

psygenet2r: An R package for querying PsyGeNET and to perform comorbidity studies in psychiatric disorders

Alba Gutierrez-Sacristan

Carles Hernandez-Ferrer
Laura I. Furlong

Juan R. Gonzalez

September 5, 2016

Contents

Introduction

The **psygenet2r** package contains functions to query PsyGeNET [?], a resource on psychiatric diseases and their genes. The **psygenet2r** package includes analysis functions to study psychiatric diseases, their genes and disease comorbidities. A special focus is made on visualization of the results, providing a variety of representation formats such as networks, heatmaps and barplots.

Background

During the last years there has been a growing interest in the genetics of psychiatric disorders, leading to a concomitant increase in the number of publications that report these studies [?]. 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.

PsyGeNET is a resource for the exploratory analysis of psychiatric diseases and their associated genes. The second release of PsyGeNET (version 2.0) contains updated information on depression, bipolar disorder, alcohol use disorders and cocaine use disorders, and has been expanded to cover other psychiatric diseases of interest: bipolar disorder, schizophrenia, substance-induced depressive disorder and psychoses and cannabis use disorder. PsyGeNET allows the exploration of the molecular basis of psychiatric disorders by providing a comprehensive set of genes associated to each disease. Moreover, it allows the analysis of the molecular mechanisms underlying psychiatric disease comorbidities.

PsyGeNET database is the result of the data extracted from the literature by text mining using BeFree [?], followed by manual curation by domain experts. A team of 22 experts participates as curators of the database. The current version of PsyGeNET (version 2.0) contains 3,771 associations between 1,549 genes and 117 psychiatric disease concepts.

With **psygenet2r** package the user will be able to submit queries to PsyGeNET from R, perform a variety of analysis on the data, and visualize the results through different types of graphical representations.

The tasks that can be performed with **psygenet2r** package are the following:

1. Retrieve Gene-Disease Associations (GDAs) from PsyGeNET using as query a gene or a disease (single or a set of genes/diseases) of interest
2. Visualize the results according to the GDAs' attributes: PsyGeNET evidence index, number of publications, sentences that report the GDA, source database
3. Visualize the results according to the disease (disease class) or gene (Panther class) attributes
4. Analyze the association between two diseases from the molecular perspective (using the Jaccard index)

In the following sections the specific functions that can be used to address each of these tasks are presented.

Installation

The package **psygenet2r** is provided through Bioconductor. To install **psygenet2r** the user must type the two following commands in an R session:

```
source( "http://bioconductor.org/biocLite.R" )
biocLite( "psygenet2r" )
```

```
library( psygenet2r )
```

DataGeNET.Psy object

DataGeNET.Psy object is obtained when **psygenetGene** and **psygenetDisease** functions are applied. This object is used as input for the rest of **psyGeNET2r** functions, like the **plot** function.

DataGeNET.Psy object contains all the information about the different diseases/genes associated with the gene/disease of interested retrieved from PsyGeNET. This object contains a summary of the search, such as the search input (gene or disease), the selected database, the gene or disease identifier, the number of associations found (N. Results) and the number of unique results obtained (U. Results).

```
t1

## Object of class 'DataGeNET.Psy'
## . Type:          gene
## . Database:      ALL
## . Term:          4852
## . N. Results:    13
## . U. Diseases:   13
## . U. Genes:      1

class( t1 )

## [1] "DataGeNET.Psy"
## attr(,"package")
## [1] "psygenet2r"
```

This object comes with a series of functions to allow users to interact with the information retrieved from PsyGeNET. These functions are **ngene**, **ndisease**, **extract** and **plot**. The first function **ngene** returns the number of retrieved genes for a given query. **ndisease** is the homologous function but for the diseases. The function **extract** returns a formatted **data.frame** with the complete set of information downloaded from PsyGeNET. Finally, the **plot** function allows the visualization of the results in a variety of ways such as gene-disease association networks or heatmaps.

PsyGeNET and psygenet2r

The PsyGeNET web interface can be explored by searching a specific gene or a specific disease, and **psygenet2r** package has the same options. Therefore, the starting point for **psygenet2r** are **psygenetGene** and **psygenetDisease** functions.

PsyGeNET data is classified according to the database used as a source of information ("source database"). Therefore, any query run on PsyGeNET requires to specify the source database using the argument called **database**. Table ?? shows the source databases in PsyGeNET and their

description. By default, the database **ALL** is used in **psygenet2r**. For illustrating purposes along the vignette, database **ALL** will be used in most of code snippets.

Table 1: Source databases included in PsyGeNET

Name	Description
psycur15	Genes associated to DEP, BD, AUD and CUD between 1980 and 2013 (PsyGeNET release v1.0)
psycur16	Genes associated to DEP, BD, AUD, CUD, SCHZ, S-DEP, CanUD and D-PSY between 1980 and 2015
ALL	All previous Databases

Retrieve gene-disease associations (GDAs) from psygenet2r

Using genes as a query

psygenet2r package allows exploring PsyGeNET information using a specific gene or a list of genes. It retrieves the information that is available in PsyGeNET (associated diseases, source database, PsyGeNET evidence index, number of publications, attributes of genes, etc) and allows to visualize the results in different ways.

Using as a query a single gene

In order to look for a single gene into PsyGeNET, we can use the **psygenetGene** function. This function retrieves PsyGeNET's information using both, the NCBI gene identifier and the official Gene Symbol from HUGO. It contains also other arguments like the database to query, the PsyGeNET evidence index (score argument).

As an example, the gene *NPY*, whose entrez id is 4852 is queried using **psygenetGene** function, and using alternatively the official HUGO Gene Symbol. In this example database "ALL".

```
t1 <- psygenetGene( gene = 4852,
                    database = "ALL")
t1

## Object of class 'DataGeNET.Psy'
## . Type:          gene
## . Database:      ALL
## . Term:          4852
## . N. Results:    13
## . U. Diseases:   13
## . U. Genes:      1
```

```
t2 <- psygenetGene( gene = "NPY", database = "ALL" )
t2

## Object of class 'DataGeNET.Psy'
## . Type:          gene
## . Database:      ALL
## . Term:          NPY
## . N. Results:    13
## . U. Diseases:   13
## . U. Genes:      1
```

Both cases result in an **DataGeNET.Psy** object:

```
class( t1 )
```

```
## [1] "DataGeNET.Psy"
## attr(,"package")
## [1] "psygenet2r"

class( t2 )

## [1] "DataGeNET.Psy"
## attr(,"package")
## [1] "psygenet2r"
```

In the particular example used, by inspecting the `DataGeNET.Psy` object, we can see that the gene *NPY* is associated to 13 different diseases in PsyGeNET (with no restriction on the PsyGeNET evidence index).

