every four people will suffer mental or neurological disorders [3]. It has been suggested that most psychiatric disorders display a strong genetic component [4, 5, 6]. During the last years there has been a growing research in psychiatric disorders' genetics [7], and therefore the number of publications that focus on psychiatric disorders have increased steadily (Figure 1).

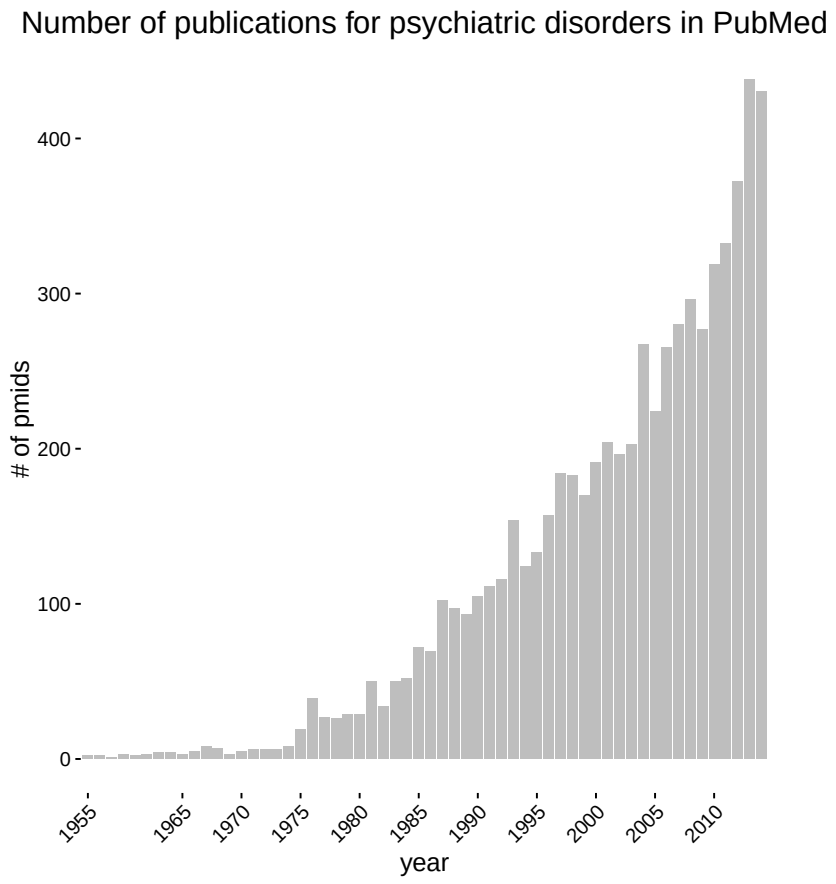

Figure 1: Number of publications for psychiatric disorders in PubMed. It has been obtained querying `'psychiatric disorder[Title/Abstract]'` from 1955 to 2014.

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 [13]. 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 bipolar disorder genes associated to Substance Use Disorders according to PsyGeNET?
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 major depression, cocaine use disorder or alcohol use disorder?

2 Implementation

2.1 psygenet2r package

PsyGeNET, a knowledge resource for the exploratory analysis of psychiatric diseases and their genes, focused on three psychiatric disorders: Major Depression (MD), Cocaine Use Disorder (CUD), Alcohol Use Disorder (AUD). PsyGeNET database is the result of data integration from DisGeNET and data extracted from the literature by text mining, followed by expert's curation [8]. The current version of PsyGeNET contains 2642 associations between 1271 genes and 37 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 [12]. 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. [13] 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"` databases are selected.

```
m1 <- psygenetGene(
    gene     = genesOfInterest,
    database = "ALL",
    verbose  = FALSE,
    warnings = FALSE,
    check = 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, 21 of the initial genes are found in PsyGeNET. These genes appear associated with 10 different disorders, involving a total of 37 gene-disease associations (GDAs).

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

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 4.1). In the GDA network, green nodes represent diseases and orange nodes represent genes.

