# exposomeShiny User's Guide

Escribà Montagut, Xavier; González, Juan R.

2021-02-16

# Contents


**1 Overview** **5**

**2 Setup** **7**
2.1 R, RStudio and Packages Versions . . . . . 7
2.2 Downloading the source files, installing the libraries and running the application . . . . . 8
2.3 Pulling the official Docker image from DockerHub . . . . . 10

**3 Data sets** **13**
3.1 Exposome dataset . . . . . 13
3.2 Plain datasets . . . . . 15
3.3 Omics dataset . . . . . 16

**4 Bioconductor packages** **17**
4.1 rexposome . . . . . 17
4.2 omicRexposome . . . . . 17
4.3 CTDquerier . . . . . 17

**5 Analysis flowcharts** **19**
5.1 Exposome analysis . . . . . 19
5.2 Exposome-Omic analysis . . . . . 47

**6 General application functionalities** **65**

**7 Methods** **69**
7.1 Missing data imputation . . . . . 69
7.2 Normality . . . . . 70
7.3 Principal component analysis (PCA) . . . . . 70
7.4 Exposures correlation . . . . . 70
7.5 Exposures clustering . . . . . 70
7.6 Exposome Association Analysis . . . . . 71
7.7 Exposome-Omic Association Analysis . . . . . 71

**8 References** **73**

# 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
- Post-omic data analysis: CTD database

To do so, exposomeShiny relies on previously existent Bioconductor packages (rexposome, omicRexposome and CTDquerier), 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 |

There are two different ways of seting 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.

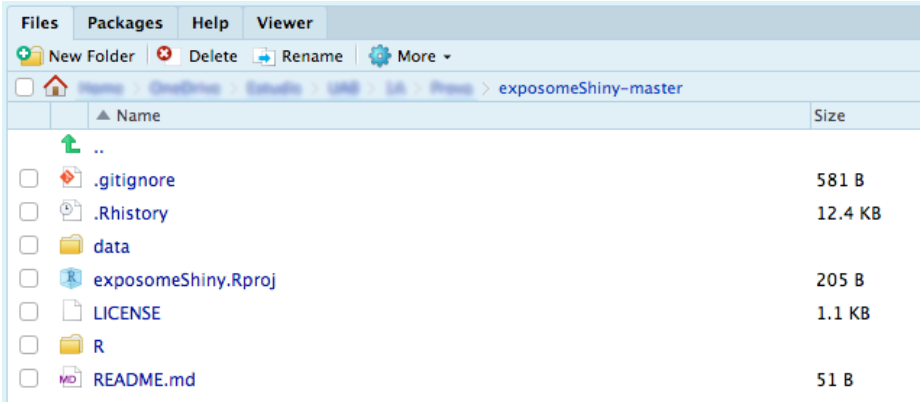
```
  # Set working directory
setwd(dir = "/some/path/")

  # 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.

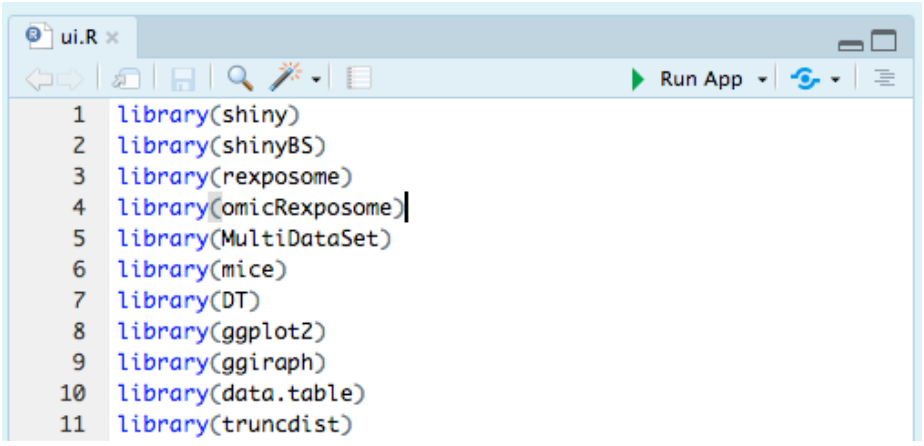



Once the project is loaded, the file found on the source folder called `installer.R` has to be sourced and run. This will install the newest versions of the packages required by Exposome Shiny on this R session. To do so, run the following code on the RStudio console.

```
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

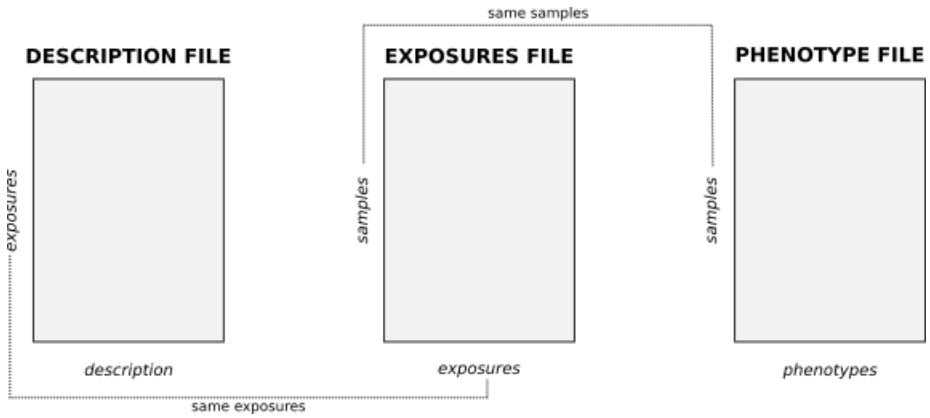
# Data sets

## 3.1 Exposome dataset

The exposome is composed of three different files (in `*.csv` 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.

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:

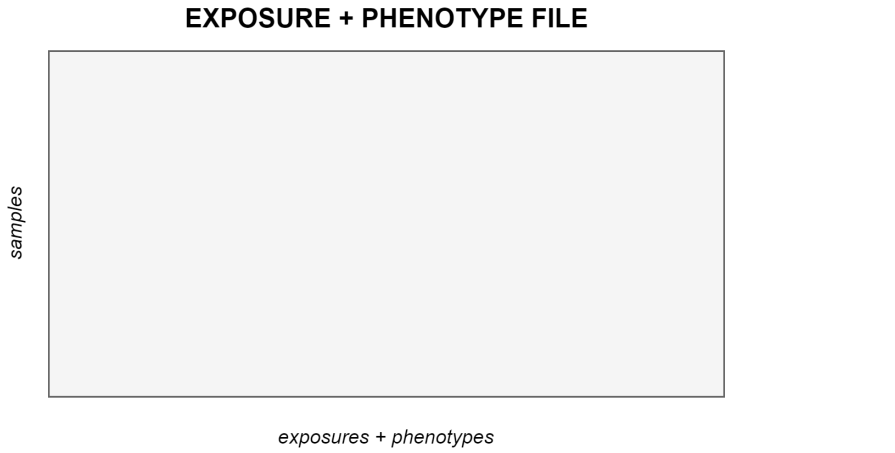
```
id    asthma   BMI      sex  age  ...
sub01 control  23.2539  boy  4    ...
sub02 asthma   24.4498  girl 5    ...
```

```
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

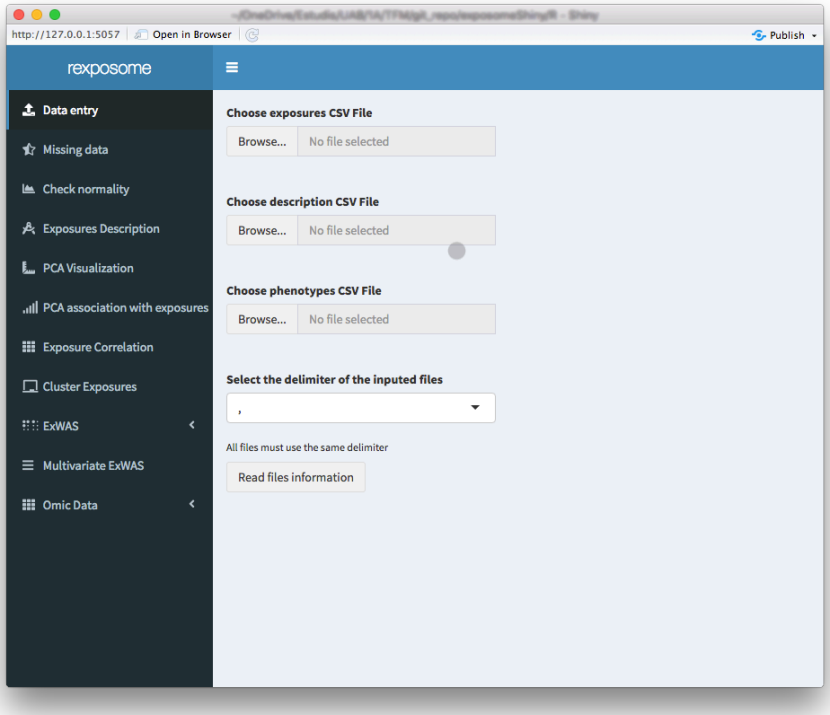
## 5.1 Exposome analysis

As any user would need to do using the Bioconductor packages (rexposome, omicRexposome and CTDquerier) when performing an analysis using an `R` script, there is some kind of flow (or pipeline) to follow in order to get to the results, this is also true on exposomeShiny, even though it's a seamless and codeless integration of the packages there's still some need for a flowchart to get the desired results. All the required flowcharts will be detailed with a box flowchart as well as screenshots of exposomeShiny in order to provide extra guidance if needed.

### 5.1.1 Data entry

#### 5.1.1.1 Exposome data

Input the exposures, description and phenotypes files and load them into the application. This files have to be provided as `csv` or `txt`, there is a selector on the UI to select the type of separation used, it can be either commas, semicolons or tabs/spaces. Excel files or R objects are not supported as inputs.

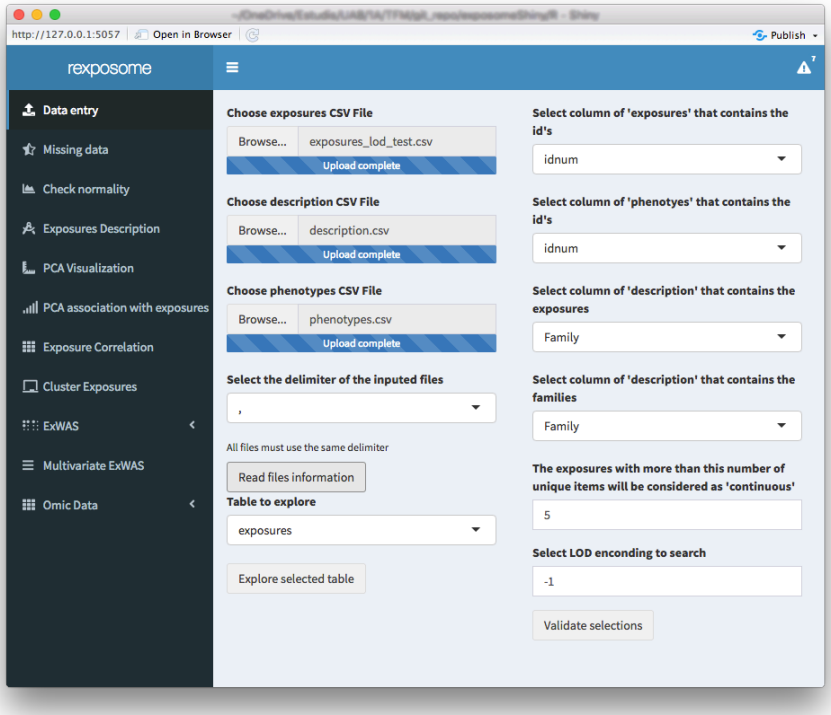



Once the tables are read two new main elements will appear, 1) the option to explore the inputted table, using a selector of the table and a button to trigger the visualization; and 2) six input fields will appear on the right side of the file browsers, they are to select the following parameters:

- 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 id's
- Column name in the *phenotypes* file that contains the id's
- The treshold of to select between continuous or factor exposures. More than this number of unique items 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`).

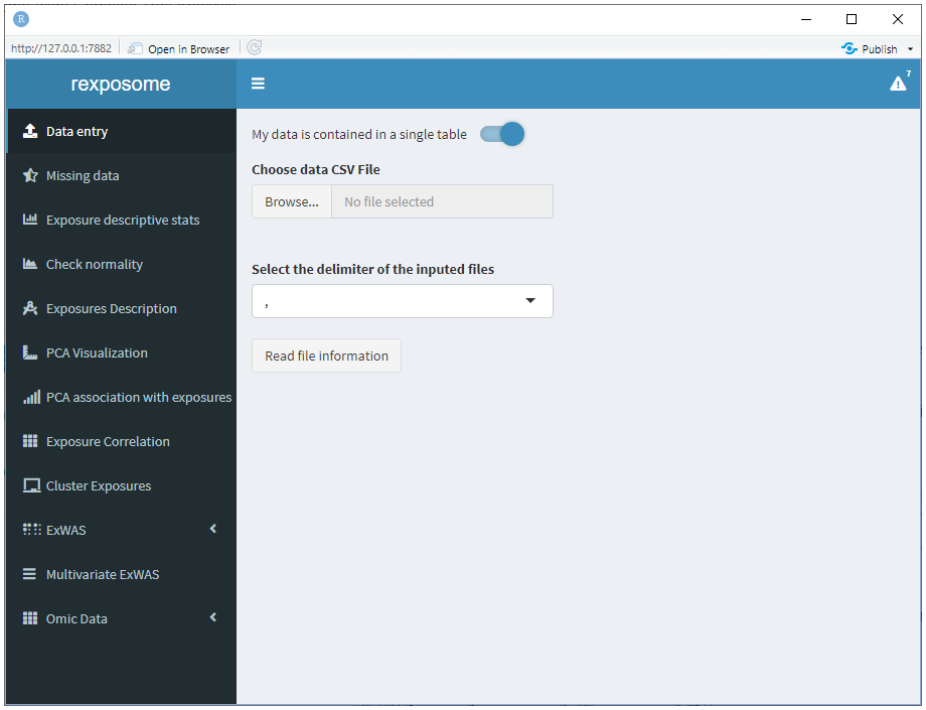
This is illustrated on the following figure.



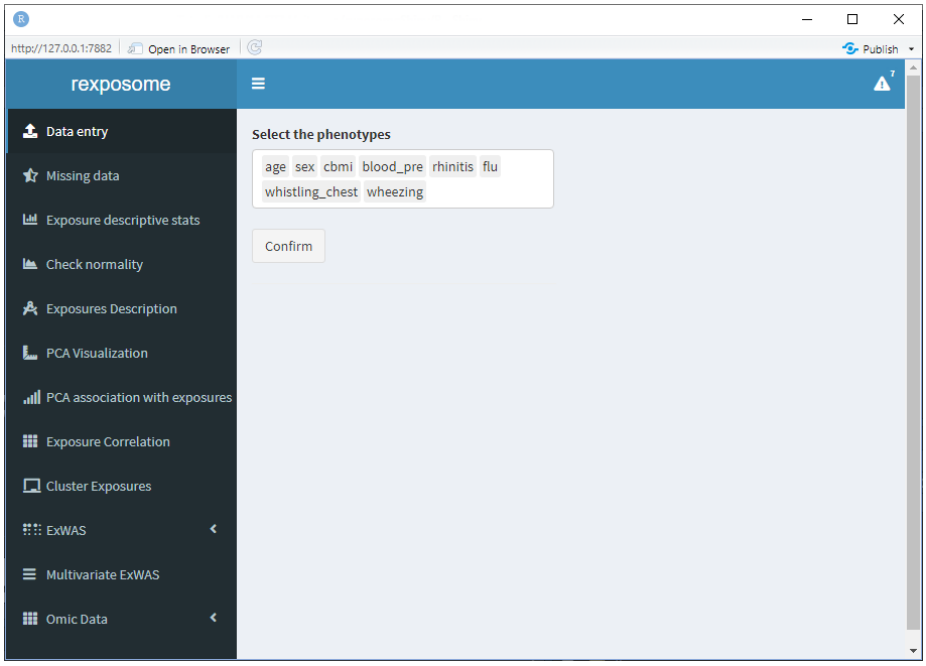