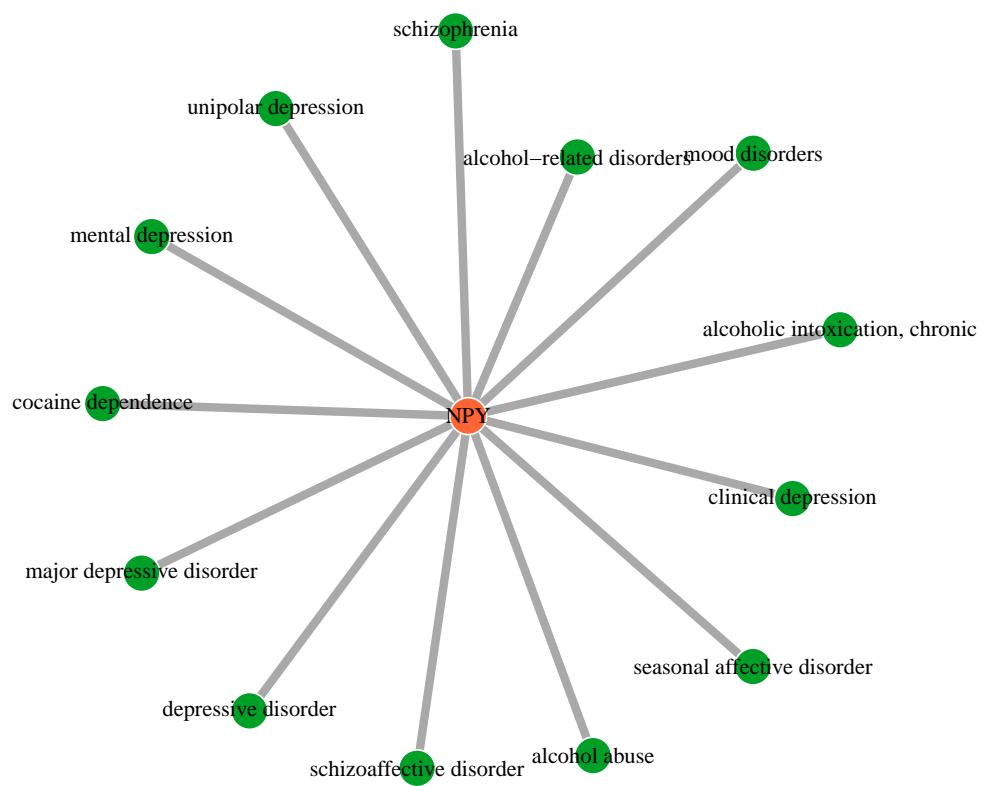
Plotting the results of a Single Gene Query

`psygenet2r` offers several options to visualize the results from PysGeNET: a network showing the diseases related to the gene of interest, or a network showing the strength of the association between the eight main psychiatric disorders in PsyGeNET and the gene of interest. Each one of these graphics can be obtained changing the `type` argument.

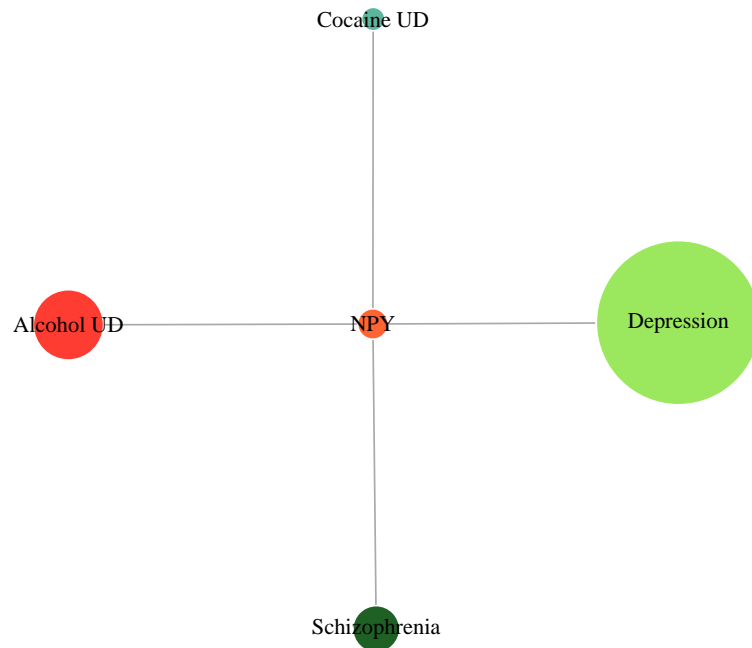
By default, `psygenet2r` shows this type of network when plotting a `DataGeNET.Psy` object obtained by a gene-query. The result is a network where green nodes are diseases and the orange node is the gene of interest.

On the other hand, results can be visualized according to the 8 psychiatric disorders classes available in PsyGeNET (depression, bipolar disorder, alcohol use disorders, cocaine use disorder, bipolar disorder, schizophrenia, cannabis use disorder, substance-induced depressive disorder and psychoses) setting the `type` argument to `"disease class"`. As a result, a network with 5 nodes is obtained. The node's size of each psychiatric disorder is proportional to the number of disease concepts that belongs to each disease class, from the total number of diseases associated to the gene.

```
plot( t1, type = "individual disease" )
```



```
plot( t1, type = "disease class" )
```



In our example, NPY is associated to four of the eight psychiatric disorders present in PsyGeNET, with an important contribution of depression.

Using as a query a list of genes

In the same way, `psygenet2r` allows to query PsyGeNET given a list of genes of interest. The same function, `psygenetGene`, accepts a vector of NCBI gene identifiers or HUGO official gene symbols.