```
plot( m1 )
```

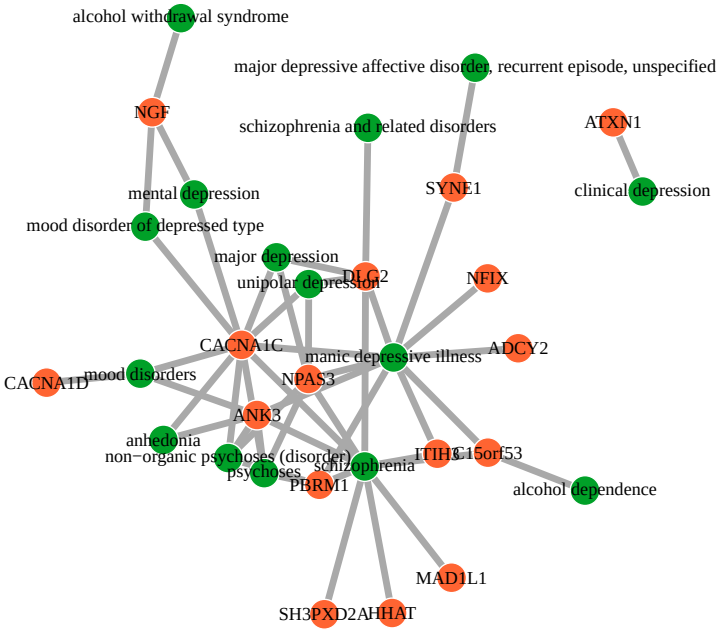

Figure 2: Gene-Disease Association Network

As shown in Figure 2, most of the genes are associated to bipolar disorder, in agreement with McCarthy et al. [13], but some of these genes are also associated to substance related disorders and different depression sub-types.

We can also visualize the gene-disease associations at the level of psychiatric disorder class using heat-maps.

In the example shown in the next figure (Figure 3), the genes associated to the main disorder classes in PysGeNET with a PsyGeNET score larger than 0.1 are shown. Eight from the 21 genes are associated with PsyGeNET diseases with a score greater than 0.1. ANK3 gene is the one with the highest mean score, approx. 0.6, and is associated to major depression class. Only one gene, NGF, passes the cut off in the cocaine use disorder group. On the other hand, none of the genes are associated to alcohol use disorders with a score higher than 0.1.

The heat-map can be obtained using the `plot` function and by setting the `type` argument to `"heatmap"` or `"heatmapScore"`. The cut-off argument can be added with the minimum score mean to be shown in the resulting graphic.

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

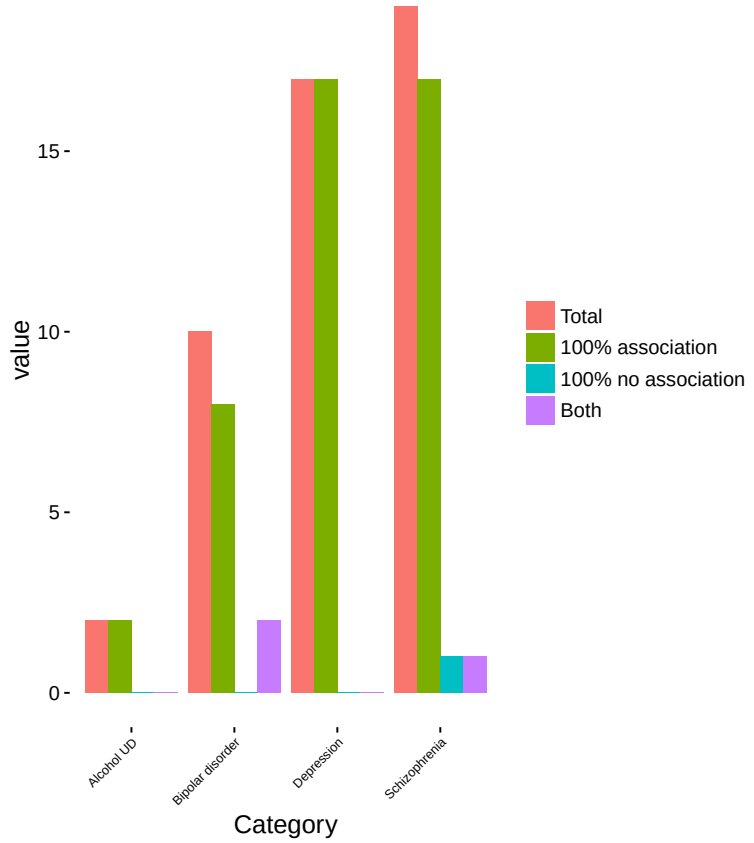

Figure 3: Gene-Psychiatric disorder association heat-map according to the PsyGeNET evidence index.

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 to which the genes belong according to their psychiatric disorder 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", check = FALSE)
```