Once the fields are completed, please press the *Validate selections* button in order to check if all the parameters allow for a successful load of the exposome dataset, if not, a pop-up will be prompted to the user. In the case that everything is correct, the UI will be updated and a button that reads *Load selected data to analyze it* will appear, by clicking this button the dataset will be loaded and the analysis can begin. The input fields are still visible past this point for the case that the user wants to load a new dataset into the application without restarting it.

#### 5.1.1.2 Plain data

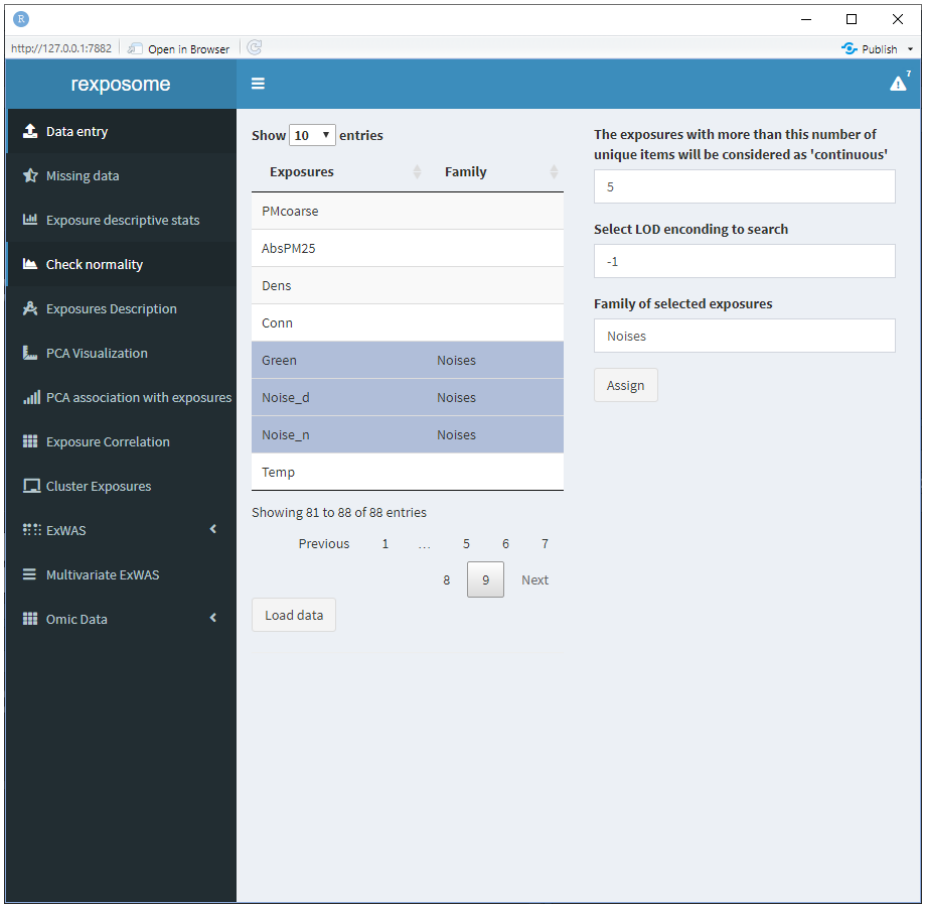
When dealing with a single file that contains the exposures and phenotype data, press the "My data is contained in a single table" toggle. This will change the interface to only show one file selector.




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 researcher can group the exposures into families, to do so, select them on the table, input the family name on the field “Family of selected exposures” and press “Assign”. All the exposures that have empty “Family” fields will be treated as they are their own family.

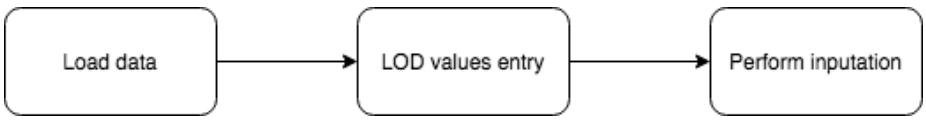



Finally, there are two configuration fields:

- The treshold of to select between continuous or factor exposures. More than this number of unique items 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`).

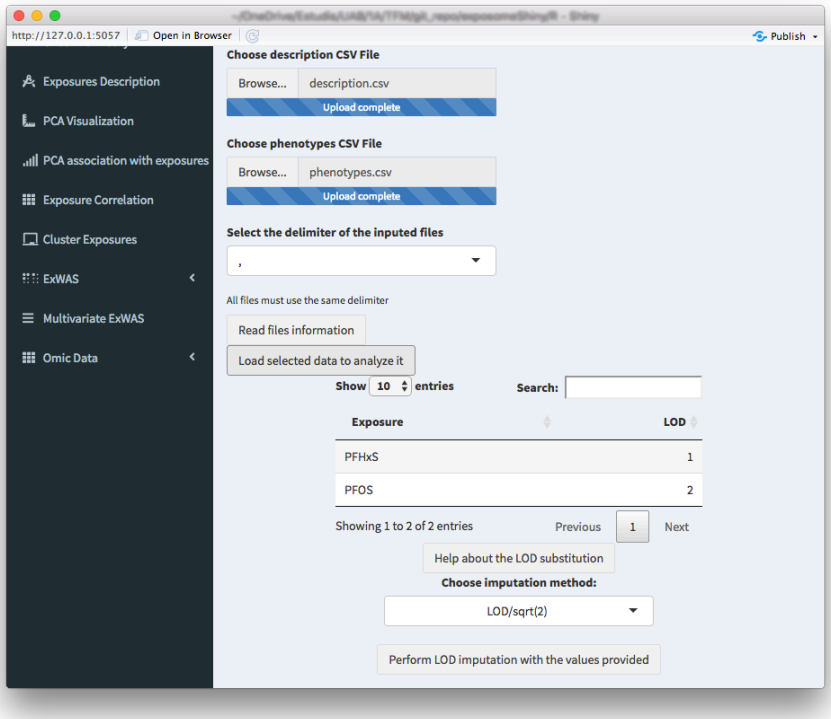
Be sure to revise them before pressing "Load data".

### 5.1.2 LOD imputation

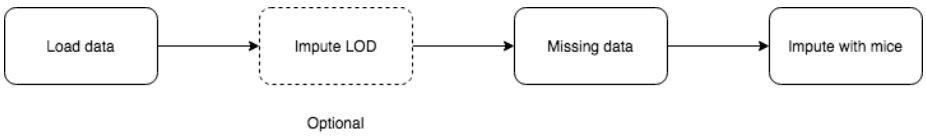



If exposomeShiny detects LODs (limit of detection) on the exposures file, it will prompt the table with the exposures with LOD and double clicking on the desired cell will enable edit mode to input the instrument LOD. There's

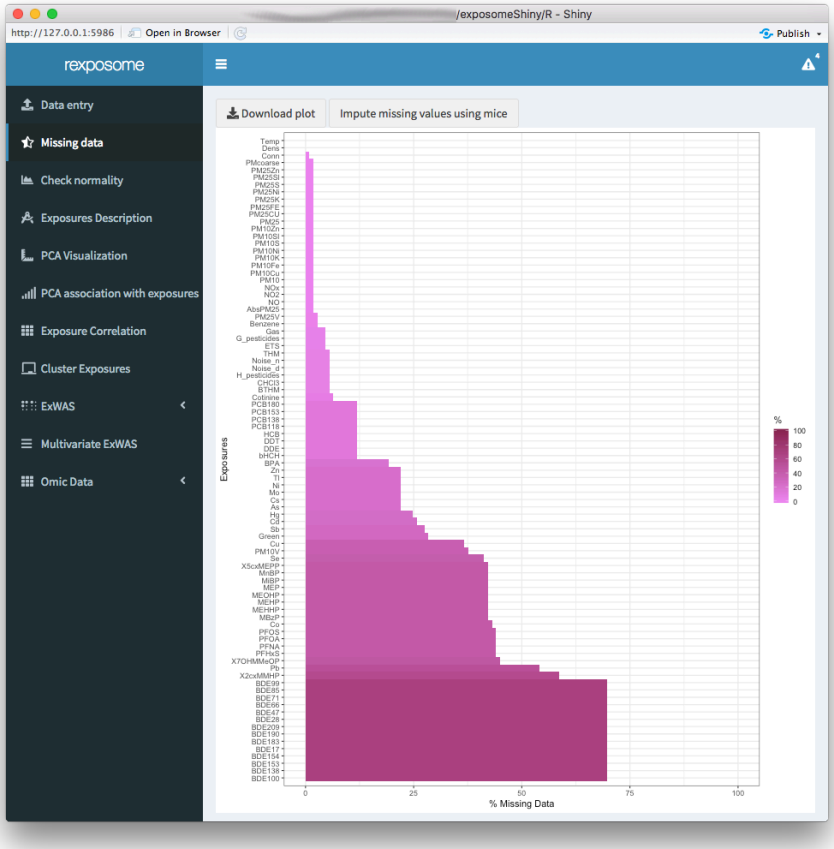
also the option of selecting "Random imputation" on the imputation method in order to imputate with random values (truncated log distribution) instead of LOD/sqrt(2). If the user imputes the LOD missings, the dataset will be imputed and the user will not need to reimport it, the loaded dataset will be updated to have the LOD missings imputed, this can be checked by clicking on the ! symbol at the top bar of the application.