To illustrate this functionality, a list of 20 genes was extracted from the article entitled "*The Genetics of Major Depression*" [?], where these genes are associated to depression. The vector of genes can be defined as follows:

```
genesOfInterest <- c( "COMT", "CLOCK", "DRD3", "GNB3", "HTR1A",  
                      "MAOA", "HTR2A", "HTR2C", "HTR6", "SLC6A4",  
                      "ACE", "BDNF", "DRD4", "HTR1B", "HTR2B",  
                      "HTR2C", "MTHFR", "SLC6A3", "TPH1", "SLC6A2",
```

```

)
" GABRA3"
)

```

Then, the function `psygenetGene` is applied. In this case an extra argument called `verbose` was set to `TRUE`. This shows some information during the query-process, for example the message informing that there are repeated genes in the list (the gene *HTR2C* was placed twice in the list to raise this message) and the message informing that one or more of the given genes is not in PsyGeNET (in this case the gene *HTR2B*).

```

m1 <- psygenetGene(
  gene      = genesOfInterest,
  database  = "ALL",
  verbose   = TRUE
)

## Warning in psygenetGene(gene = genesOfInterest, database = "ALL", verbose = TRUE):
## Removing duplicates from input genes list.
## Starting querying PsyGeNET for COMT, CLOCK, DRD3, GNB3, HTR1A, MAOA, HTR2A, HTR2C,
## HTR6, SLC6A4, ACE, BDNF, DRD4, HTR1B, HTR2B, MTHFR, SLC6A3, TPH1, SLC6A2, GABRA3 in
## ALL database.
## Warning in psygenetGene(gene = genesOfInterest, database = "ALL", verbose = TRUE):
## One or more of the given genes is not in PsyGeNET ( 'ALL' ):
##      - HTR2B

```

A `DataGeNET.Psy` object is obtained. In this particular example, 19 genes are present in PsyGeNET and are associated to 42 diseases, involving 212 GDAs.

```

m1

## Object of class 'DataGeNET.Psy'
## . Type:          gene
## . Database:      ALL
## . Term:          COMT ... GABRA3
## . N. Results:    212
## . U. Diseases:   42
## . U. Genes:      19

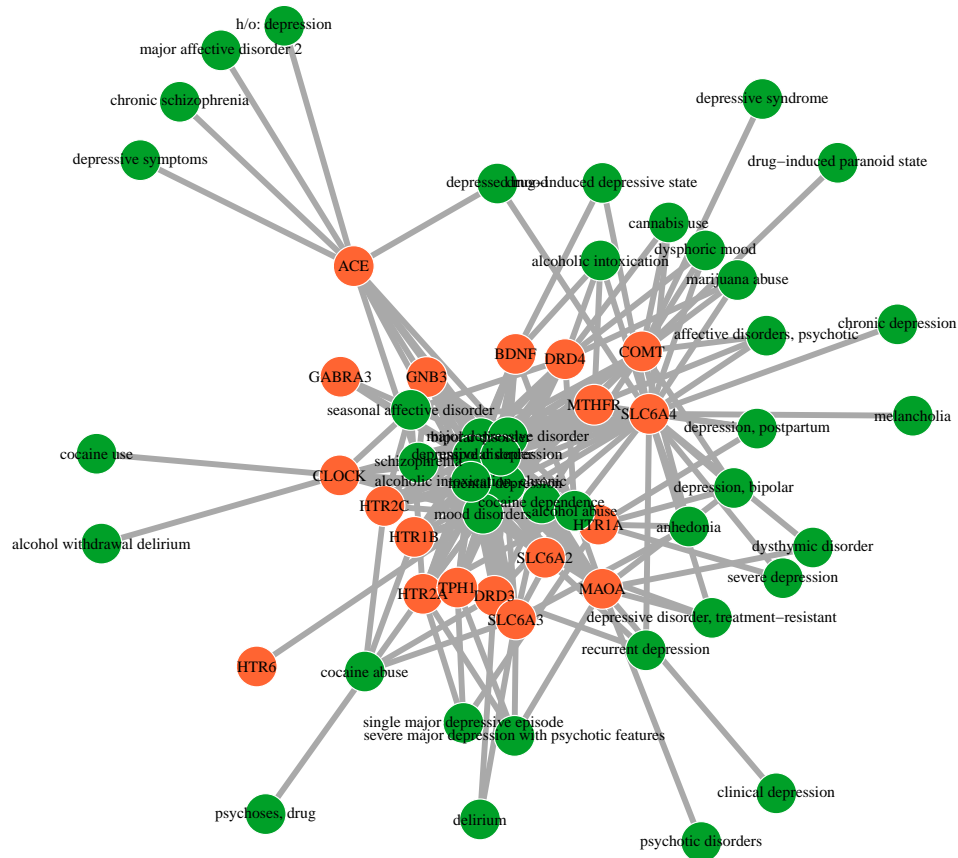
```

Plotting the results of the query using a list of genes

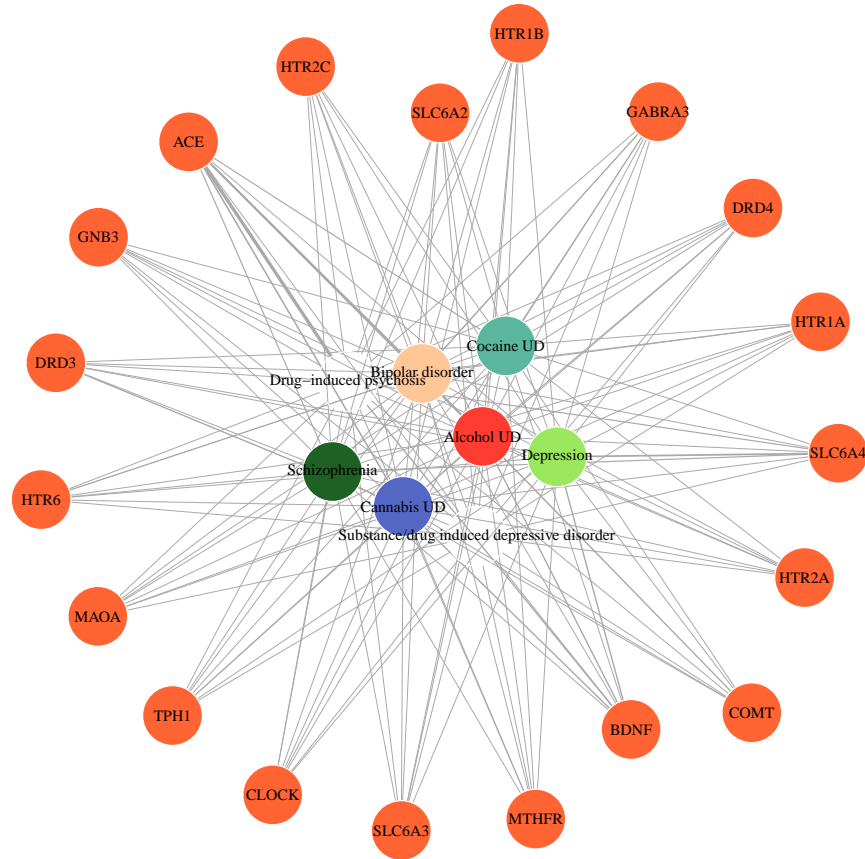
`psygenet2r` provides several options to visualize the results of these queries, such as networks, heatmaps and barplots.