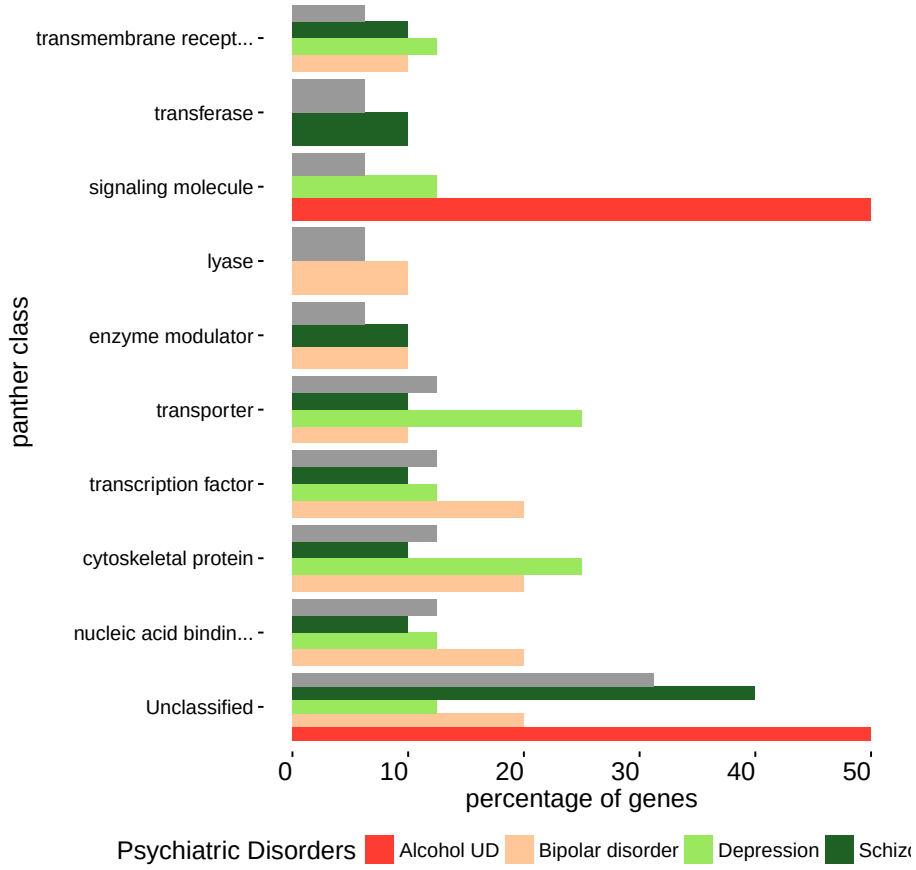

Figure 4: Panther graphic analysis of the genes of interest.

The bar-plot shown in Figure 4 is obtained from the gene-list of interest. All the genes in the list that are associated with the psychiatric disease class of cocaine use disorders are signaling molecules. On the other hand, it can be observed that those genes associated with alcohol use disorder are 25% receptors and 25% phosphatases. Finally, those genes associated with major depression present the highest range of possible panther class.