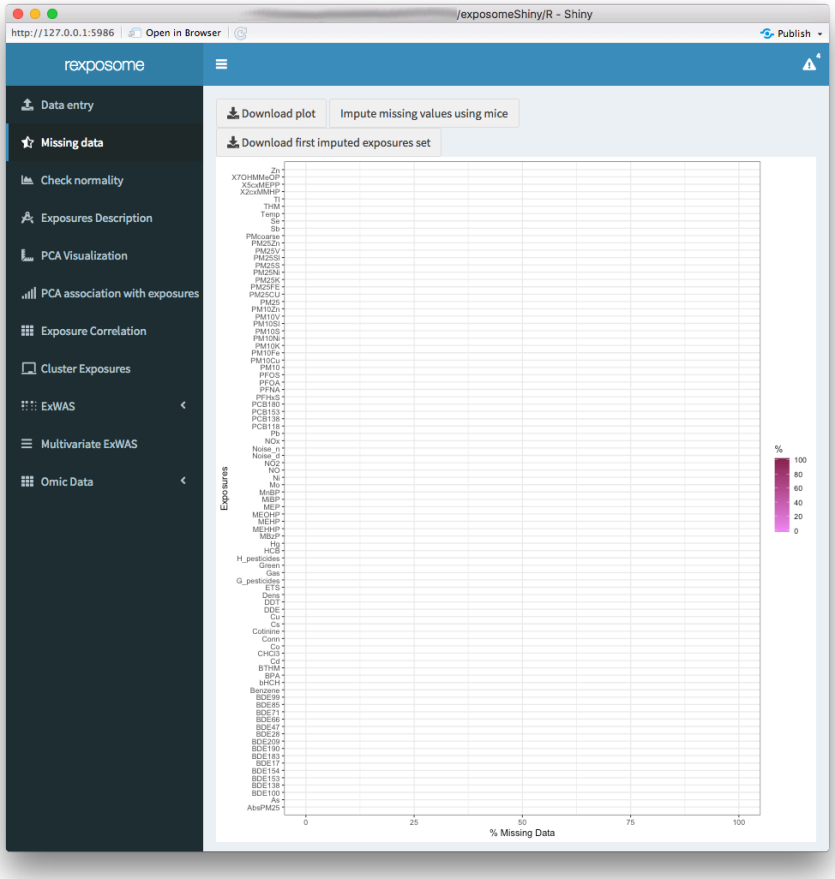

### 5.1.3 Missing imputation




Once the dataset is loaded into the Shiny, look at the "Missing Data" tab to check the percentages of missing data for each exposure present.




To impute the missing values select "Impute missing values using mice" (Multiple Imputation by Chained Equations), no other method of imputation is available in the software at the moment. After the process finishes, the expect output should be a new missing data graph where there's no missing of any exposure.




The new imputed exposures set can be downloaded as a `*.csv` file, please note that the downloaded file just assigns numbers to the `idnum` column, if the data you are using has different `idnum` format it's needed to format it properly so that it matches the `idnum` on the phenotypes input file when inputting it to the Shiny.

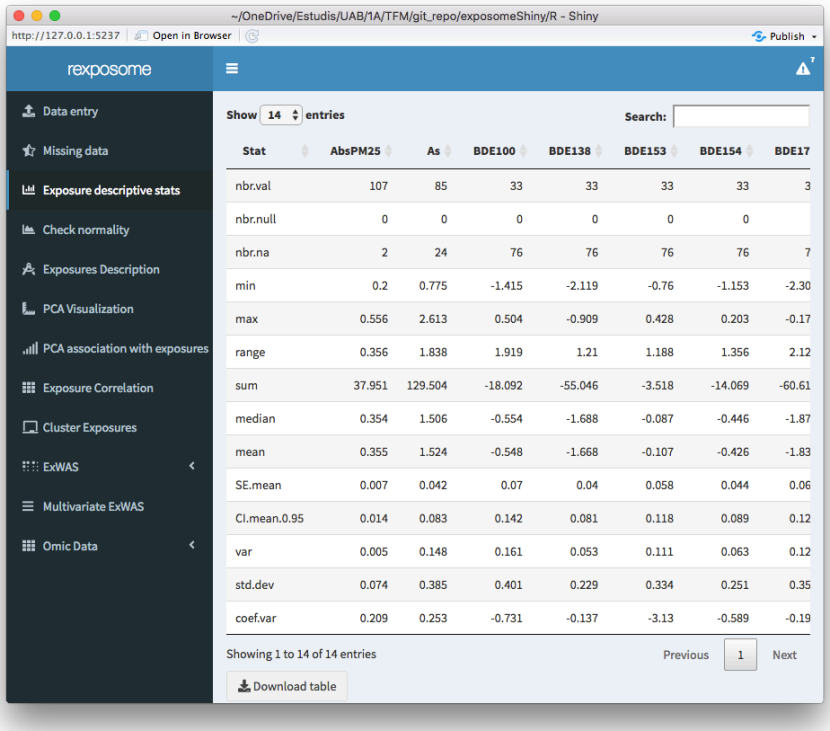
### 5.1.4 Exposures description

There's the option of visualizing the main descriptive stats of the exposures dataset, available for quantitative exposure varibables. The descriptive stats (per exposure) included on the table are:

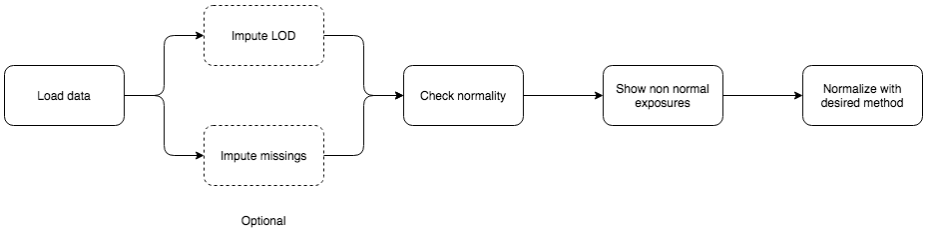
- Number of values
- Number of NULLs
- Number of NAs
- Minimum
- Maximum

- Range of values
- Sum of values
- Median
- Mean
- Standard Error of mean
- 0.95 confidence interval of the mean
- Variance
- Standard deviation
- Variance coefficient

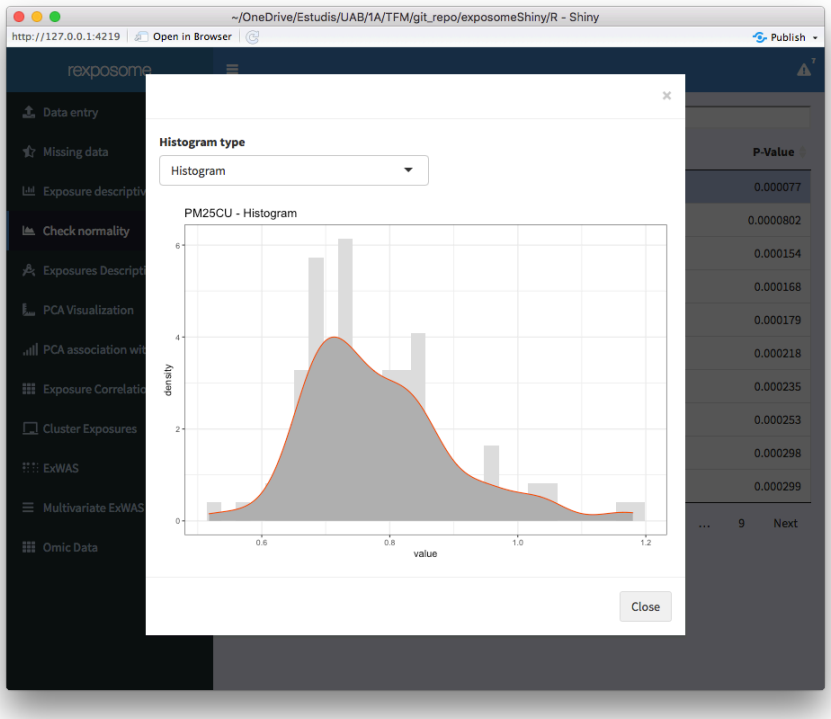
This table will have the descriptive stats of the loaded dataset, this means that if the user has imputed the missings (remember that after imputing the missings, the imputed dataset becomes active) it will be reflected on the table as it will show 0 NAs.



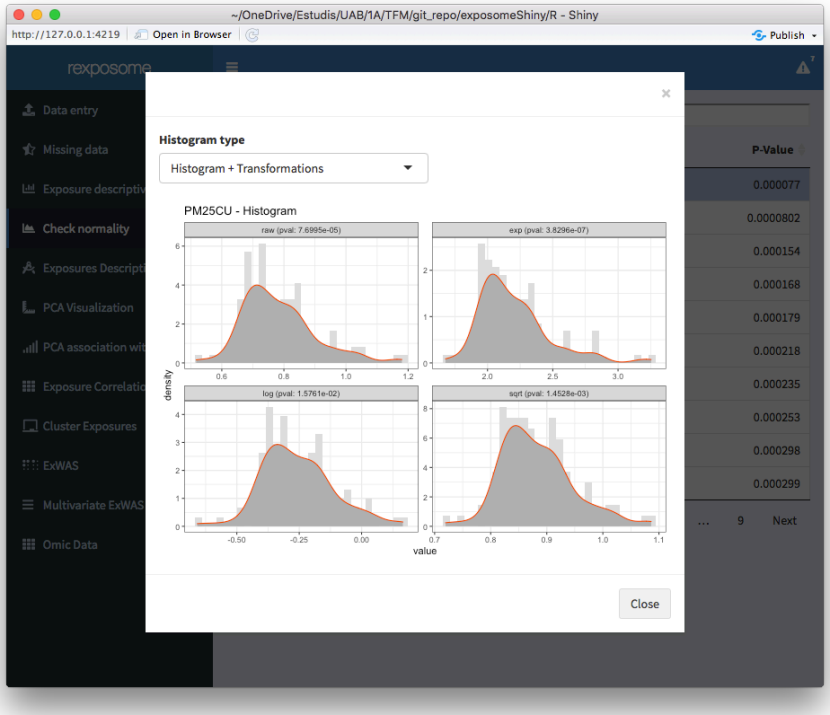
### 5.1.5 Normality correction



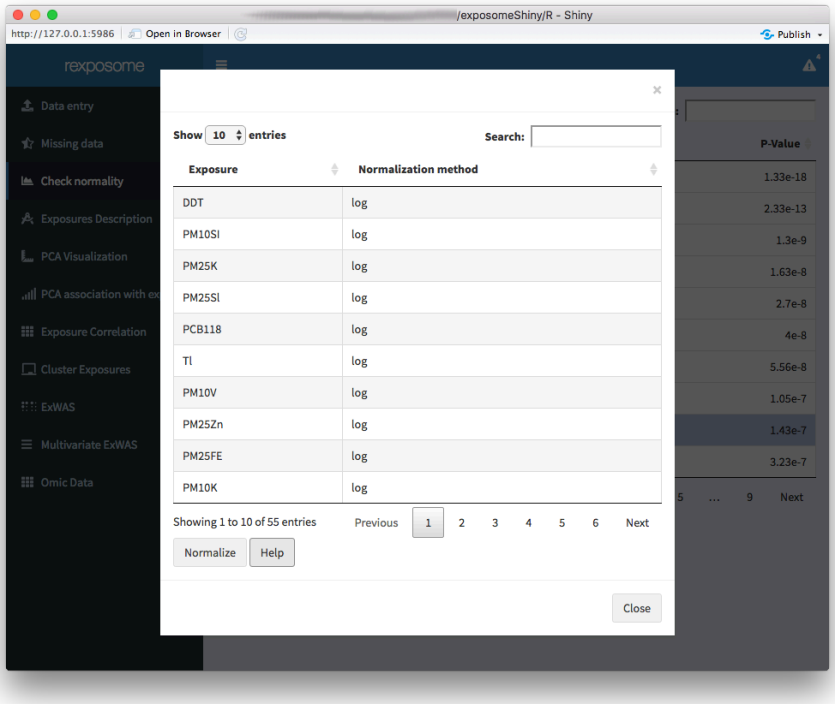