As for the single gene example, the default option in `psygenet2r` results is a network chart, where the green nodes represent diseases and the oranges nodes represent genes. It is also possible to visualize the results by grouping the diseases according to the psychiatric disorders present in PsyGeNET specifying the argument type.


```
plot( m1 )
```



```
plot( m1, type = "disease class" )
```

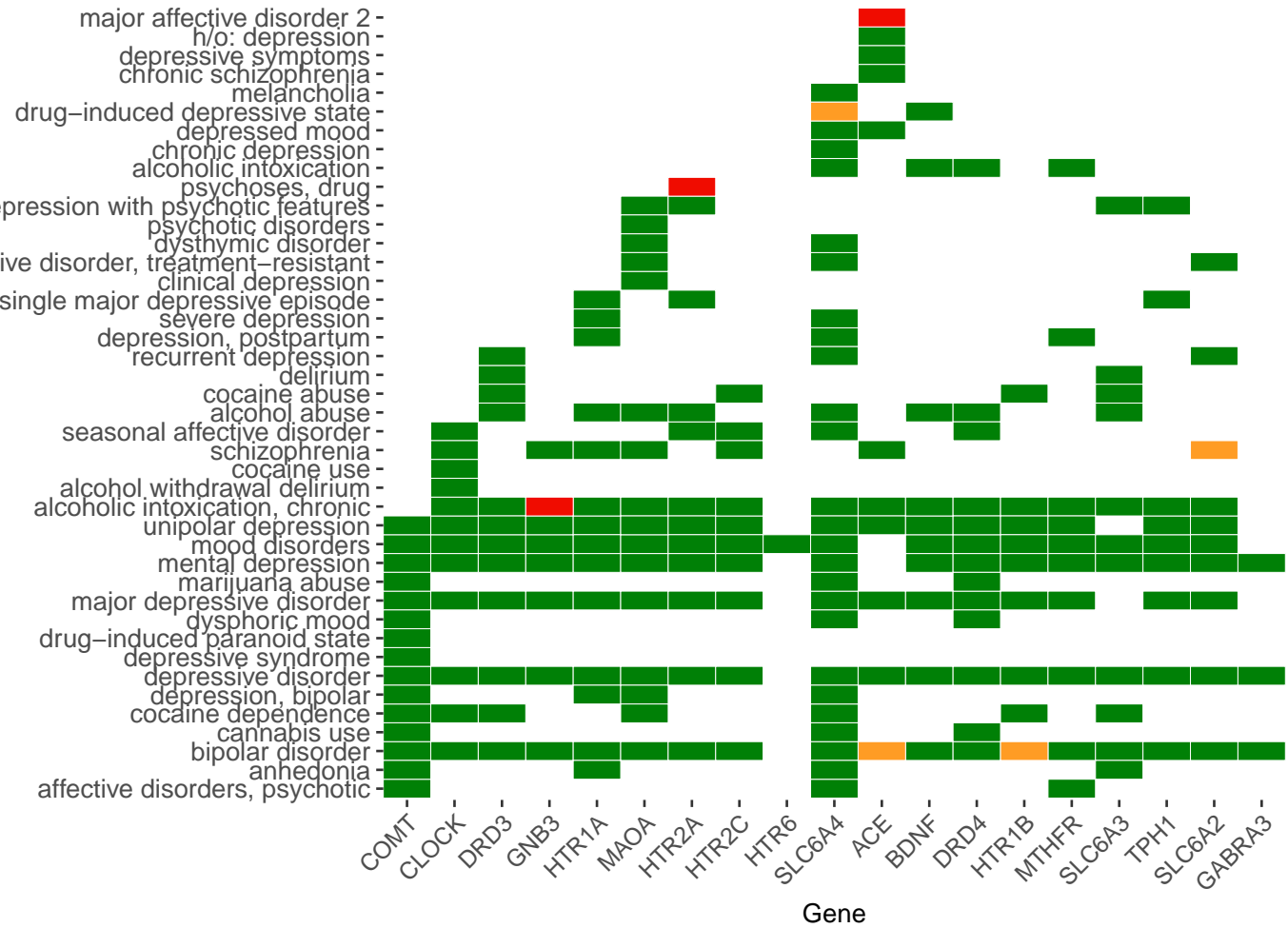


psygenet2r package allows to visualize the GDAs attributes in a heatmap. The argument **type** must be "heatmapGenes" and the PsyGeNET evidence index can also be determined by the user setting the **cut-off** argument to the evidence index of interest. In this example, the cut-off is set to 0, in order to obtain all the results. If we set the cut-off to 0.5, only those associations with at least half of the publications supporting the association will be shown. In this kind of representation we can identify genes that are associated to several diseases (e.g. SLC6A4), others that are associated only to one disease (e.g. HTR6) and we can visualize the evidence index for each association.