3.4 What is the level of evidence for each GDA?

In PsyGeNET, each GDA is ranked with the PsyGeNET score, that reflects the level of evidence for each association. We can use PsyGeNET score to visualize the level of evidence in a heat-map.

In order to obtain the heat-map, the `plot` function can be applied to the `DataGeNET.Psy` object (`m1`). The argument type must to be set to `"heatmapGenes"` and the score cut-off can also be determined by the user. Notice that from the more than 100 associations, only the 100 with the highest scores will be shown.

To visualize those GDAs with a PsyGeNET score > 0.3, the `cutOff` argument is set to 0.3. As a result a heat-map is obtained with genes in the X axis and disorders in the Y axis. The green rank color is proportional to the PsyGeNET score, being the darkest one the association with the highest score.

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

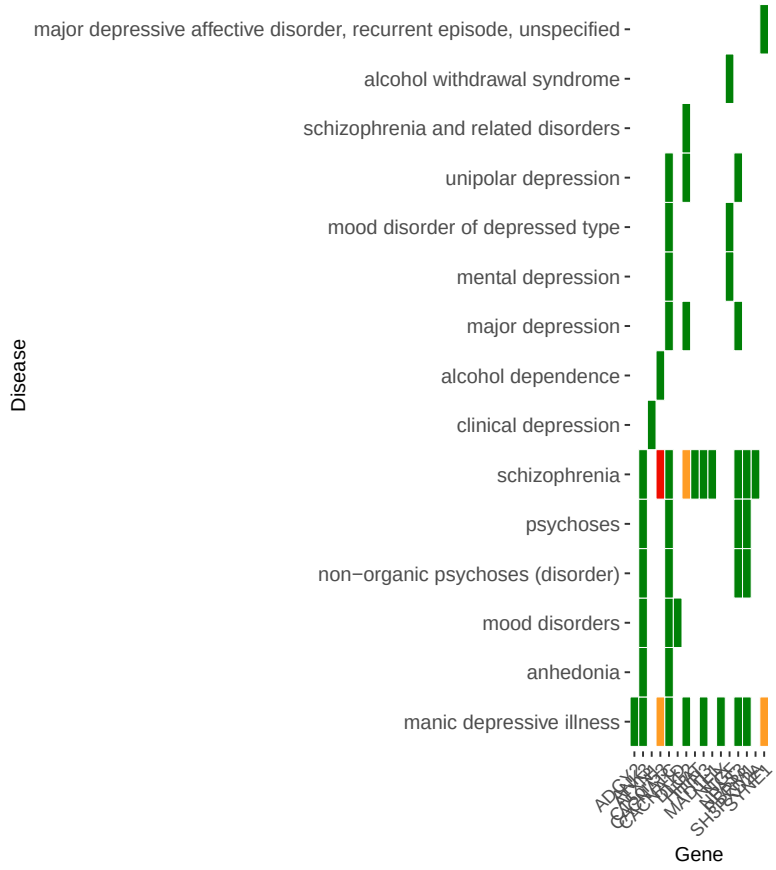

Figure 5: Gene-Disease Association Heat-map for GDAs with PsyGeNET score greater than 0.3.

Figure 5 shows that those GDA with score greater than 0.3 are related to major depression disorders, specially with bipolar disorder. The genes with the greatest score (approximately 0.8) are ANK3 and CACNA1C genes, which appear strongly associated with bipolar disorder. These results are in agreement with the ones obtained in a meta-analysis of GWAS where these genes were identified to contain the higher risk alleles to bipolar disorder [11].