Once the dataset is loaded into the Shiny, look at the "Check Normality" tab to check which exposures are not normal (Normality = false), this are the results of Shapiro–Wilk testing. By selecting from the table the desired exposure and clicking the "Plot histogram of selected exposure", as the label of the button implies, a histogram of the selected exposure from the table can be seen.




There is also the option of visualizing the histogram for the implemented transformations along the normality test p-value for said transformations.

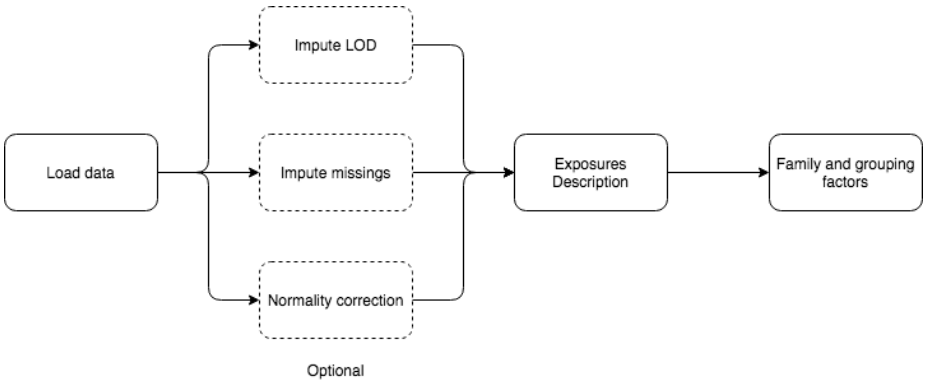



By clicking the "Show false" button, all the non normal exposures are listed with the method that will be applied to normalize, this table can be edited (the "Normalization method" column) by double clicking on the desired row. There are three possible methods to use, "log" (default, natural logarithm), "^1/3" and "sqrt". If no method is desired to be applied to an exposure (keep the original variable) input "none". The normalization method refers to applying X function to a column of the exposures tables, to transform it.



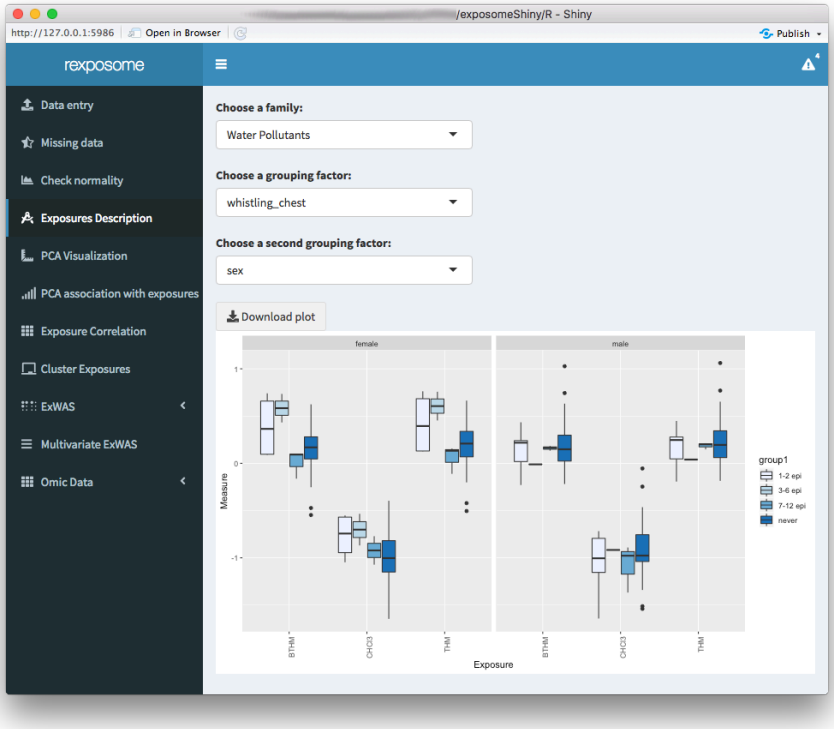

Click "Normalize" and the normalization method selected will be applied, the table on the "Check Normality" tab will be updated with the results of the normalization.

### 5.1.6 Exposures description

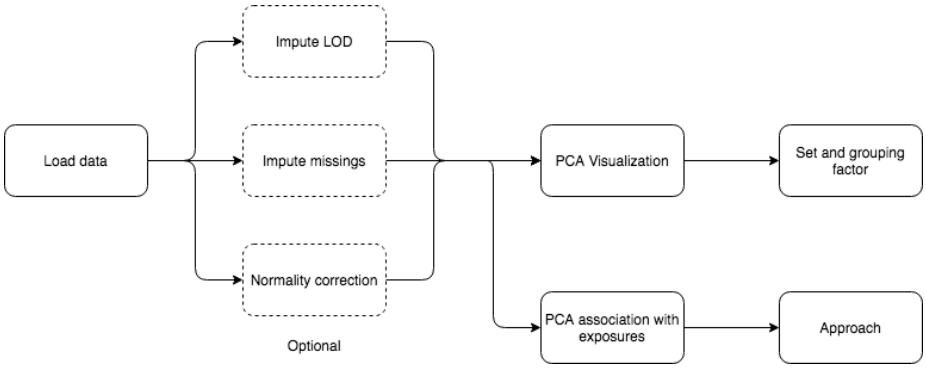



To see all the insights of the exposures dataset loaded into the Shiny, once loaded it check the exposures description tab, there are three options to dig into

the dataset, the family (family of the exposure) to visualize and two grouping factors (phenotypes).

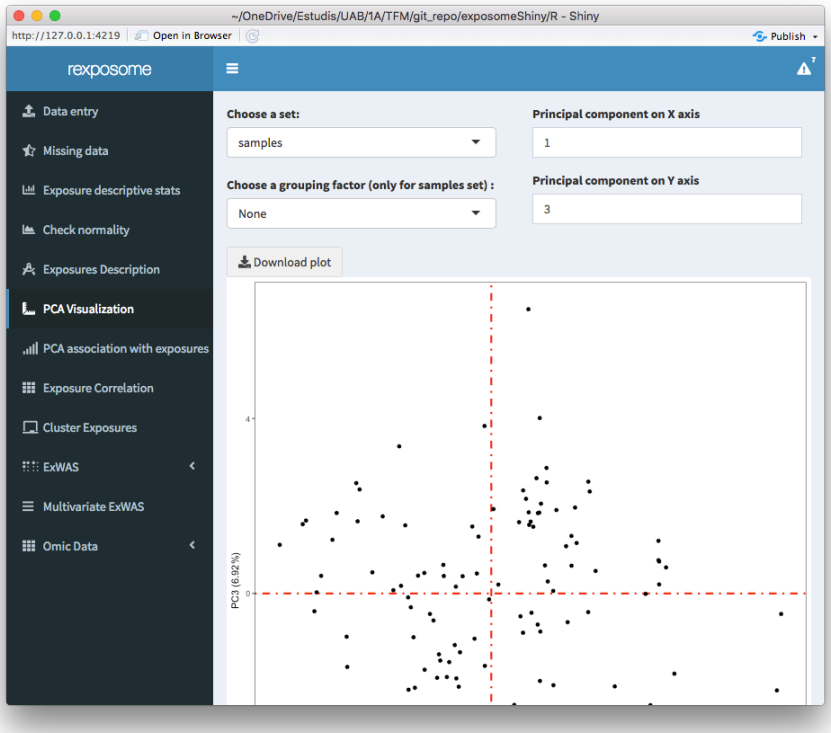



### 5.1.7 PCA Analysis

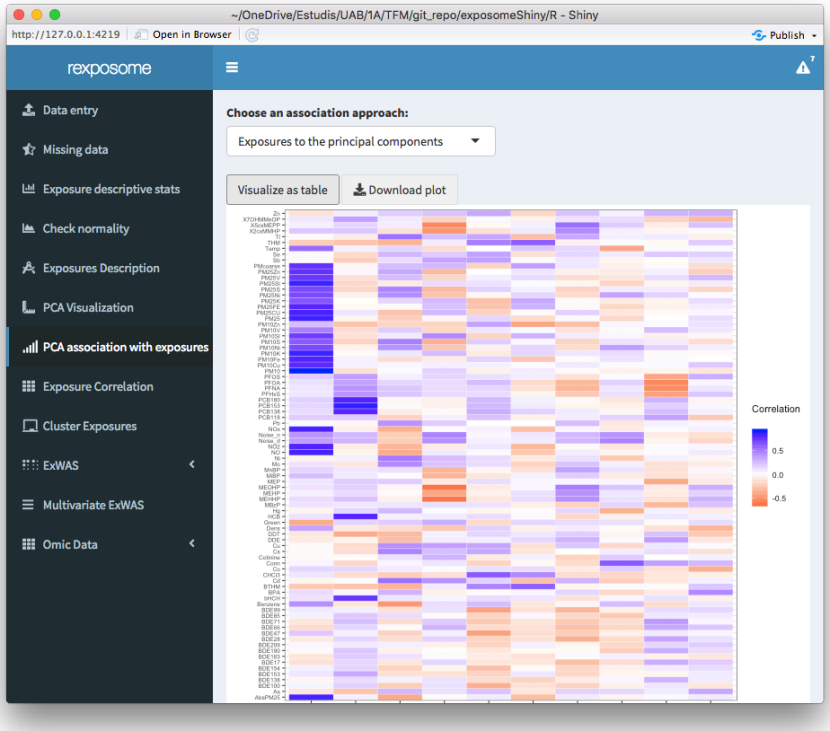



To see the results of a PCA (principal component analysis) study, load the data and check the PCA Visualization tab, there a set and grouping factor can be choose, it's important noting (as it's already stated on the Shiny) that the

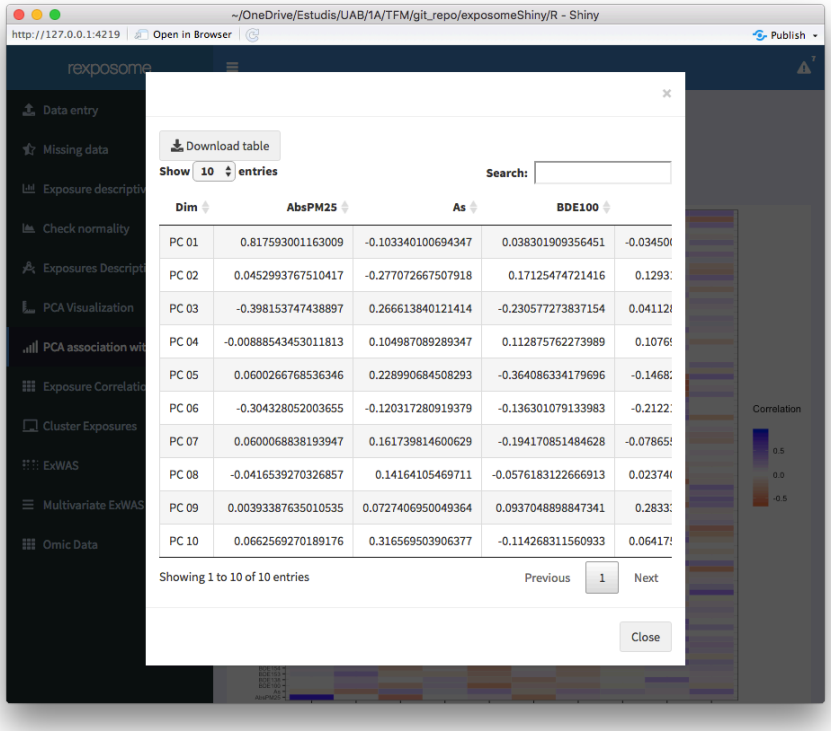
grouping parameter only works when the set is selected to "samples". There are also selectors to choose which principal component to visualize on each axis, 10 principal components are computed.




