exposomeShiny User's Guide

Escribà Montagut, Xavier; González, Juan R.

2021-02-26

Contents

Chapter 1

Overview

exposomeShiny is a data analysis toolbox with the following features:

- Data handling: imputation, LOD, transformation, ...
- Exposome characterization
- Exposome-wide association analysis
- Multivariate association
- Omic data integration
- Omic data association
- Post-omic data analysis: CTD database
- Post-omic data analysis: Enrichment analysis

To do so, exposomeShiny relies on previously existent Bioconductor packages (rexposome, omicRexposome and CTDquerier among others), it uses them in a seamless way so the final user of exposomShiny can perform the same studies that would conduct using the Bioconductor packages but without writing a single line of code.

Chapter 2

Setup

2.1 R, RStudio and Packages Versions

If the user chooses to install and use exposomeShiny using RStudio instead of Docker, the list of package versions used for the development of the application is provided for stability purposes. When using the Docker version of the application, all of the following is bundled on the image so the user does not have to deal with the installation of any package.

Software:

R software	Version
R	4.0.2
RStudio	1.4.1103

R packages:

R Packages	Version
shiny	1.5.0
shinyBS	0.61
rexposome	1.12.2
omicRexposome	1.12.0
MultiDataSet	1.18.0
mice	3.11.0
DT	0.16
ggplot2	3.3.2
data.table	1.13.2
truncdist	1.0
shinyalert	2.0.0

R Packages	Version
shinydashboard	0.7.1
shinyjs	2.0.0
TxDb.Hsapiens.UCSC.hg19.knownGene	3.2.2
org.Hs.eg.db	3.12.0
GenomicRanges	1.42.0
CTDquerier	1.4.3
shinycssloaders	1.0.0
pastecs	1.3.21
shinyWidgets	0.5.4
clusterProfiler	3.18.1
enrichplot	1.10.2
ggupset	0.3.0
imputeLCMD	2.0
pls	2.7-3

There are two different ways of setting up and using exposomeShiny

2.2 Downloading the source files, installing the libraries and running the application

The user can choose to download the source code of the shiny application and install all the required libraries on their local R installation. Make sure Rtools is installed to use this method.

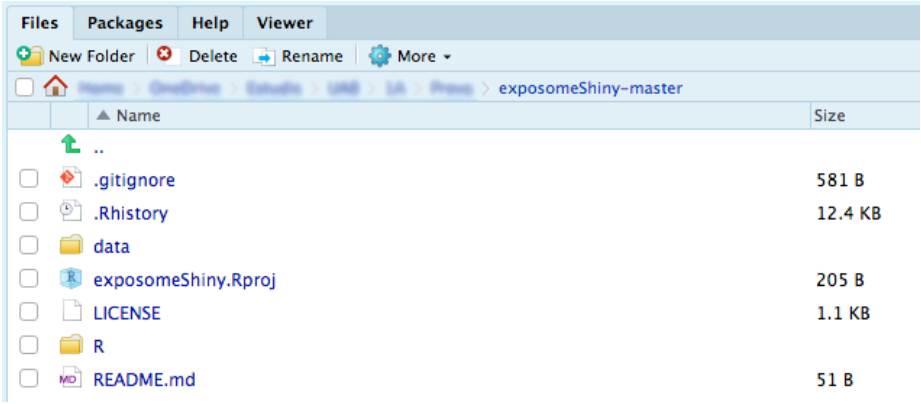
```
  # Set working directory
setwd(dir = "/some/path/")

  # Download zip
download.file(url = "https://github.com/isglobal-brge/exposomeShiny/archive/master.zip"

  # Unzip the .zip to the working directory
unzip(zipfile = "master.zip")

  # Set the working directory inside the downloaded folder
setwd(dir = "/some/path/exposomeShiny-master")
```

Now all the source files are downloaded to the location of chose and the working directory moved to the correct folder, to start the project, open the `Rproj` file by clicking it on the Files explorer of RStudio.

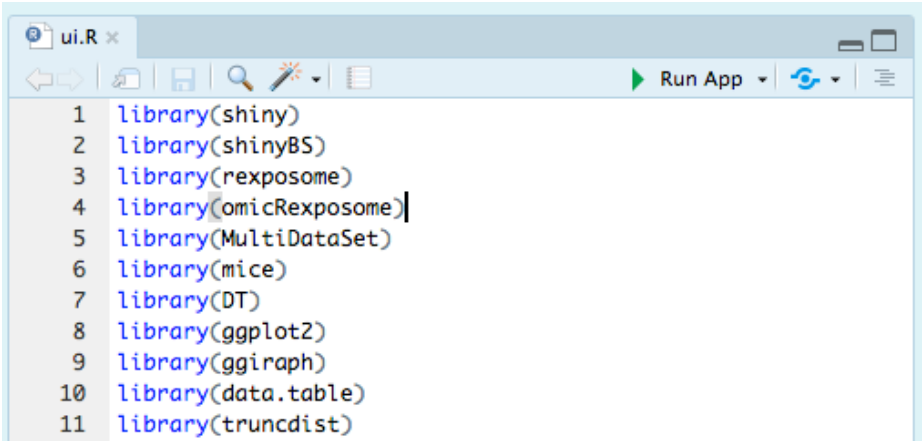


Once the project is loaded, the file found on the source folder called `installer.R` has to be sourced and run. This will install the newest versions of the packages required by Exposome Shiny on this R session. To do so, run the following code on the RStudio console.

```
source("installer.R")
```

This is only needed on the first run, once completed it doesn't need to be done prior to launching the application itself any other time.

Now everything is ready to launch the Shiny application. To do so there a two approaches, one is to open the `ui.R` or the `server.R` files that are inside the `R` folder and press `Run App`.



Or the other option is to input the following command on the console.

```
shiny::runApp('R')
```

2.3 Pulling the official Docker image from DockerHub

If there's any troubble downloading the required R packages to make exposomeShiny work, there's the option of using Docker. It has the disadvantage of being a little bit difficult to install on a Windows machine, however, it's extremely simple on a Mac OS X / Linux environment. For the Windows users refer to the following links for instructions on how to install Docker and setup you machine to run WSL2 and launch bash commands on Windows 1, 2, 3.

To download and launch exposomeShiny, execute the following command on a bash terminal(make sure Docker is running, if not search for the `Docker Desktop` app and launch it).

```
docker run --rm -p 80:80 brgelab/exposome-shiny
```

This command will download the Docker image of exposomeShiny (be aware it weights ~ 3 GB, so if your internet connection is slow it may take a while) and run a container with it. The container will be exposed on the local port 80 and it will render on that port the application itself, so to start using exposomeShiny open your web browser and go to the site

```
localhost:80
```

At the beginning it may take some time for the application to render, this is because all the needed R libraries are being loaded, to be sure the container is actually working, take a look at the terminal where you inputed the Docker command, there you will see all the R verbose stating the libraries are being loaded.

Once the user has finished using exposomeShiny, the container needs to be stopped to avoid wasting CPU resources, to do so, input the following command on a bash terminal (the command needs to be inputed on a new bash window):

```
docker container ls
```

This will prompt all the running containers, find the one with the NAMES `brgelab/exposome-shiny` and copy it's CONTAINER ID, then input the following bash command:

```
docker stop xxxxxxxxxxxx
```

Where xxxxxxxxxxxx is the CONTAINER ID.

To run the application again, just enter the first bash command (`docker run --rm -p 80:80 brgelab/exposome-shiny`), since it has already been downloaded, the application is cached on the computer and it will launch straight

away. If the user wants to remove the Docker image from the computer, input the following bash command:

```
docker image rm brgelab/exposome-shiny
```

Chapter 3

Data sets

3.1 Exposome dataset

The exposome is composed of three different files (in `*.csv`, `*.tsv` or `*.txt` format). Those files are refered inside the Shiny as exposures, description and phenotypes. Their content is the following:

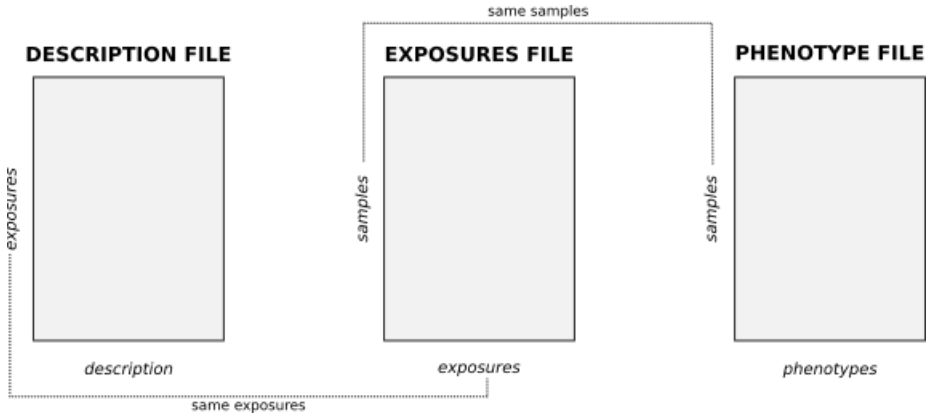
- The `exposures` file contains the measures of each exposure for all the individuals included on the analysis. It is a matrix-like file having a row per individual and a column per exposures. It must includes a column with the subject's identifier.
- The `description` file contains a row for each exposure and, at last, defined the families of exposures. Usually, this file incorporates a description of the exposures, the matrix where it was obtained and the units of measurement among others.
- The `phenotypes` file contains the covariates to be included in the analysis as well as the health outcomes of interest. It contains a row per individual included in the analysis and a column for each covariate and outcome. Moreover, it must include a column with the individual's identifier.

Some remarks regarding this files:

- All three files have to share the same separator element, for `*.csv` files is typical to use a comma (,) but it could also be a semicolon (;).
- The exposure names have to start with a character [a-z/A-Z], leading special characters will cause the data entry to return errors.
- Exactly the same exposures have to be present on the description and exposures files.
- Exactly the same samples have to be present on the exposures and phenotypes files.
- The exposures and phenotypes files have an ID column, the description

file does not have an ID column nor row.

A visual representation of the three matrices and how they correlate is the following.



Exposures data file example:

```
id    bde100  bde138  bde209  PFOA    ...
sub01  2.4665  0.7702  1.6866  2.0075 ...
sub02  0.7799  1.4147  1.2907  1.0153 ...
sub03 -1.6583 -0.9851 -0.8902 -0.0806 ...
sub04 -1.0812 -0.6639 -0.2988 -0.4268 ...
sub05 -0.2842 -0.1518 -1.5291 -0.7365 ...
...   ...     ...     ...     ...
```

Description data file example:

```
exposure  family  matrix         description
bde100    PBDEs   colostrum       BDE 100 - log10
bde138    PBDEs   colostrum       BDE 138 - log10
bde209    PBDEs   colostrum       BDE 209 - log10
PFOA      PFAS    cord blood      PFOA - log10
PFNA      PFAS    cord blood      PFNA - log10
PFOA      PFAS    maternal serum  PFOA - log10
PFNA      PFAS    maternal serum  PFNA - log10
hg        Metals  cord blood      hg - log 10
Co        Metals  urine           Co (creatinine) - log10
Zn        Metals  urine           Zn (creatinine) - log10
Pb        Metals  urine           Pb (creatinine) - log10
THM       Water   ---             Average total THM uptake - log10
CHCL3     Water   ---             Average Chloroform uptake - log10
BROM      Water   ---             Average Brominated THM uptake - log10
NO2       Air     ---             NO2 levels whole pregnancy- log10
Ben       Air     ---             Benzene levels whole pregnancy- log10
```

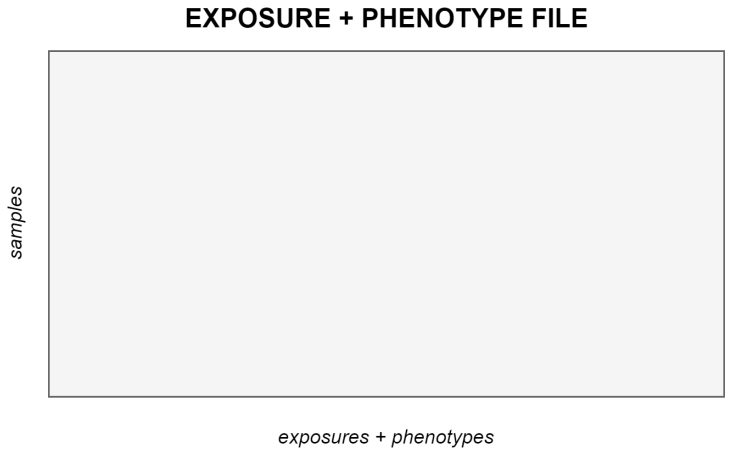
Phenotypes data file example:

```
id    asthma   BMI      sex  age  ...
sub01 control  23.2539  boy  4    ...
sub02 asthma   24.4498  girl 5    ...
sub03 asthma   15.2356  boy  4    ...
sub04 control  25.1387  girl 4    ...
sub05 control  22.0477  boy  5    ...
...   ...      ...      ...  ...
```

3.2 Plain datasets

If the researcher has gathered all the data on a single file which contains both phenotype and exposure data, this file can be used too. The user interface has a selector for it, more information on the correspondent section.

A visual representation of a plain dataset is the following.



Plain dataset example (3 exposures + 2 phenotypes):

```
id    bde100  bde138  bde209    asthma   BMI      ...
sub01  2.4665  0.7702  1.6866   control  23.2539  ...
sub02  0.7799  1.4147  1.2907   asthma   24.4498  ...
sub03 -1.6583 -0.9851 -0.8902   asthma   15.2356  ...
sub04 -1.0812 -0.6639 -0.2988   control  25.1387  ...
sub05 -0.2842 -0.1518 -1.5291   control  22.0477  ...
...   ...     ...      ...      ...      ...
```

3.3 Omics dataset

The omics data inputed to the Shiny must be provided as an `*.RData`. This file has to contain an ExpressionSet, which is an S4 object. This object is a data container of the Bioconductor toolset.

For further information on ExpressionSet and how to create and manipulate them, please visit the official documentation and this selected vignette.

Chapter 4

Bioconductor packages

This Shiny application is a front end support for other Bioconductor packages in order to provide a comfortable environment on to conduct different analysis with those packages. In concrete the packages are rexposome, omicRexposome and CTDquerier.

4.1 rexposome

Rexposome is a package that allows to explore the exposome and to perform association analyses between exposures and health outcomes.

4.2 omicRexposome

OmicRexposome is a package that systematizes the association evaluation between exposures and omic data, taking advantage of MultiDataSet for coordinated data management, rexposome for exposome data definition and limma for association testing. Also to perform data integration mixing exposome and omic data using multi co-inherent analysis (omicade4) and multi-canonical correlation analysis (PMA).

4.3 CTDquerier

CTDquerier is a package to retrieve and visualize data from the Comparative Toxicogenomics Database. The downloaded data is formated as DataFrames for further downstream analyses.

Chapter 5

Analysis flowcharts

On this section, a detailed guide on how to perform different analyses using exposomeShiny will be provided. The guide contains screenshots of the analysis steps as well as some flowcharts.

5.1 Exposome health analysis

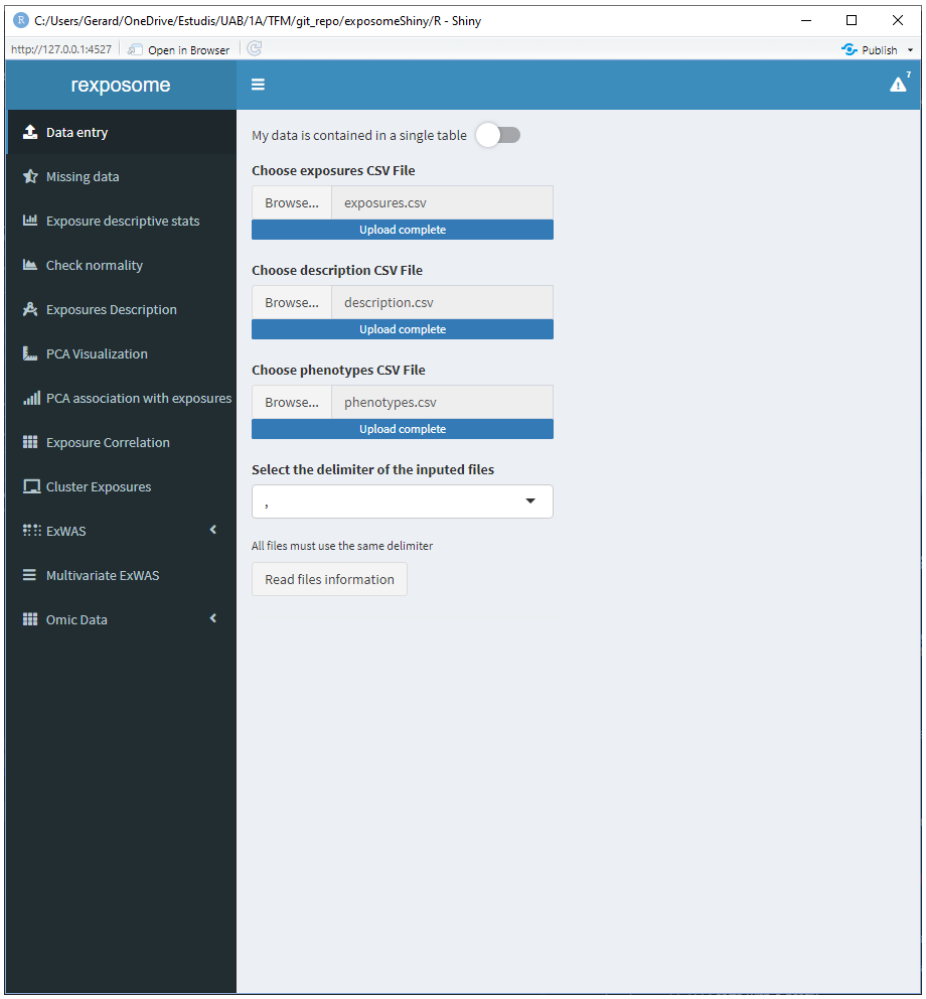
The exposome health analysis corresponds to the study of the relation between exposures (exposome) and health outcomes (phenotypes). Information about exposome data exploration and pre-processing can also be found on this subsection. Along this section, the data files used as examples are: `exposures.csv`, `exposures_lod.csv`, `description.csv`, `phenotypes.csv` and `exposome_plain.csv` which are available here.

5.1.1 Data entry

There are two different data entry methods. Exposome data which uses the three tables and plain data which only uses one.

5.1.1.1 Exposome data

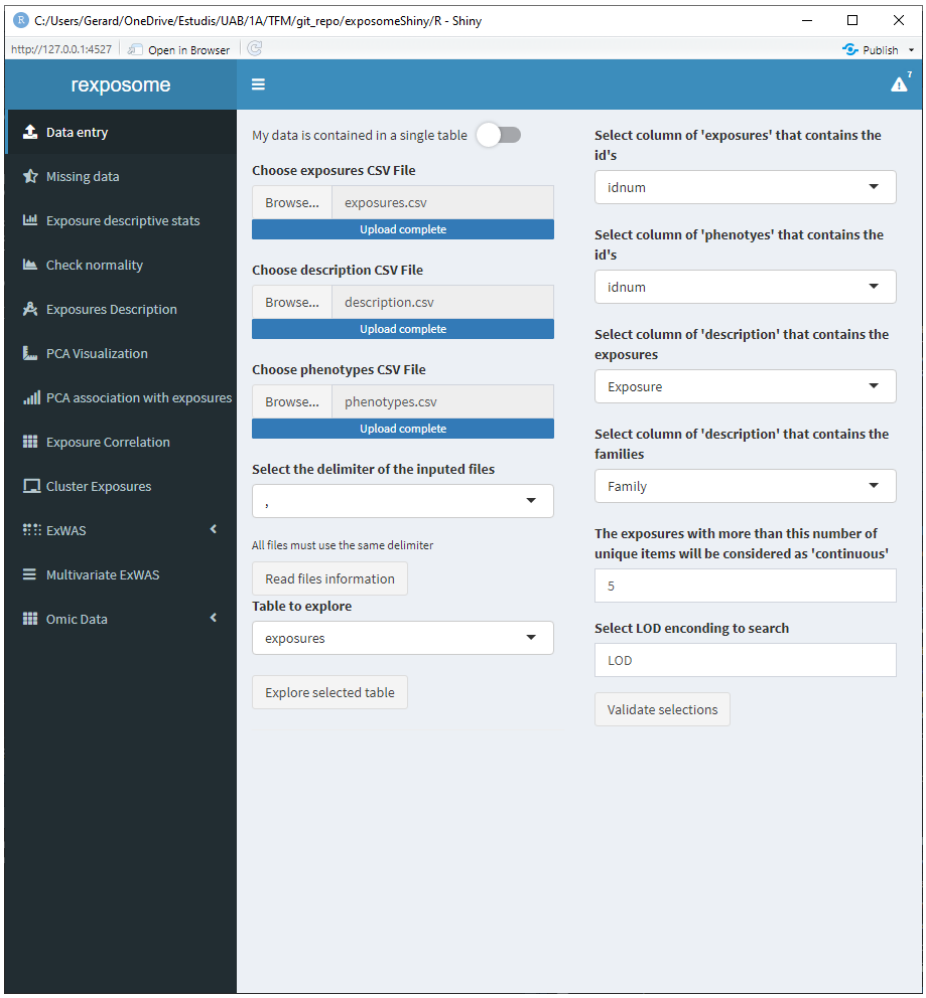
Three tables are needed to input an exposome dataset: the exposures, description and phenotypes. This files have to be provided as `csv` or `txt`. Different separator formats are supported (',', ';', tabs and spaces). Excel files or R objects are not supported as inputs. Be sure all three files share the same separator.



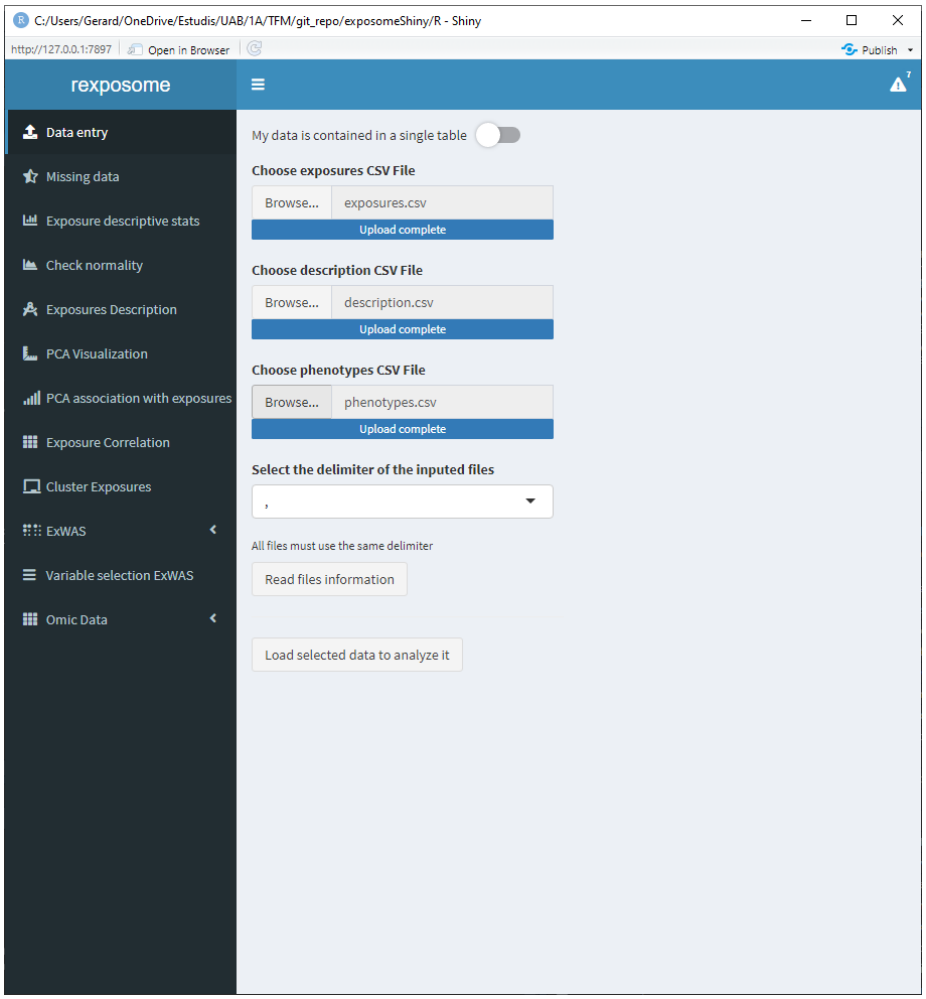
Once the tables are read two new main elements will appear, 1) the option to explore the inputted tables; and 2) six input fields which control the following file parameters:

- Column name in the *description* file that contains the exposures
- Column name in the *description* file that contains the families
- Column name in the *exposures* file that contains the identifier
- Column name in the *phenotypes* file that contains the identifier
- The threshold to select between continuous or factor exposures. More than this number of unique items on an exposure will be considered as 'continuous'
- The encoding to search for limit of detection (LOD) missings. It can be either a number (example: `-1`) or a string (example: `LOD`). All the cells that contain this encoder will be considered LOD missings.

This is illustrated on the following figure.

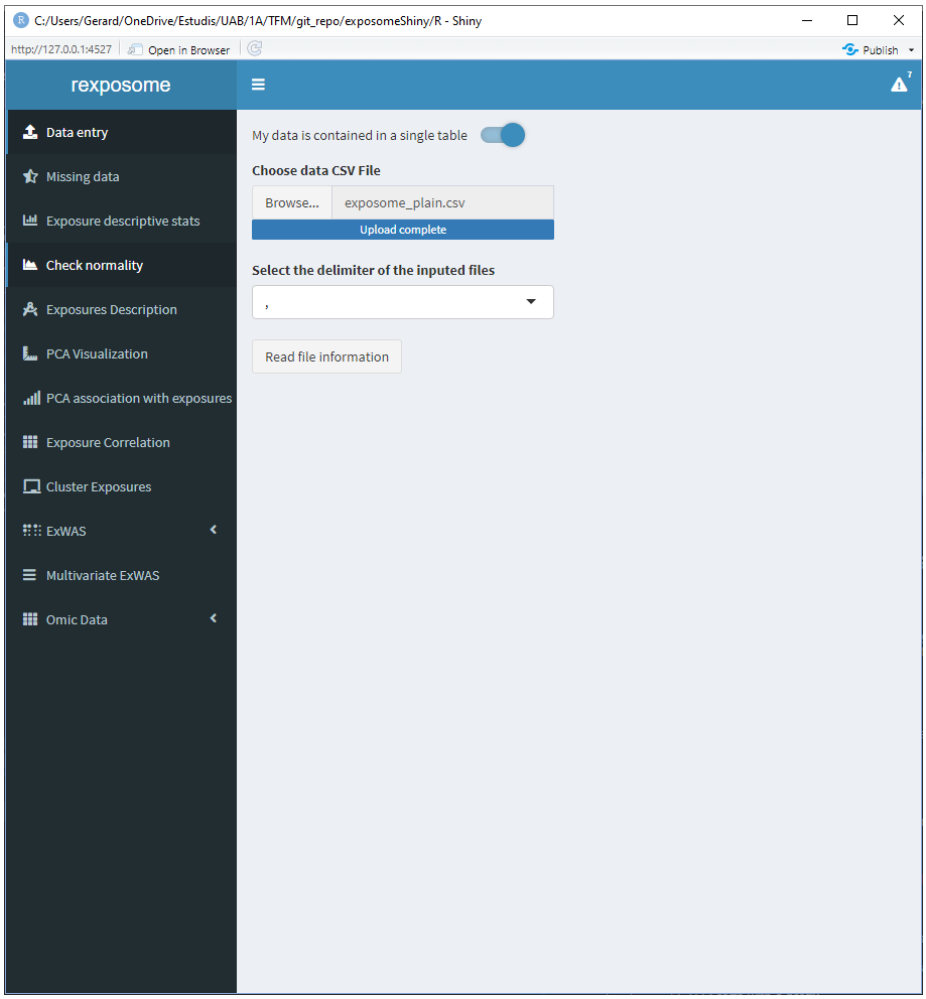


Once the fields are completed, press the *Validate selections* button in order to check if all the parameters allow for a successful load of the exposome dataset, if not, a pop-up will be prompted to the user with the R error message that stopped the execution. In the case that everything is correct, the interface will be updated and a button that reads *Load selected data to analyze it* will appear, by clicking this button the dataset will be loaded.

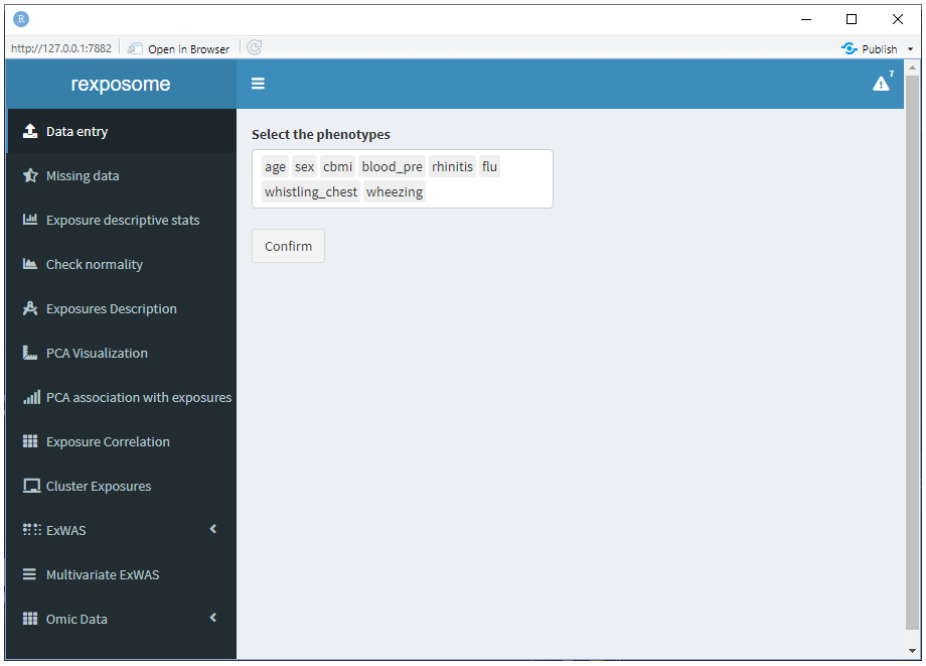


5.1.1.2 Plain data

When dealing with a single file that contains the exposures and phenotype data, press the "My data is contained in a single table" toggle. This will change the interface to only show one file selector.



Select a file and press “Read file information”. This will change the interface to show a multiple selector input, all the available columns will be listed, select the ones which correspond to phenotypes.

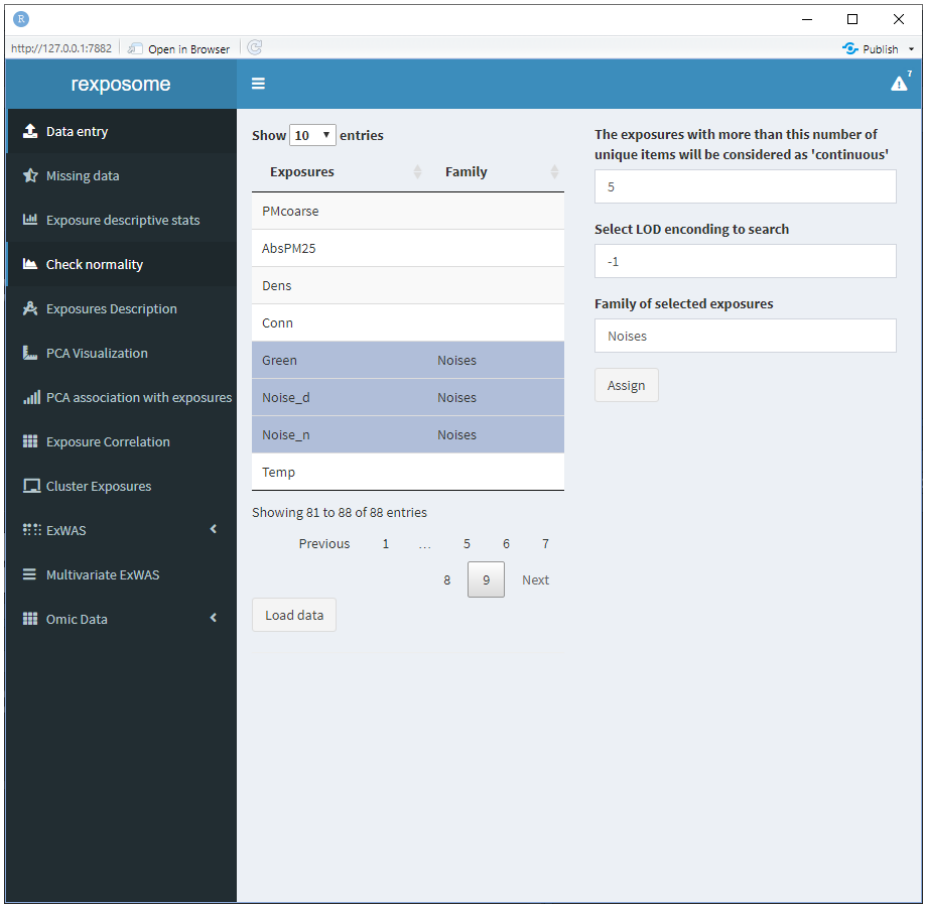


When all the phenotypes are selected, press “Confirm”. Now the exposures can be grouped into families, to do so:

- Select all the exposures from the same family
- Write the family name on the box “Family of selected exposures”
- Press the “Assign” button

The table will be updated to visualize the action performed.

Exposures left with an empty “Family” field will be treated as they are their own family.

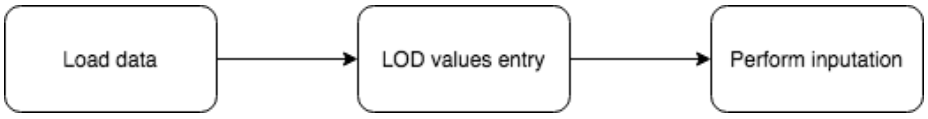


Finally, there are two configuration fields:

- The threshold to select between continuous or factor exposures. More than this number of unique items on an exposure will be considered as 'continuous'
- The encoding to search for limit of detection (LOD) missings (example: `LOD`). All the cells that contain this encoder will be considered LOD missings.