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

In addition to the PsyGeNET score, 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 6 shows an example, with the genes in the X-axis and the number of PNIDs in the y-axis.

```
plot( m1, name="manic depressive illness", type="barplot")
```

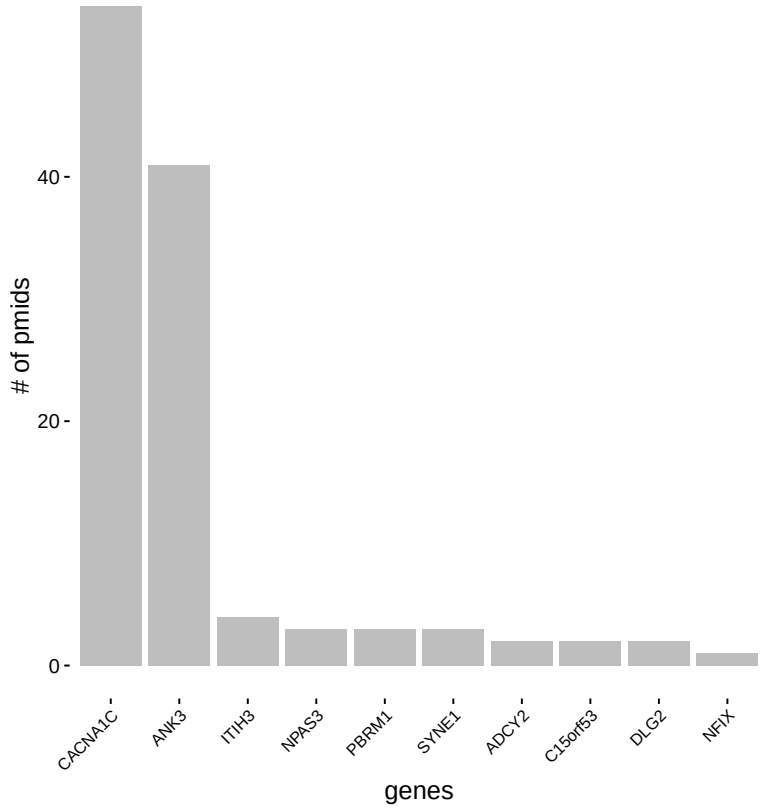

Figure 6: PubMed Ids that report each gene association with bipolar disorder (manic depressive illness).

The results show that the CACNAC1 gene is associated with bipolar disorder in more than 40 publications, followed by ANK3 gene with more than 30 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"
)

## Warning in psygenetGeneSentences(geneList = genesOfInterest, database
= "ALL"):  One or more of the given genes is not in PsyGeNET ( 'ALL'
). Genes:  AKAP13, ANKS1A, ATP6V1G3, C11orf80, CACNB3, CROT, DNAJB4,
DUSP22, FAM155A, FLJ16124, FSTL5, GATA5, GNA14, GPR81, IFI44, KDM5B,
KIF1A, LOC150197, MAPK10, MCM9, MSI2, ODZ4, PAPOLG, PAX1, PTPRE, PTPRT,
RASIP1, RIMBP2, RXRG, SGCG, SIPA1L2, SNX8, SPERT, STK39, THSD7A, TNR,
TRANK1, TRIM9, UBE2E3, UBR1, ZMIZ1, ZNF274

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, disease, original db, the pmid and the sentence. As an example the first pmids are shown.

```
sentences <- extractSentences( m2,
  disorder = "manic depressive illness" )
head(sentences$PUBMED_ID)

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

3.7 Is bipolar disorder significantly associated with other diseases?

An interesting calculation 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 the Jaccard Index, this function calculates an associated p-value to the index.

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

(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).

(C) See how many of these genes are in common and divide them for the total length to calculate their Jaccard Index. Let's call it *iJI*.

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

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

```
xx <- jaccardEstimation( genesOfInterest, "manic depressive illness",
  database = "ALL", nboot = 500 )
xx

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

The result shows the number of interactions 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 manic depressive illness      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*: 0.02; *pval*: 0.002). 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 )

## Warning in singleInput.genes(diseases$diseases$`gene-list`$genes,
psy, universe, :  Jaccard Index for all diseases in PsyGeNET will be
calculated.
```

The result from the Jaccard Index estimation can be plot 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, in this case it is not necessary:

```
plot( xx )
```

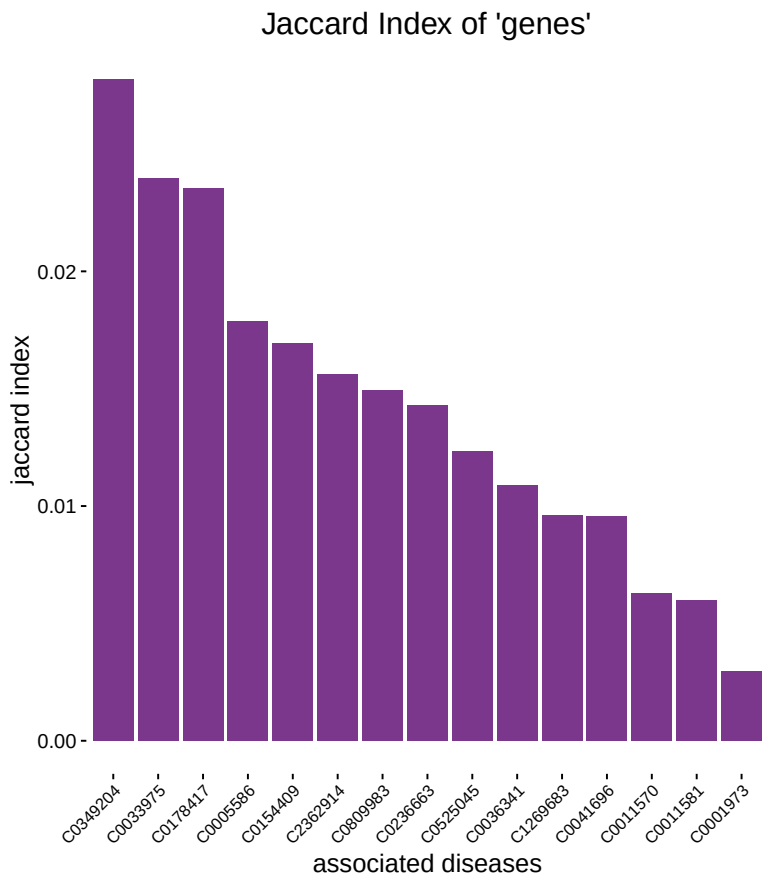

Figure 7: Bar-plot where the Jaccard Index of each comparison between the list of genes of interest and PsyGeNET's diseases is shown.

The similarity between diseases according to shared genes can be visualized using Venn diagrams. `psygenet2r` package allows to plot two types of Venn diagrams: the conventional one and a variation of it. In the variation of the Venn diagram, a triangle is drawn, representing each vertex the three psychiatric disorders from PsyGeNET (MD, AUS and CUD). In the middle of the triangle, those genes that are in common for the three psychiatric disorders will appear, while in the vertex will be represent those that are exclusively to each one of them.

In order see the similarity between our genes of interest and PsyGeNET's psychiatric disorder, the `type` argument of `plot` function is set to `VennA`. This will allow to see how many of the 21 genes, that appear in PsyGeNET, are associated with each psychiatric disorder:

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

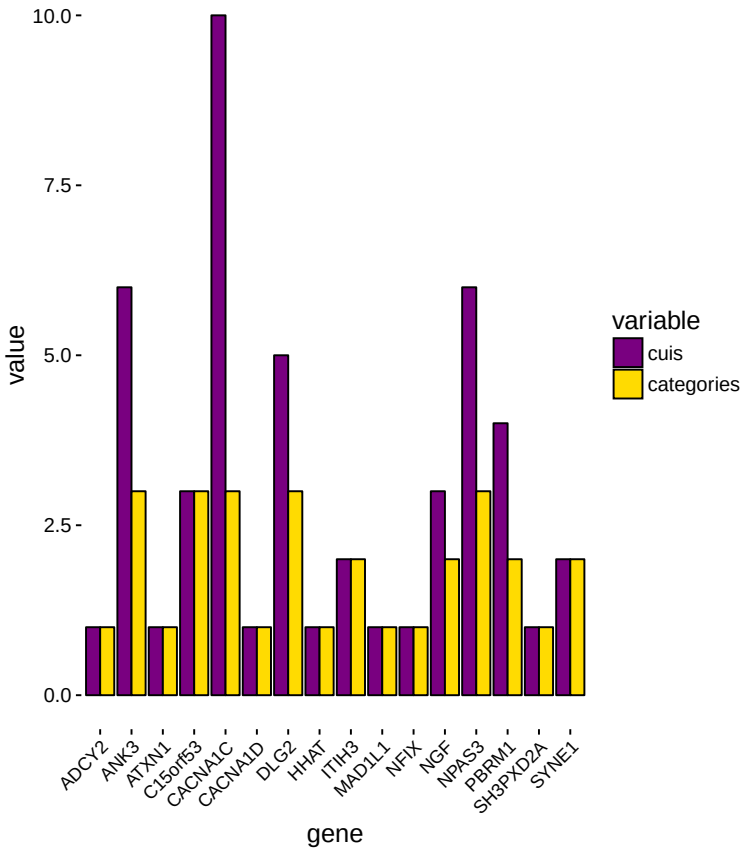

Figure 8: Venn Diagram: Genes shared between the three main psychiatric disorders

Figure 9 clearly shows that none of our genes of interest are associated to the three disorders. There are 14 exclusively associated to major depression and 4 to alcohol use disorder. The gene that appears associated to cocaine use disorder is also associated to major depression. Alcohol use disorder and major depression share two genes.