If the association of the PCA analysis with the exposures is desired to visualize, check the "PCA association with exposures" tab, there are two grouping methods to visualize, the phenotypes to principal components and the exposures to principal components.



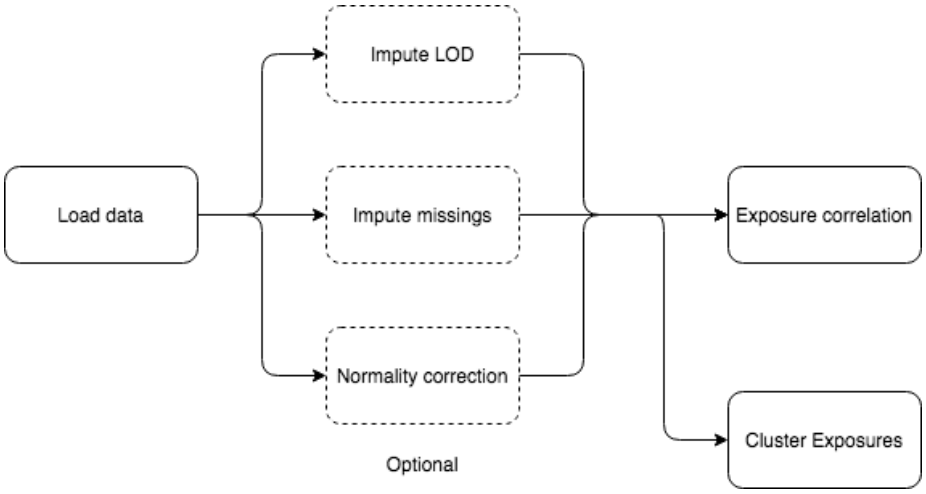

By clicking on "Visualize as table", the table corresponding to the selected association will be prompted on a pop-up window, this table can be downloaded as a csv.



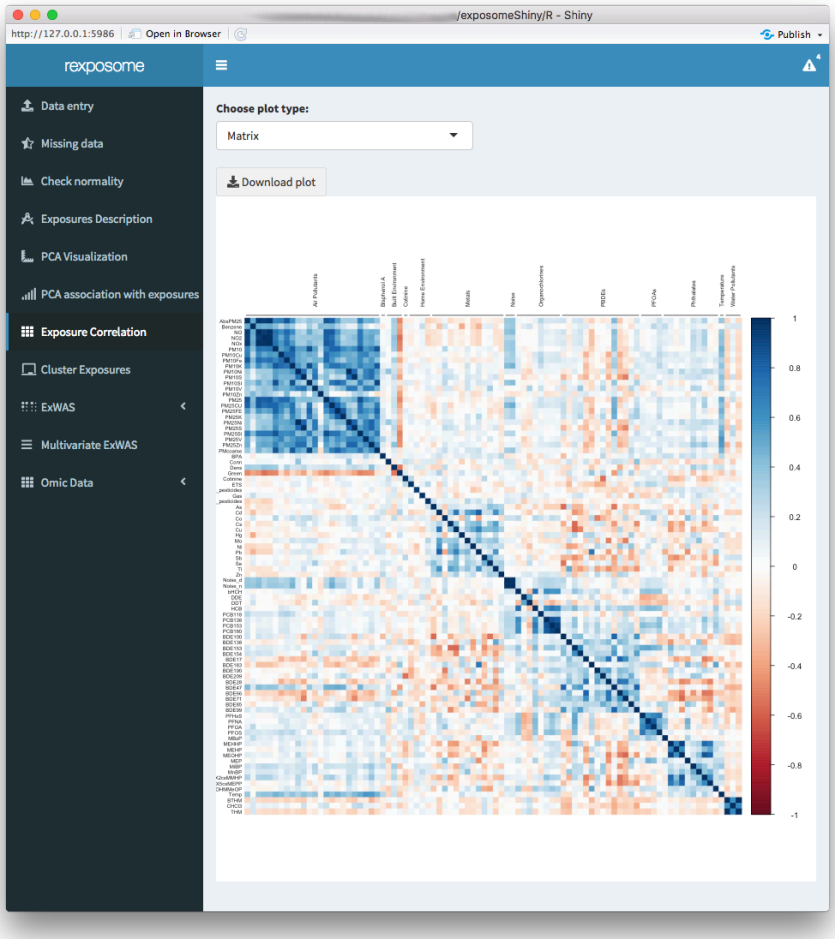

### 5.1.8 Clusterization and correlation of exposures

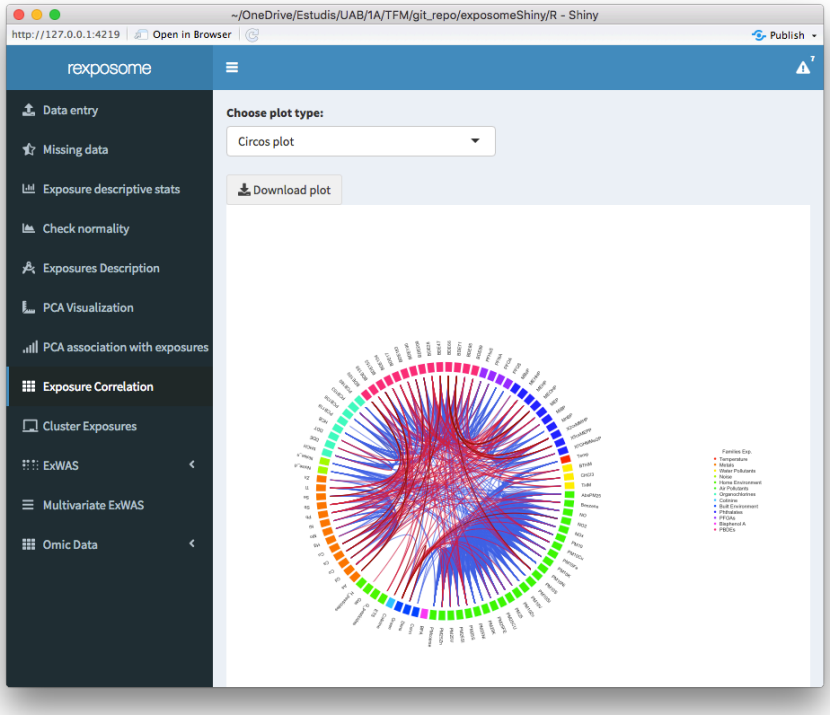
Displaying the correlation of the exposures can help to visualize intra and inter family relations between the exposures, for that reason there are two different visualization options, the circos and the matrix.

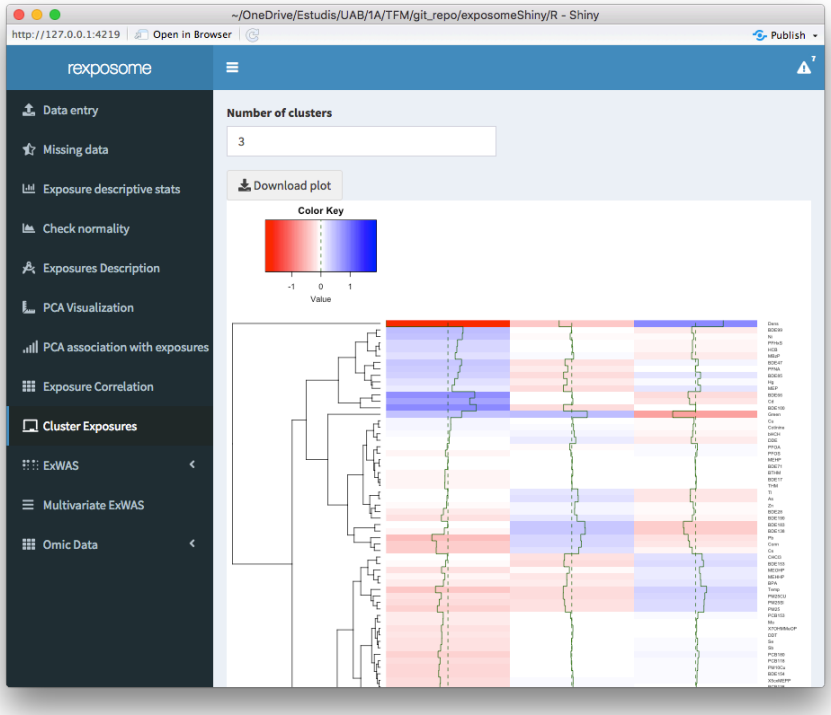
The clusterization of exposures uses a hierarchical clustering algorithm to classify the individuals profiles of exposures in k groups, where k can be selected by the user. The plot shows the profile for each group of individuals.



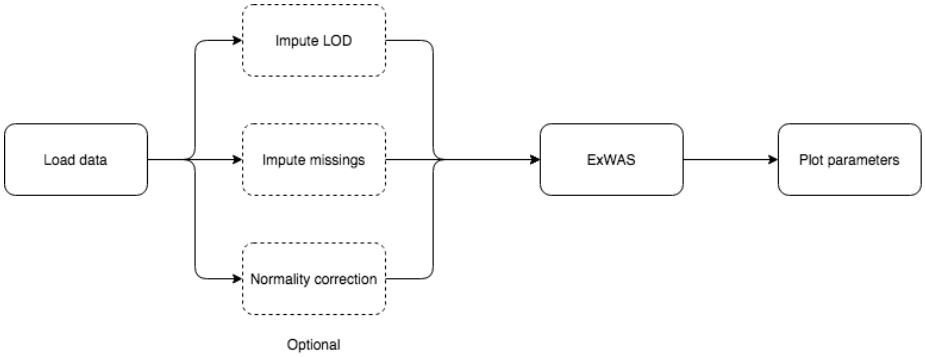

To see the results of the exposure correlation and clustering, select the corresponding tab to each analysis. For the exposure correlation analysis there are two visualizations, the matrix representation and the circos. The correlation uses Pearson method for numerical-to-numerical correlation, Cramer's V for categorical-to-categorical correlation and linear models for categorical-to-numerical.


/exposomeShiny/R - Shiny
http://127.0.0.1:5986
Open in Browser
Publish
rexposome
Data entry
Missing data
Check normality
Exposures Description
PCA Visualization
PCA association with exposures
Exposure Correlation
Cluster Exposures
ExWAS
Multivariate ExWAS
Omic Data
Choose plot type:
Matrix
Download plot
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


~/OneDrive/Estudis/UAB/1A/TFM/git_repo/exposomeShiny/R - Shiny
http://127.0.0.1:4219
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 plot type:
Circos plot
Download plot