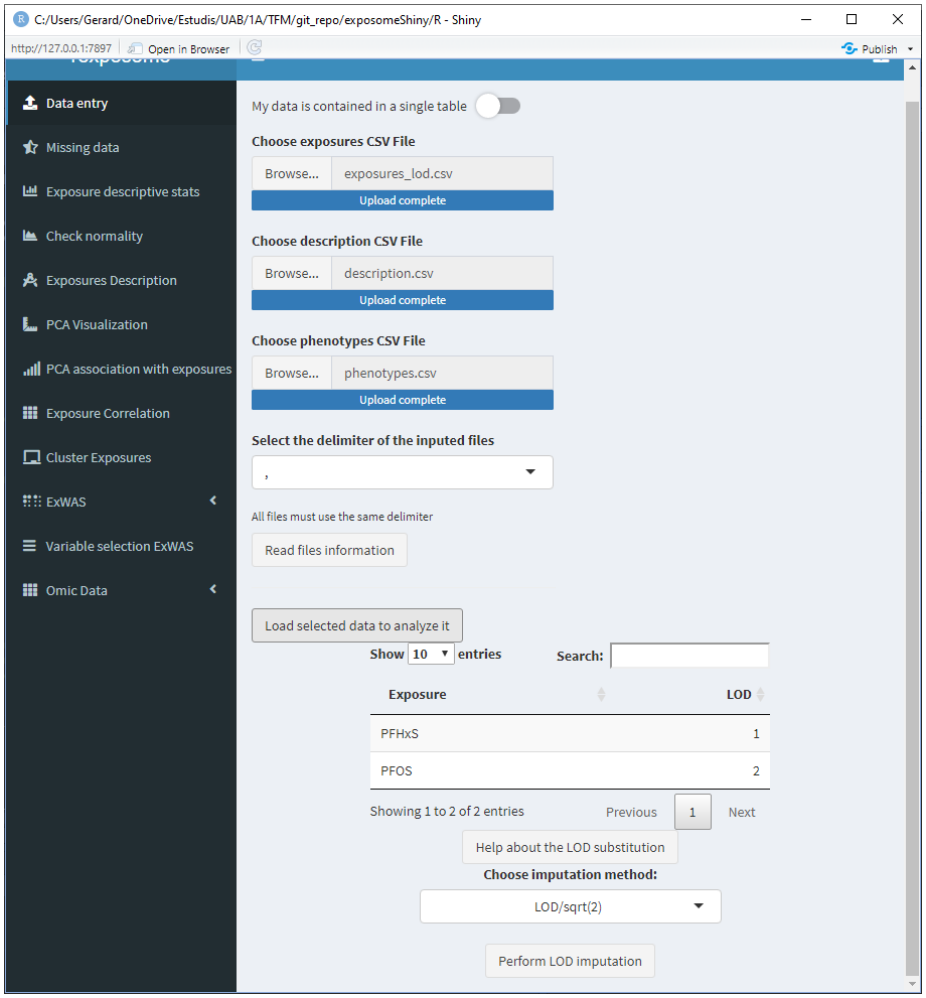
Be sure to revise them before pressing "Load data".

5.1.2 LOD imputation



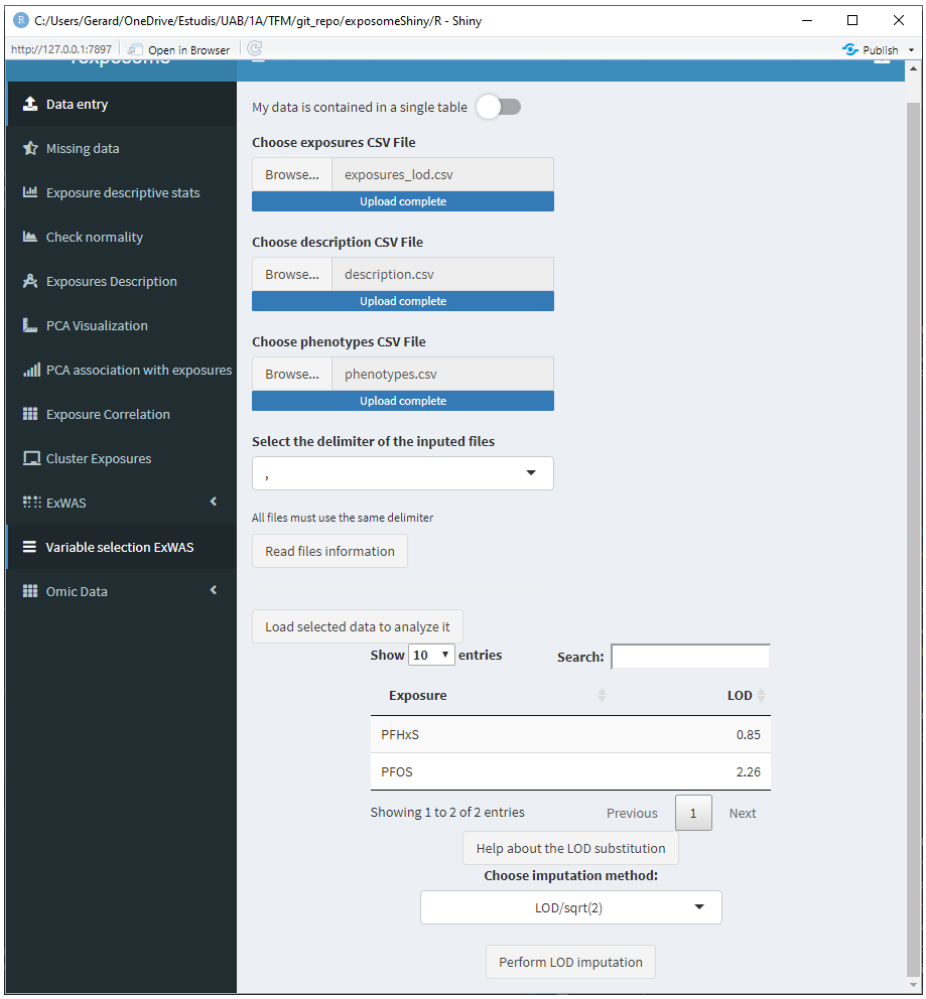
When loading the data, if LOD missings are detected a small table with the

exposures that have LOD missings will appear at the bottom of the "Data entry" page.

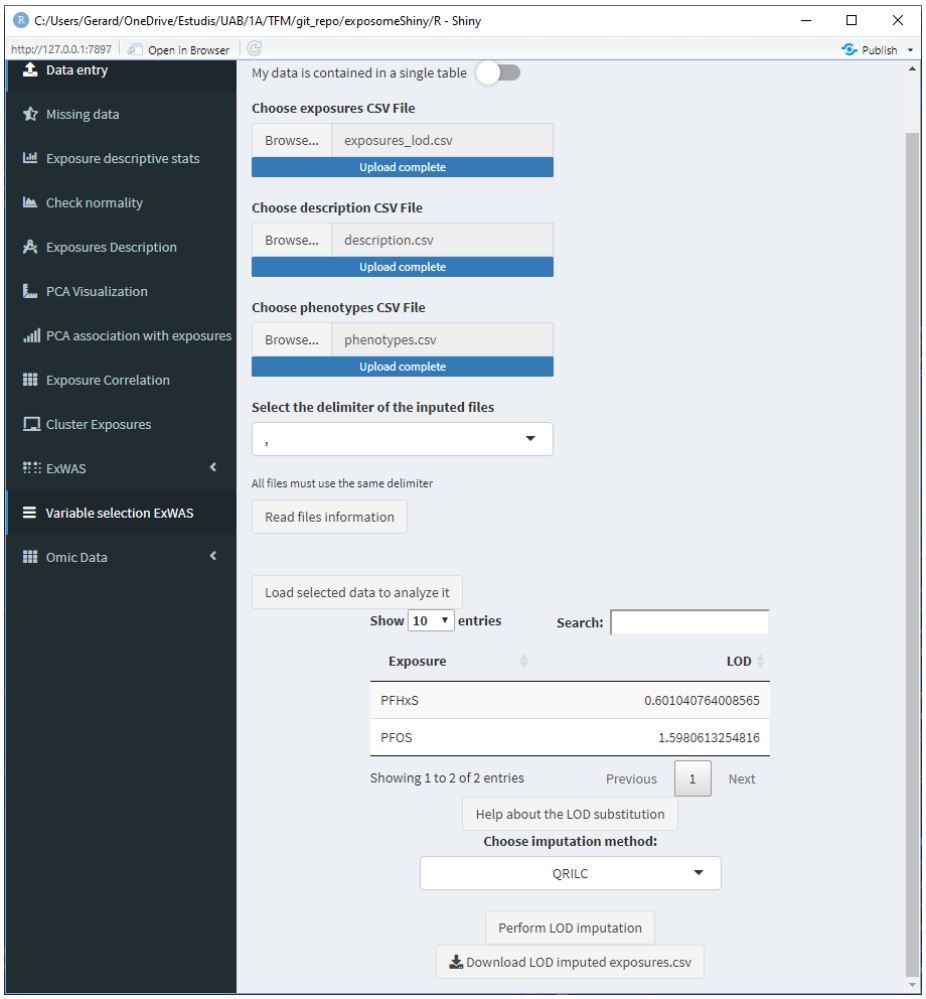


This table has a column named "LOD", which by default reads as 1, 2, 3, ... Those values are meant to be modified by the user and input the limit of detection values for the different exposures. Those values will be used if the LOD imputation method is LOD/sqrt(2). If the imputation method used is QRILC there is no need to modify those values, as they won't be used.

To modify the LOD values of the table, double click on the cell of interest and type the value (use points for decimal separation e.j. 1.156).

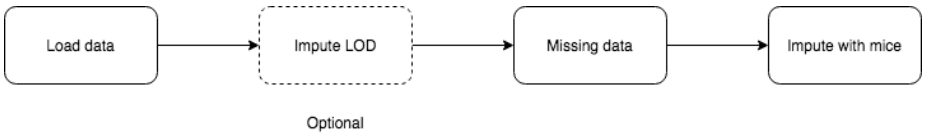


After clicking "Perform LOD imputation", a new button to download the imputed dataset will appear, this will download the exposures data with the LOD missings imputed.

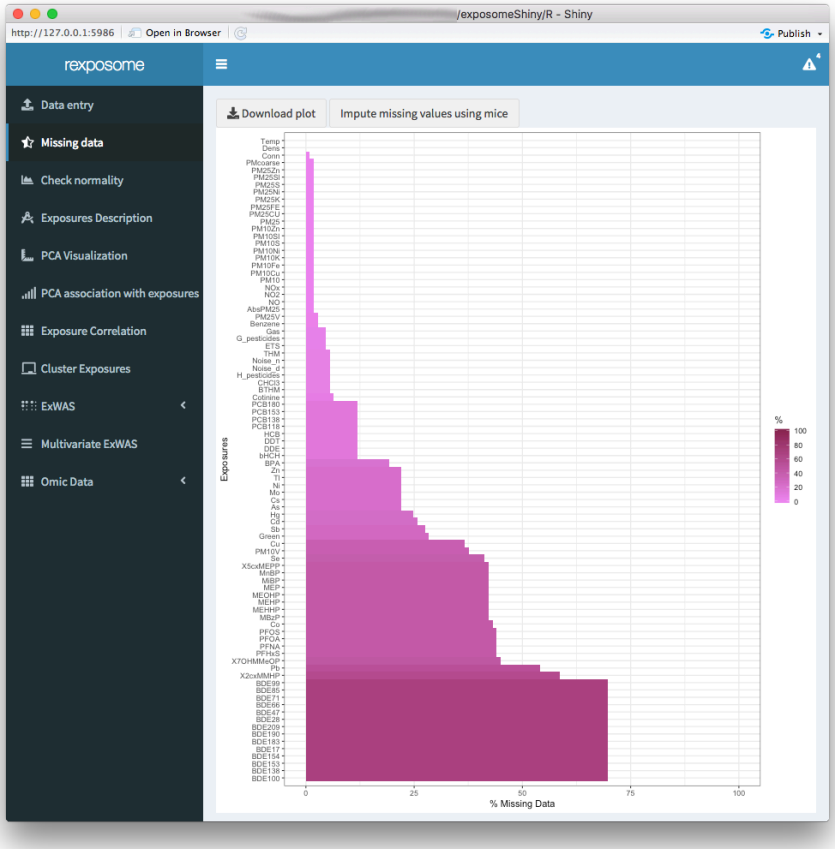


Once the LOD imputation is completed, the exposome dataset that will be used on the following steps is imputed.

5.1.3 Missing imputation



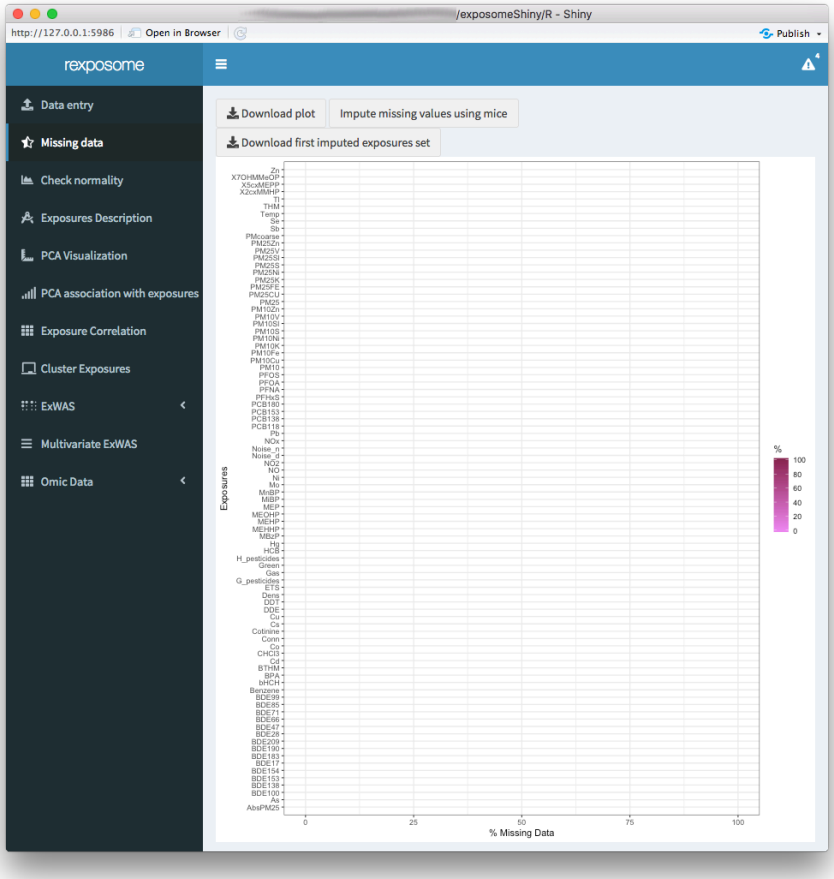
The "Missing data" tab displays a plot with the percentages of missing data for each exposure.



The missings can be imputed using Multiple Imputation by Chained Equations (MICE) by clicking the "Impute missing values using mice" button.

Please note that missings imputation might require other methodologies or more fine control. If that is the case, we encourage users to impute the data beforehand and input to our application data that is already imputed.

Once the missings imputation is completed, the exposome dataset that will be used on the following steps is imputed. This is reflected by refreshing the plot, which should read 0% missings for all exposures.



The imputed exposures set can be downloaded as a `*.csv` file.

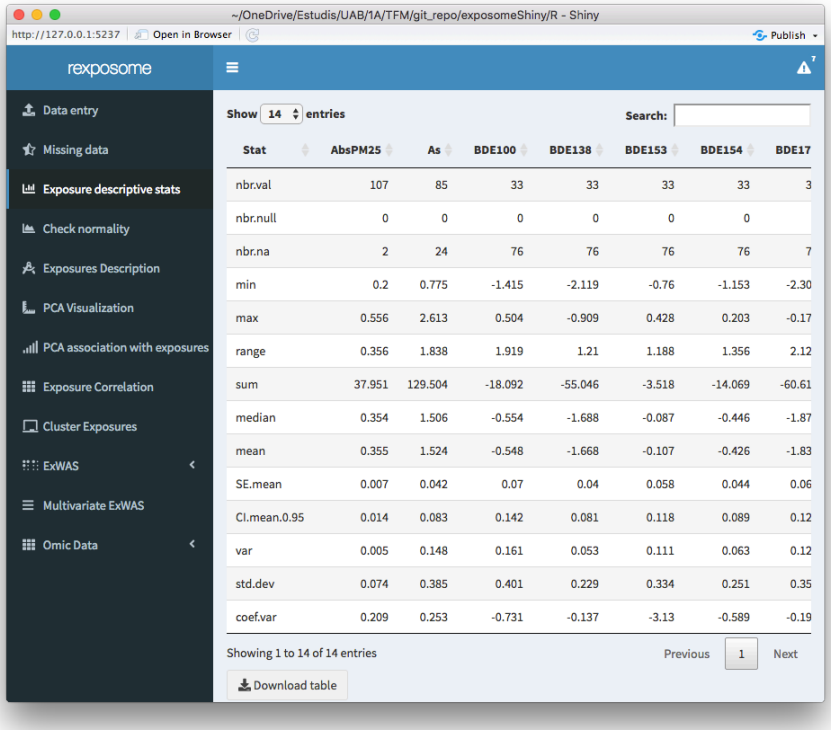
5.1.4 Exposures description

The exposure descriptive stats tab, provides a table with the main descriptive stats of the quantitative exposures. The descriptive stats (per exposure) included on the table are:

- Number of values
- Number of NULLs
- Number of NAs
- Minimum
- Maximum
- Range of values
- Sum of values
- Median

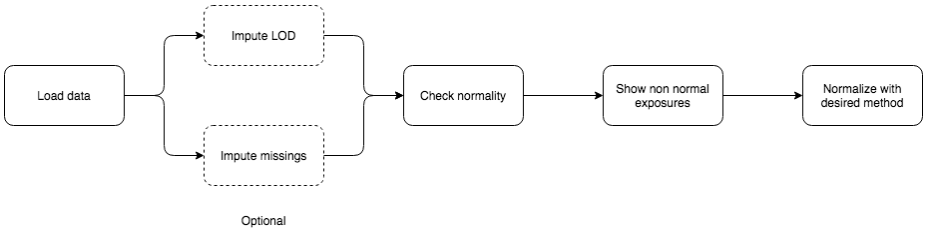
- Mean
- Standard Error of mean
- 0.95 confidence interval of the mean
- Variance
- Standard deviation
- Variance coefficient

Remember that after imputing the missings, the imputed dataset becomes active, this will be reflected showing 0 NAs for example.

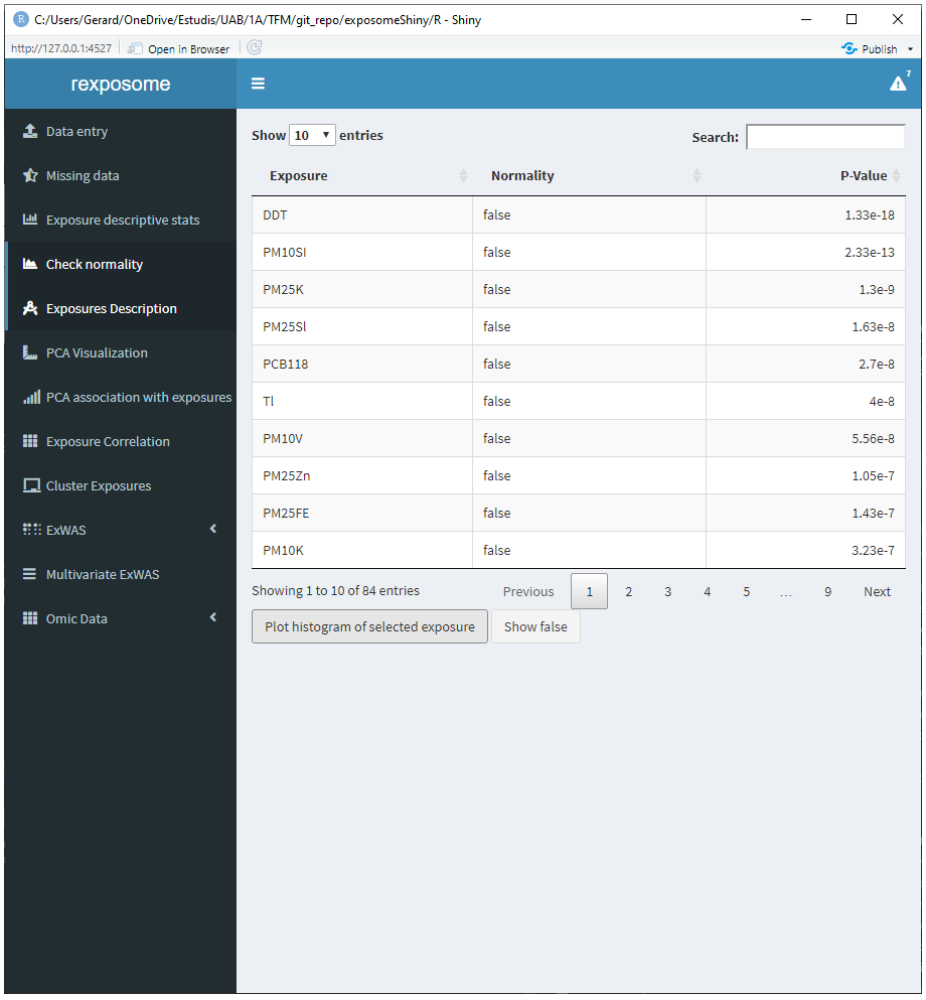


Stat	AbsPM25	As	BDE100	BDE138	BDE153	BDE154	BDE17
nbr.val	107	85	33	33	33	33	3
nbr.null	0	0	0	0	0	0	
nbr.na	2	24	76	76	76	76	7
min	0.2	0.775	-1.415	-2.119	-0.76	-1.153	-2.30
max	0.556	2.613	0.504	-0.909	0.428	0.203	-0.17
range	0.356	1.838	1.919	1.21	1.188	1.356	2.12
sum	37.951	129.504	-18.092	-55.046	-3.518	-14.069	-60.61
median	0.354	1.506	-0.554	-1.688	-0.087	-0.446	-1.87
mean	0.355	1.524	-0.548	-1.668	-0.107	-0.426	-1.83
SE.mean	0.007	0.042	0.07	0.04	0.058	0.044	0.06
CI.mean.0.95	0.014	0.083	0.142	0.081	0.118	0.089	0.12
var	0.005	0.148	0.161	0.053	0.111	0.063	0.12
std.dev	0.074	0.385	0.401	0.229	0.334	0.251	0.35
coef.var	0.209	0.253	-0.731	-0.137	-3.13	-0.589	-0.19

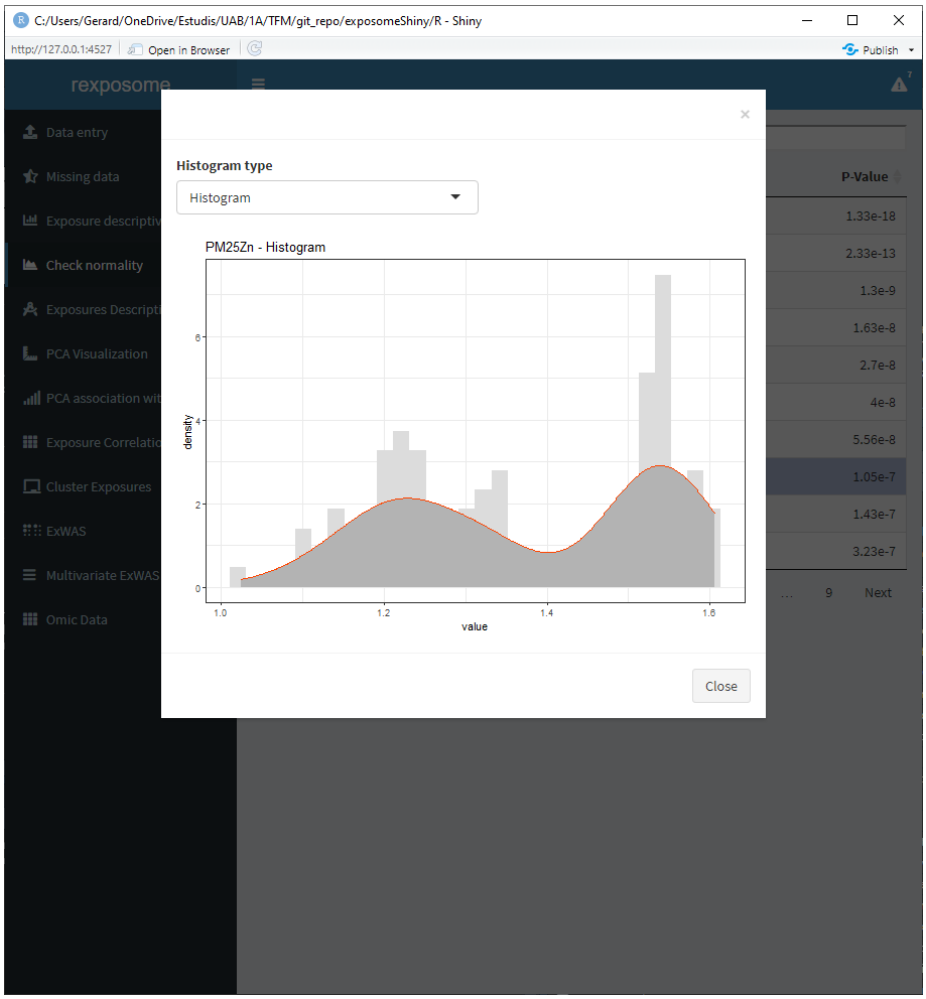
5.1.5 Normality correction



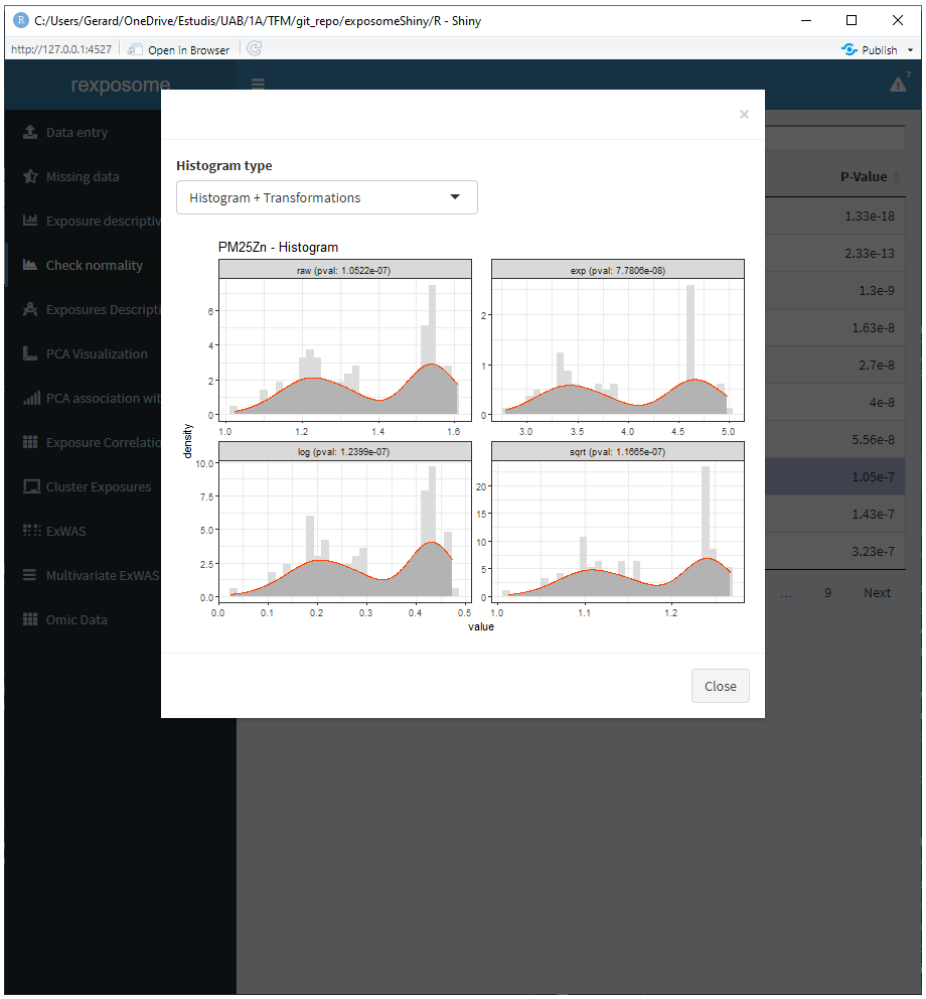
The table shown on the 'Check normality' tab contains all the exposures, if they can be considered normal distributed and the p-value of the normality test (Shapiro-Wilk).



Select an exposure from the table and click on 'Plot histogram of selected exposure' to see the histogram of that exposure.

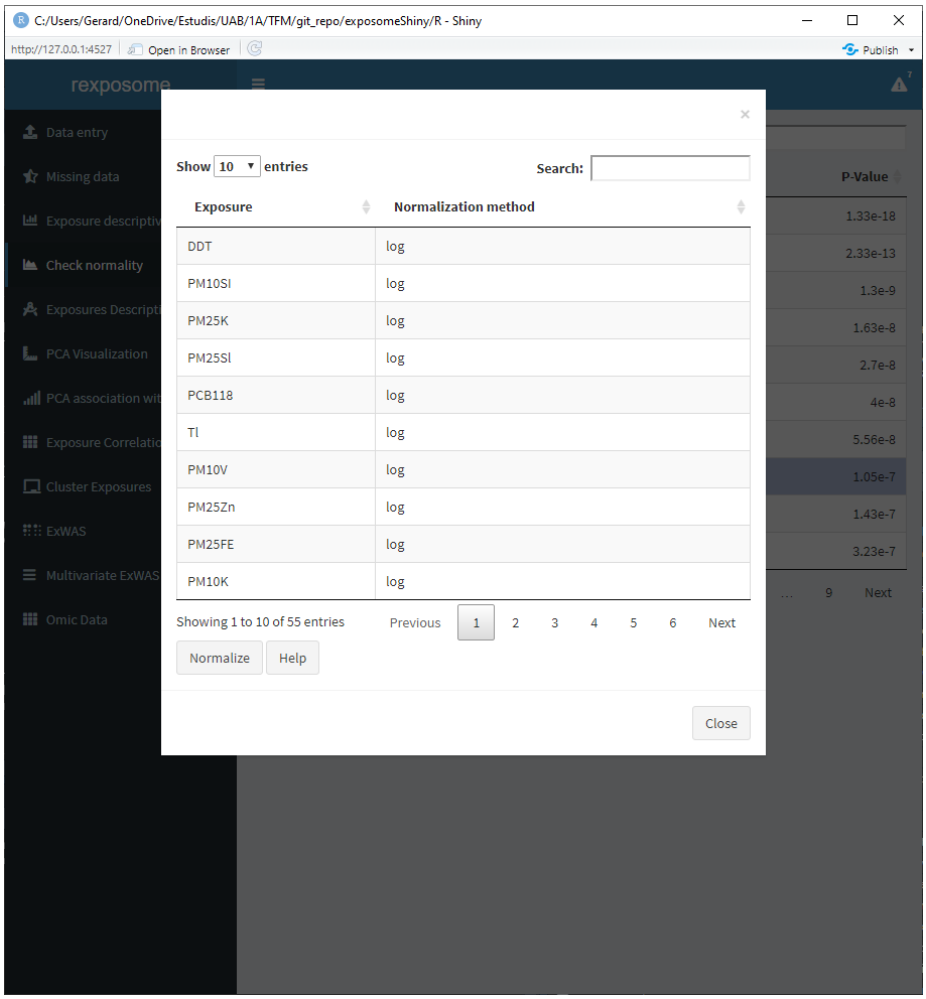


There is the option to visualize the histograms with the available transformations applied, in order to see how an exposure will be affected by them. To see it select 'Histogram + transformations' on the Histogram type.

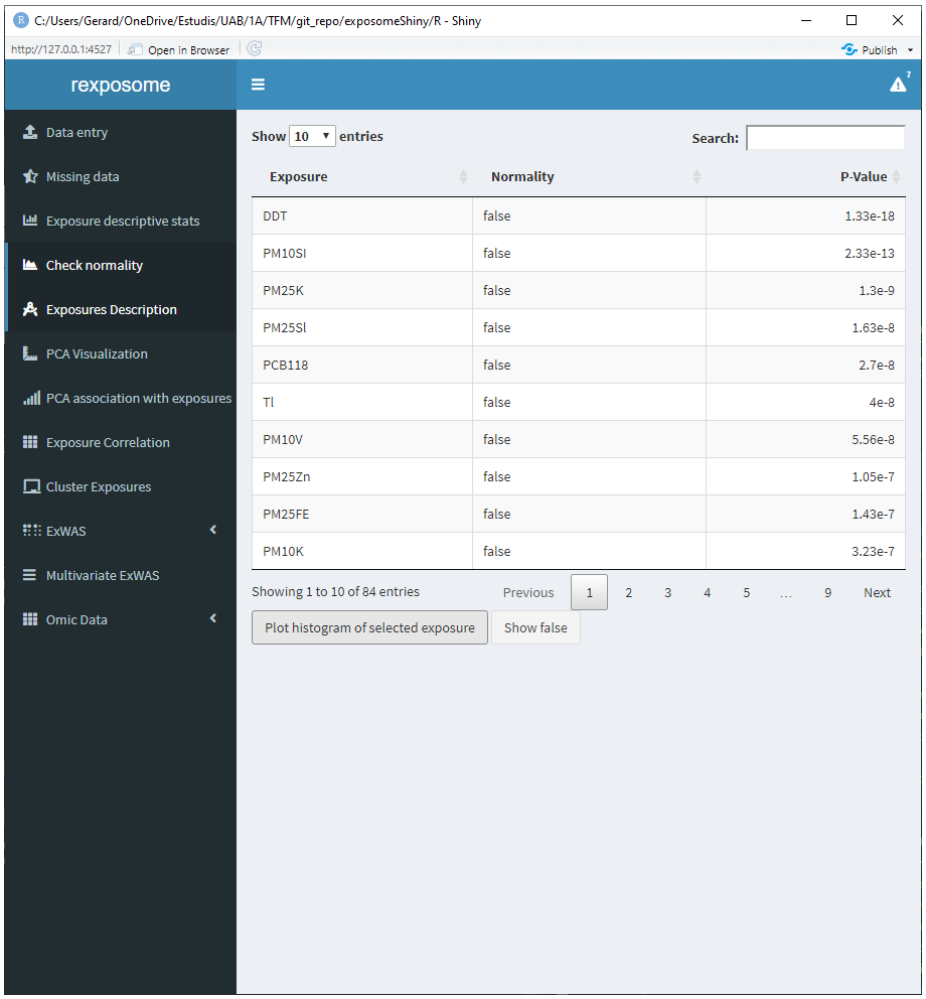


Aside from visualizing if exposures are normal, this tab can be used to apply transformations to the exposures. The available transformations are "log" (default, natural logarithm), "^1/3" and "sqrt". To do so, the procedure is the following:

- Press the 'Show false' button. A pop-up will appear with a table. This table contains all the exposures that did not pass the normality test, next to the exposures the transformation to be applied is shown.