Note that heatmap cells can be coloured in green, yellow or red. 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, $1 > EI > 0$), 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" )
```

Disease



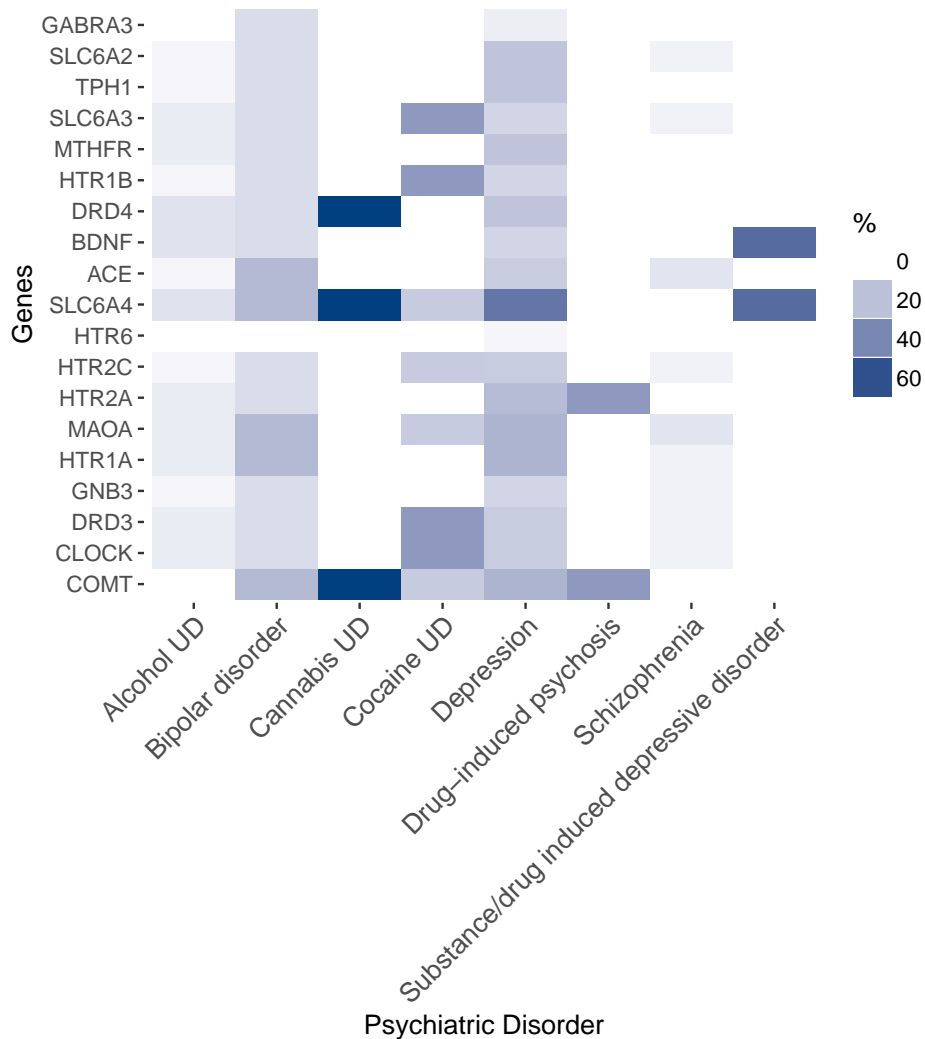
Gene

In this example we can see that there are 3 GDAs in red (GNB3-chronic alcoholic intoxication; HTR2A-drug psychoses and ACE-major affective disorder 2), indicating that all publications report that there is no association.

An alternative graphical representation is a psychiatric disorder heatmap related to the percentage of diseases to which a gene appear associated. It allows to analyze if the genes that are being studied present a specific association with a subtype of disorder or if they are associated with several of them in the same psychiatric disorder class. The percentage of diseases to which a gene appear associated with each psychiatric disorder is estimated. This percentage is relative to the total possible subtypes of disorders present in PsyGeNET (33 for alcohol UD, 9 for bipolar disorder, 37 for depression, 24 for schizophrenia, 6 for cocaine UD, 3 for cannabis UD, 3 for DI-Psychosis and 2 for SI-Depression). The resultant values are represented in a heatmap according to a blue color scale.

In order to obtain this graphic, **type** must be set to **"heatmap"**. The resulting heatmap shows which are the genes that are higher or lower associated to each one of the psychiatric disorders.

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



As it could be expected for the genes that are used as genes of interest, all of them are associated to depression, to a greater or lesser extent. Some of them are also associated to the other six

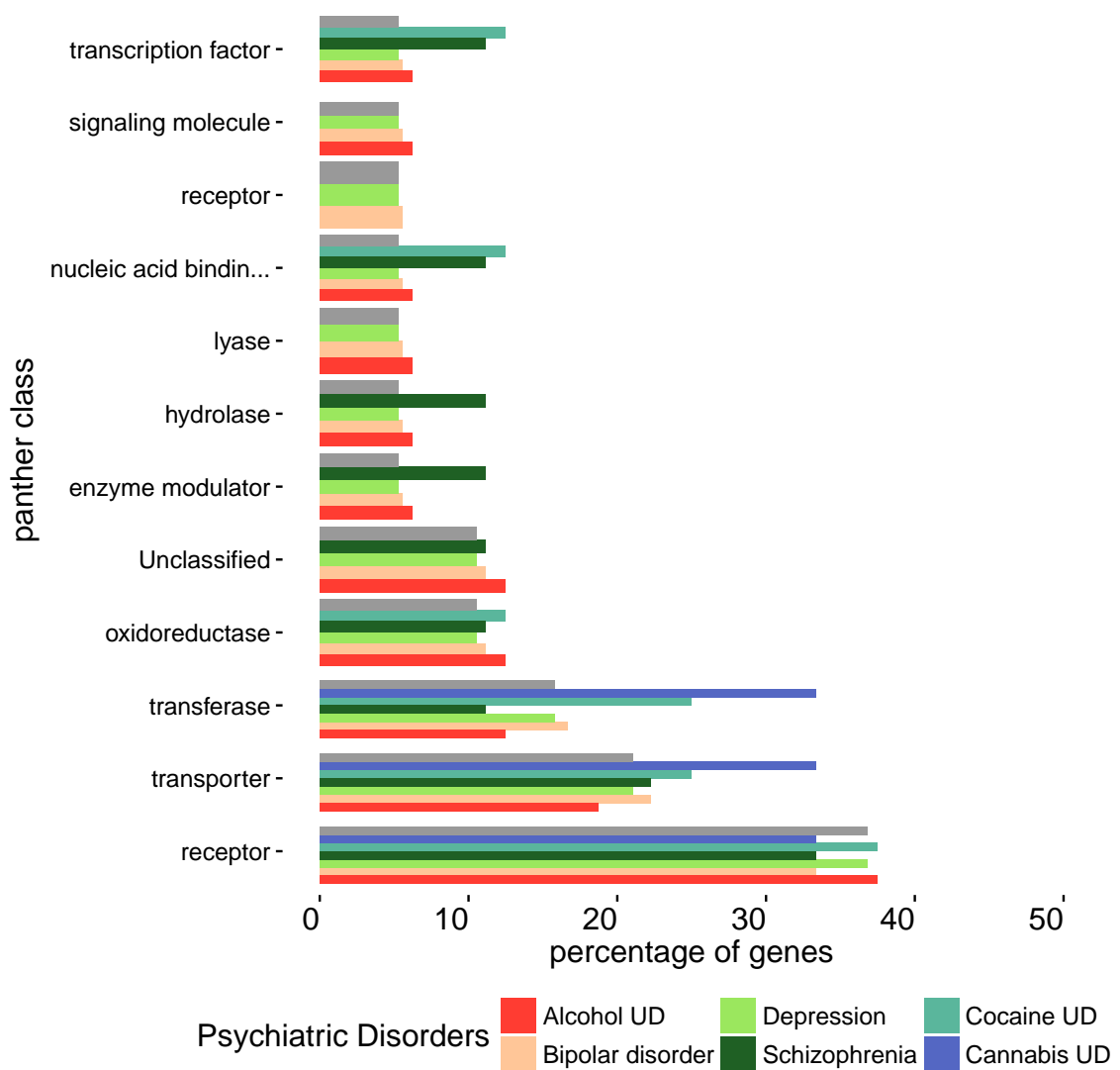
psychiatric disorders, following a similar pattern alcohol UD and bipolar disorder.

`psygenet2r` package also allows to analyze a gene list according to the function of the proteins encoded by these genes. The PANTHER Protein Class Ontology classifies proteins according to their function.

The `pantherGraphic` function shows the Panther class to which the proteins belong according to their associated psychiatric disorder. It provides a graphic with the results of these analysis, being the input a list of genes and the database (`ALL`, `psycur15`, `psycur16`). The input genes can be from a vector that contains the genes of interest, or from the genes obtained in the `DataGeNET.psy` object in a disease or disease-list query. An score argument can be added to filter results. It can also be done given a `DataGeNET.psy` object obtained by querying with a single gene.