```
geneAttrPlot( m1, type = "pie" )
```

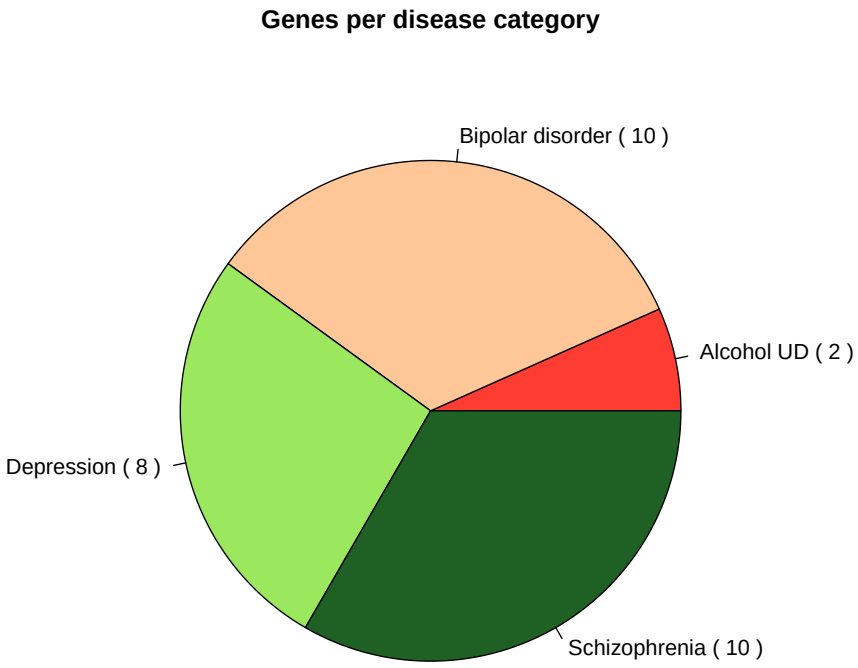

Figure 9: Venn Diagram: Genes shared between the three main psychiatric disorders

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] Black Donald W., Andreasen Nancy C. **Introductory Textbook of Psychiatry (5th ed.)**. American Psychiatric Publishing, Inc., pg. 158 ISBN-13: 978-1585624003

[10] McGuffin Peter, Rijsdijk Fruhling, Andrew Martin, Sham Pak, Katz Randy, Cardno Alastair. **The heritability of bipolar affective disorder and the genetic relationship to unipolar depression**. JAMA Psychiatry 2003. doi:10.1001/archpsyc.60.5.497

[11] 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.

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

[13] 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