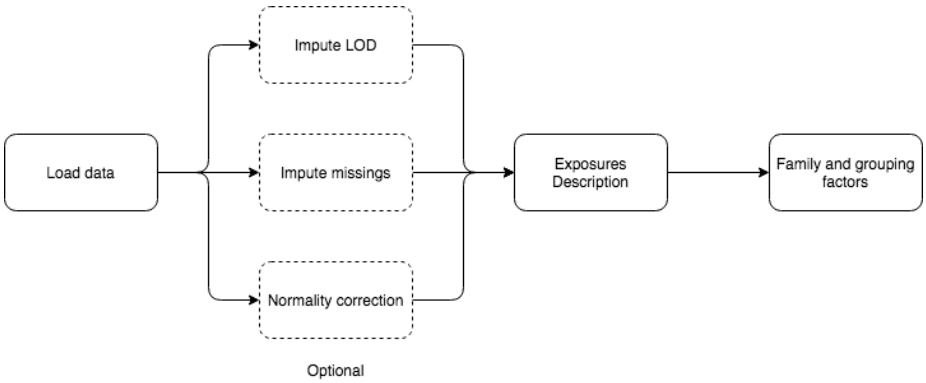
- Modify the transformation method by double clicking on the cell to be changed and typing the required method. Input 'none' if no transformation is desired for a certain exposure.



- Press ‘Normalize’ to apply the transformations.

After the transformations are applied, the table of normality tests will be updated. The dataset used after transforming is updated with the new values.

5.1.6 Exposures description



On the 'Exposures description' tab, exploratory analysis of the exposures can be performed. Again, the data analyzed by this tab will be imputed or normalized if the user has performed it before.

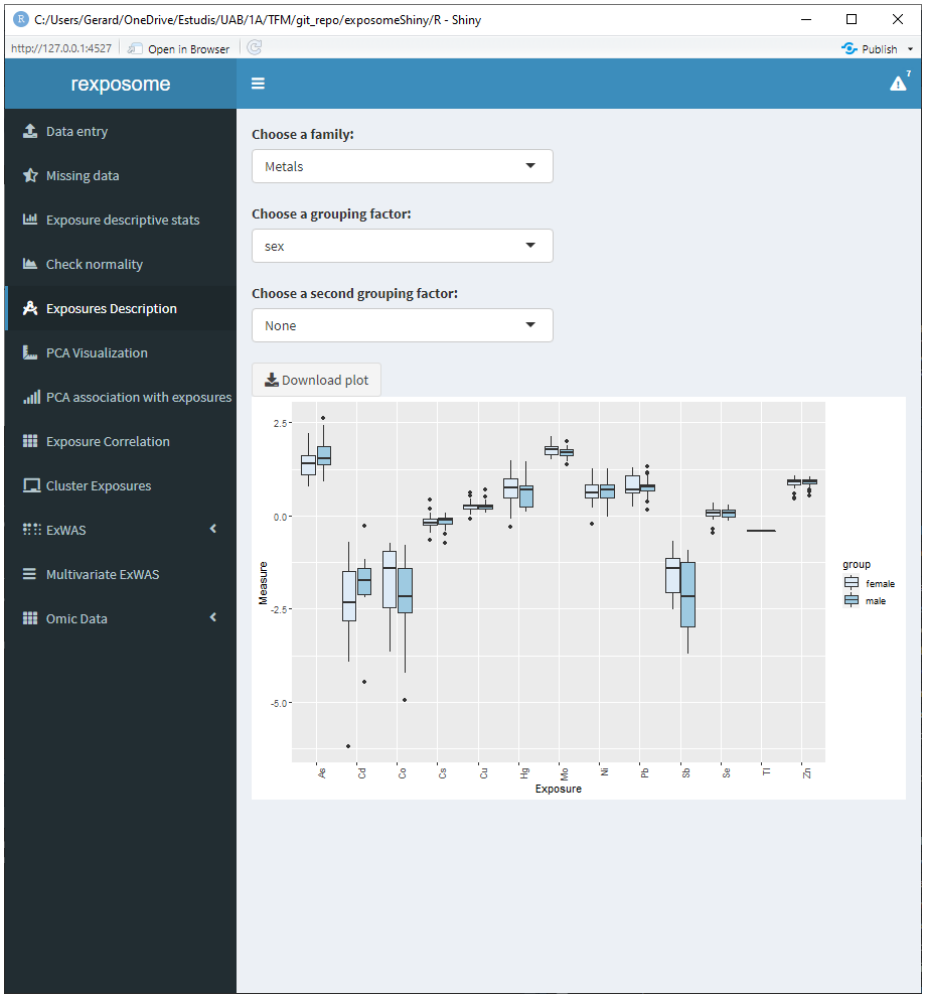
There are three input fields to obtain different plots:

- Family: To define which family of exposures wil be plotted
- Grouping factor: Qualitative phenotype to group the exposures
- Second grouping factor: Second qualitative phenotype to group the exposures

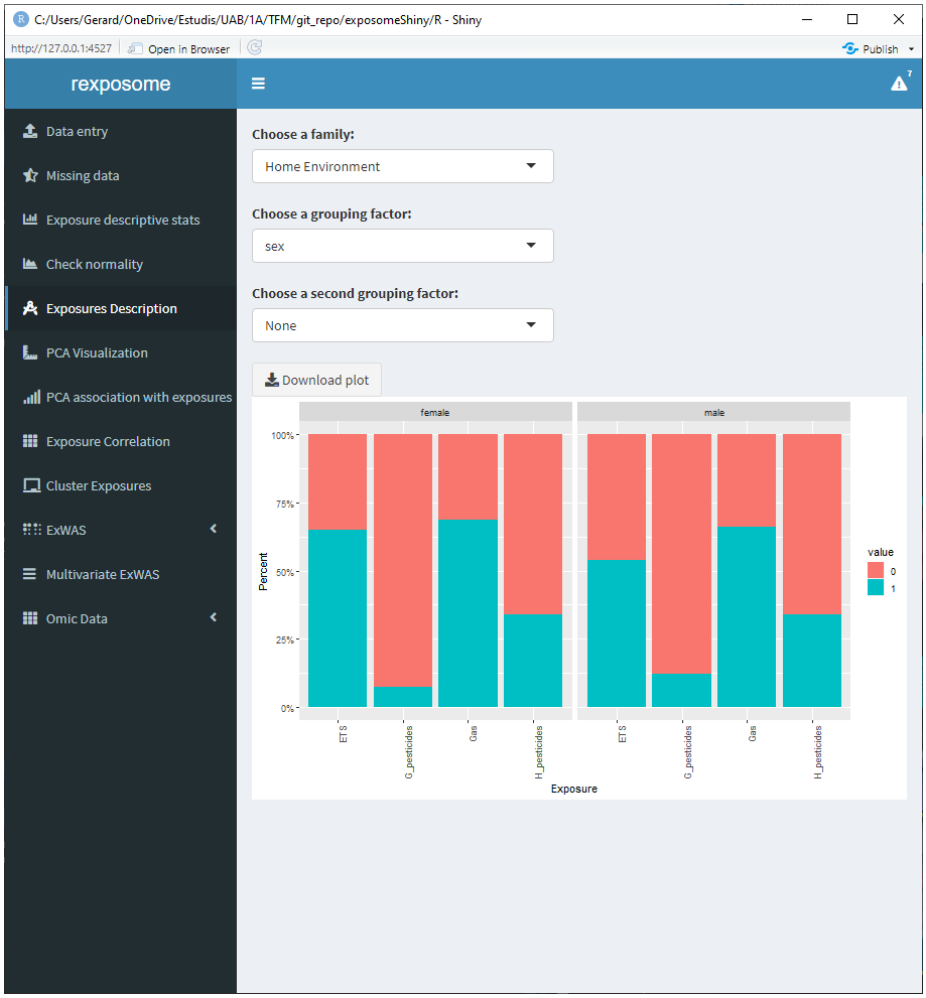
The visualization is different for categorical and quantitative exposures.

Some examples:

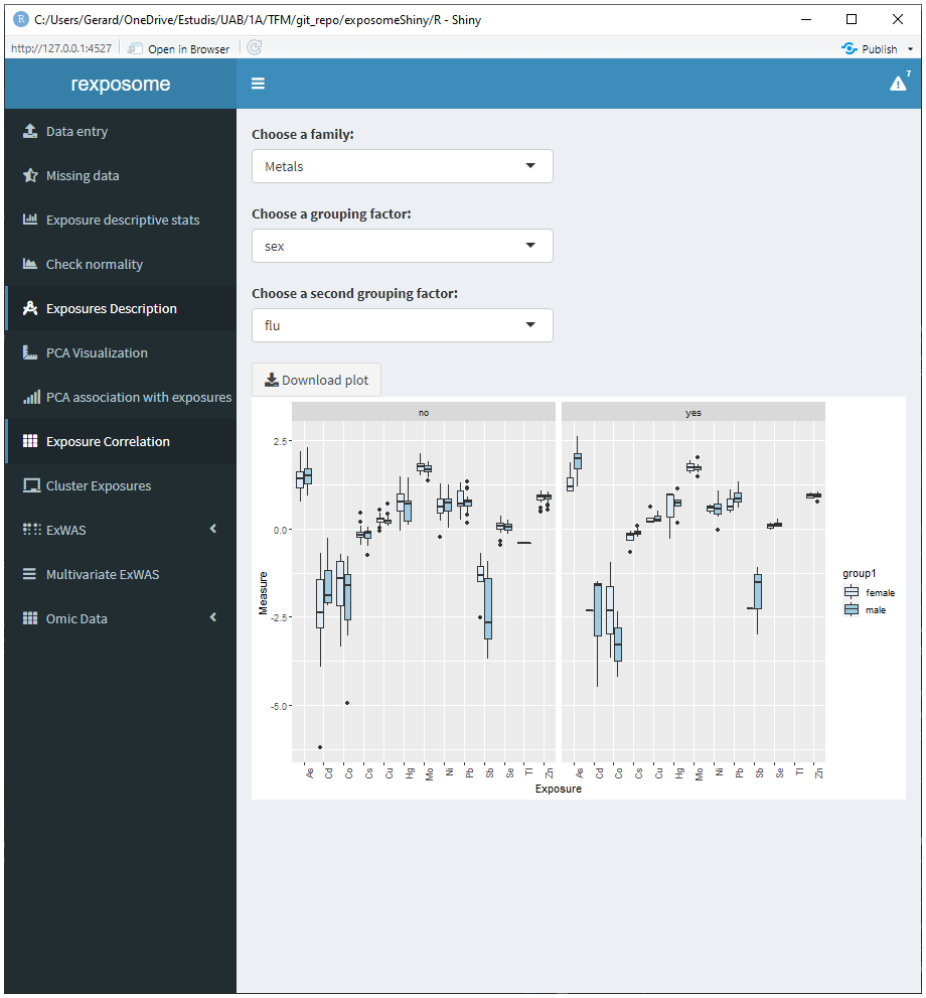
5.1.6.1 Quantitative family, grouped by sex



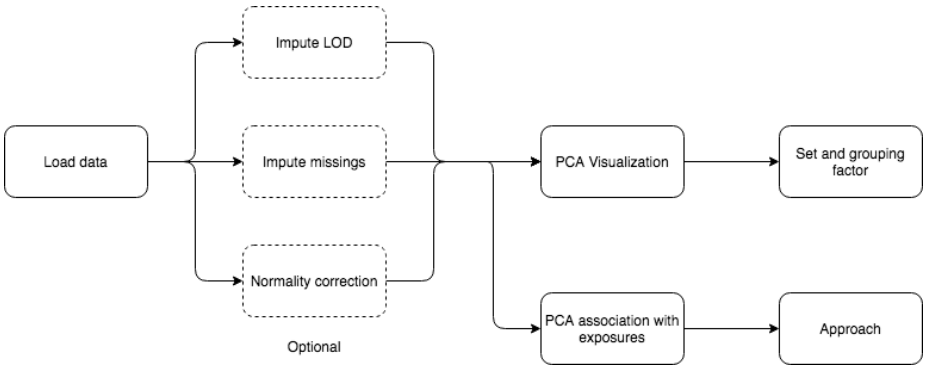
5.1.6.2 Qualitative family, grouped by sex



5.1.6.3 Quantitative family grouped by sex and flu diagnosis (binomial factor)



5.1.7 PCA Analysis

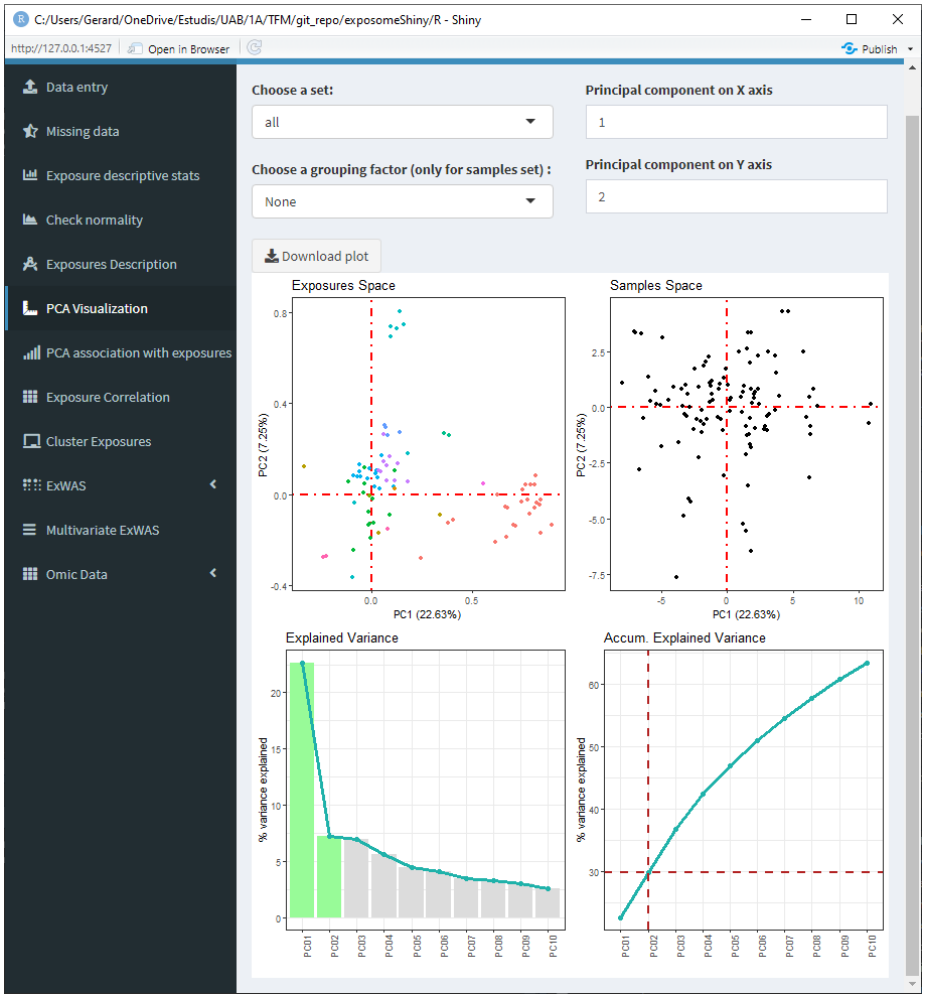


On the 'PCA visualization' tab, the results of a PCA analysis on the exposome data is displayed.

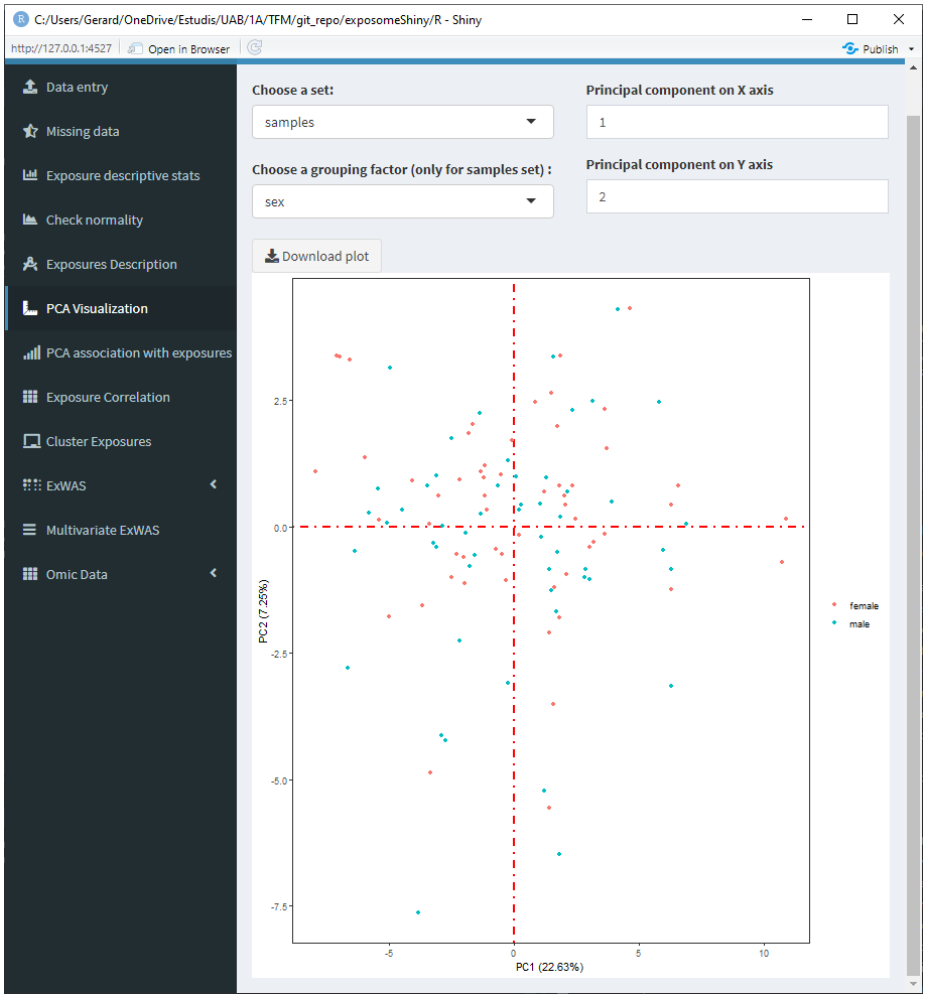
There are some input fields to modify the plot:

- Set: Select 'all' to see an array of four plots with different PCA related information. There is also the option of selecting 'samples' and 'exposures' to only display those two plots on a bigger scale.
- Grouping factor: To group the 'samples' set using a qualitative phenotype.
- Principal components: By default the first and second principal components are used for the plots. They can be changed using those inputs, which affect all the plots.

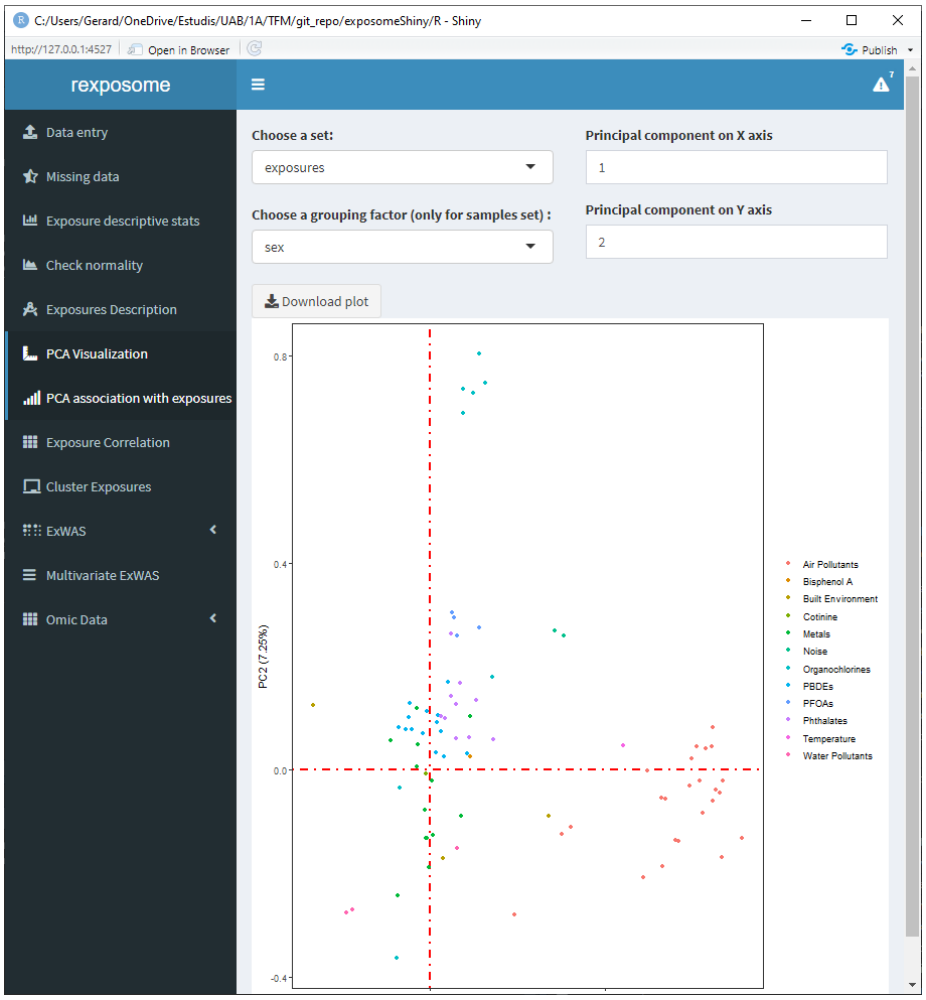
5.1.7.1 Set: 'all'


C:/Users/Gerard/OneDrive/Estudis/UAB/1A/TFM/git_repo/exposomeShiny/R - Shiny
http://127.0.0.1:4527
Open in Browser
Publish
Data entry
Missing data
Exposure descriptive stats
Check normality
Exposures Description
PCA Visualization
PCA association with exposures
Exposure Correlation
Cluster Exposures
ExWAS
Multivariate ExWAS
Omic Data
Choose a set:
all
Choose a grouping factor (only for samples set) :
None
Principal component on X axis
1
Principal component on Y axis
2
Download plot
Exposures Space
Samples Space
PC2 (7.25%)
PC1 (22.63%)
Explained Variance
Accum. Explained Variance
% variance explained
PC01
PC02
PC03
PC04
PC05
PC06
PC07
PC08
PC09
PC10

5.1.7.2 Set: 'samples' grouped by sex

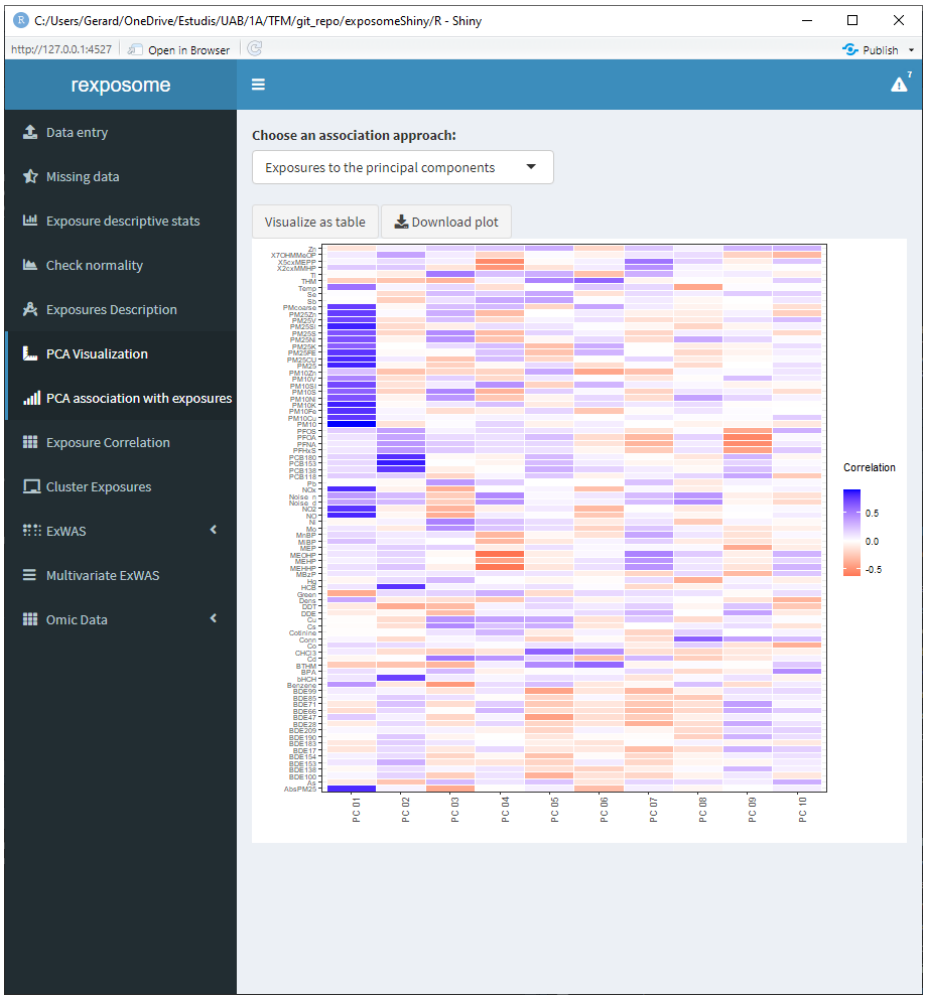


5.1.7.3 Set: 'exposures'

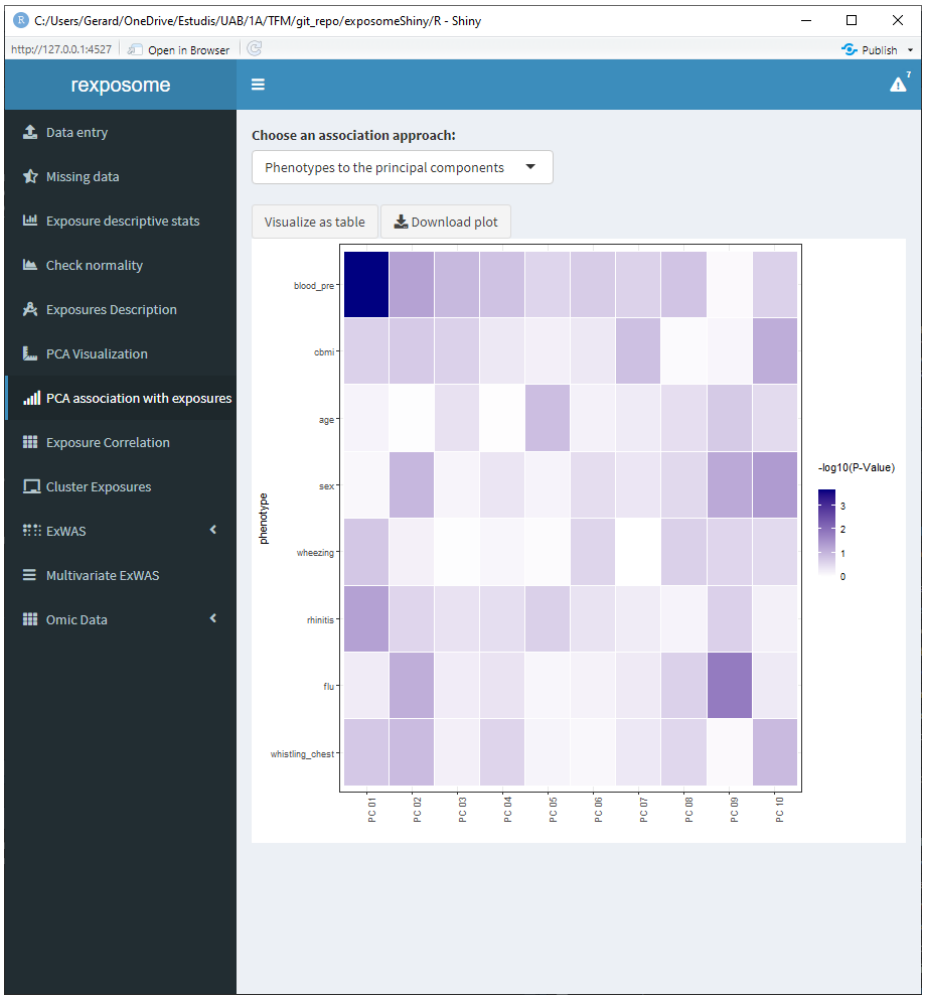


Moreover, there is an additional PCA visualization tab named 'PCA association with exposures'. On this tab there are two heatmap plots available.

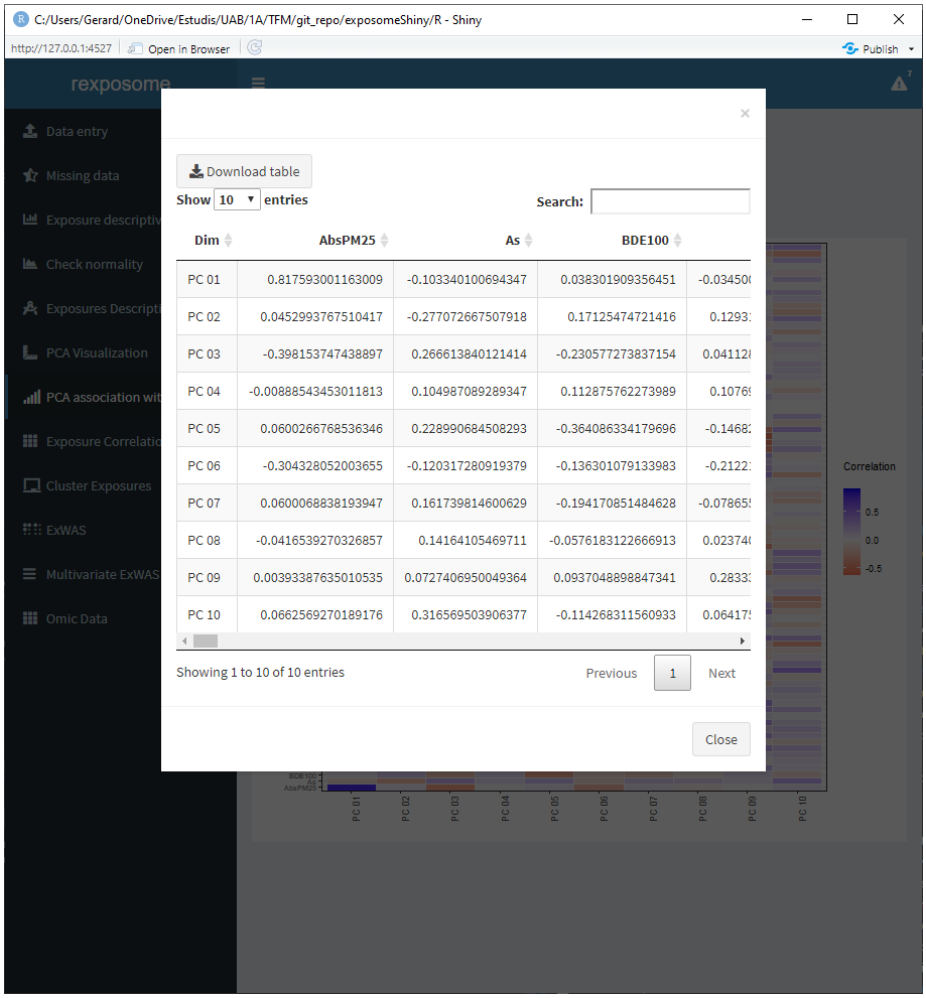
- Association of the exposures to the principal components



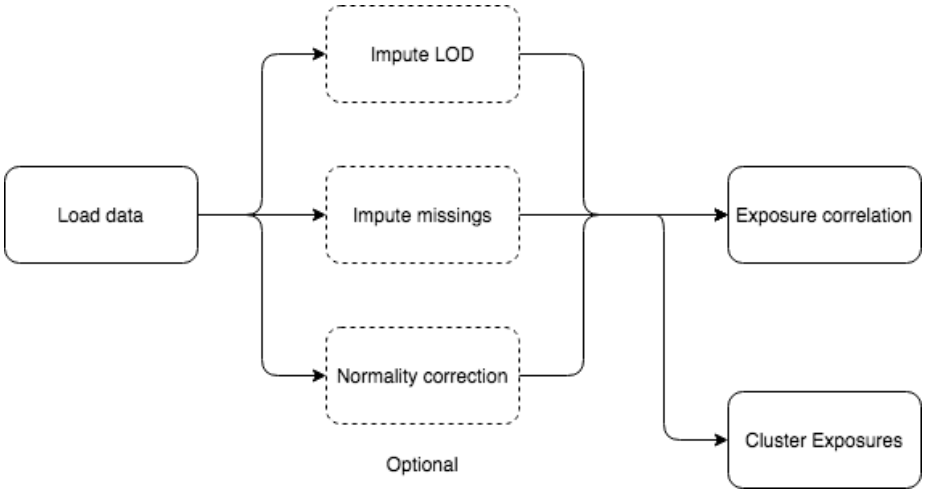
- Association of the phenotypes to the principal components



The values of this heatmaps can be visualized as tables by clicking on 'Visualize as table' and be downloaded as `*.csv`.