```
genesOfInterest <- unique( genesOfInterest )
pantherGraphic( genesOfInterest, "ALL" )

## Warning in psygenetGene(geneList, database, verbose = verbose): One or more of
the given genes is not in PsyGeNET ( 'ALL' ):
##      - HTR2B
```

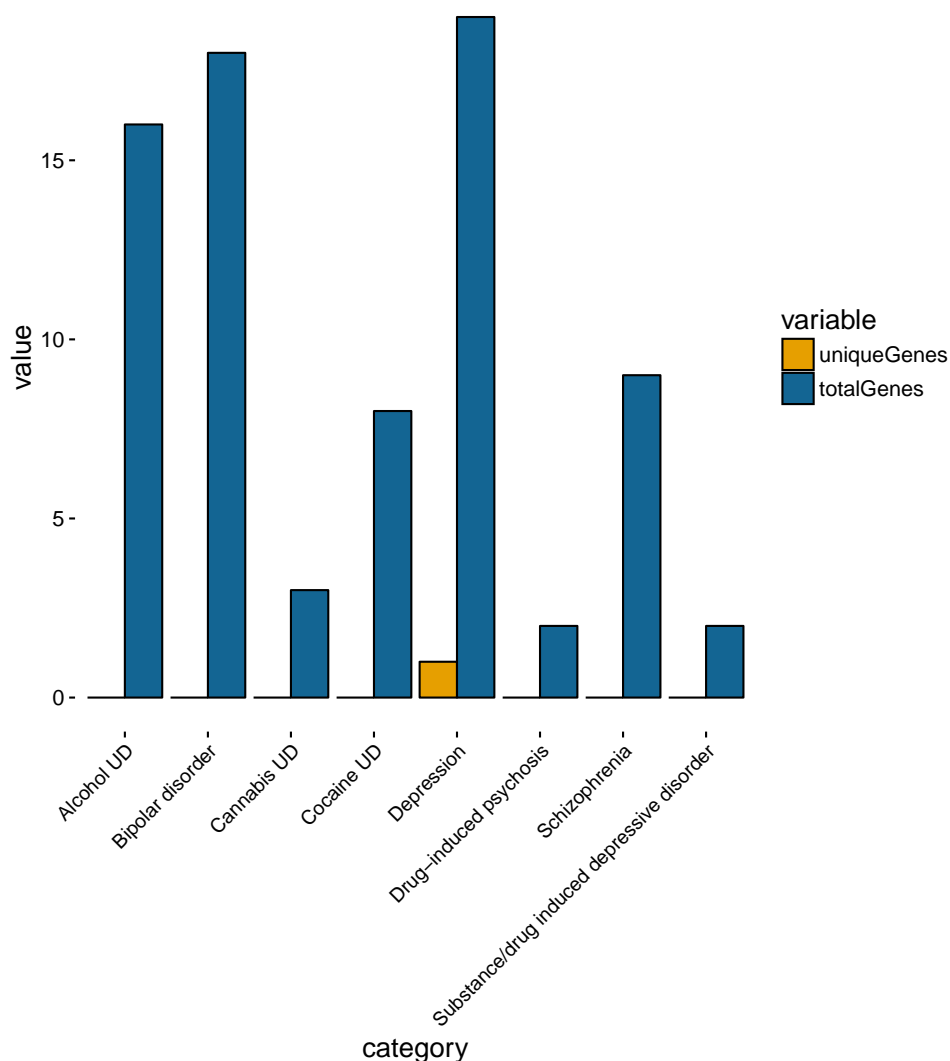


Plotting results of the multiple gene query

psygenet2r offers several options to visualize the results of a query using a list of genes or from the genes obtained in the **DataGeNET.psy** object in a gene or gene-list query. Barplots and pie charts showing the gene attributes can be obtained by applying the **geneAttrPlot** function.

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. This can be done applying the **geneAttrPlot** function and setting the **type** argument to "category".

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

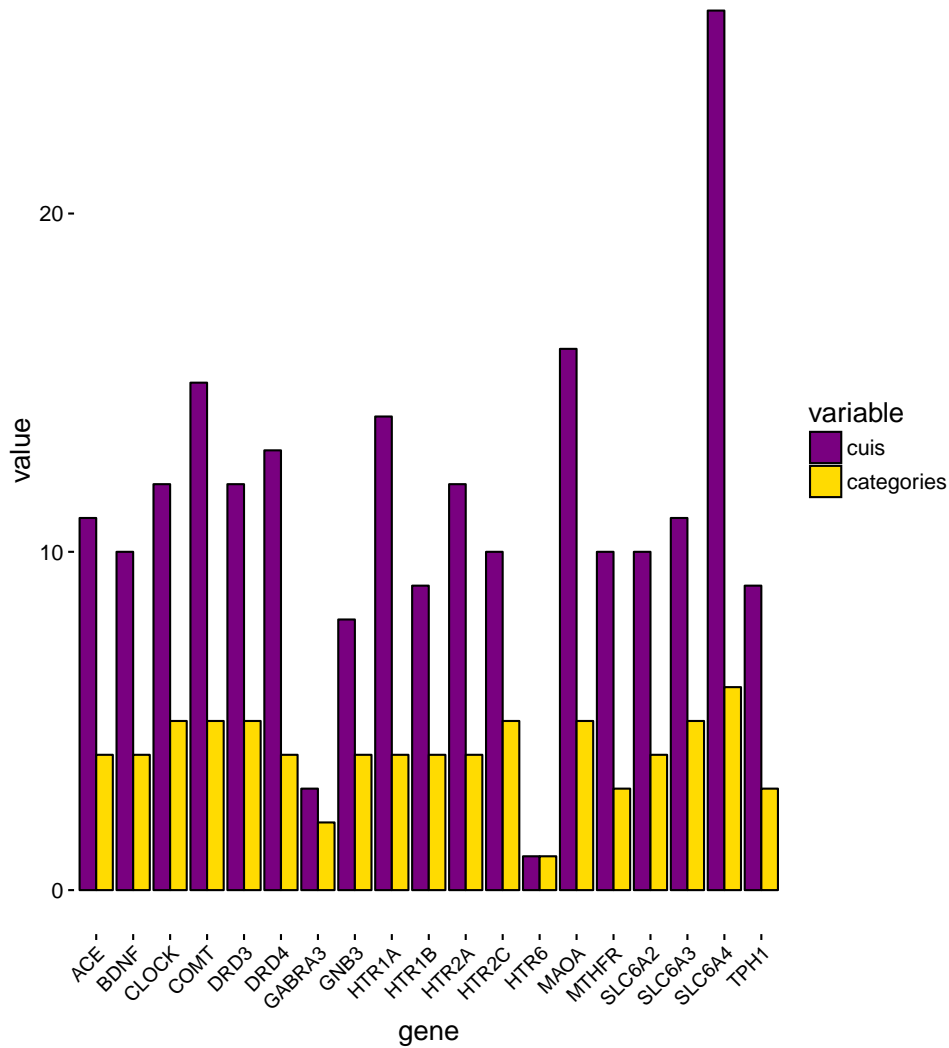


As a result, a barplot is obtained. The X axis contains the psychiatric disorders, sorted alphabetically and the number of genes are represented in the Y axis. For each psychiatric disorder a maximum of two bars can be plotted, the blue bar represents the total number of genes for the psychiatric disorder. For those cases in which some of these genes are specific for the disorder a second bar, coloured in orange, will be displayed.

Alternatively, **psygenet2r** package allows to visualize for each gene, how many disease concepts and

how many psychiatric categories are associated to it. This can be done applying the `geneAttrPlot` function and setting the `type` argument to "gene".

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



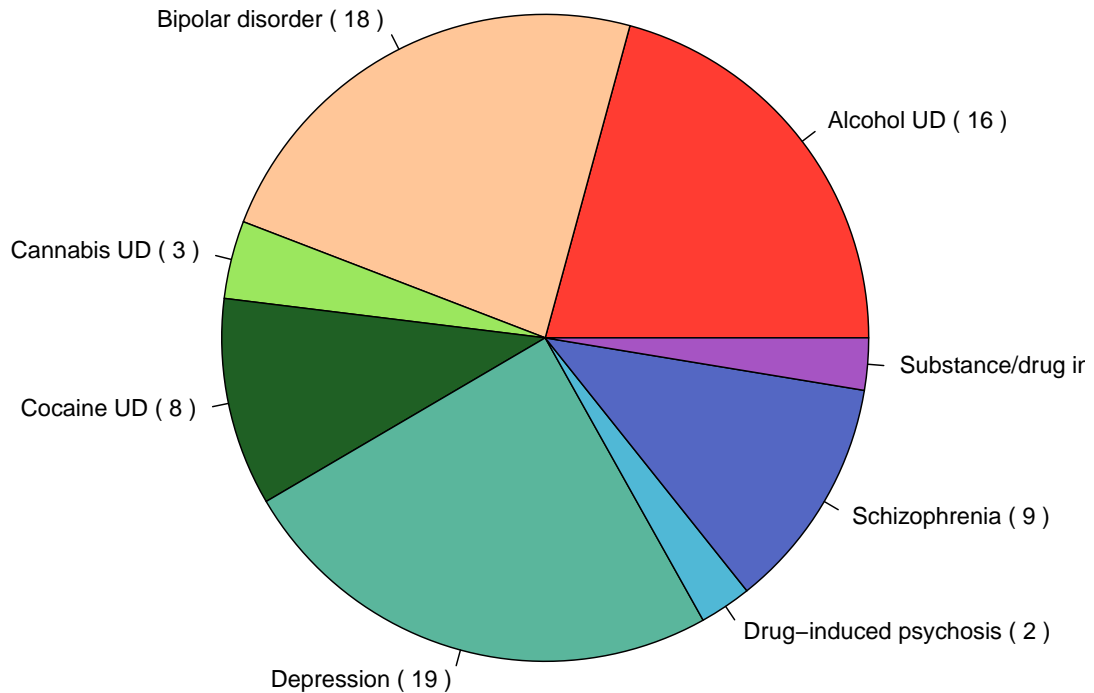
As a result, a barplot is obtained. The X axis contains the genes, sorted alphabetically and the number of cuis and psychiatric categories are represented in the Y axis. For each psychiatric disorder two bars are plotted, the purple bar represents the number of cuis and the yellow one belongs to the number of categories.