C:/Users/Gerard/OneDrive/Estudis/UAB/1A/TFM/git_repo/exposomeShiny/R - Shiny
http://127.0.0.1:4527
Open in Browser
Publish
rexposome
Data entry
Missing data
Exposure descriptiv
Check normality
Exposures Descript
PCA Visualization
PCA association wit
Exposure Correlatio
Cluster Exposures
ExWAS
Multivariate ExWAS
Omic Data
Download table
Show 10 entries
Search:
Dim
AbsPM25
As
BDE100
PC 01 0.817593001163009 -0.103340100694347 0.038301909356451 -0.03450
PC 02 0.0452993767510417 -0.277072667507918 0.17125474721416 0.1293
PC 03 -0.398153747438897 0.266613840121414 -0.230577273837154 0.04112
PC 04 -0.00888543453011813 0.104987089289347 0.112875762273989 0.1076
PC 05 0.0600266768536346 0.228990684508293 -0.364086334179696 -0.1468
PC 06 -0.304328052003655 -0.120317280919379 -0.136301079133983 -0.2122
PC 07 0.0600068838193947 0.161739814600629 -0.194170851484628 -0.07865
PC 08 -0.0416539270326857 0.14164105469711 -0.0576183122666913 0.02374
PC 09 0.00393387635010535 0.0727406950049364 0.093704889847341 0.2833
PC 10 0.0662569270189176 0.316569503906377 -0.114268311560933 0.06417
Showing 1 to 10 of 10 entries
Previous
1
Next
Close
Correlation
0.5
0.0
-0.5
PC 01
PC 02
PC 03
PC 04
PC 05
PC 06
PC 07
PC 08
PC 09
PC 10

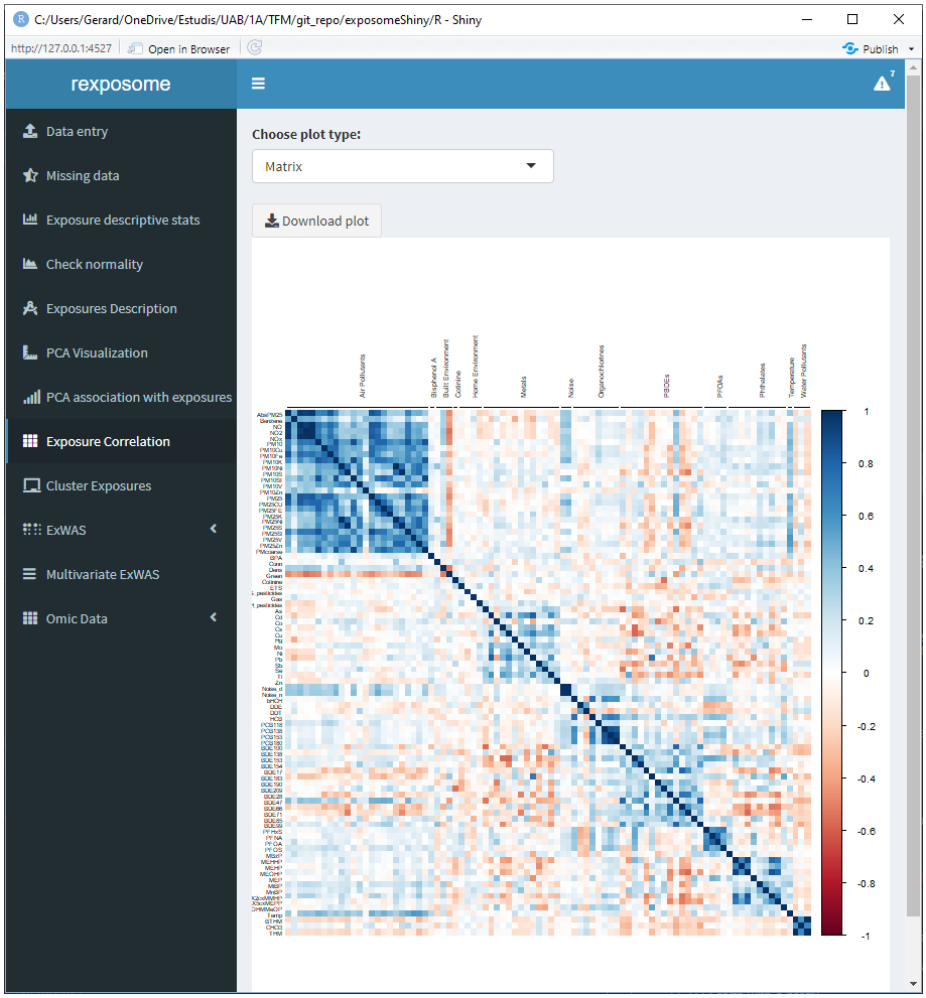
5.1.8 Correlation of exposures



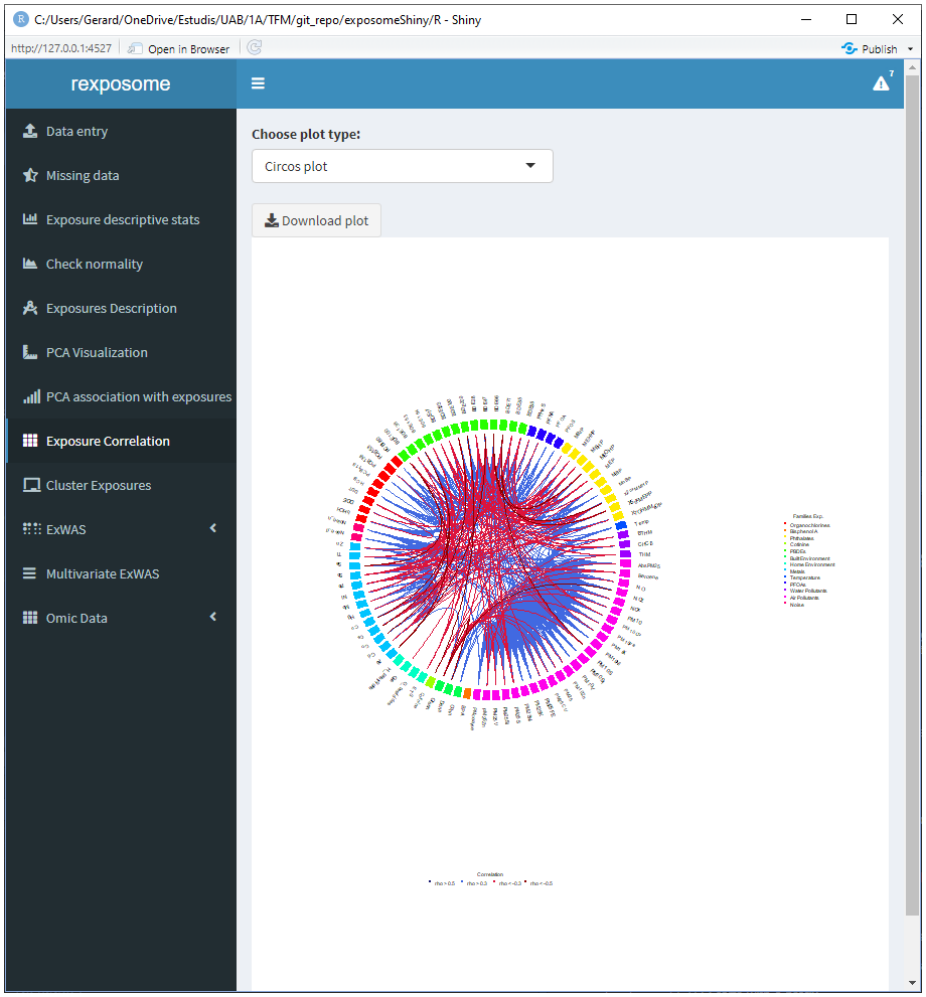
Displaying the correlation of the exposures can help to visualize intra and inter family relations between the exposures, for that reason there are two different visualization options, the circos and the matrix.

The correlation is computed using Pearson method for numerical-to-numerical correlation, Cramer's V for categorical-to-categorical correlation and linear models for categorical-to-numerical.

5.1.8.1 Matrix plot

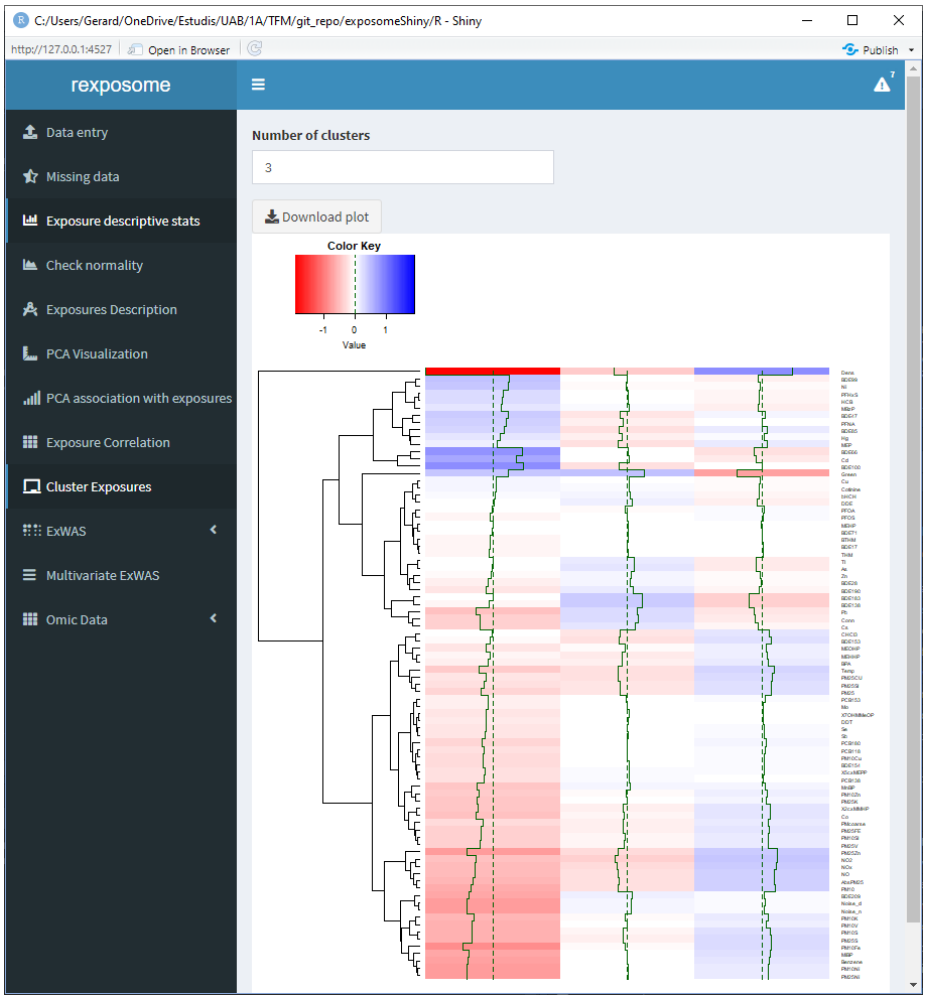


5.1.8.2 Circos plot

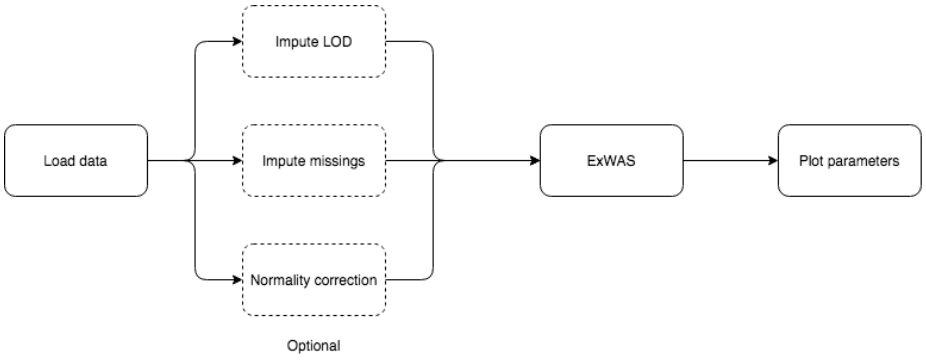


5.1.9 Clusterization of exposures

The clusterization of exposures uses a hierarchical clustering algorithm to classify the individuals profiles of exposures in k groups, where k can be selected by the user. The plot shows the profile for each group of individuals.


C:/Users/Gerard/OneDrive/Estudis/UAB/1A/TFM/git_repo/exposomeShiny/R - Shiny
http://127.0.0.1:4527
Open in Browser
Publish
rexposome
Data entry
Missing data
Exposure descriptive stats
Check normality
Exposures Description
PCA Visualization
PCA association with exposures
Exposure Correlation
Cluster Exposures
ExWAS
Multivariate ExWAS
Omic Data
Number of clusters
3
Download plot
Color Key
-1
0
1
Value

5.1.10 ExWAS

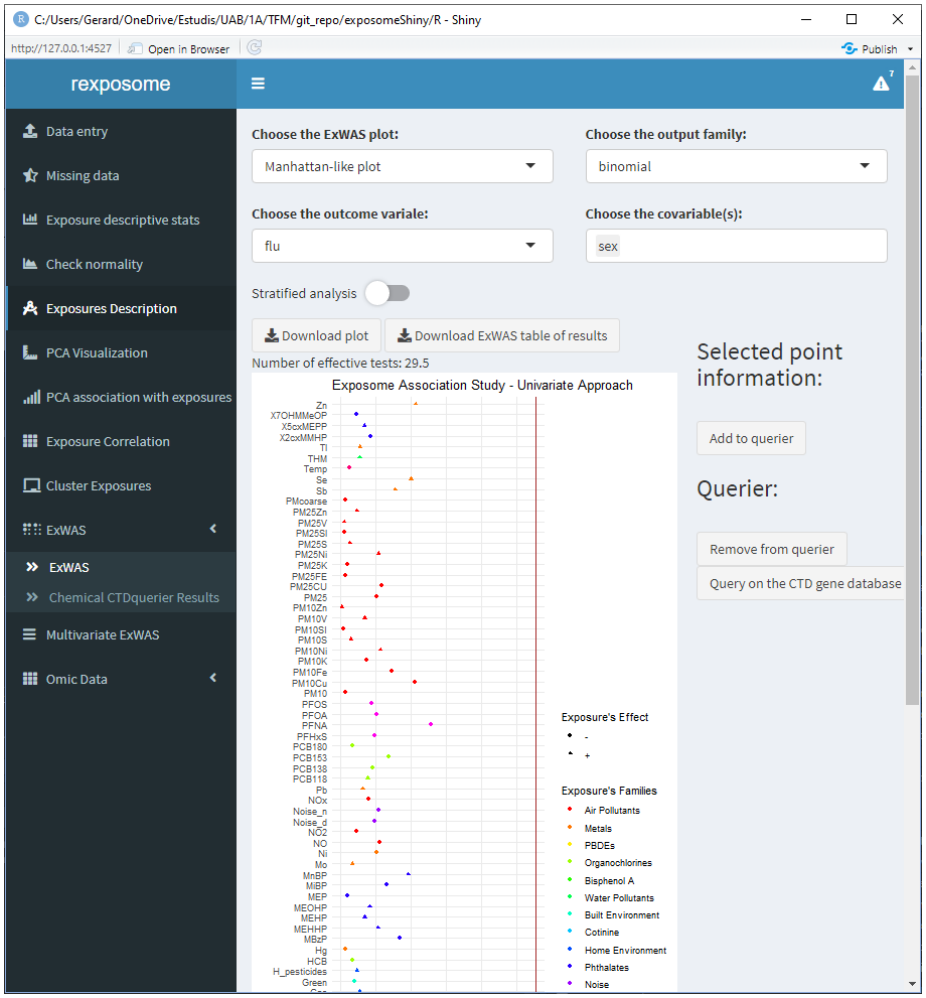

Impute LOD
Load data
Impute missings
ExWAS
Plot parameters
Normality correction
Optional

There are two ExWAS methods implemented, the regular and the stratified. The difference is that the stratified performs multiple ExWAS subsetting the data with a quantitative phenotype (example: ExWAS with male data and ExWAS with female data).

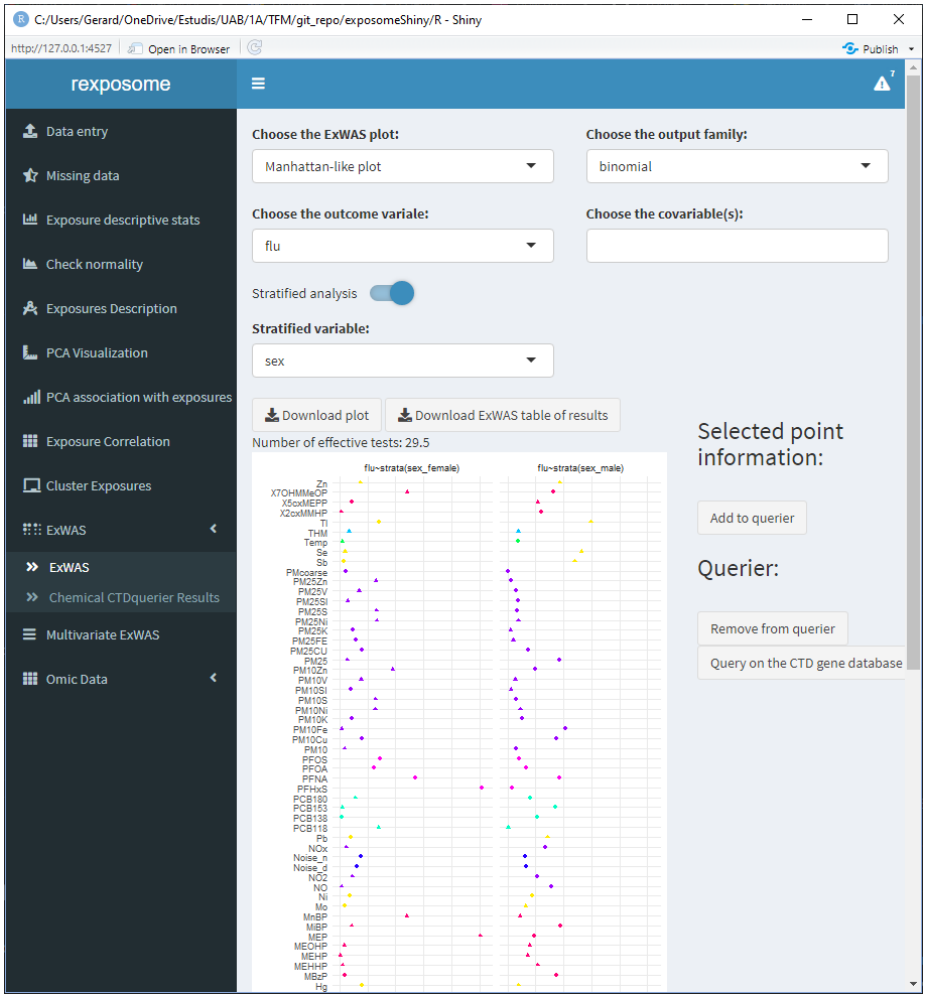
To perfom an ExWAS enter the required fields:

- Outcome variable: Output phenotype
- Covariable(s): Adjusting phenotype(s)
- Output family: Family of the outcome variable (Gaussian/Binomial/Poisson)
- Type of plot: Manhattan / Effects

5.1.10.1 Regular ExWAS. Outcome: flu, Covariable: Sex, Family: Binomial

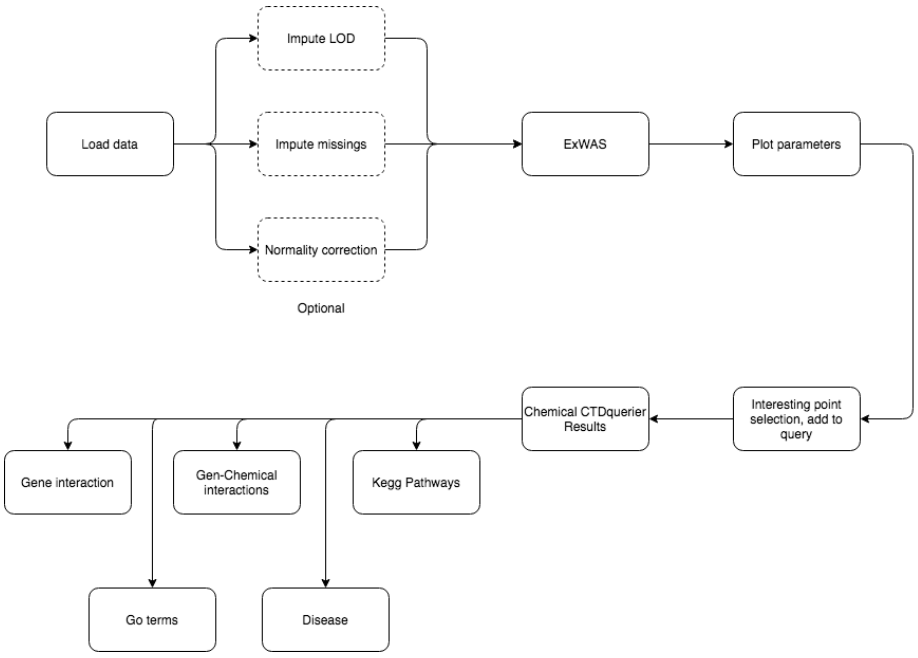


5.1.10.2 Stratified ExWAS. Outcome: flu, Cobariable: None, Family: Binomial, Stratifying variable: Sex



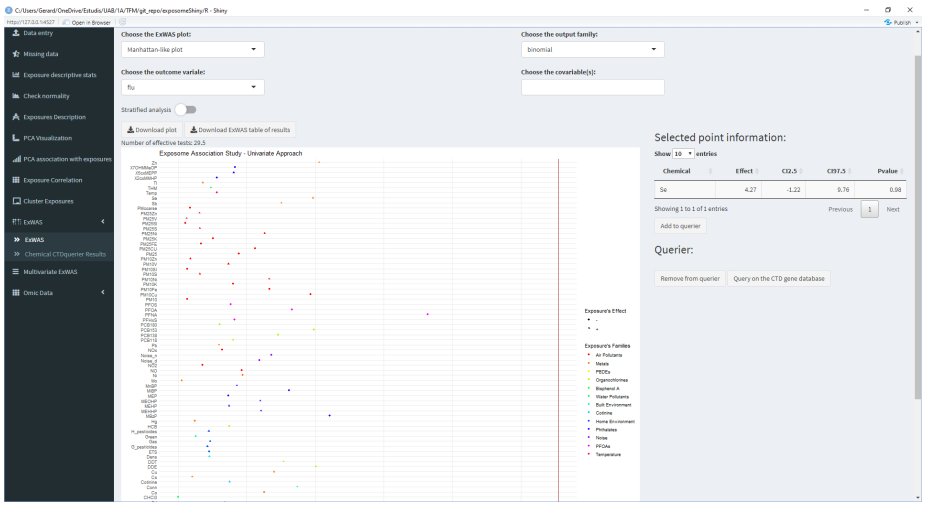
The plots and results table can be downloaded. The results table contains the exposure names, association pvalue and effect (with CI 95% for the effect). For the stratified ExWAS, the results table is actually a `*.zip` file with a file for each subset.

5.1.11 ExWAS - CTDquerier

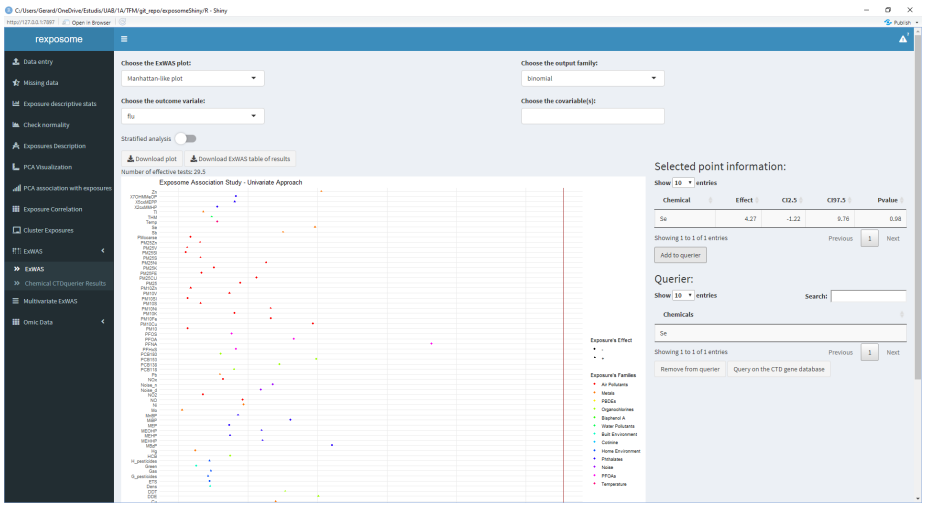


The ExWAS analysis can reveal high association between health outcomes and exposures. To obtain further information (from the CTD database) on exposures of interest:

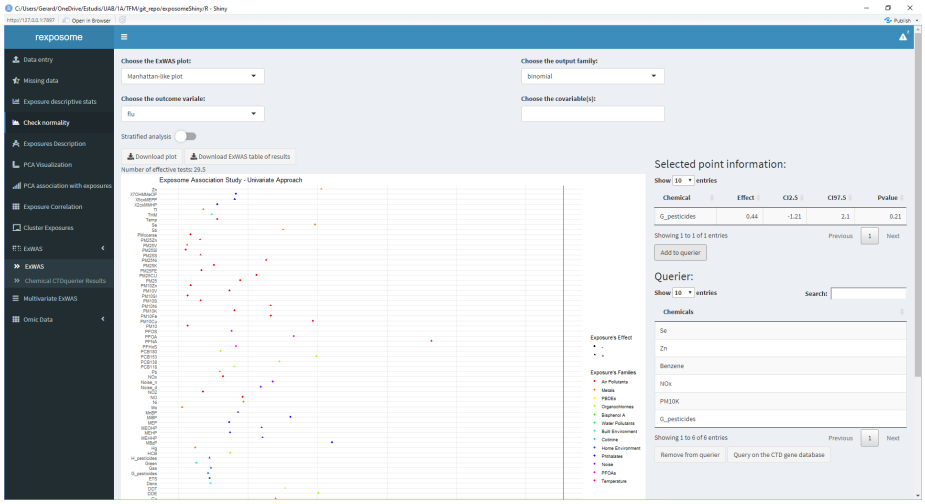
1. Perform the ExWAS
2. Click on the exposure of interest on the plot
3. A table with information about the selected exposure will appear on the right



4. If the exposure shown on the right table is of interest, click on 'Add to querier'



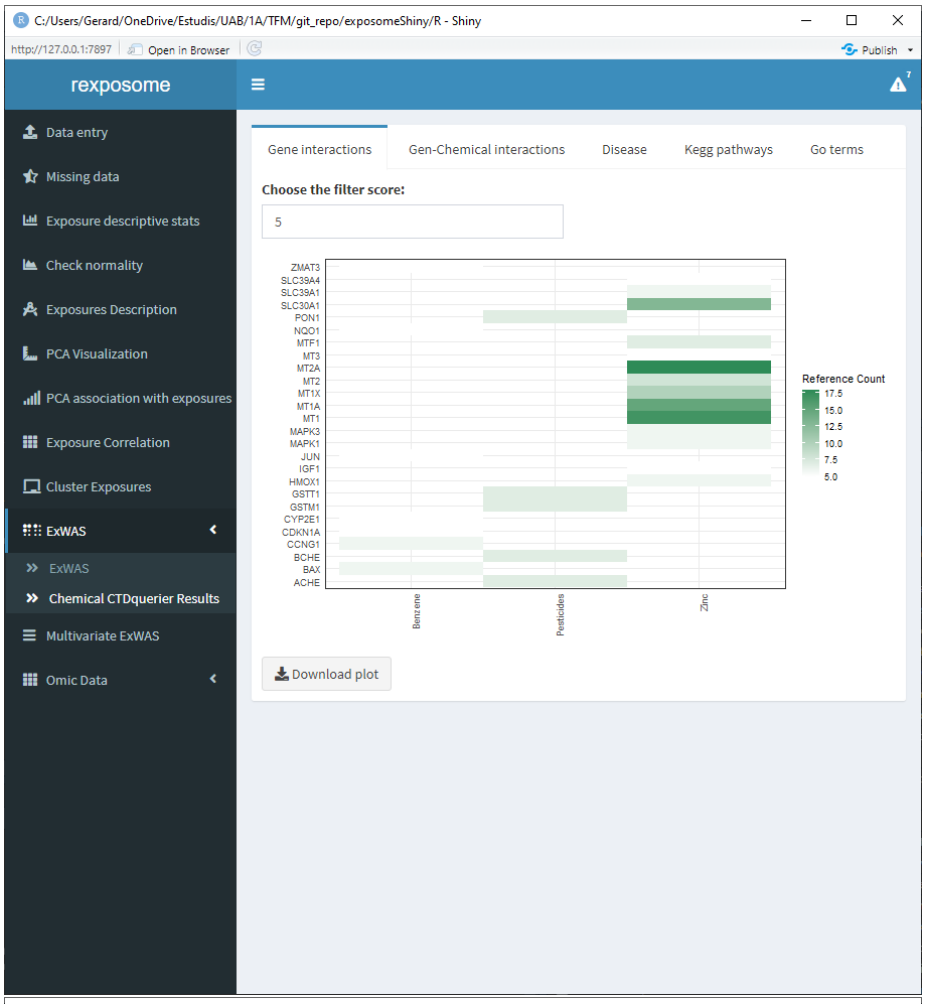
5. Repeat 2/3/4 for all the exposures of interest

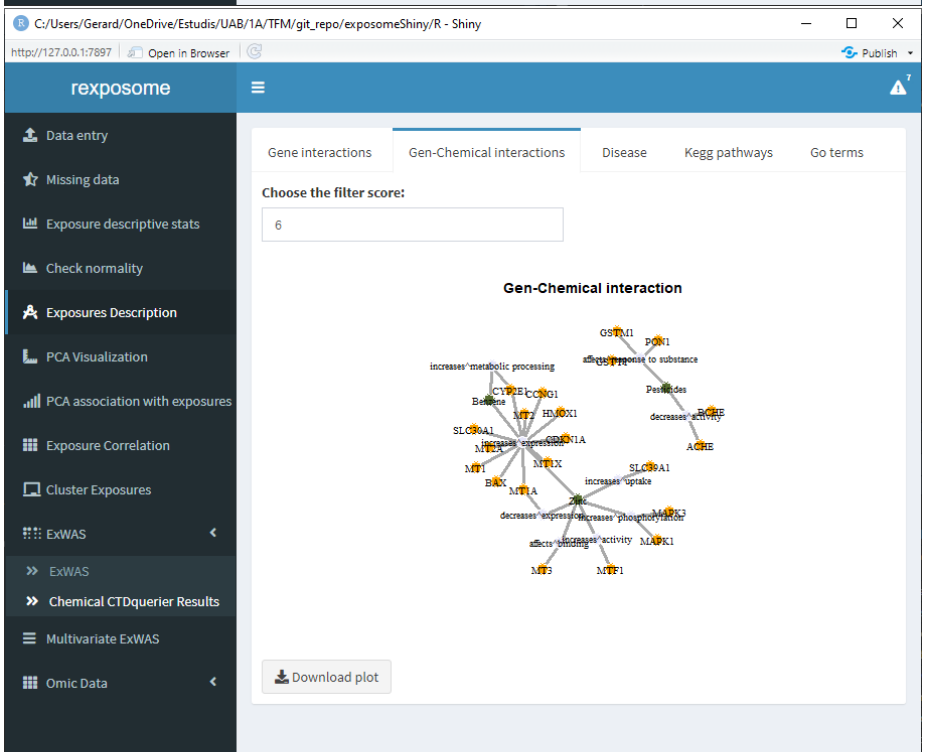


6. Once all the exposures of interest are on the 'Querier' table, click 'Query on the CTD gene database'

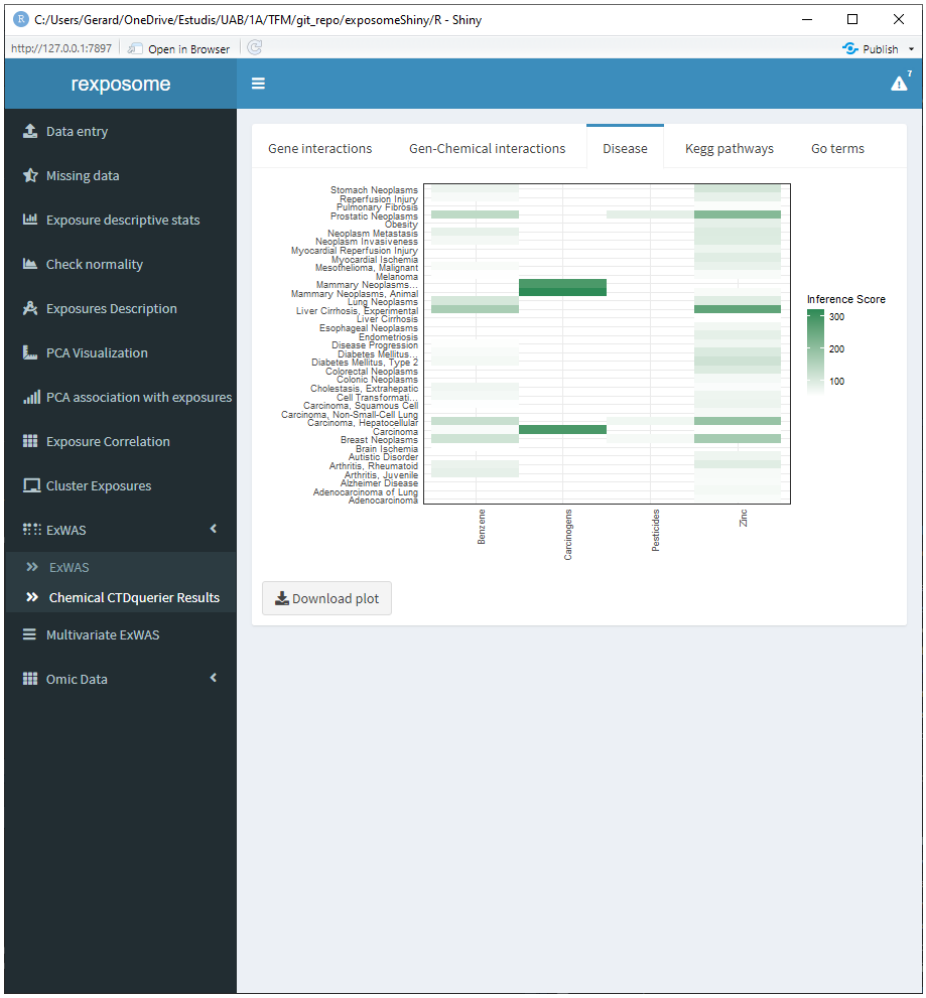
The results of this query can be visualized on the 'Chemical CTDquerier Results'. There are five visualization options:

5.1.11.1 Gene - chemical interactions

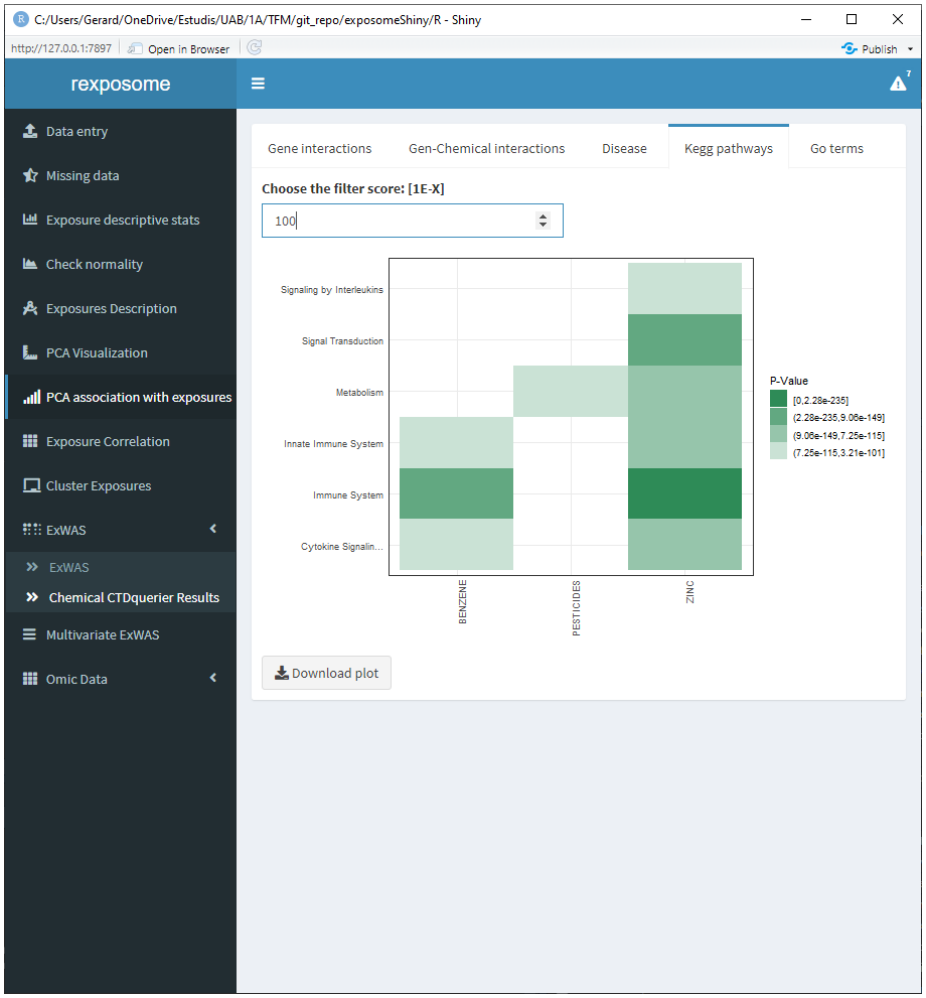




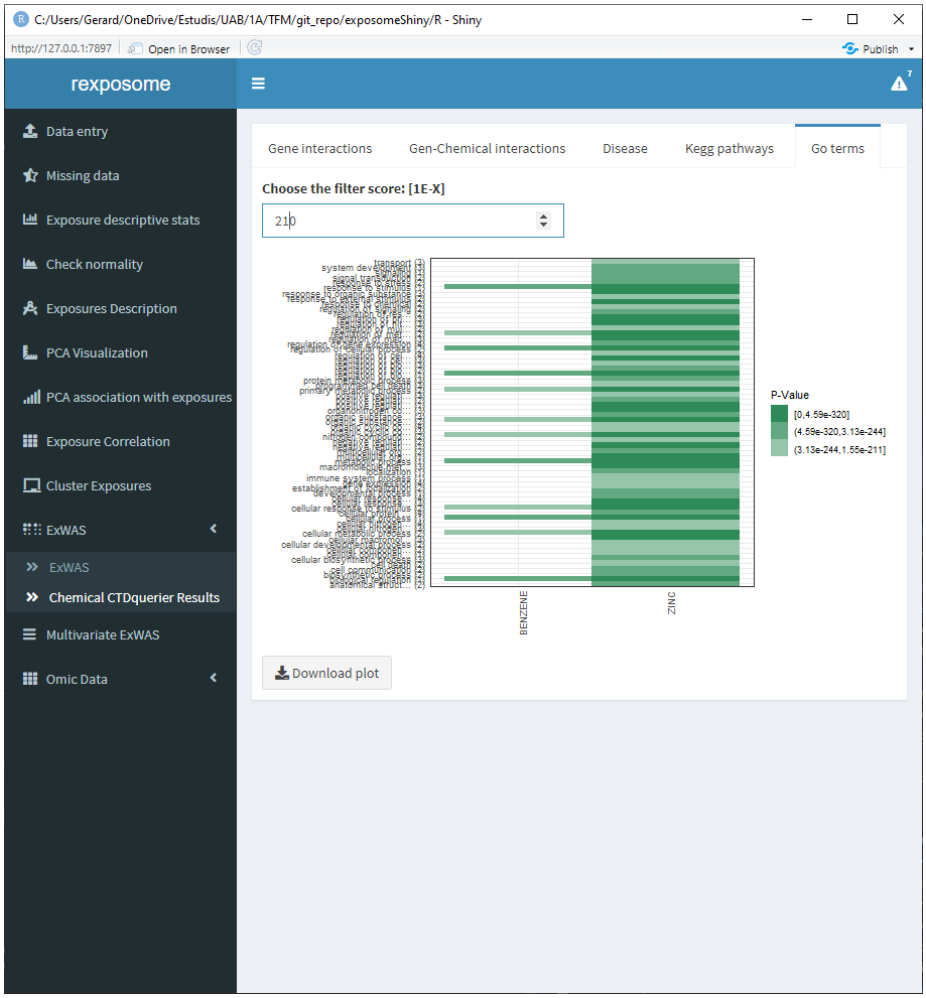
5.1.11.2 Disease relation



5.1.11.3 Kegg pathways

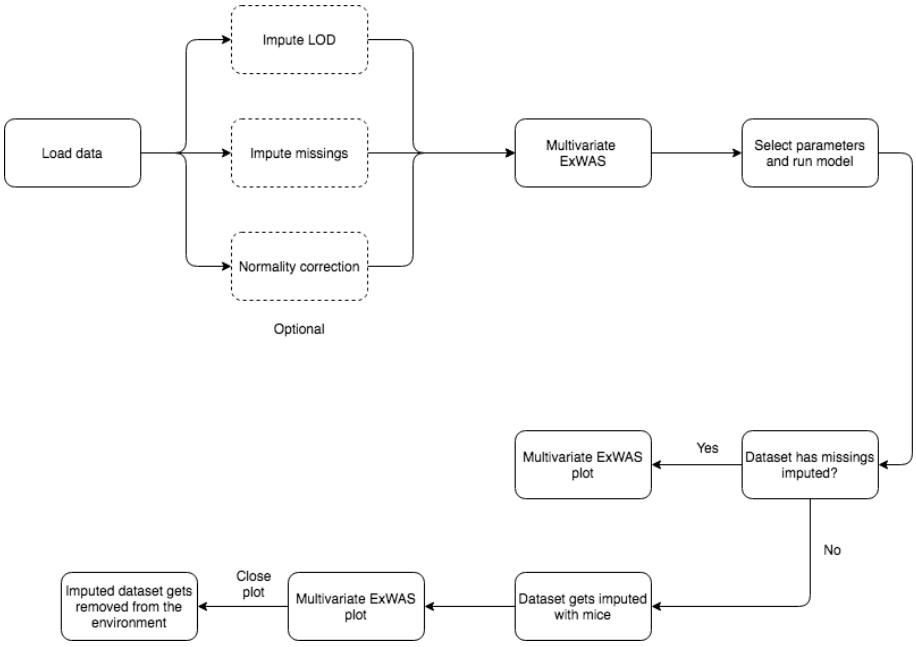


5.1.11.4 Go terms



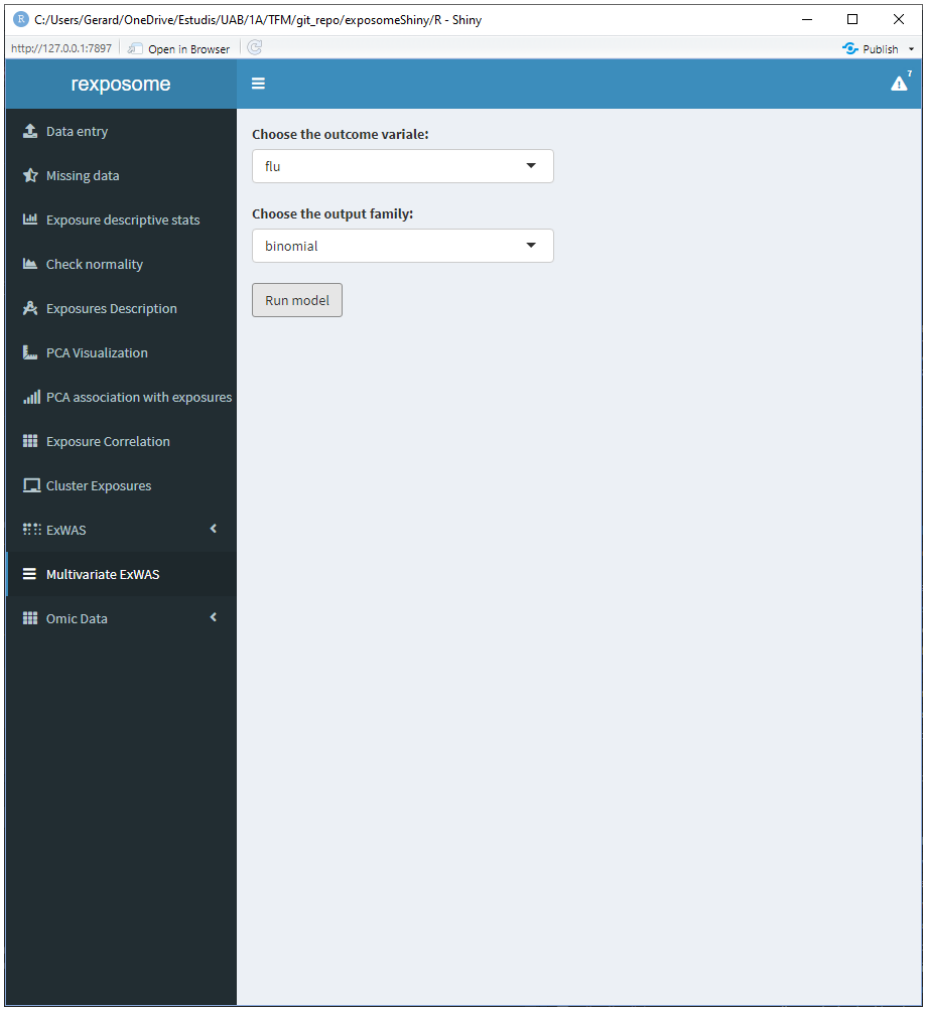
5.1.12 Variable selection ExWAS

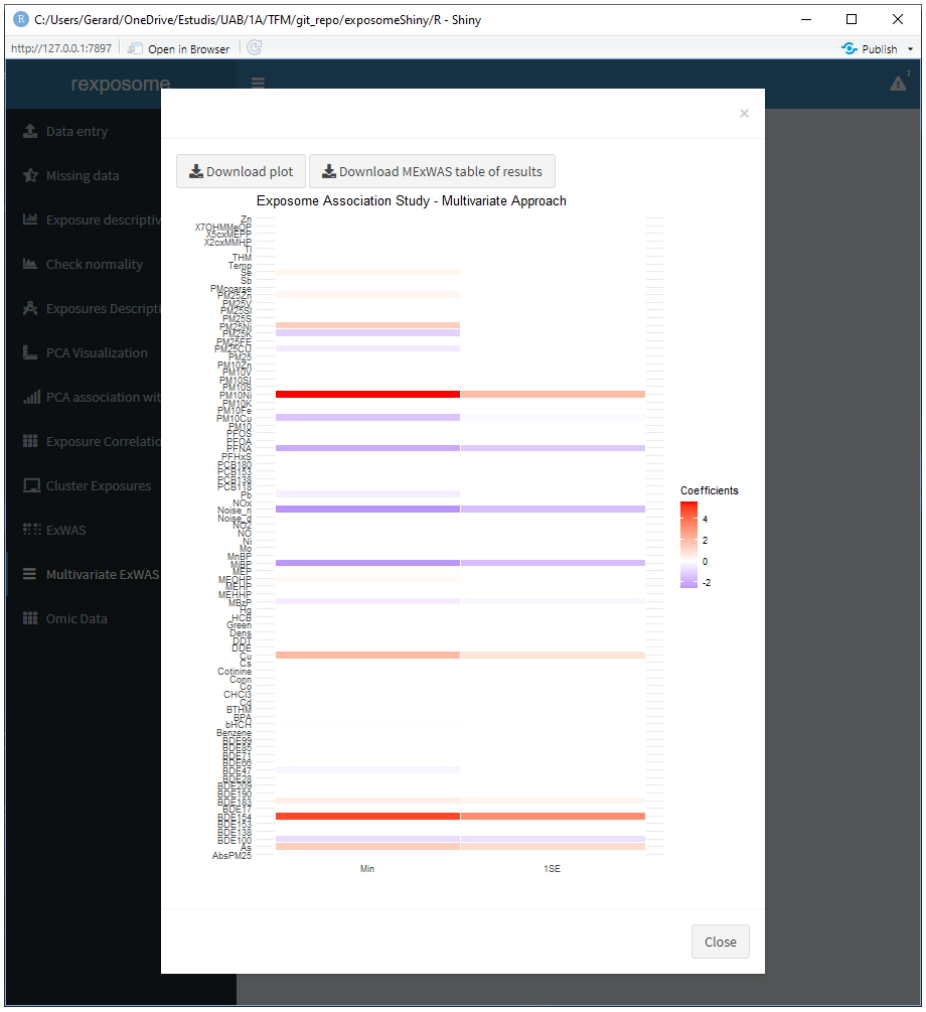
Variable selection ExWAS applies elastic net (LASSO regression) to the exposures given a health outcome of interest. The resulting heat map is coloured with the coefficient of each exposure in relation to the health outcome, so the ones in white are not associated. The two columns of the heat map correspond to the minimum lambda (`Min`) and to the lambda which gives the most regularized model such that error is within one standard error of the minimum (`1SE`).



To perform a variable selection ExWAS study, check the Variable selection ExWAS tab and select the desired output parameter, click on run model to generate the plot.

This analysis requires data without missings, so if there are missings MICE imputation is applied beforehand. After the model is fitted, the imputed dataset will be removed.


C:/Users/Gerard/OneDrive/Estudis/UAB/1A/TFM/git_repo/exposomeShiny/R - Shiny
http://127.0.0.1:7897
Open in Browser
Publish
rexposome
Data entry
Missing data
Exposure descriptive stats
Check normality
Exposures Description
PCA Visualization
PCA association with exposures
Exposure Correlation
Cluster Exposures
ExWAS
Multivariate ExWAS
Omic Data
Choose the outcome variale:
flu
Choose the output family:
binomial
Run model



The table of results and plot can be downloaded.

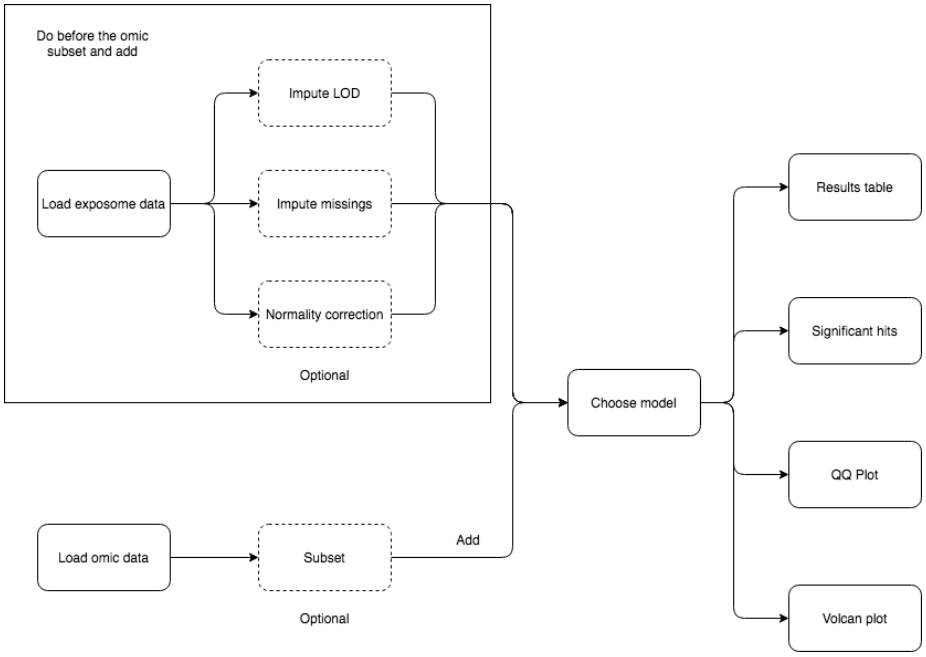
5.2 Exposome-Omic analysis

The aim of this analysis is to perform an association test between the gene expression levels and the exposures. The datasets used in this section are `exposures.csv`, `description.csv`, `phenotypes.csv` and `gene_exp.Rdata` which are available here.

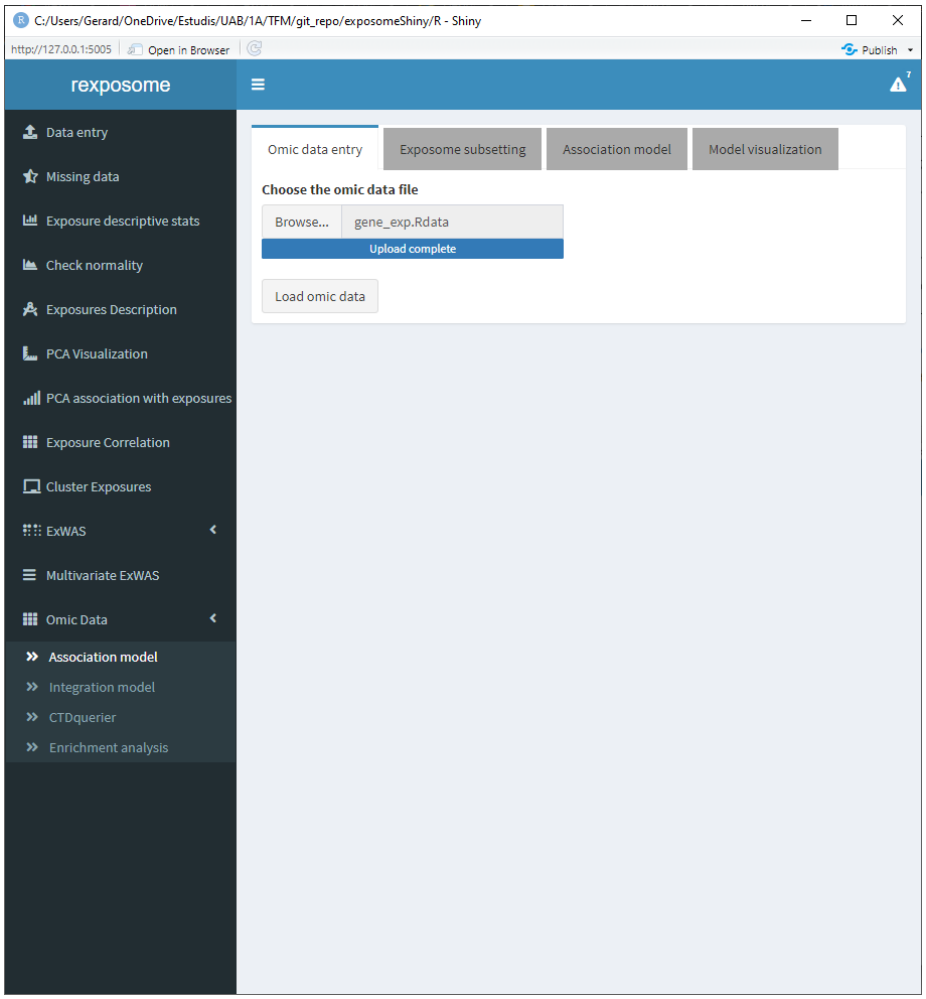
It's important noting that (by default) the maximum size of the omics data is 30 MB, if the omics file to be analyzed is bigger, change the line number 2 of the `server.R` file.

```
  # the "30" refers to 30MB, change as needed
options(shiny.maxRequestSize=30*1024^2)
```

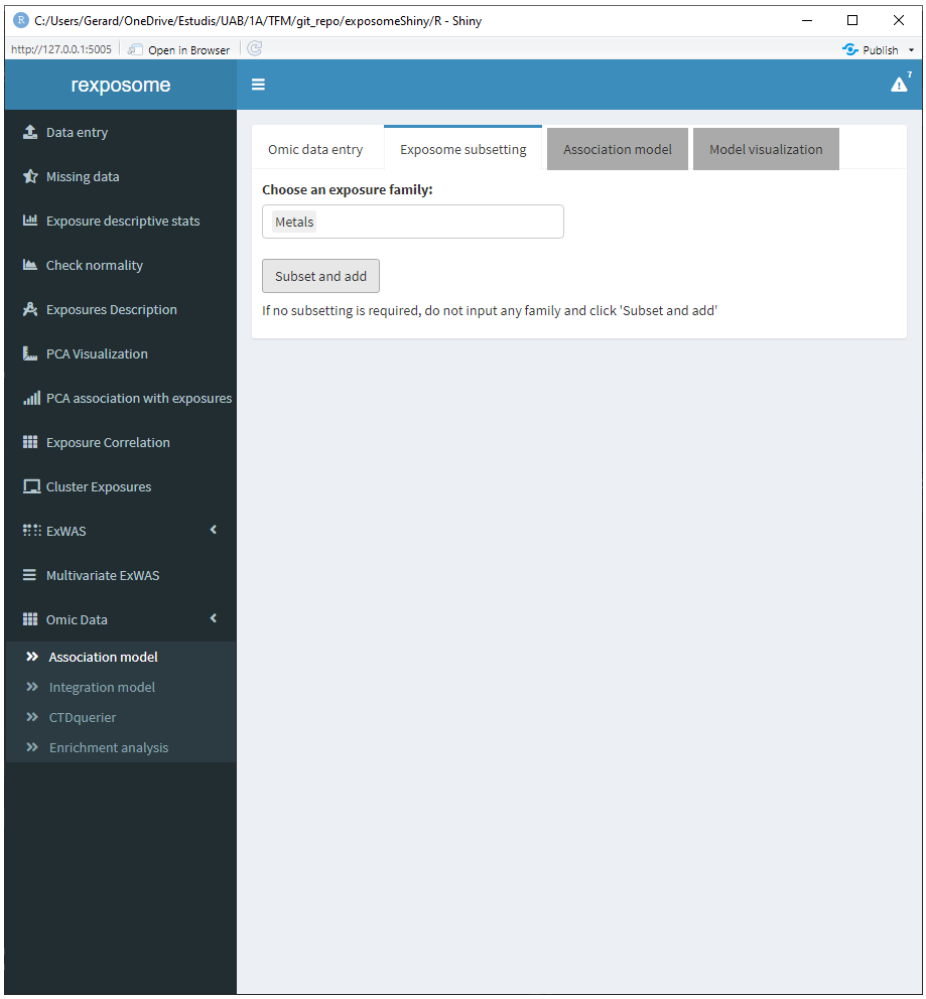
5.2.1 Association analysis



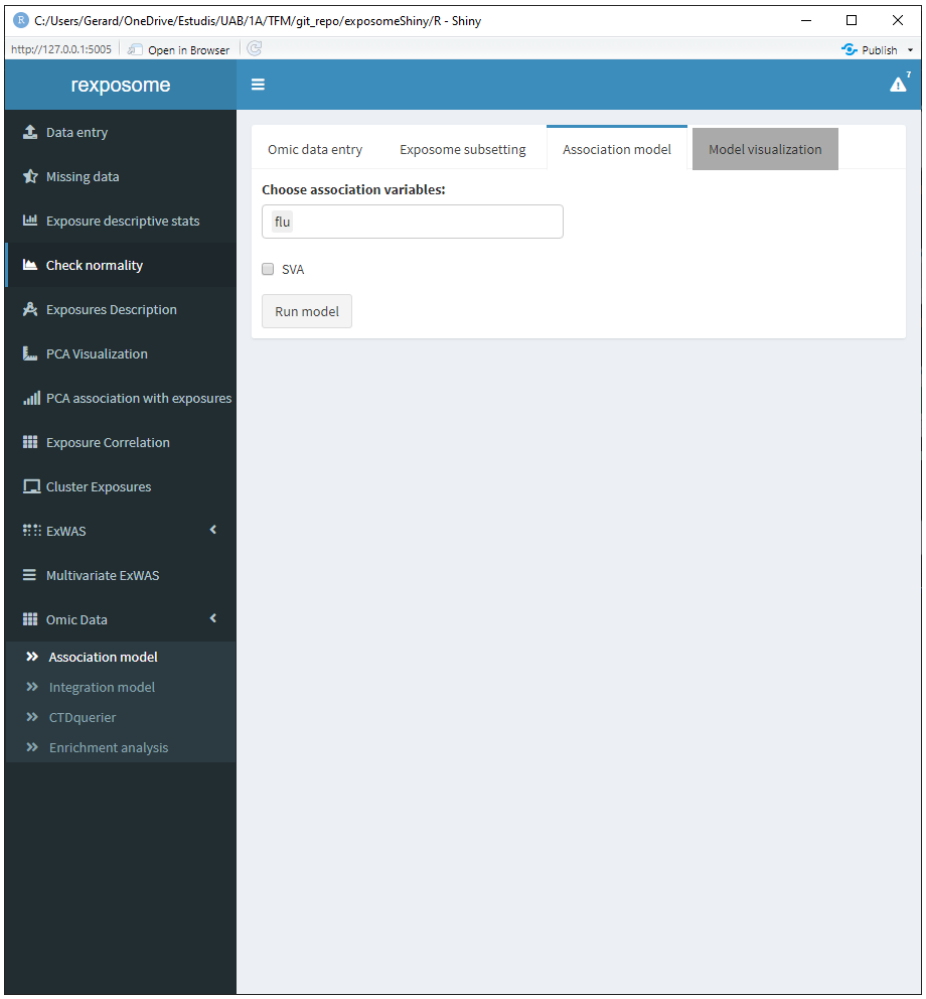
First, make sure that the exposome dataset is loaded. Then, proceed to loading the omics data, which should be providad as a `*.RData`.



This type of analysis is usually very computationally intensive, so it's useful to be able to limit the scope. The exposome data can be subsetted by families for that reason. If all the families are desired don't input any and proceed to click the "Subset and add".

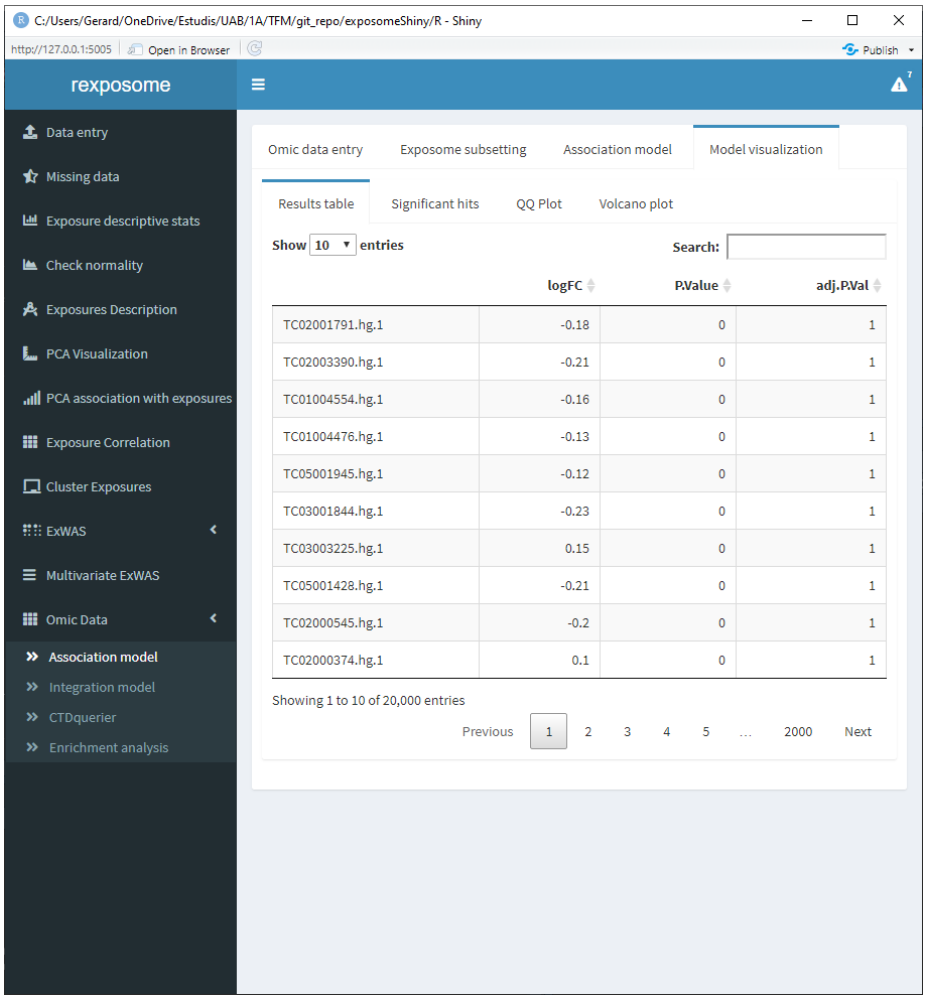


Select the variables for the association analysis (linear models) and if SVA (surrogate variable analysis) is wanted on the "Association model" subtab.

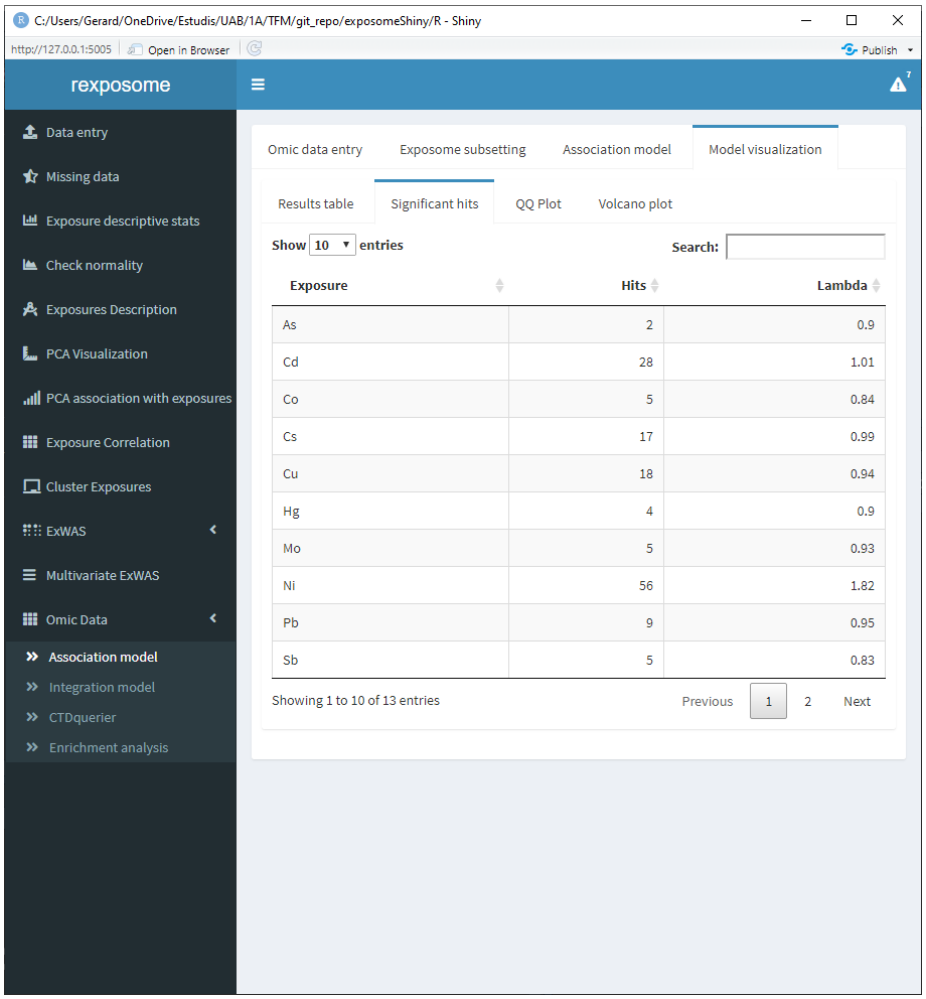


Once complete, there are various tabs to visualize the results.

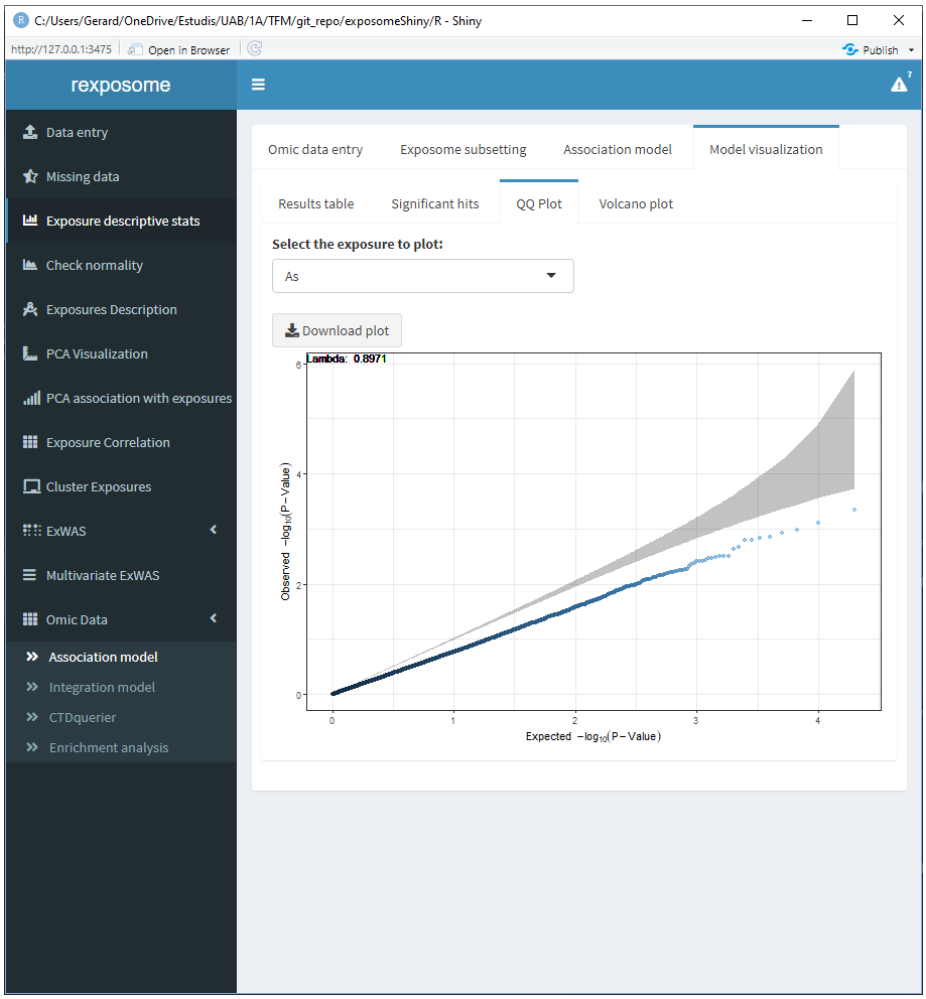
The “Results table” shows the gene, log of the fold change, p-value and adjusted p-value.



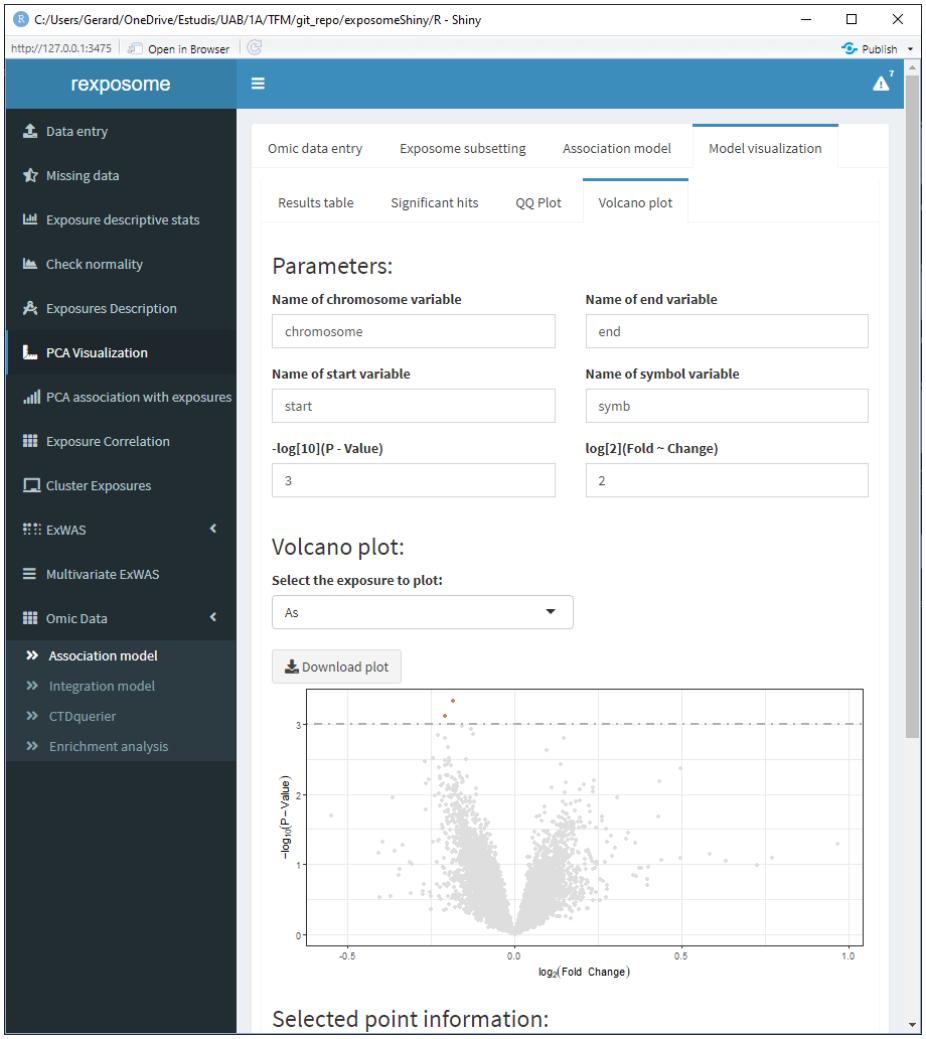
The “Significant hits” shows the exposure, hits and lambda.



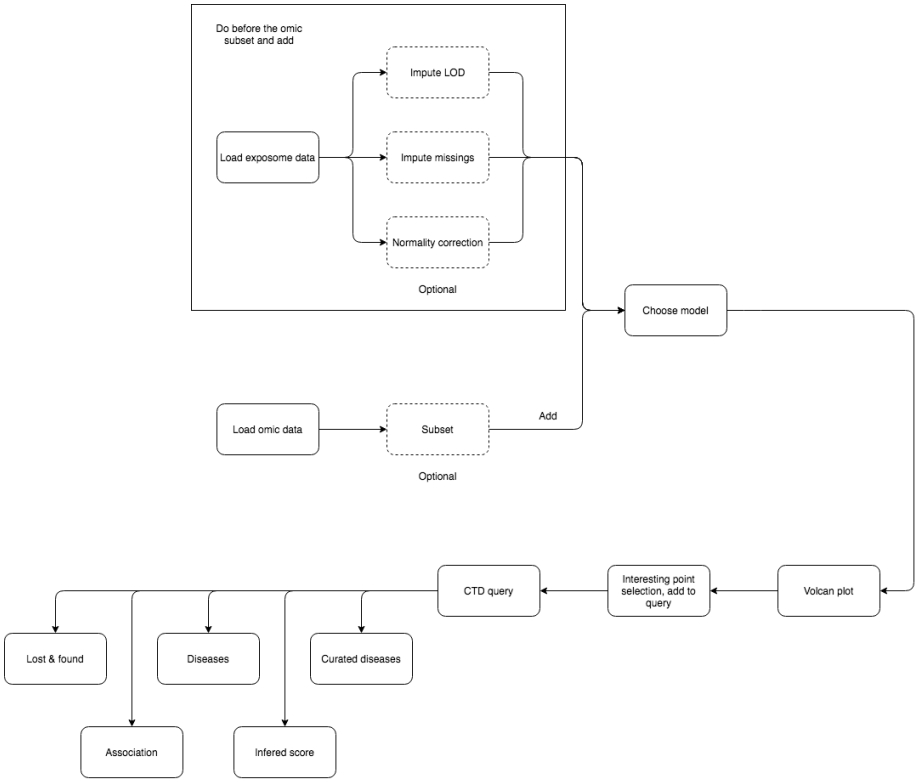
The QQ Plot shows a QQ plot (expected vs. observed -lo10(p-value)) for the selected exposure.



The Volcan plot shows a volcano plot (log2(fold change) vs -log10(p-value)). For this plot there are two input cells to adjust the horizontal and verital limit lines to filter out the results.


C:/Users/Gerard/OneDrive/Estudis/UAB/1A/TFM/git_repo/exposomeShiny/R - Shiny
http://127.0.0.1:3475
Open in Browser
Publish
rexposome
Data entry
Missing data
Exposure descriptive stats
Check normality
Exposures Description
PCA Visualization
PCA association with exposures
Exposure Correlation
Cluster Exposures
ExWAS
Multivariate ExWAS
Omic Data
Association model
Integration model
CTDquerier
Enrichment analysis
Omic data entry
Exposome subsetting
Association model
Model visualization
Results table
Significant hits
QQ Plot
Volcano plot
Parameters:
Name of chromosome variable
chromosome
Name of end variable
end
Name of start variable
start
Name of symbol variable
symb
-log[10](P - Value)
3
log[2](Fold ~ Change)
2
Volcano plot:
Select the exposure to plot:
As
Download plot
−log10(P − Value)
log2(Fold Change)
Selected point information:

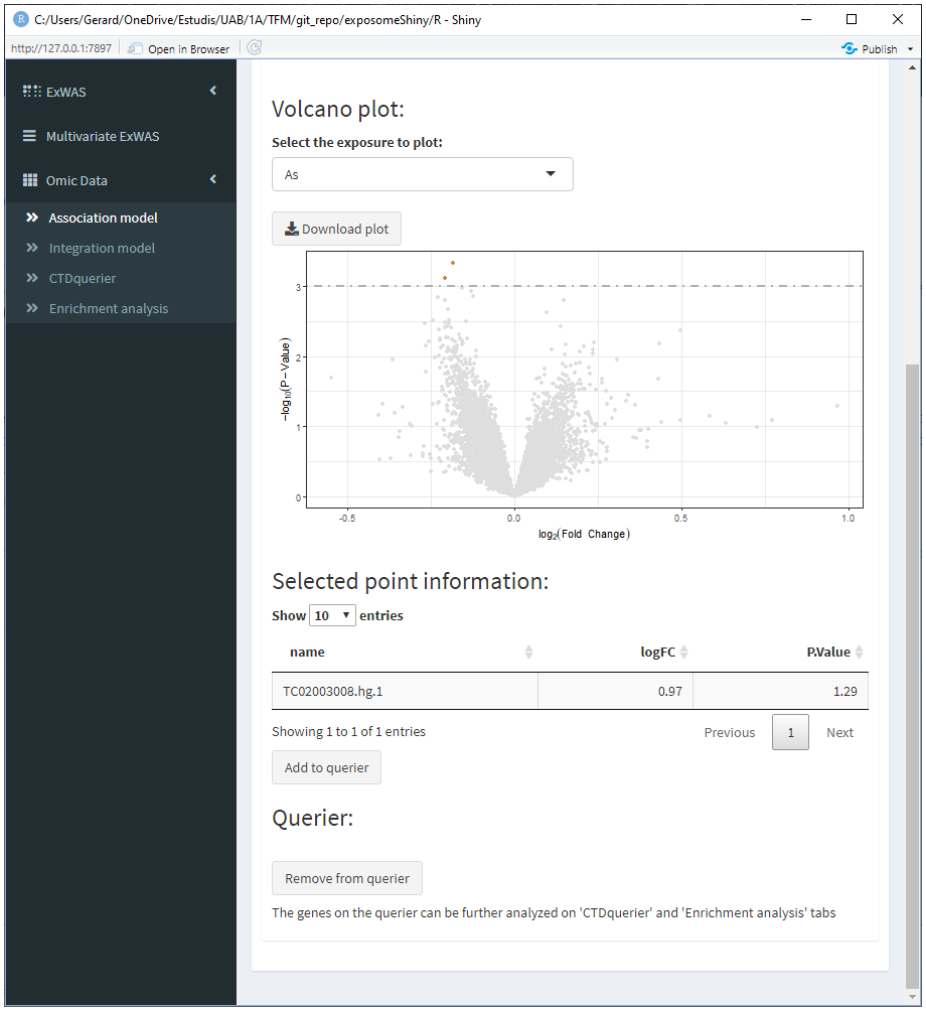
5.2.2 CTD querier



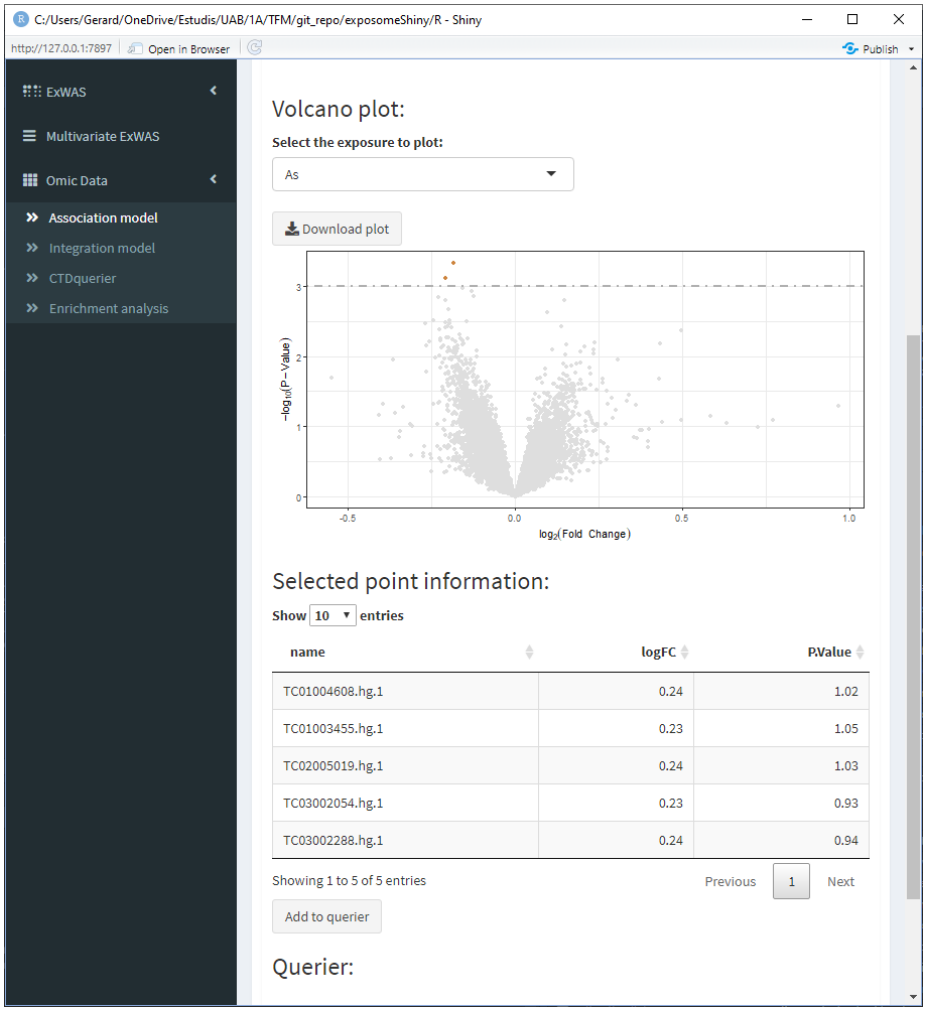
The association analysis of exposome data and omic data may point some genes of interest. Information about these genes can be queried on the CTD database.

To do so:

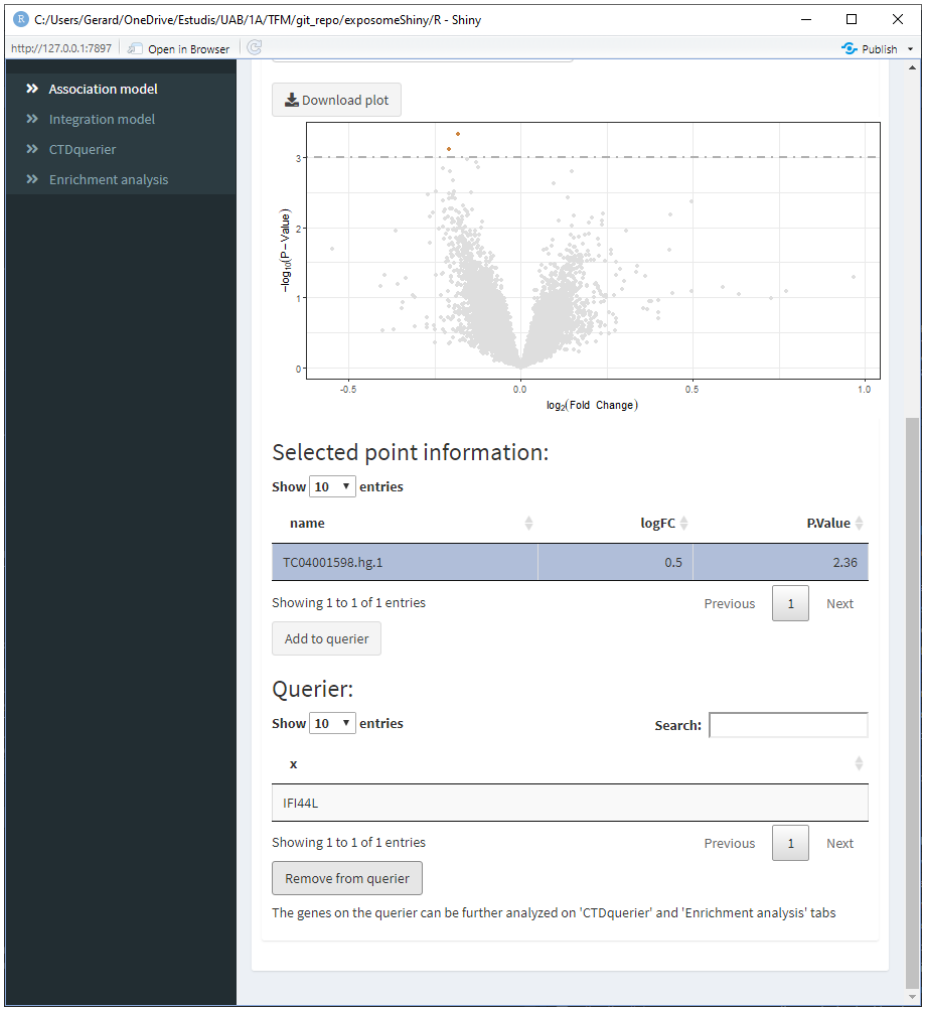
1. Perform association analysis
2. Go to the volcano plot visualization
3. Select one point. This will update the 'Selected point information' table at the bottom



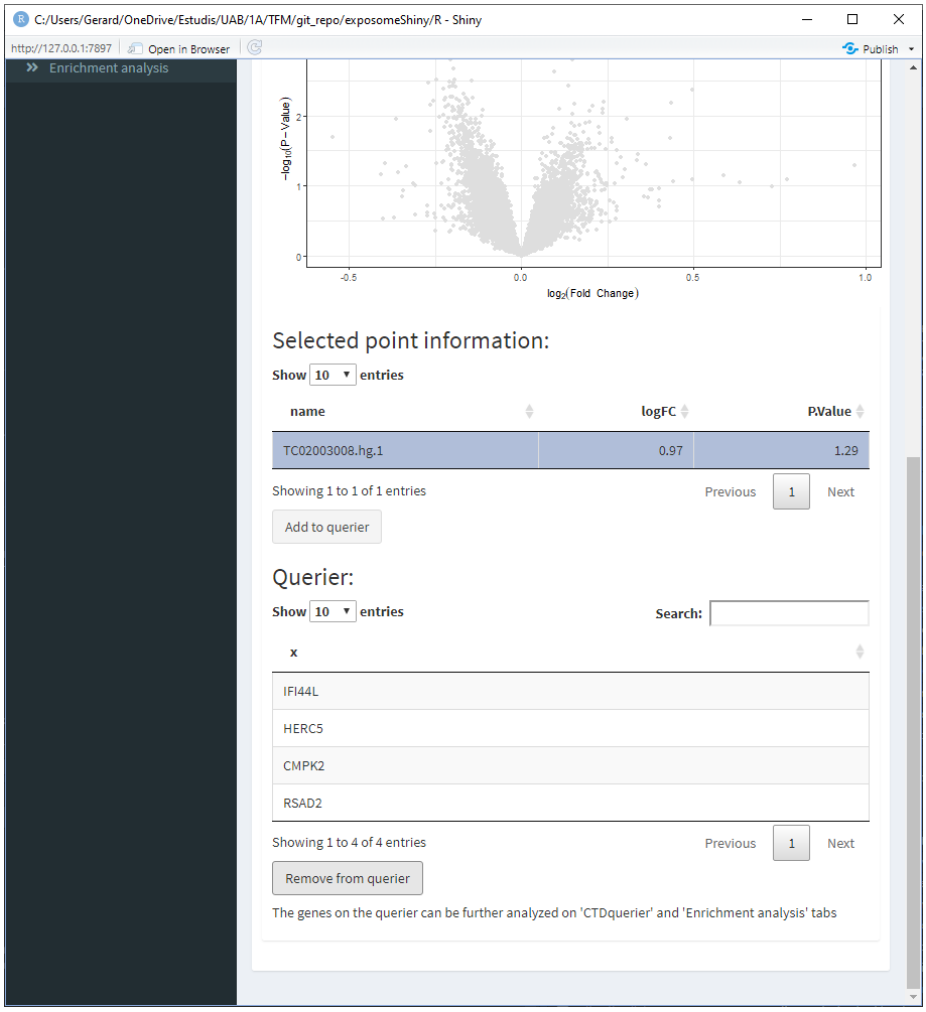
Sometimes multiple points may be very close, so this table will show information about them



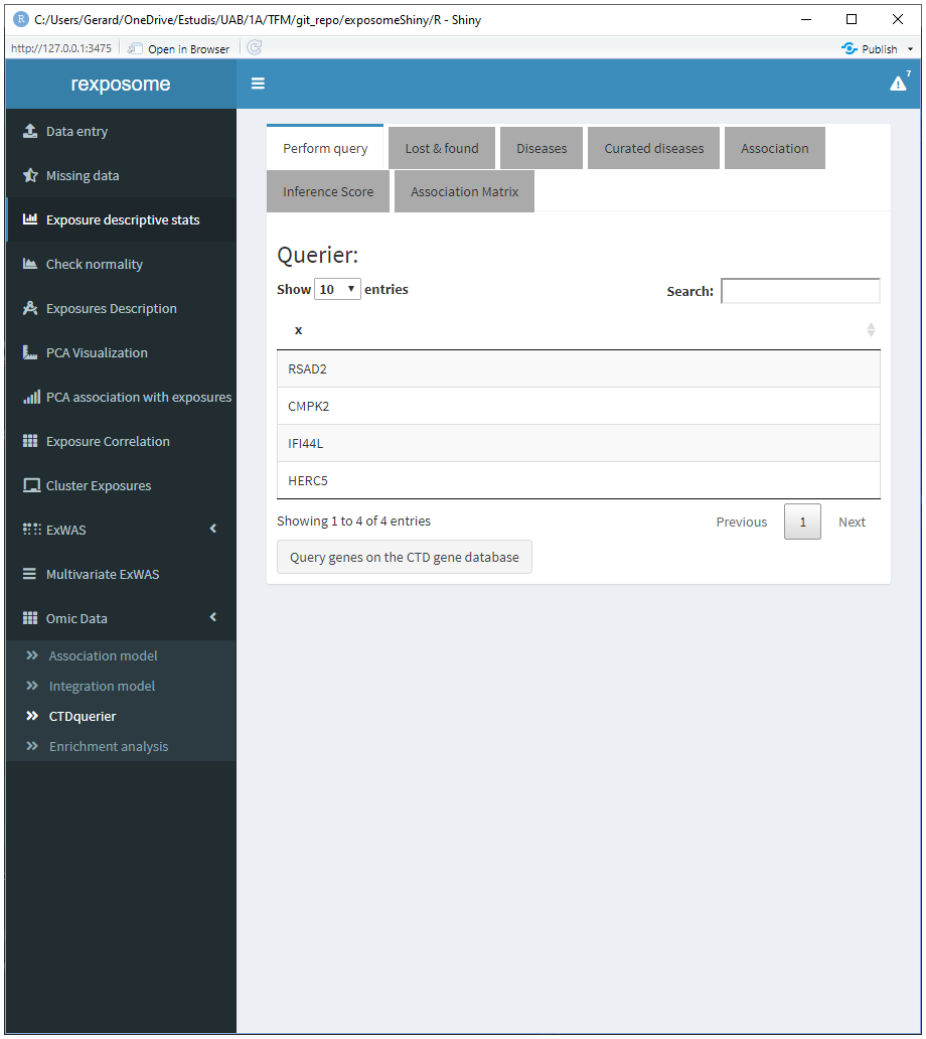
4. Select the point of interest from the 'Selected point information' table and click on 'Add to querier'. This will the search of the gene related to the selected SNP. Sometimes the search may not yield a resulting gene.



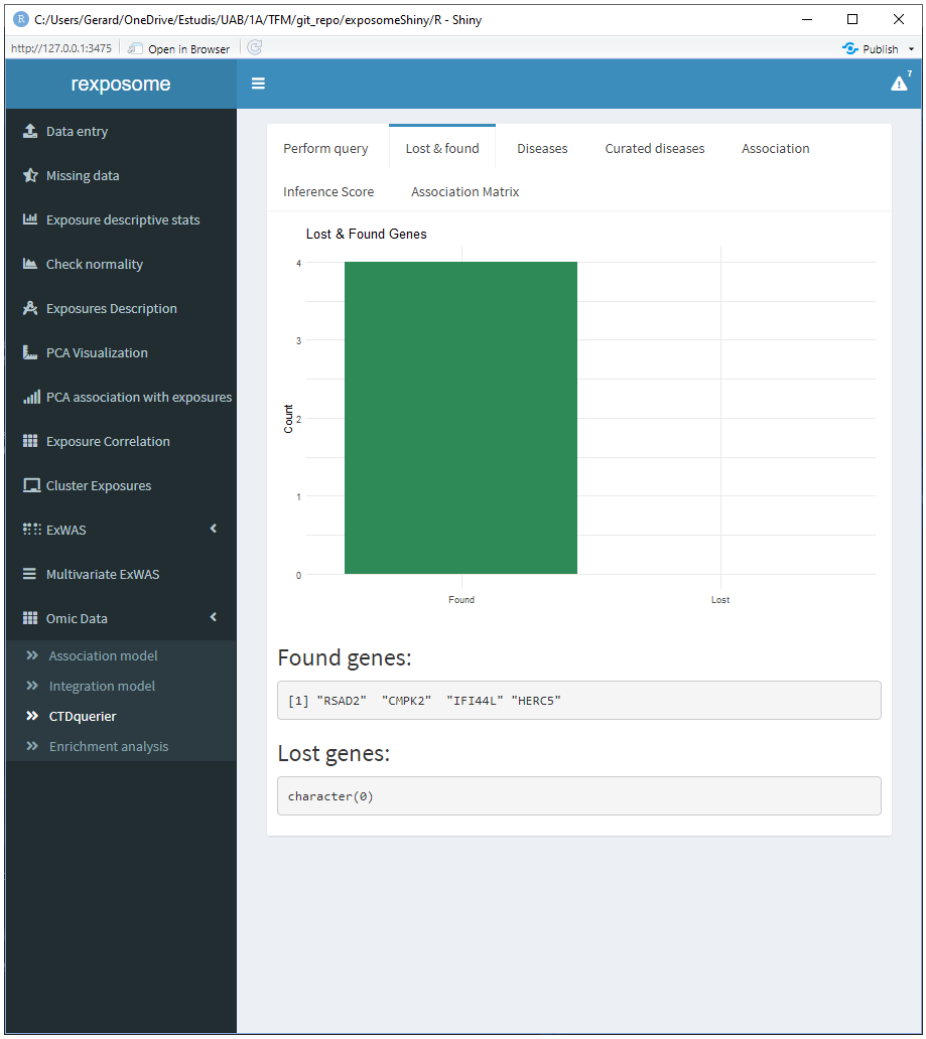
5. Repeat 3/4 for all the points of interest



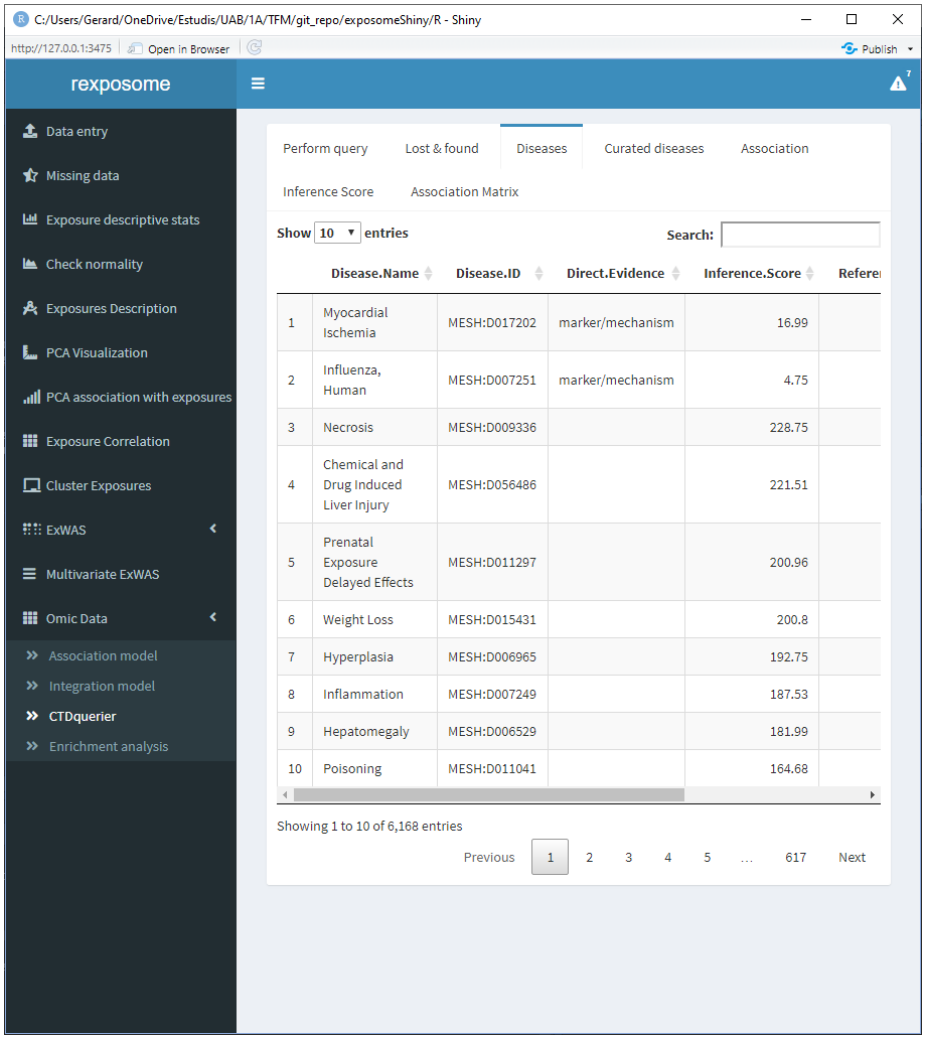
When all the genes of interest are on the querier, move to the 'CTDquerier' tab. On it, the genes to be queried can be visualized before performing the query. Once the query is performed, the results can be visualized on the different tabs.



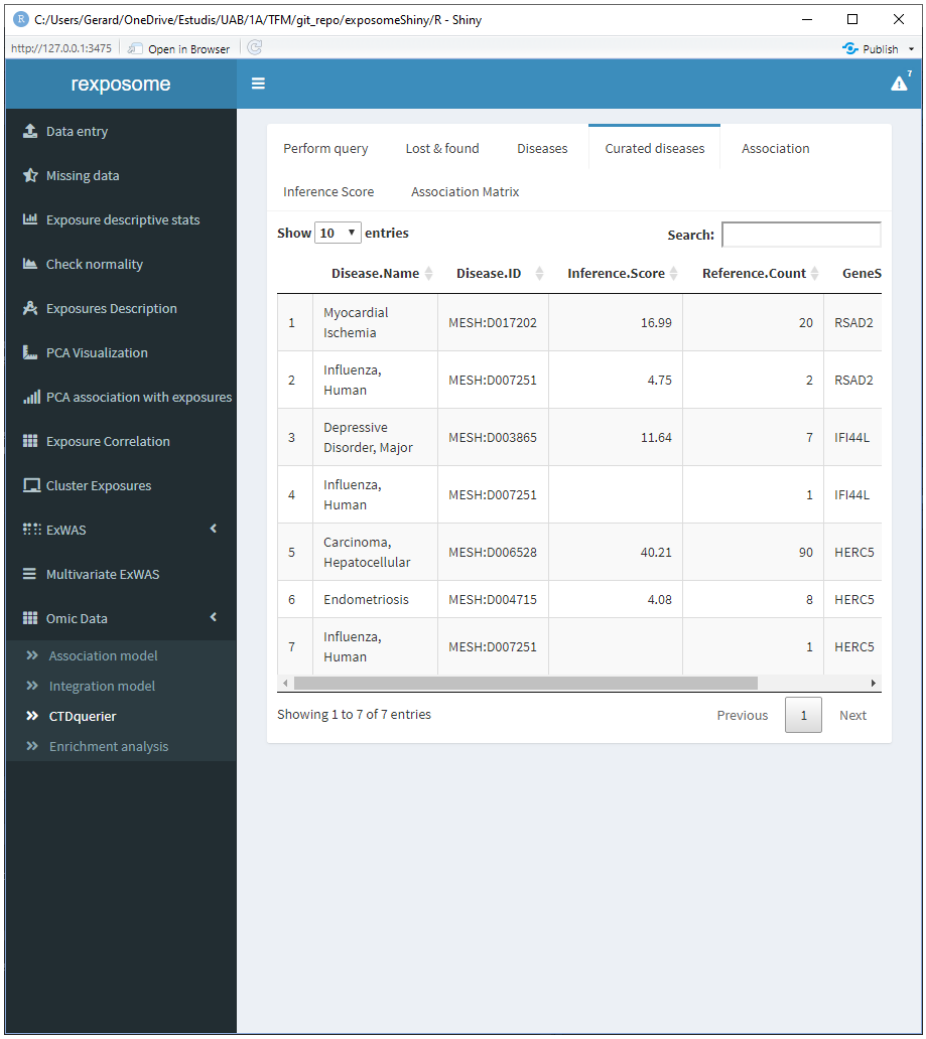
There are six tabs showing different results interpretations. First there's the "Lost & found" tab which a plot to see the amount of genes found on the CTD database and the ones that were not found , there's also two lists stating the names of them.



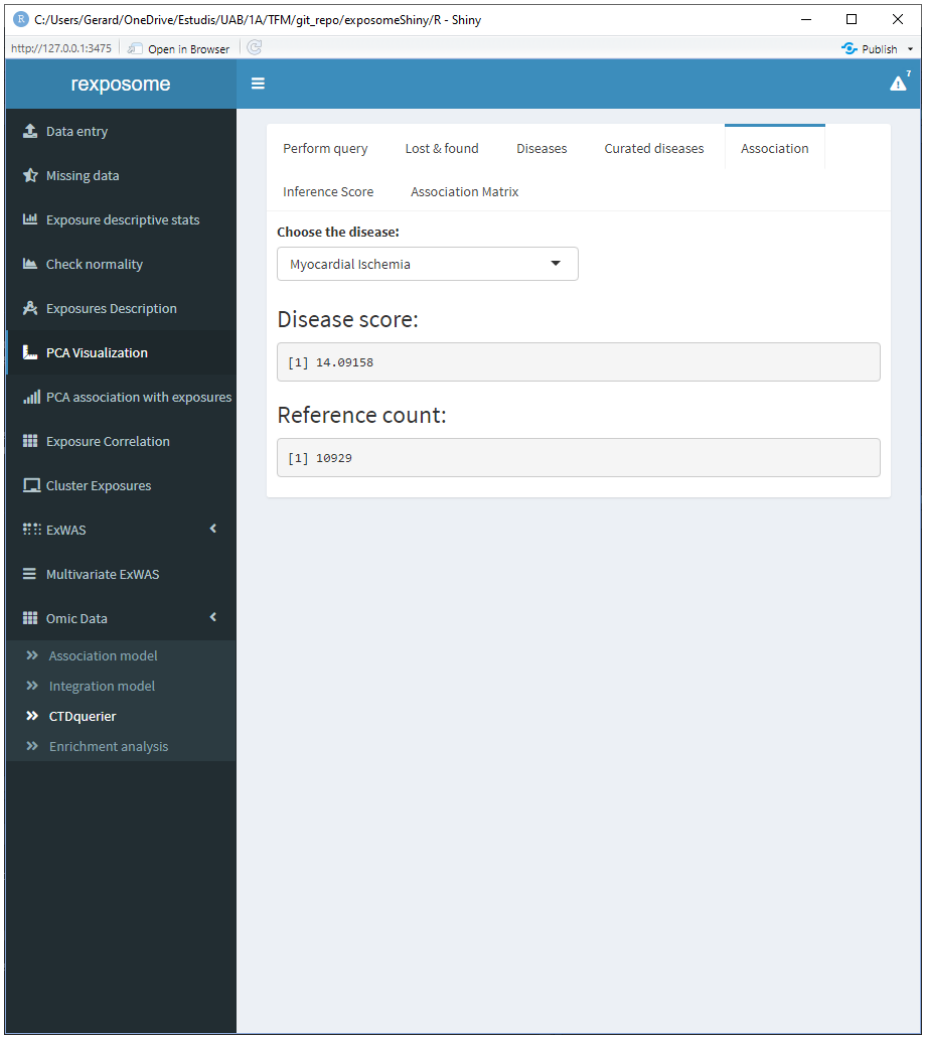
The diseases tab shows a table of all the associated diseases found on the CTD database.



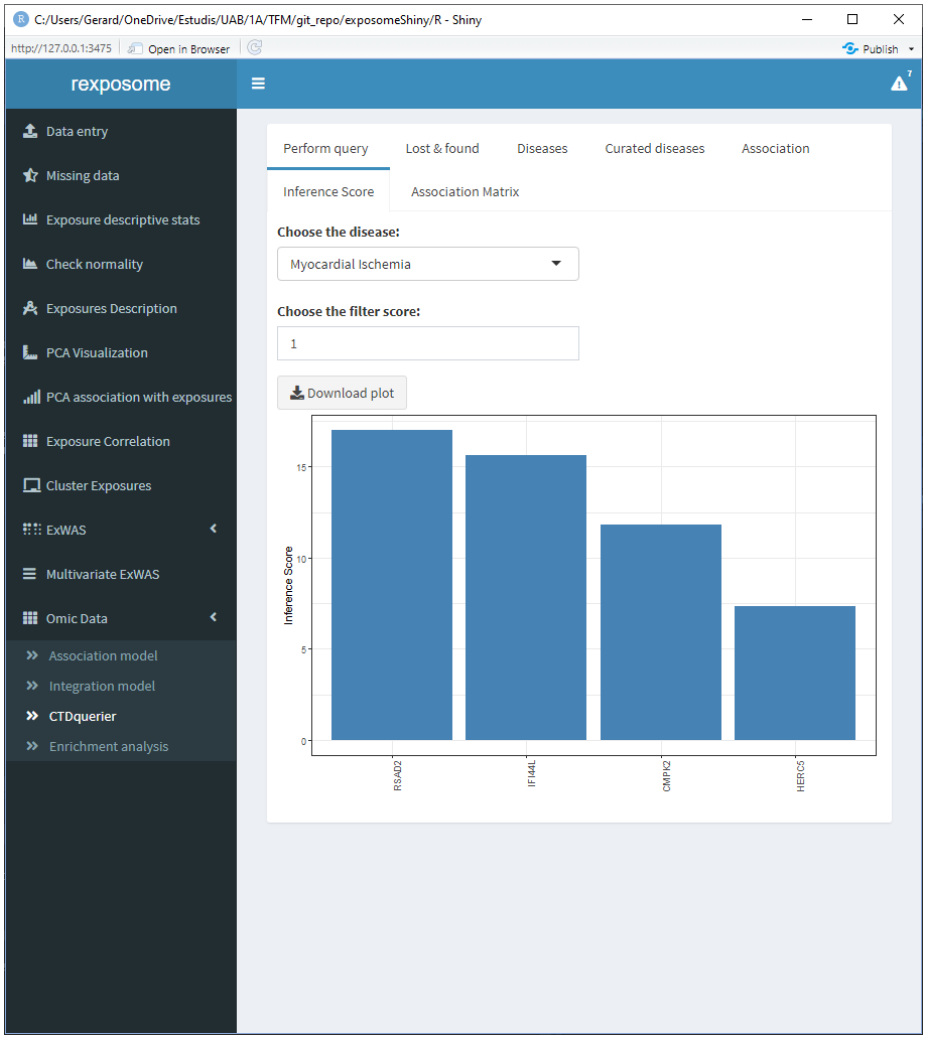
The curated diseases tab shows the table of associated diseases but only shows the ones with direct evidence.