In our example, the gene SLC6A4 is the one with more associated cuis (26) and categories (6).

If we are only interested in how many of our input genes are associated to each disease category, `psygenet2r` package allows to visualize it in a pie chart by applying the `geneAttrPlot` function and setting up the `type` argument to `pieChart`.

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

Genes per disease category

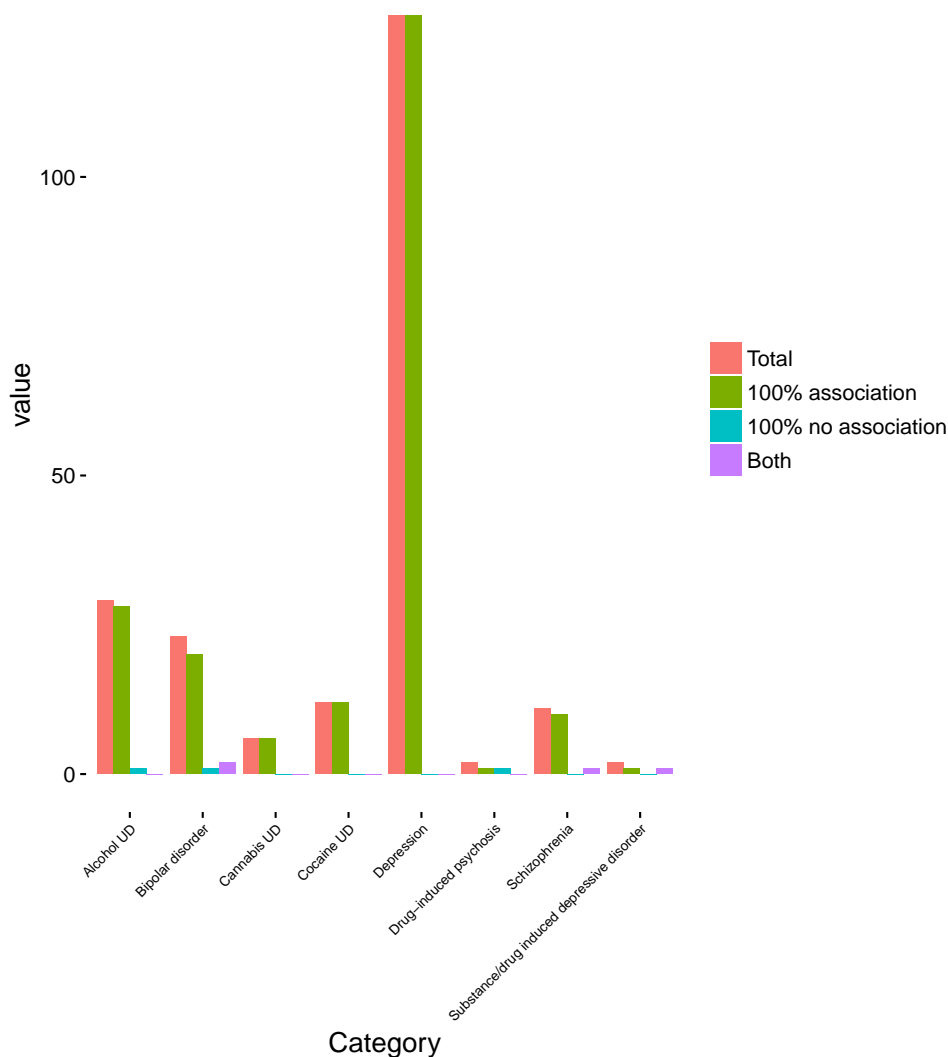


In our example, depression has 19 associated genes, and the majority of them are also associated to other psychiatric disorders, like bipolar disorder (18 genes) and alcohol UD (16 genes).

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. Thus, for each GDA and each supporting publication, the association type information is provided. According to the evidence, there are two types: “Association” and “No Association” (e.g. the “negative findings”).

`psygenet2r` package allows to visualize this information in a representative barplot. It can be done by applying the `geneAttrPlot` function and setting the `type` argument to `"index"`. As a result a barplot showing, for each psychiatric disorder, how many gene-disease association are in total (red bar), how many of them are 100% association (green bar), 100% no association (blue bar) and how many GDAs are supported by both association types (purple bar).

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

In our example, the barplot shows that depression is the psychiatric disorder with more gene-disease associations (127 GDAs), and there is no negative evidence for any of the associations.

Enrichment analysis for a list of genes

The R package `psygenet2r` allows to perform an enrichment analysis on a list of genes with PsyGeNET diseases. It is done by using the function `enrichedPD`. In order to illustrate this function, the previous list of 20 genes associated to depression [?] will be used.