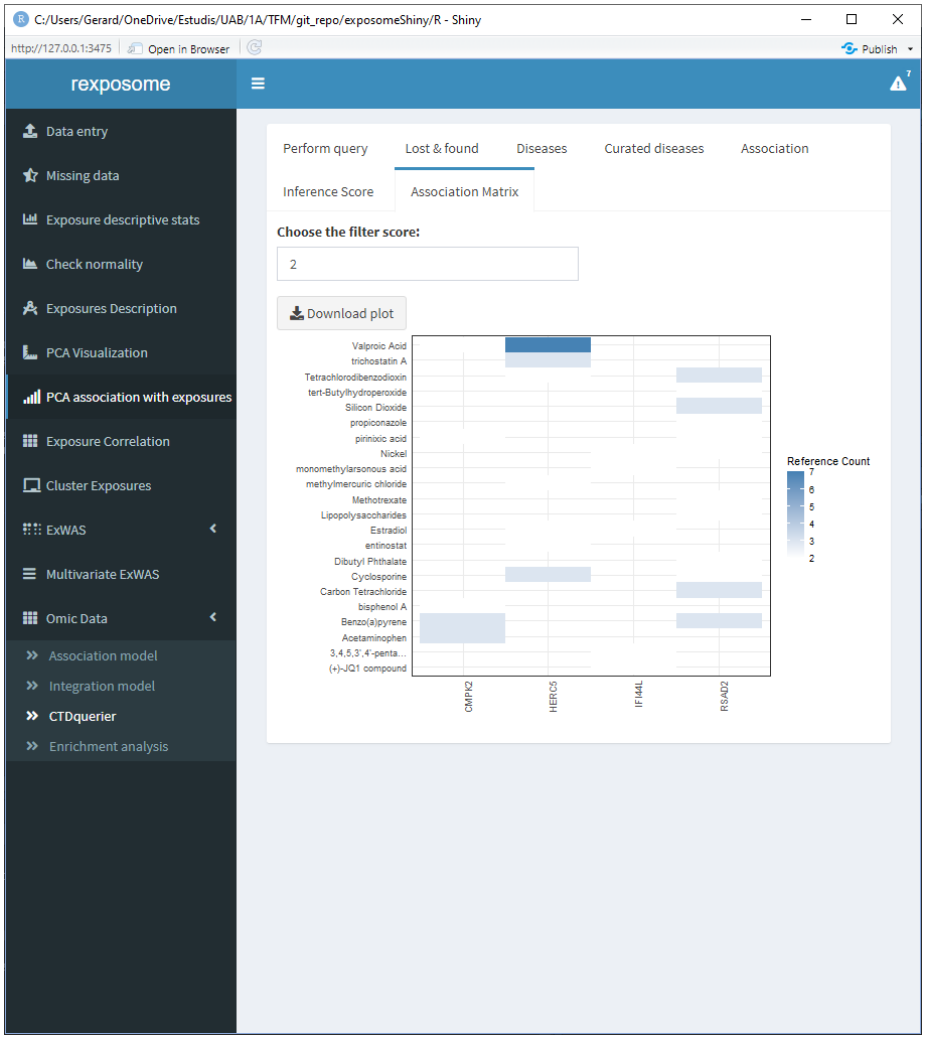
The association tab shows information about all the direct evidence associated diseases. Select the disease of interest to see the score and reference count of it.



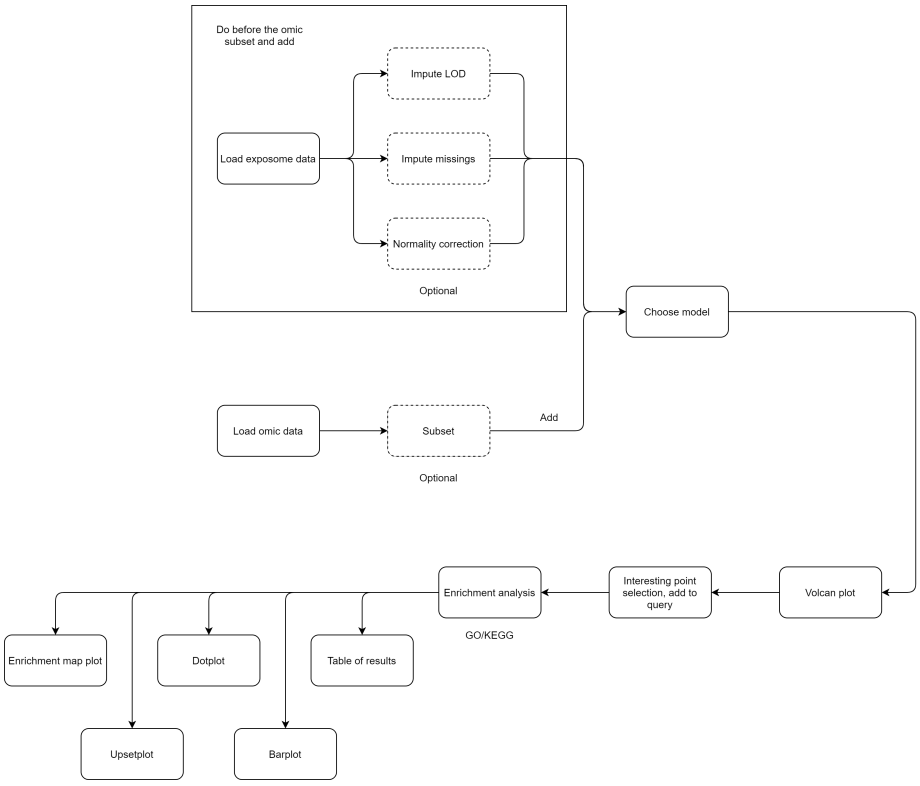
The inference score tab shows the inference score for each gene for a selected disease, the filter parameters puts out the genes with an inference score lower than the selected filter.



The association matrix tab shows a matrix of genes vs. chemicals with a heatmap representing the existing papers (references) providing evidence about the association between chemicals and genes.

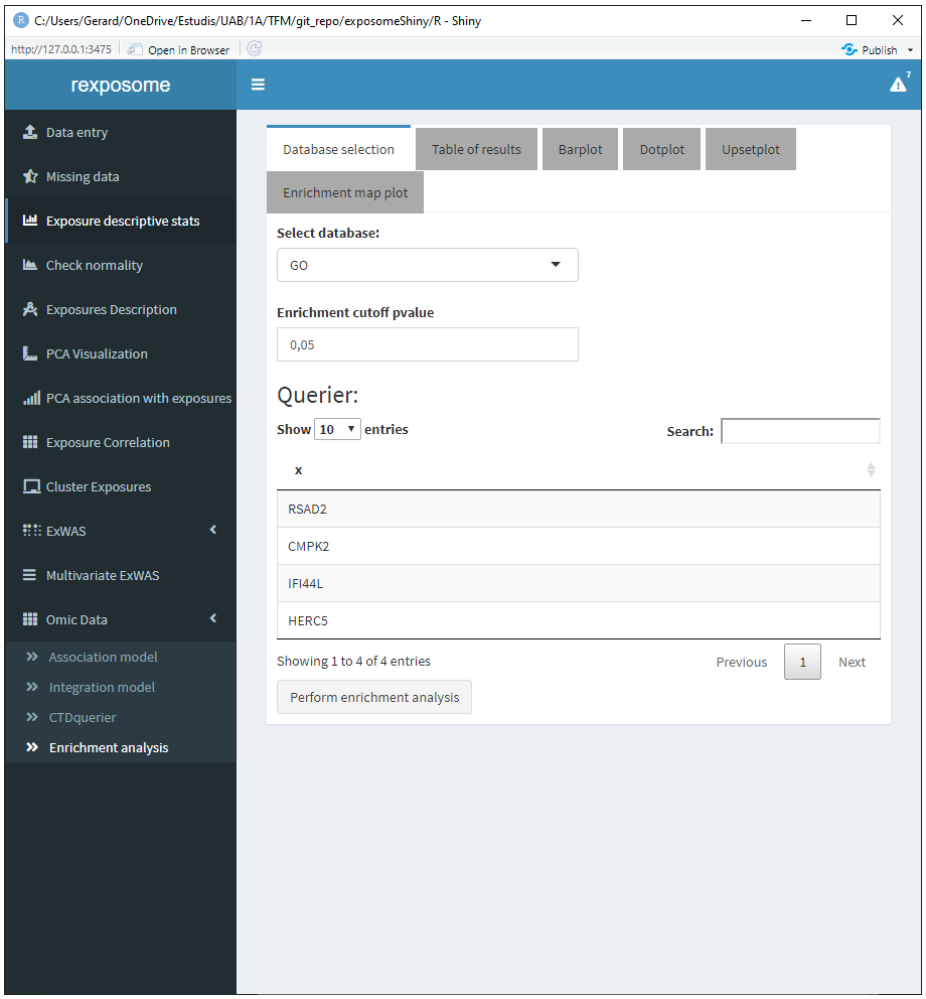

C:/Users/Gerard/OneDrive/Estudis/UAB/1A/TFM/git_repo/exposomeShiny/R - Shiny
http://127.0.0.1:3475
Open in Browser
Publish
rexposome
Data entry
Missing data
Exposure descriptive stats
Check normality
Exposures Description
PCA Visualization
PCA association with exposures
Exposure Correlation
Cluster Exposures
ExWAS
Multivariate ExWAS
Omic Data
Association model
Integration model
CTDquerier
Enrichment analysis
Perform query
Lost & found
Diseases
Curated diseases
Association
Inference Score
Association Matrix
Choose the filter score:
2
Download plot
Valproic Acid
trichostatin A
Tetrachlorodibenzodioxin
tert-Butylhydroperoxide
Silicon Dioxide
propiconazole
pirinixic acid
Nickel
monomethylarsonous acid
methylmercuric chloride
Methotrexate
Lipopolysaccharides
Estradiol
entinostat
Dibutyl Phthalate
Cyclosporine
Carbon Tetrachloride
bisphenol A
Benzo(a)pyrene
Acetaminophen
3,4,5,3',4'-penta...
(+)-JQ1 compound
CMPK2
HERC5
IFI44L
RSAD2
Reference Count
7
6
5
4
3
2

5.2.3 Enrichment analysis

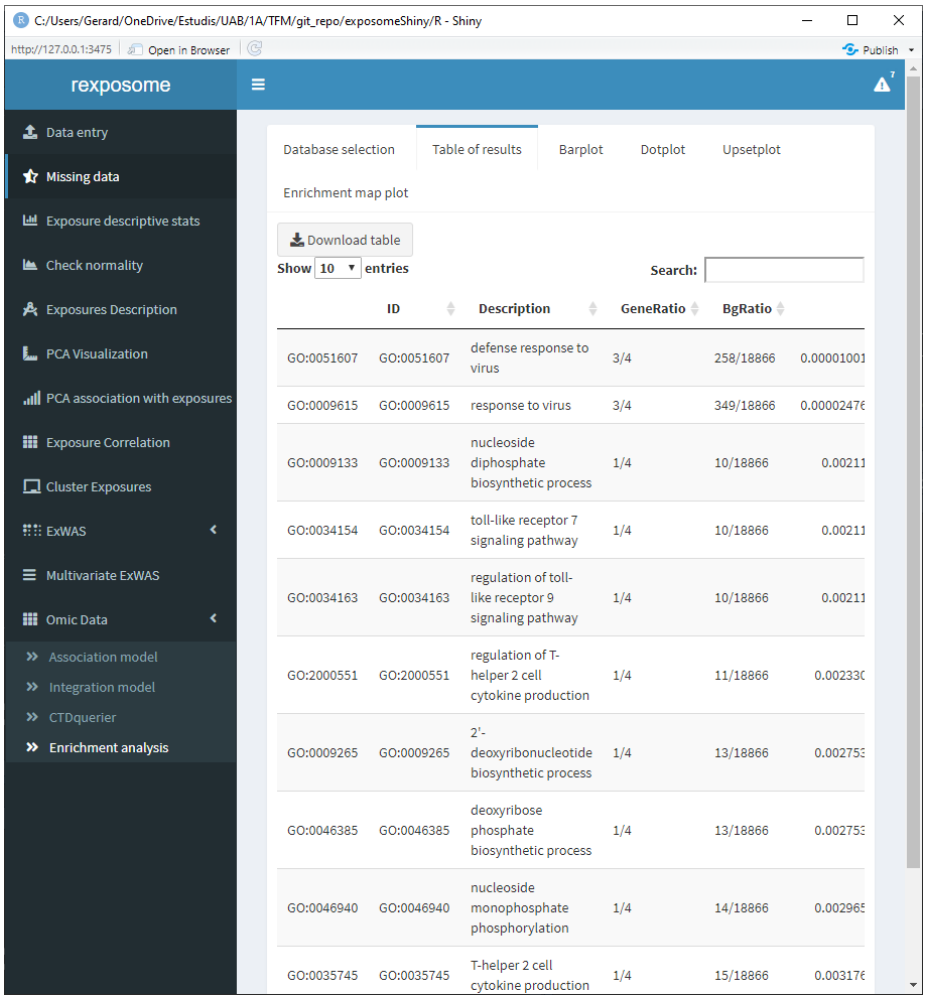


To perform an enrichment analysis, select the genes of interest by following the same procedure as with the CTDquerier.

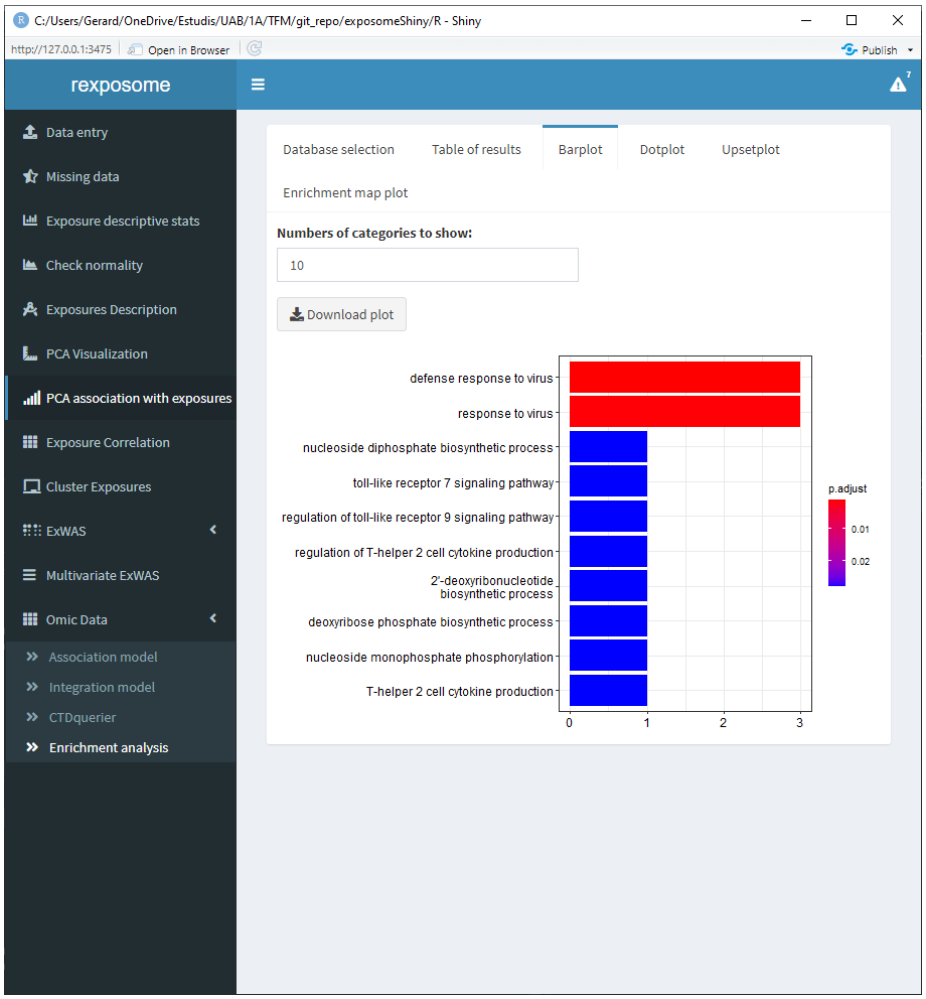
After selecting them, go to the “Enrichment analysis” tab, on it there is a selector to choose between GO and KEGG databases and a threshold selector for the cutoff pvalue of the enrichment analysis.

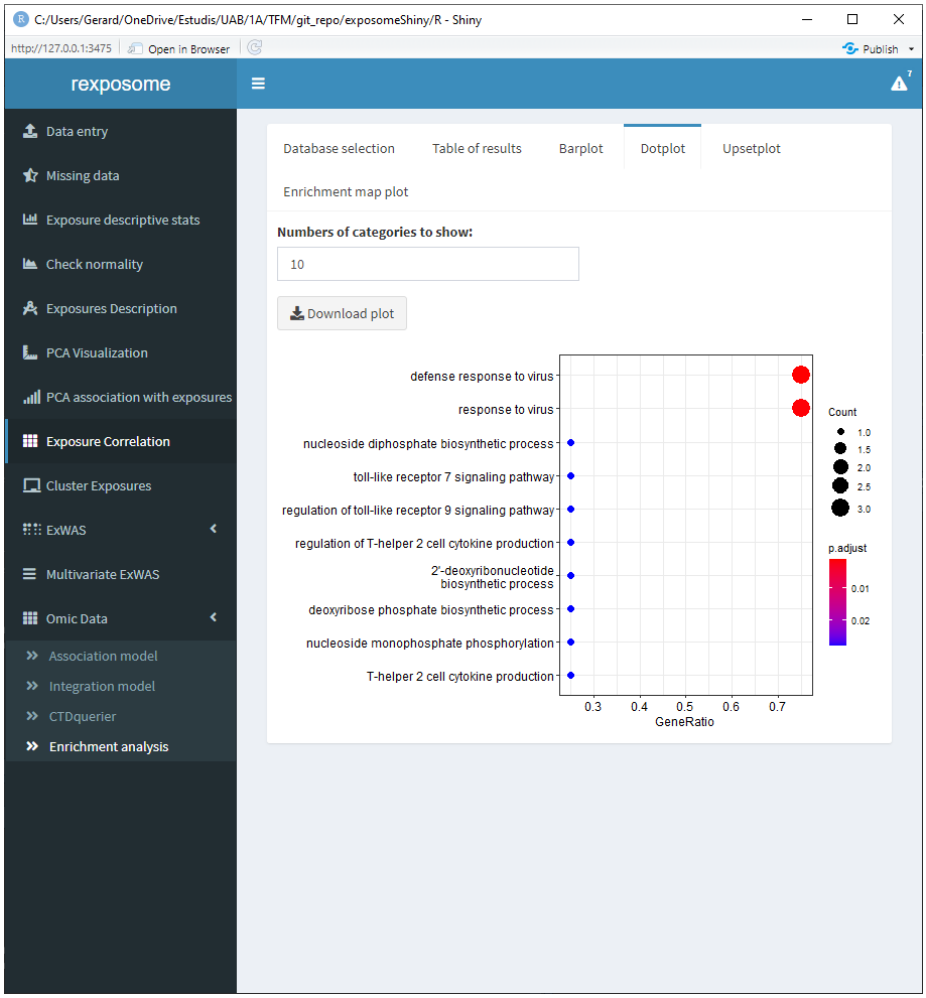


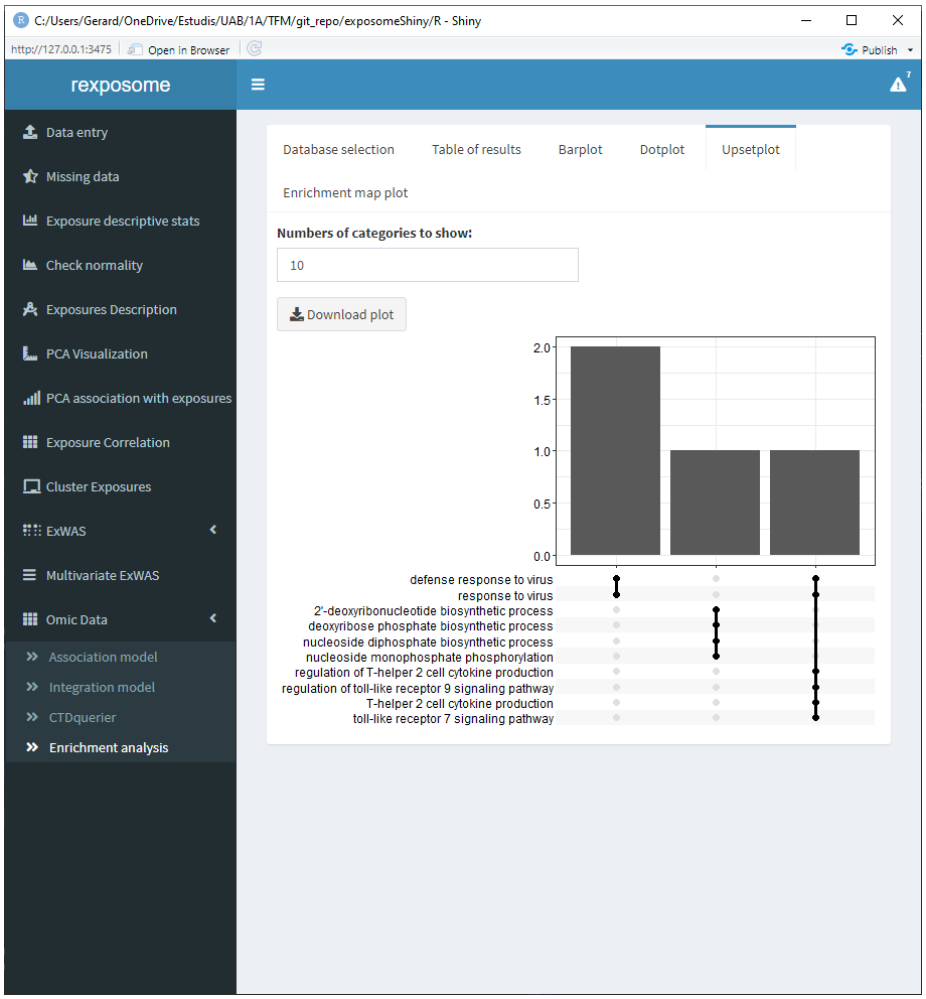
When the enrichment analysis is performed, there are multiple visualization tabs for the results, the first one is a table with the plain results, which can be downloaded.

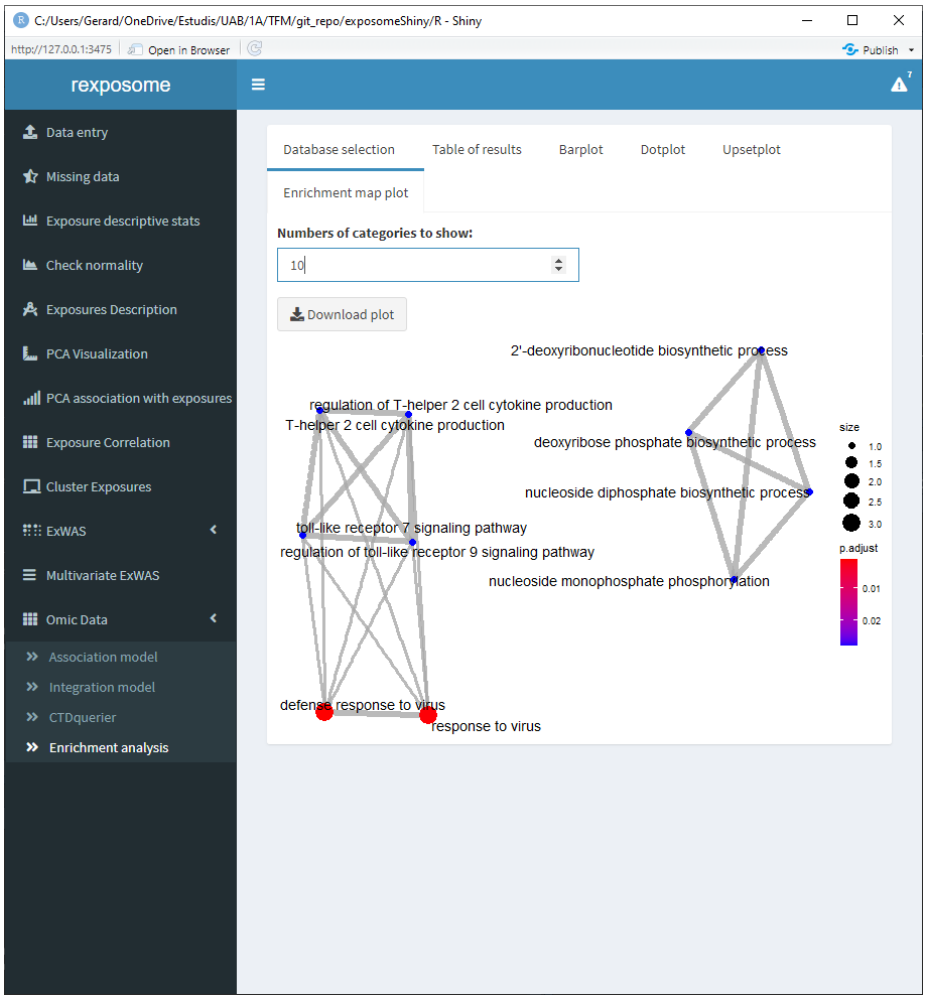


The following four tabs correspond to different visualization options for the results. All of them can be downloaded.


C:/Users/Gerard/OneDrive/Estudis/UAB/1A/TFM/git_repo/exposomeShiny/R - Shiny
http://127.0.0.1:3475
Open in Browser
Publish
rexposome
Data entry
Missing data
Exposure descriptive stats
Check normality
Exposures Description
PCA Visualization
PCA association with exposures
Exposure Correlation
Cluster Exposures
ExWAS
Multivariate ExWAS
Omic Data
Association model
Integration model
CTDquerier
Enrichment analysis
Database selection
Table of results
Barplot
Dotplot
Upsetplot
Enrichment map plot
Numbers of categories to show:
10
Download plot
defense response to virus
response to virus
nucleoside diphosphate biosynthetic process
toll-like receptor 7 signaling pathway
regulation of toll-like receptor 9 signaling pathway
regulation of T-helper 2 cell cytokine production
2'-deoxyribonucleotide biosynthetic process
deoxyribose phosphate biosynthetic process
nucleoside monophosphate phosphorylation
T-helper 2 cell cytokine production
0
1
2
3
p.adjust
0.01
0.02

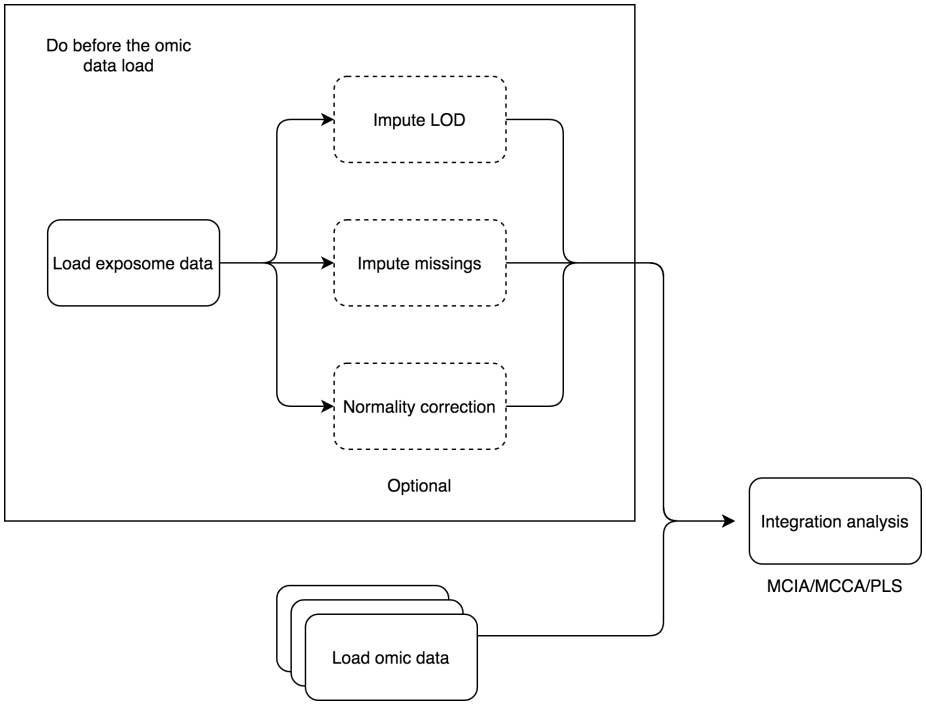

C:/Users/Gerard/OneDrive/Estudis/UAB/1A/TFM/git_repo/exposomeShiny/R - Shiny
http://127.0.0.1:3475
Open in Browser
Publish
rexposome
Data entry
Missing data
Exposure descriptive stats
Check normality
Exposures Description
PCA Visualization
PCA association with exposures
Exposure Correlation
Cluster Exposures
ExWAS
Multivariate ExWAS
Omic Data
Association model
Integration model
CTDquerier
Enrichment analysis
Database selection
Table of results
Barplot
Dotplot
Upsetplot
Enrichment map plot
Numbers of categories to show:
10
Download plot
defense response to virus
response to virus
nucleoside diphosphate biosynthetic process
toll-like receptor 7 signaling pathway
regulation of toll-like receptor 9 signaling pathway
regulation of T-helper 2 cell cytokine production
2'-deoxyribonucleotide biosynthetic process
deoxyribose phosphate biosynthetic process
nucleoside monophosphate phosphorylation
T-helper 2 cell cytokine production
0.3
0.4
0.5
0.6
0.7
GeneRatio
Count
1.0
1.5
2.0
2.5
3.0
p.adjust
0.01
0.02


C:/Users/Gerard/OneDrive/Estudis/UAB/1A/TFM/git_repo/exposomeShiny/R - Shiny
http://127.0.0.1:3475
Open in Browser
Publish
rexposome
Data entry
Missing data
Exposure descriptive stats
Check normality
Exposures Description
PCA Visualization
PCA association with exposures
Exposure Correlation
Cluster Exposures
ExWAS
Multivariate ExWAS
Omic Data
Association model
Integration model
CTDquerier
Enrichment analysis
Database selection
Table of results
Barplot
Dotplot
Upsetplot
Enrichment map plot
Numbers of categories to show:
10
Download plot
2.0
1.5
1.0
0.5
0.0
defense response to virus
response to virus
2'-deoxyribonucleotide biosynthetic process
deoxyribose phosphate biosynthetic process
nucleoside diphosphate biosynthetic process
nucleoside monophosphate phosphorylation
regulation of T-helper 2 cell cytokine production
regulation of toll-like receptor 9 signaling pathway
T-helper 2 cell cytokine production
toll-like receptor 7 signaling pathway



5.3 Exposome-Omic integration (e.g. crossomics)

The datasets used in this section are `exposures_2.csv`, `description_2.csv`, `phenotypes_2.csv`, `brge_gexp.rda` and `brge_prot.rda` which are available here.



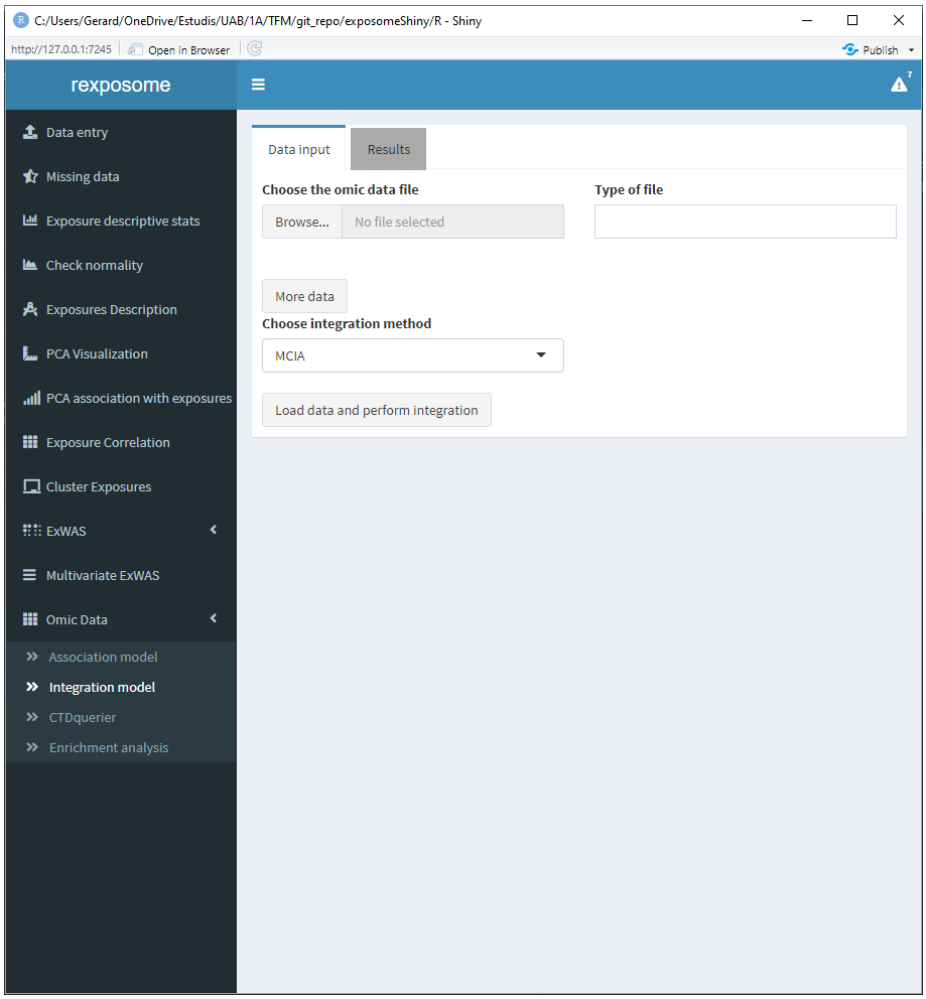
The relation between exposures and omic-features can be studied from another perspective, different from the association analyses. The integration analysis can be done, using multi canonical correlation analysis, multiple co-inertia analysis or partial least squares (this method only is supported with one omics dataset). The first method is implemented in R package `PMA` (CRAN) and the second in `omicade4` R package (Bioconductor). The PLS uses the `mixOmics` R package (Bioconductor).