```
tbl <- enrichedPD( genesOfInterest, database = "ALL")
tbl
```

##		MPD	p.value
## 1	Alcohol use disorders	1.873960e-08	
## 2	Bipolar disorders and related disorders	2.559998e-08	
## 3	Depressive disorders	6.286367e-09	
## 4	Schizophrenia spectrum and other psychotic disorders	8.854607e-01	
## 5	Cocaine use disorders	3.697393e-07	
## 6	Substance/drug induced depressive disorder	1.361419e-02	

```
## 7 Cannabis use disorders 3.406556e-03
## 8 Drug-induced psychosis 1.737779e-02
```

The result is a table with a p-value of the enrichment of the given list of genes for each psychiatric disorder in PsyGeNET. As we can see, if we put a p-value cut-off of 0.01, these genes are enriched in 5 of the 8 psychiatric disorders, being alcohol UD (p-val 9e-10) and depression (p-val 6e-9), the ones with the lowest p-value.

Enrichment analysis based on anatomical terms (TopAnat) for a list of genes

psygenet2r package allows to perform gene set enrichment test based on expression of genes in anatomical structures, importing data from the Bgee database [?] and importing functions from BgeeDB R package [?].

It is done by using the function `topAnatEnrichment`. This function perform the enrichment analysis using a list of genes (NCBI gene identifier or official Gene Symbol from HUGO). It contains also other arguments like the `dataType` (rna_seq or affymetrix), `statistics`, that by default is `fisher` and the `cutOff` argument.

```
tpAnat <- topAnatEnrichment( genesOfInterest, cutOff = 1 )

##
## Warning: an organ relationships file was found in the working directory and will be used as is.
##
## Warning: an organ names file was found in the working directory and will be used as is. Please
##
## Warning: a gene to organs mapping file was found in the working directory and will be used as i
##
## Parsing the results.....
##
## Adding BGEE:0 as unique root of all terms of the ontology.....
##
## Done.
##
## Checking topAnatData object.....
##
## Checking gene list.....
## Warning: Some genes in your gene list have no expression data in Bgee, and will not be included
##
## Building 'most specific' Terms..... ( 22 Terms found. )
##
## Build DAG topology..... ( 178 terms and 293 relations. )
##
## Annotating nodes (Can be long)..... ( 1577 genes annotated to the nodes. )
##
## Building the results table for the 178 significant terms at FDR threshold of 1... Done
```

Warnings

heatmap type argument do not allow queries for single gene:

```
> plot( t1, type = "heatmap" )
```

=> Error: For this type of chart, a multiple gene query created with 'psygenetGene' is required.

References

- [1] Alba Gutierrez-Sacristan; Solene Grosdidier; Olga Valverde; Marta Torrens; Alex Bravo; Janet Pinero; Ferran Sanz; Laura I. Furlong. **PsyGeNET: a knowledge platform on psychiatric disorders and their genes** Bioinformatics 2015 doi: 10.1093/bioinformatics/btv301
- [2] Sullivan, Patrick F; Daly, Mark J; O'Donovan, Michael. **Genetic architectures of psychiatric disorders: the emerging picture and its implications** Nature reviews. Genetics (2012) vol. 13 (8) p. 537-51
- [3] Janet Piñero, Núria Queralt-Rosinach, Àlex Bravo, Jordi Deu-Pons, Anna Bauer-Mehren, Martin Baron, Ferran Sanz, Laura I Furlong. **DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes** Database (2015) Vol. 2015: article ID bav028; doi:10.1093/database/bav028
- [4] Bravo, À.; Piñero, J.; Queralt, N.; Rautschka, M.; Furlong, L.I. **Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research** BMC Bioinformatics 2015, 16:55 doi:10.1186/s12859-015-0472-9
- [5] Flint, Jonathan; Kendler, Kenneth S. **The genetics of major depression** Neuron (2014) Vol. 81 (3) p. 484-503
- [6] Jens Treutlein, Sven Cichon et al. **Genome-wide association study of alcohol dependence**. Archives of general psychiatry (2009) vol.66(7) p.773 doi: 10.1001/archgenpsychiatry.2009.83.
- [7] Komljenovic A and Roux J **BgeeDB: an R package for annotation and gene expression data retrieval from Bgee database** (2016)
- [8] Bastian F **Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species** Data Integration Life Sci. Lecture Notes in Computer Science (2008), pp. 124-31.