Before conducting an integration model, an exposome dataset has to be loaded into the Shiny application.

For the MCIA and MCCA options, the exposome dataset can't contain missings, so it needs to be imputed beforehand. For the PLS there can be missings. Different visualizations are provided for each method, as the used R packages provide different visualization tools, all three methods generate different plots. The Rdata object that contains the results can be downloaded to be explorated by the user on a separate R session.

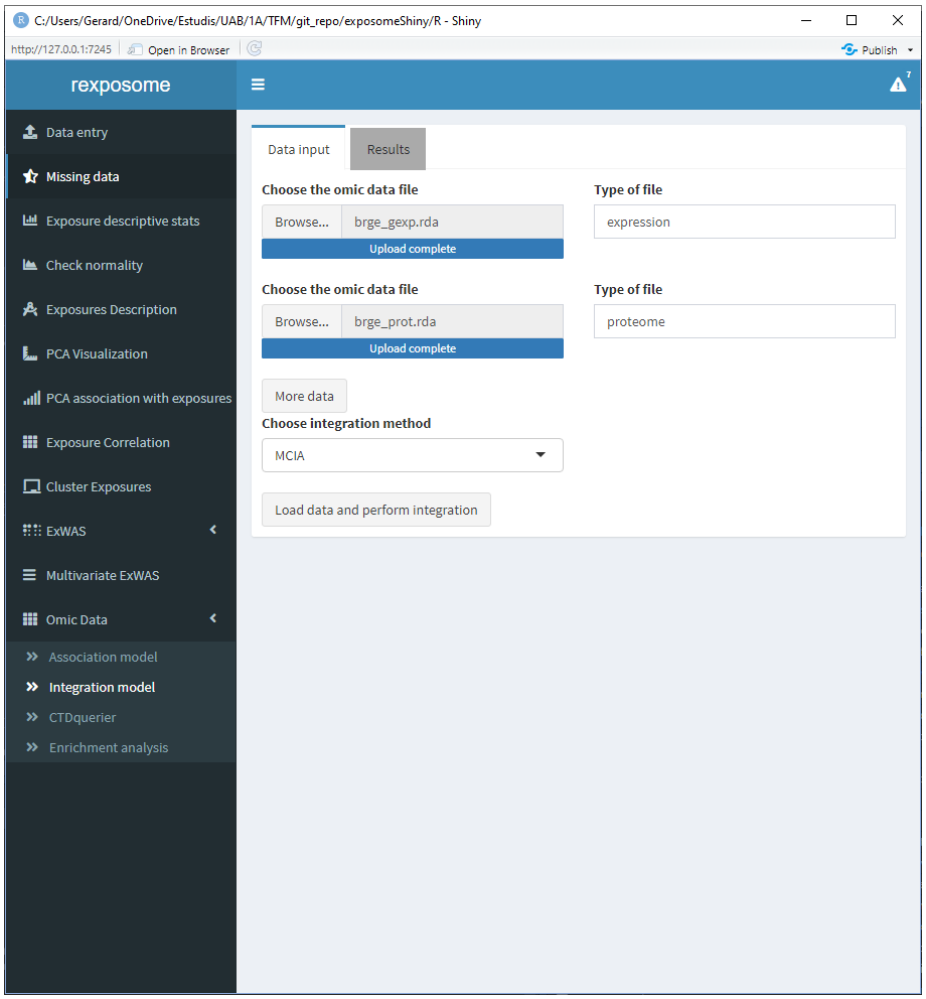
The differences between association and crossomics are that the first method test association between two complete data-sets, by removing the samples having missing values in any of the involved data-sets, and the second try to find latent relationships between two or more sets.

The initial page of the integration is the following.



The user has a field to select a omics file and a 'Type of file' field. This text field is to input whether the selected data is expression/proteome/... this field will be used for informative purposes when displaying the results.

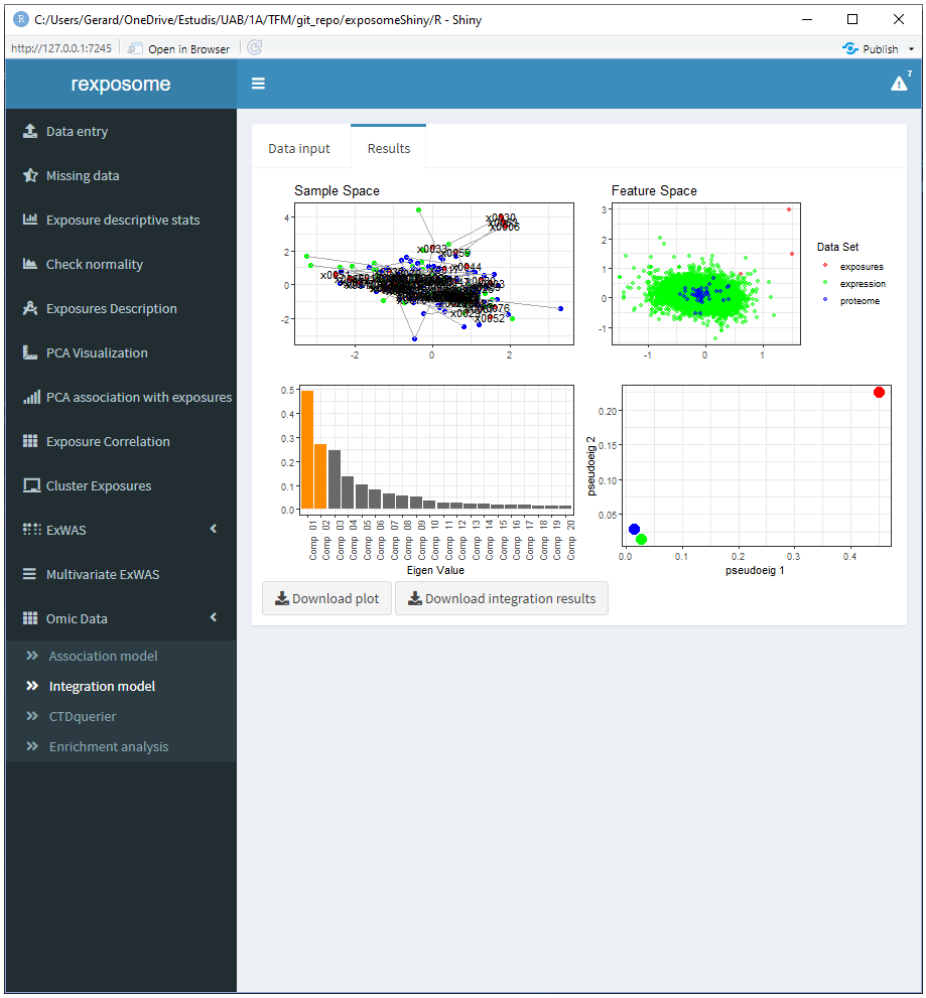
If more than one omics dataset is going to be used on the integration, click on 'More data' to get additional input fields, remember that only MCIA and MCCA accept more than one omics dataset.



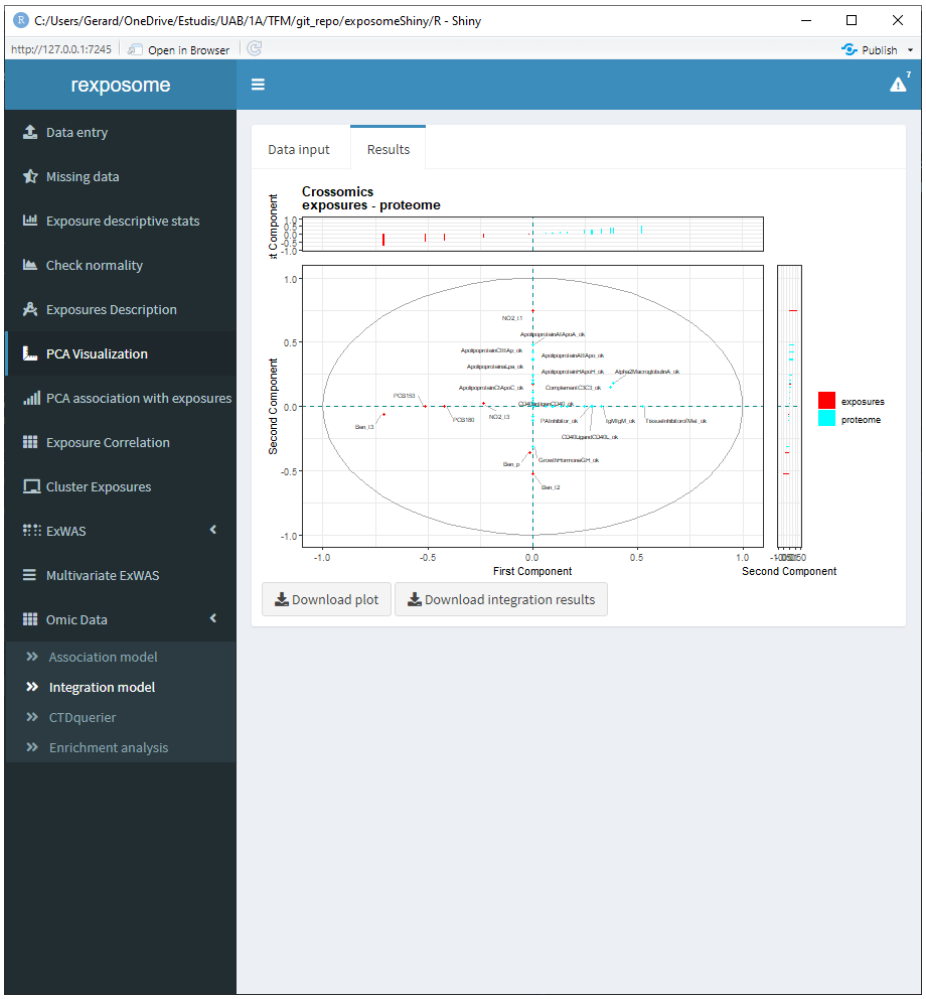
Once the data entry is completed, click on 'Load data and perform integration'. Go to the results page to see the plots of the integration.

The following screenshots are the results of an integration analysis using the exposome dataset (`exposures_2.csv`, `description_2.csv`, `phenotypes_2.csv`) with the missings imputed (using the 'Missing data' tab of the Shiny) expression omics (`brge_gexp.rda`) and proteome omics (`brge_prot.rda`). All this mentioned datasets can be found here.

5.3.1 Integration using MCIA



5.3.2 Integration using MCCA (using only the proteome omics)

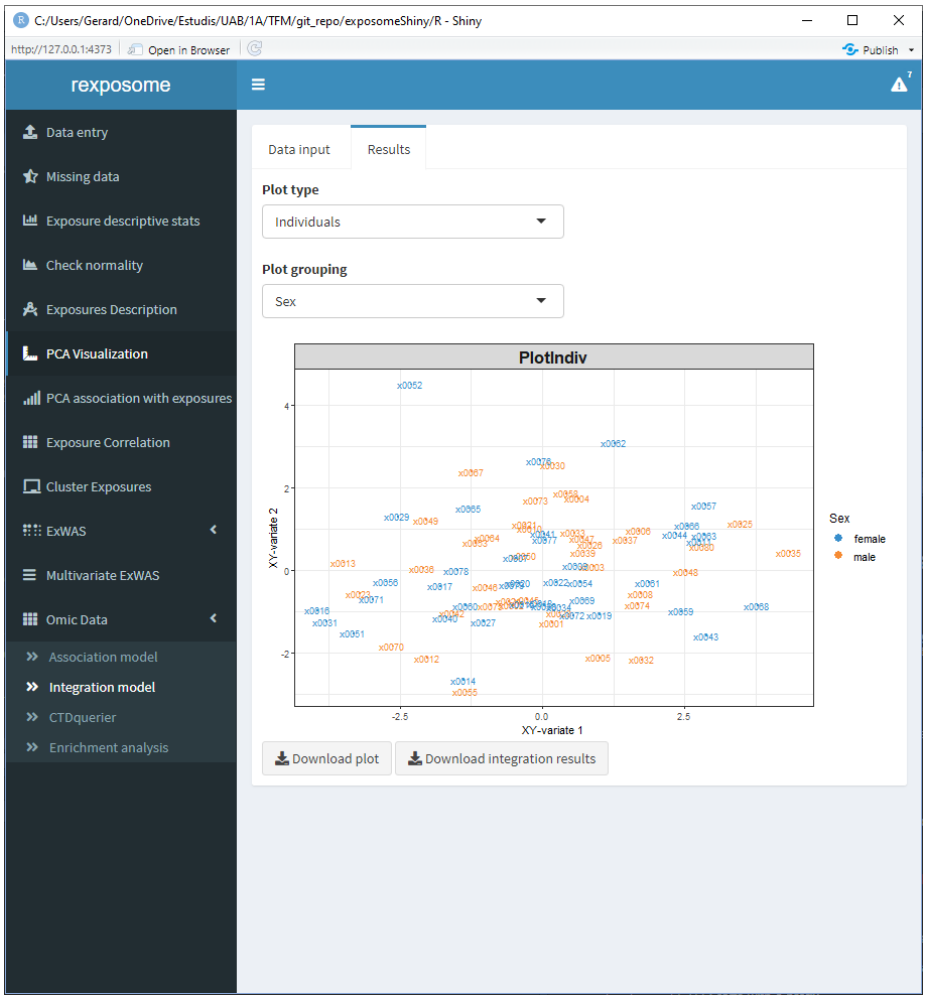


5.3.3 Integration using PLS (using only the proteome omics)

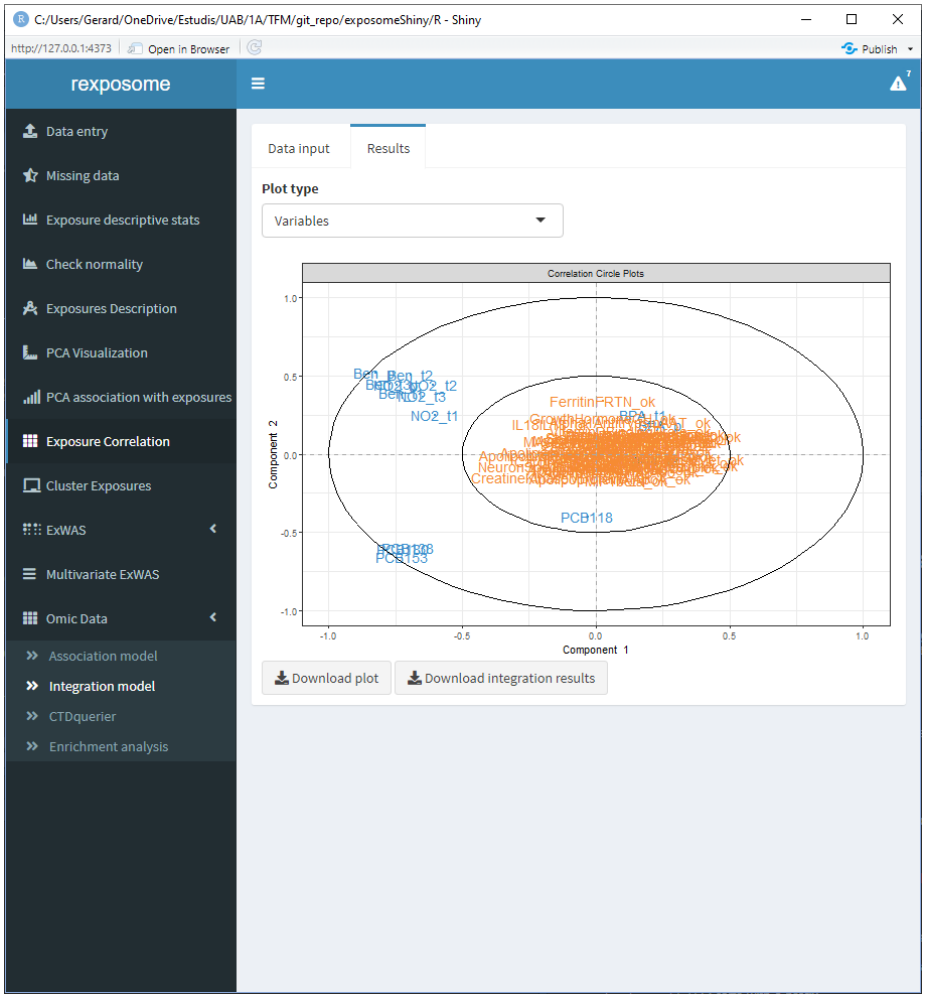
There are multiple visualization options for the PLS integration.

5.3.3.1 The exposures space

Can be grouped using the categorical phenotypes of the exposome data

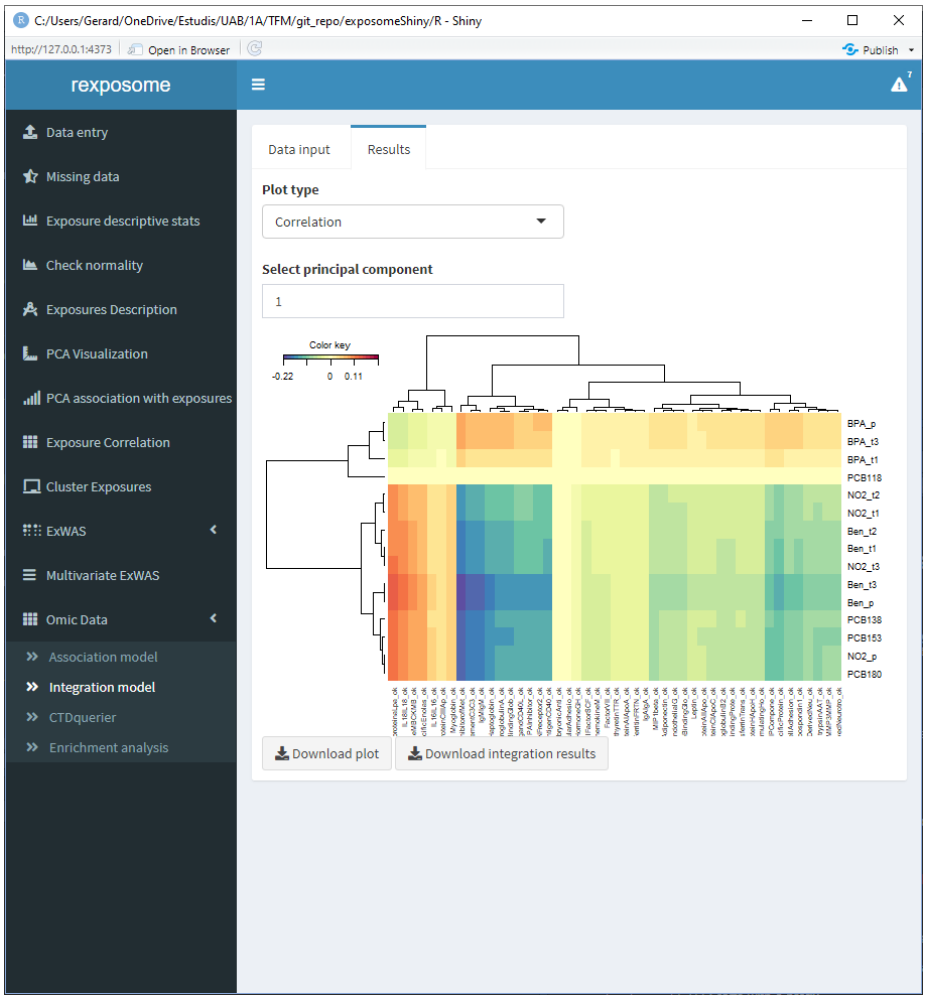

C:/Users/Gerard/OneDrive/Estudis/UAB/1A/TFM/git_repo/exposomeShiny/R - Shiny
http://127.0.0.1:4373
Open in Browser
Publish
rexposome
Data entry
Missing data
Exposure descriptive stats
Check normality
Exposures Description
PCA Visualization
PCA association with exposures
Exposure Correlation
Cluster Exposures
ExWAS
Multivariate ExWAS
Omic Data
Association model
Integration model
CTDquerier
Enrichment analysis
Data input
Results
Plot type
Individuals
Plot grouping
Sex
PlotIndiv
XY-variate 2
XY-variate 1
Sex
female
male
Download plot
Download integration results

5.3.3.2 The variables space



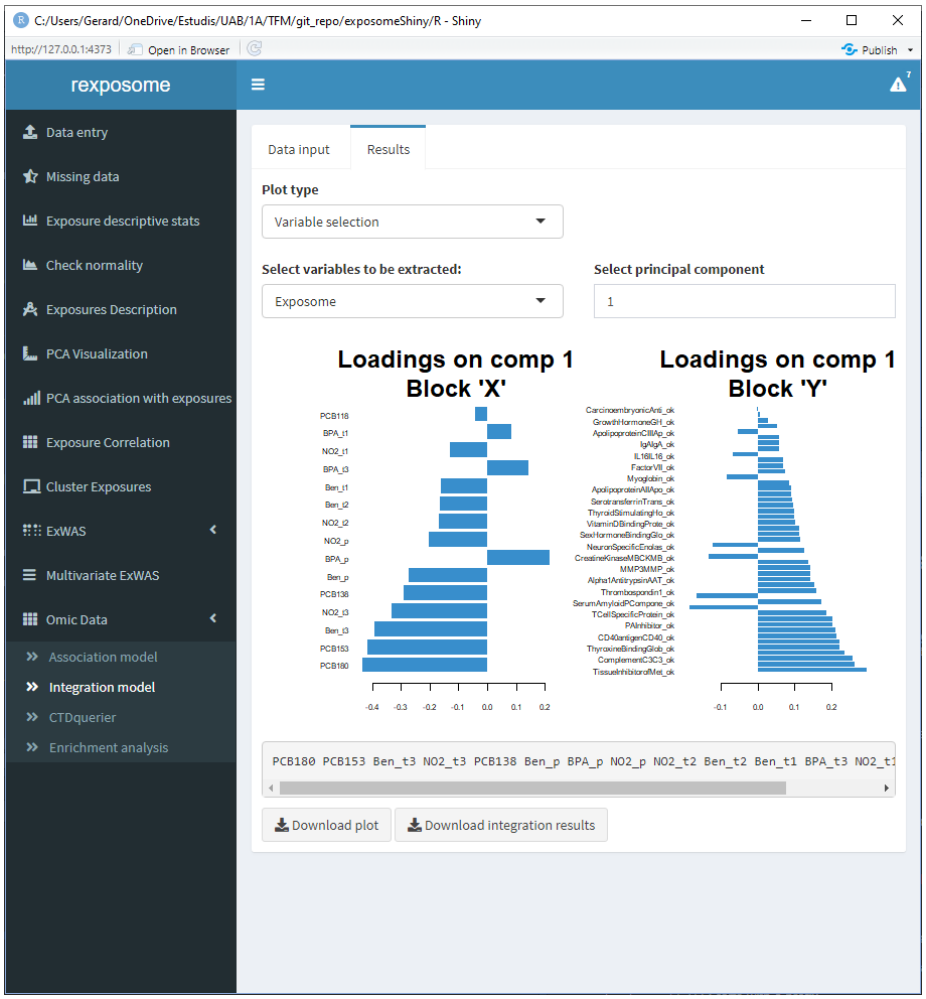
5.3.3.3 A correlation heatmap

Available for the first three principal components



5.3.3.4 Loading plots (coefficients of each variable)

Available for the first three principal components

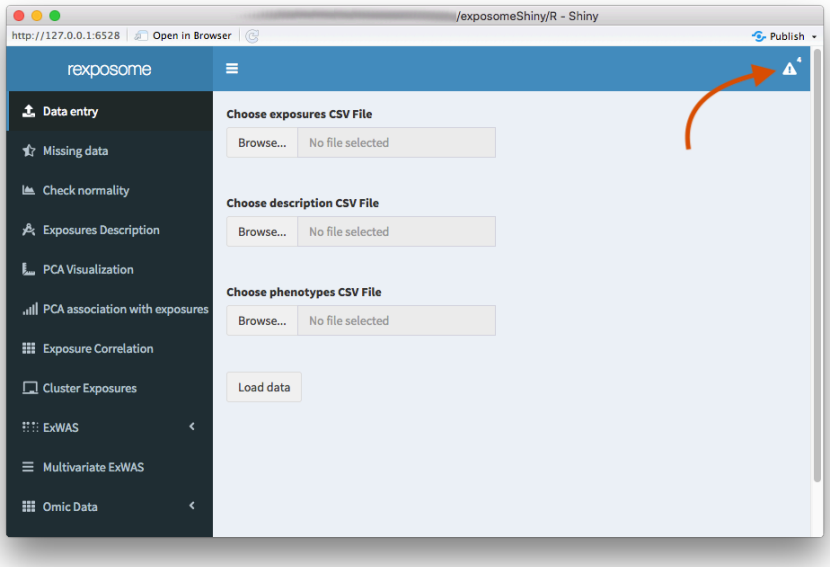


All plots can be downloaded individually, also, the Rdata file that contains the results of the integration can be downloaded.

Chapter 6

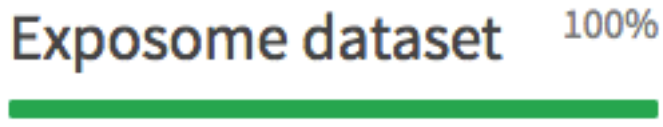
General application functionalities

This whole Shiny application serves the purpose to perform a various number of exposome and omic analysis with an input set of data. To perform this analysis some operations can performed on the inputed dataset prior to the analysis and so in order to follow track of what exactly has been done or what is loaded on the current session, there's implemented some sort of state tracker inside the Shiny application. In order to access it, press the icon on the top right of the application



When clicking it a dropdown menu appears, inside there are seven different notifications:

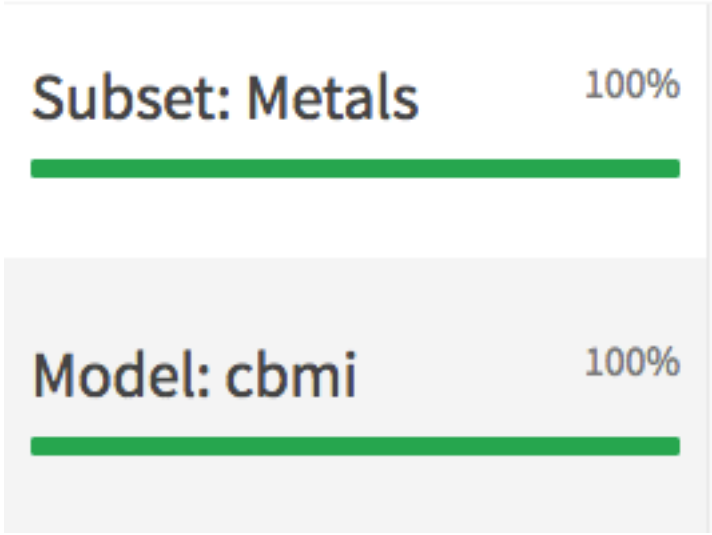
- Exposome dataset: Turns to 100% when an exposome dataset is loaded in the environment. Here's a graphical example.



- LOD imputed: Turns to 100% if the exposome dataset is LOD imputed.
- Missing imputed: Turns to 100% if the exposome dataset has the missings imputed.
- Normality corrected: Turns to 100% if the exposome dataset is normality

corredted.

- Omics dataset: Turns to 100% when an omics dataset is loaded in the environment.
- Subset: Turns to 100% (and displays the subset family(ies)) when the exposome dataset is subseted.
- Model: Turns to 100% (and displays the association variable(s)) when a model is performed for an omics association analysis. Here's an example of a subset and model information.



Chapter 7

Methods

7.1 Missing data imputation

LOD missings are discovered through a encoding provided by the user, there is no method implemented to separate missing values between missing at random at LOD, meaning that all NA values are considered missing at random.

7.1.1 Limit of detection (LOD) missing

LOD missings can be imputed using two methodologies:

- LOD value / sqrt(2) : Use a LOD value provided by the user (one value per exposures) divided by the square root of two. Richardson and Ciampi (2003)
- QRILC: a quantile regression approach for the imputation of left-censored missing data Lazar (2015).

7.1.2 Missing at random

Multiple imputation chained equations (MICE) is used to impute missing at random data. The *mice* package is used to do so. A brief explanation on the algorithm:

1. Imputation of the variable (exposure) xn with the mean of all it's values.
2. Perform 1 for all the variables.
3. Set the mean imputed values from one variable back to missing.
4. Perform a regression model and fill those missings.
5. Repeat 3 and 4 for all the variables.
6. Repeat 3, 4 and 5 until the imputed values obtained are stabilized.

7.2 Normality

7.2.1 Normality testing

To test the normality of a variable, a Shapiro-Wilks test is used. The Shapiro-Wilks test, tests the null hypothesis of a sample (variable of the dataset) is normally distributed, to perform the test it calculates the W statistic.

$$W = \frac{\left(\Sigma_{i=1}^{n} a_i x_{(i)}\right)^2}{\Sigma_{i=1}^{n}(x_i - \overline{x})^2}$$

To perform this test exposome uses the *shapiro.test* function from the *base* package of R.

7.2.2 Normalization

A user selected function can be applied to exposures (selected by the user) to normalize them. The available functions are: *log*, *sqrt* and *^1/3*.

7.3 Principal component analysis (PCA)

Rexposome contains two PCA methodologies

- Regular PCA Jolliffe and Cadima (2016) (only numerical exposures)
- FAMD Chavent et al. (2014) (numerical and categorical)

exposomeShiny uses regular PCA from the *FactoMineR* package. A toggle to select between the two may be added in future releases.

7.4 Exposures correlation

The correlation method takes into account the nature of each pair of exposures: continuous vs. continuous uses cor function from R *base*, categorical vs. categorical uses cramerV function from *lsr* R package and categorical vs. continuous exposures correlation is calculated as the square root of the adjusted r-square obtained from fitting a lineal model with the categorical exposures as dependent variable and the continuous exposure as independent variable.

7.5 Exposures clustering

Clustering analysis on samples can be performed to cluster individuals having similar exposure profiles. This is done using hierarchical clustering using the function *hclust* from the *stats* R package. The results this analysis yields are the exposure profiles of a selected number of groups.

7.6 Exposome Association Analysis

7.6.1 Single Association Analysis

Exposome-Wide Association Study (ExWAS) is equivalent to a Genome-Wide Association Study (GWAS) in genomics or to Epigenetic-Wide Association Study (EWAS) in epigenomics. The ExWAS was first described by Patel et al. Patel et al. (2010) . ExWAS are based on generalized linear models using any formula describing the model that should be adjusted for (following standard formula options in R). That is, continuous or factor variables can be incorporated in the design, as well as interaction or splines using standard R functions and formulas. Multiple comparisons in the ExWAS analysis is addressed by computing the number of effective (Neff) tests as described by Li and Ju Li and Ji (2005) . The method estimates Neff by using the exposure correlation matrix that is corrected when it is not positive definite by using *nearPD* R function. The significant threshold is computed as 1-(1-0.05)Meff. This threshold is added to the Manhattan plots. When using imputed data, analysis is done for each imputed set and P-Values are pooled to obtain a global association score.

7.6.2 Stratified Single Association Analysis

The stratified analysis option for the ExWAS corresponds to applying the same method as regular ExWAS to subsetted datasets. As example, a stratified analysis with the `sex` variable stratified corresponds to performing two ExWAS, one to the `male` and one for the `female` group.

7.6.3 Variable selection ExWAS

There are some authors that proposed to perform association analysis in a multivariate fashion, just to take into account the correlation across exposures Agier et al. (2016) . A Lasso regression is implemented using Elastic-Net regularized generalized linear models implemented in *glmnet* R package.

7.7 Exposome-Omic Association Analysis

Perform association analyses between exposures and omic data bt fitting linear models as described in the *limma* R package Ritchie et al. (2015) . The pipeline implemented in association allows performing surrogate variable analysis in order to correct for unwanted variability. This adjustment is provided by *SVA* R package Leek et al. (2020) .

7.8 Integration analysis

There are three different methodologies to perform the integration analysis:

- Multiset canonical correlation analysis (MCCA). Implemented using the `MultiCCA` function of `PMA` R package Witten et al. (2020) .
- Multiple co-inertia analysis (MCIA). Implemented using the `mcia` function of `omicade4` R package Meng et al. (2013) , Min and Long (2020) .
- Partial least squares (PLS). Implemented using the `plsr` function of `pls` R package Mevik and Wehrens (2015) .

7.9 Enrichment analysis

Functional profiles of selected genes are obtained using the Bioconductor package `clusterProfiler` Yu et al. (2012) . The available enrichment databases are GO and KEGG.

Chapter 8

References

Bibliography

Agier, L., Portengen, L., Chadeau-Hyam, M., Basagaña, X., Giorgis-Allemand, L., Siroux, V., Robinson, O., Vlaanderen, J., González, J. R., Nieuwenhuijsen, M. J., et al. (2016). A systematic comparison of linear regression–based statistical methods to assess exposome-health associations. *Environmental health perspectives*, 124(12):1848–1856.

Chavent, M., Kuentz-Simonet, V., Labenne, A., and Saracco, J. (2014). Multivariate analysis of mixed data: The r package pcamixdata. *arXiv preprint arXiv:1411.4911*.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Lazar, C. (2015). CRAN - Package imputeLCMD.

Leek, J. T., Johnson, W. E., Parker, H. S., Fertig, E. J., Jaffe, A. E., Zhang, Y., Storey, J. D., and Torres, L. C. (2020). *sva: Surrogate Variable Analysis*. R package version 3.38.0.

Li, J. and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95(3):221–227.

Meng, C., Kuster, B., Culhane, A., and Gholami, A. M. (2013). A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*.

Mevik, B.-H. and Wehrens, R. (2015). Introduction to the pls package. *Help Section of The "Pls" Package of R Studio Software; R Foundation for Statistical Computing: Vienna, Austria*, pages 1–23.

Min, E. J. and Long, Q. (2020). Sparse multiple co-inertia analysis with application to integrative analysis of multi-omics data. *BMC bioinformatics*, 21:1–12.

Patel, C. J., Bhattacharya, J., and Butte, A. J. (2010). An environment-wide association study (ewas) on type 2 diabetes mellitus. *PloS one*, 5(5):e10746.

Richardson, D. B. and Ciampi, A. (2003). Effects of exposure measurement

error when an exposure variable is constrained by a lower limit. *American Journal of Epidemiology*, 157(4):355–363.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47.

Witten, D., Tibshirani, R., Gross, S., Narasimhan, B., and Witten, M. D. (2020). Package 'pma'. *Genetics and Molecular Biology*, 8(1):28.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287. PMID: 22455463.