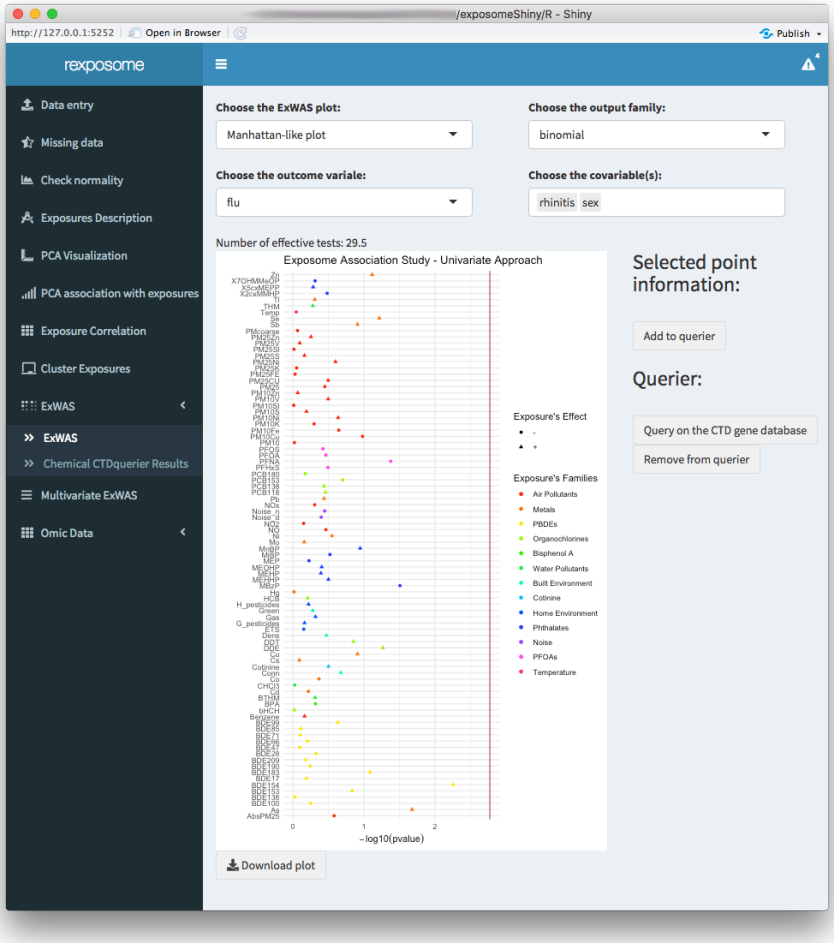



### 5.1.9 ExWAS



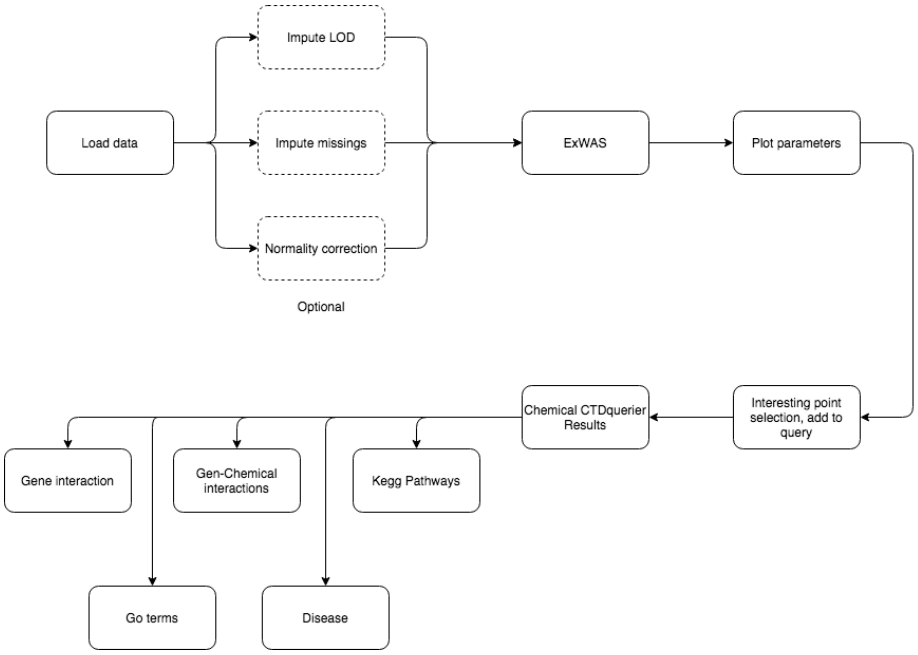

To perform an ExWAS (exposome-wide association, univariate test of the association between exposures and health outcomes using generalized linear models) study, check the ExWAS tab and select the addecuate parameters for the ExWAS plot, there are two different plot representations, the output variable to choose (phenotype), the output family and as many covariables (phenotypes) as the user wants. There are internal checks to advise the user on which parameters

to select depending if the selected outcome is numerical or bionomial.

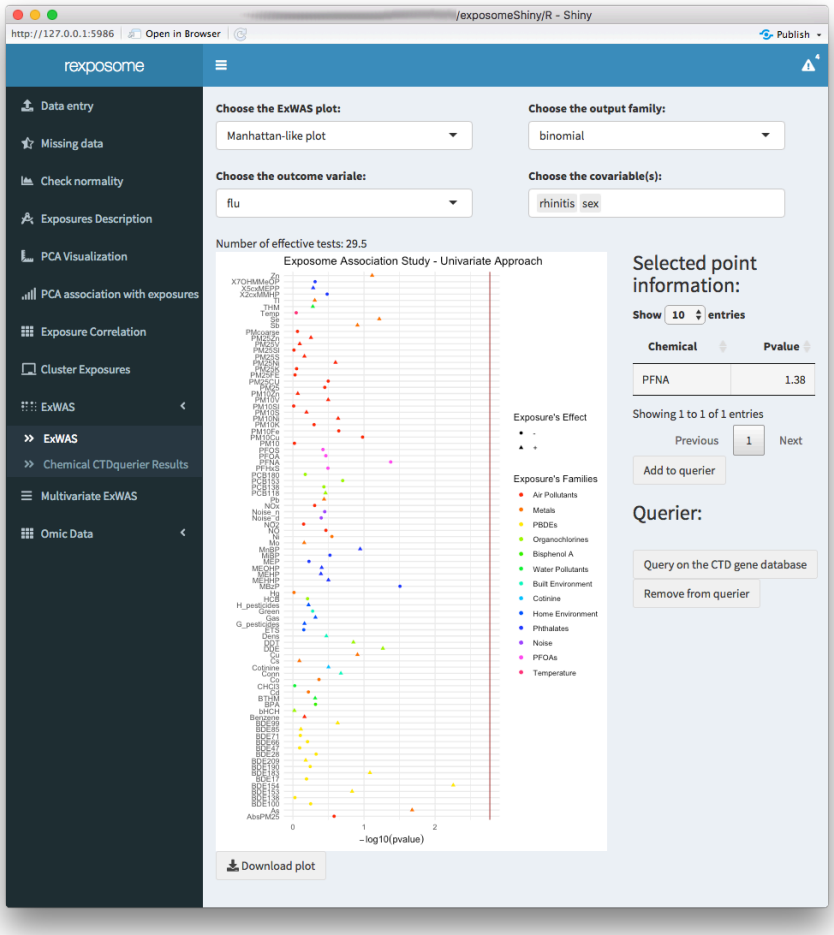


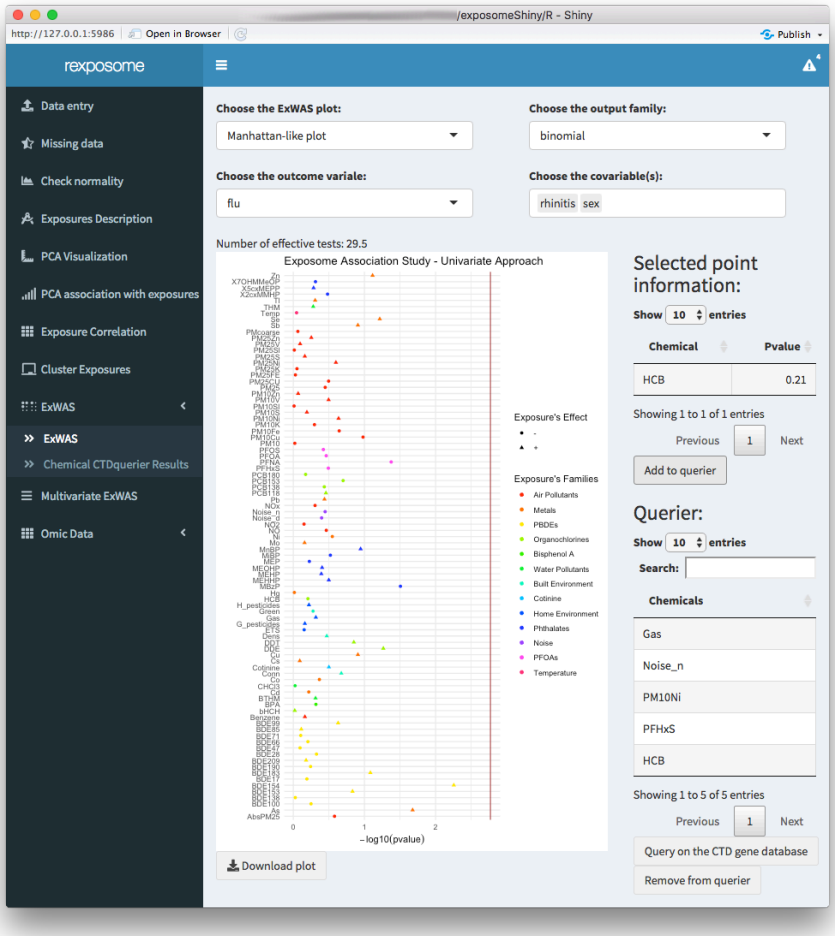

### 5.1.10 ExWAS - CTDquerier

The ExWAS tab also is able to perform a CTD query of the desired chemicals.

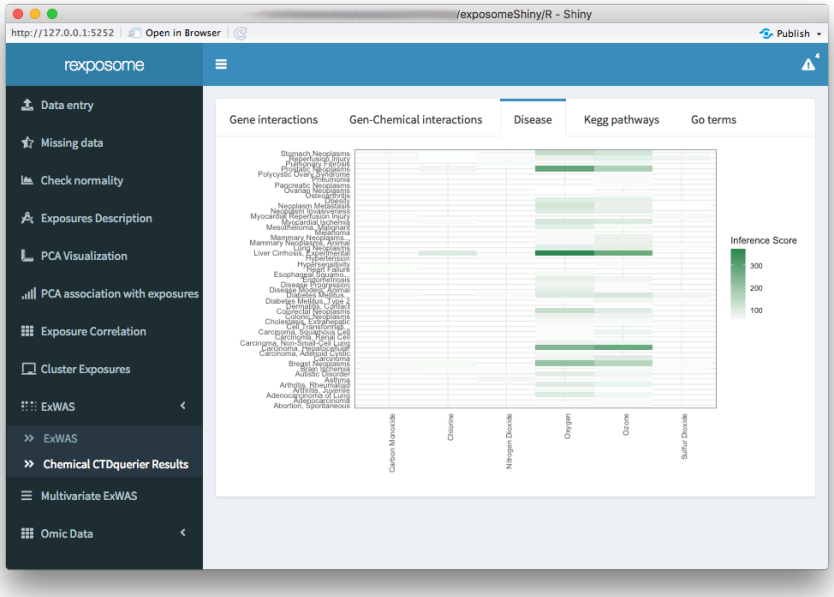



To perform a CTDquerier of chemicals with the results of the ExWAS, click on the desired exposure to preload it into the query, when clicked, a chemical name with it's associated P-Value will appear on the table on the right, if that's the desired chemical to add to the query list click "Add to querier". In the case of adding an unwanted chemical to the query list, select it (or them) by clicking on the Querier list and click on "Remove from querier".


/exposomeShiny/R - Shiny
http://127.0.0.1:5986
Open in Browser
Publish
rexposome
Data entry
Missing data
Check normality
Exposures Description
PCA Visualization
PCA association with exposures
Exposure Correlation
Cluster Exposures
ExWAS
ExWAS
Chemical CTDquerier Results
Multivariate ExWAS
Omic Data
Choose the ExWAS plot:
Manhattan-like plot
Choose the output family:
binomial
Choose the outcome variale:
flu
Choose the covariable(s):
rhinitis
sex
Number of effective tests: 29.5
Exposome Association Study - Univariate Approach
−log10(pvalue)
Exposure's Effect
Exposure's Families
Air Pollutants
Metals
PBDEs
Organochlorines
Bisphenol A
Water Pollutants
Built Environment
Cotinine
Home Environment
Phthalates
Noise
PFOAs
Temperature
Download plot
Selected point information:
Show 10 entries
Chemical
Pvalue
PFNA
1.38
Showing 1 to 1 of 1 entries
Previous
1
Next
Add to querier
Querier:
Query on the CTD gene database
Remove from querier

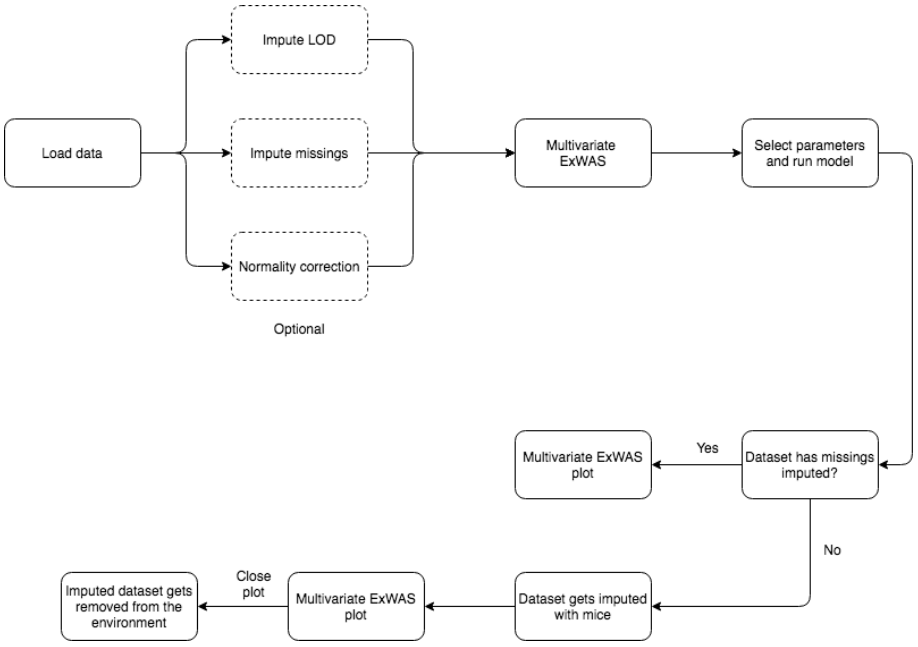



To do the query of the chemicals to de CTD database click on “Query on the CTD gene database” and see the results on the “Chemical CTDquerier Results” subtab. It’s important noting that on the “Kegg pathways” and “Go terms” the input field corresponds to the negative exponent of the filter.

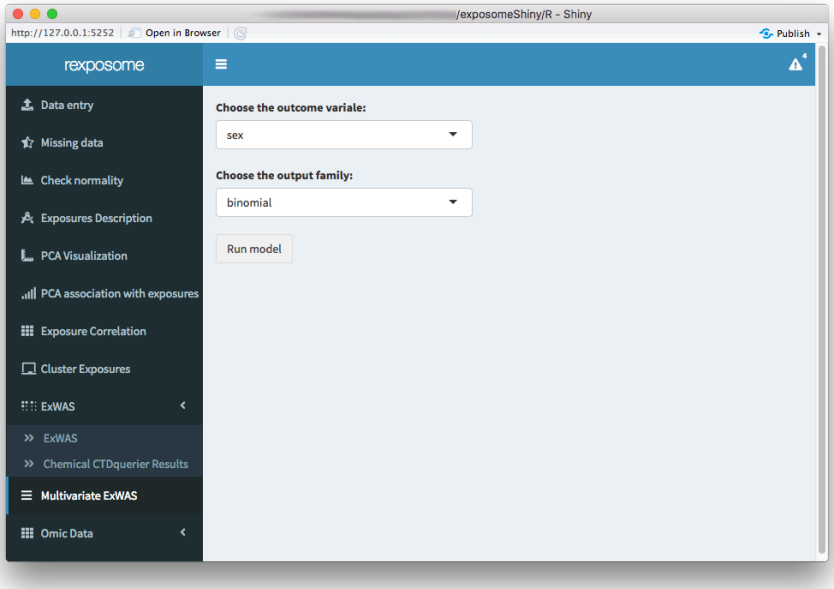


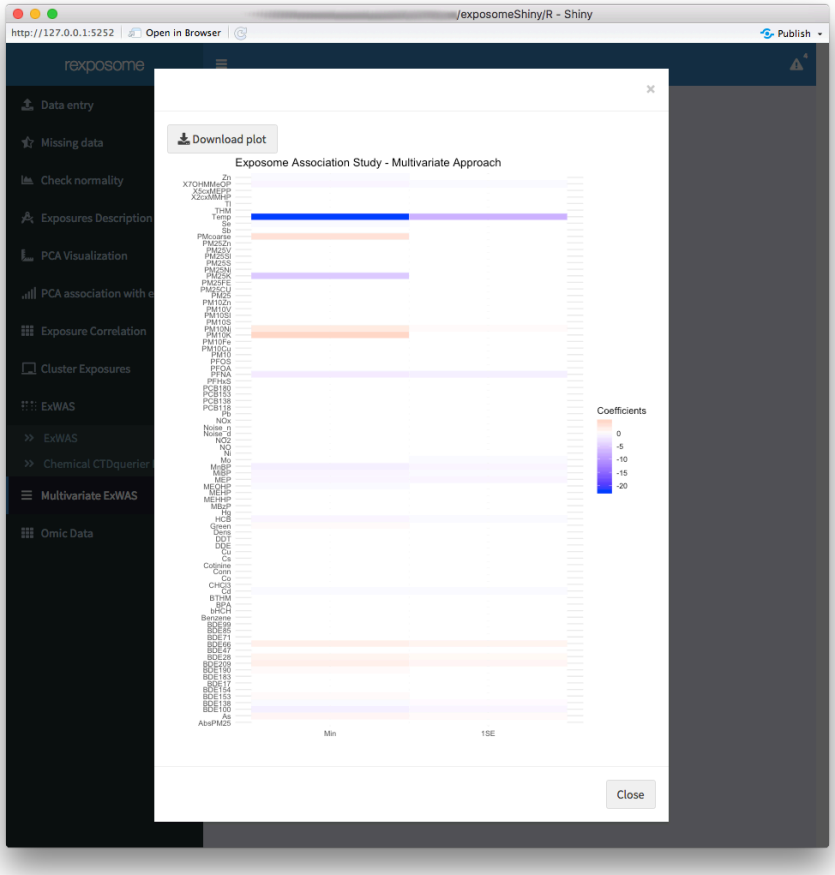

### 5.1.11 Multivariate ExWAS

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




To perform a multivariate ExWAS study, check the Multivariate ExWAS tab and select the desired output parameter, click on run model to generate the plot. As on the ExWAS plot options there's implemented an internal check to advise the user on which parameters to select depending if the selected outcome is numerical or bionomial, as the diagrams states if the dataset has not been imputed the missings, it will automatically do it to perform the Multivariate ExWAS, however when closing the plot the imputed dataset will be removed from the environment, so all the other studies performed afterwards will not be altered.


/exposomeShiny/R - Shiny
http://127.0.0.1:5252
Open in Browser
Publish
rexposome
Data entry
Missing data
Check normality
Exposures Description
PCA Visualization
PCA association with exposures
Exposure Correlation
Cluster Exposures
ExWAS
ExWAS
Chemical CTDquerier Results
Multivariate ExWAS
Omic Data
Choose the outcome variale:
sex
Choose the output family:
binomial
Run model

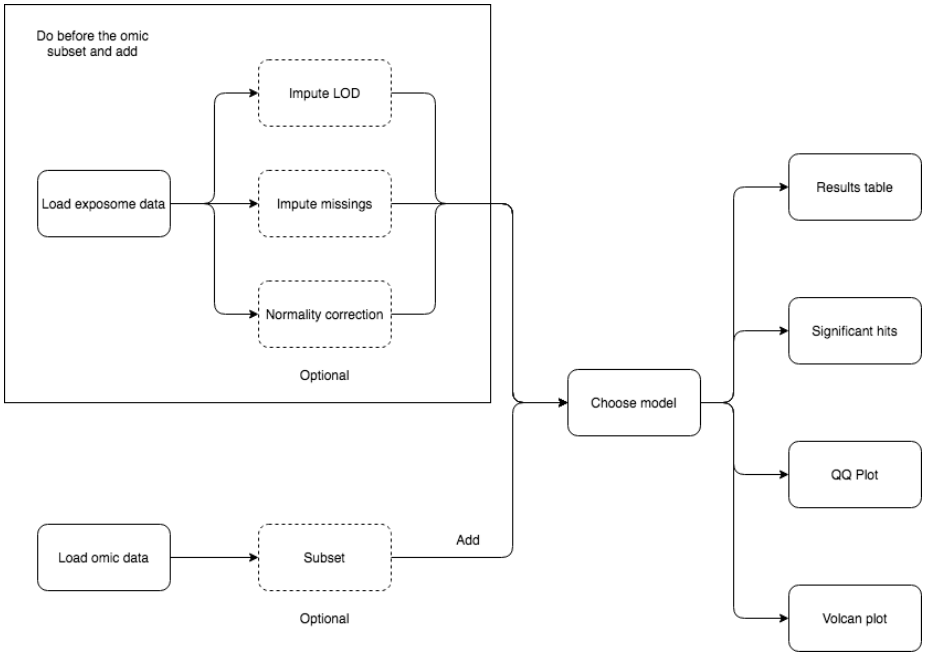



## 5.2 Exposome-Omic analysis

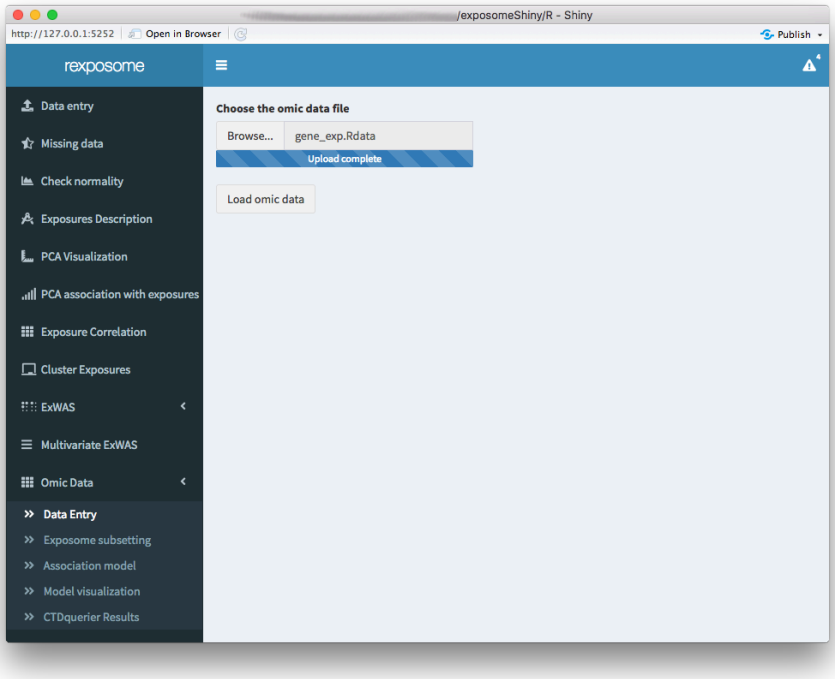
It's important noting that the maximum size of the omics data is 30 MB, if the omics file to be analyzed is bigger, change the line number 2 of the `server.R` file.

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

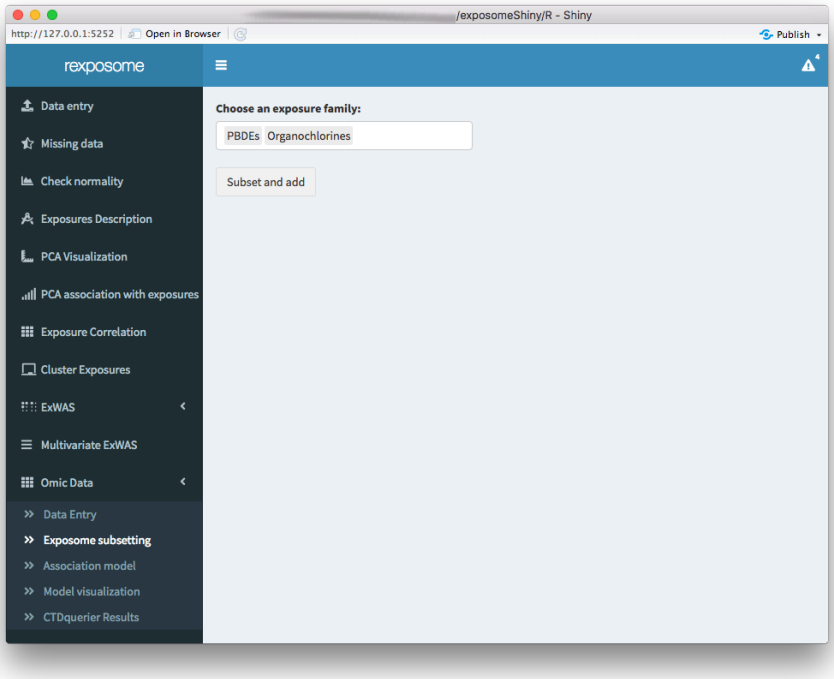
### 5.2.1 Association analysis



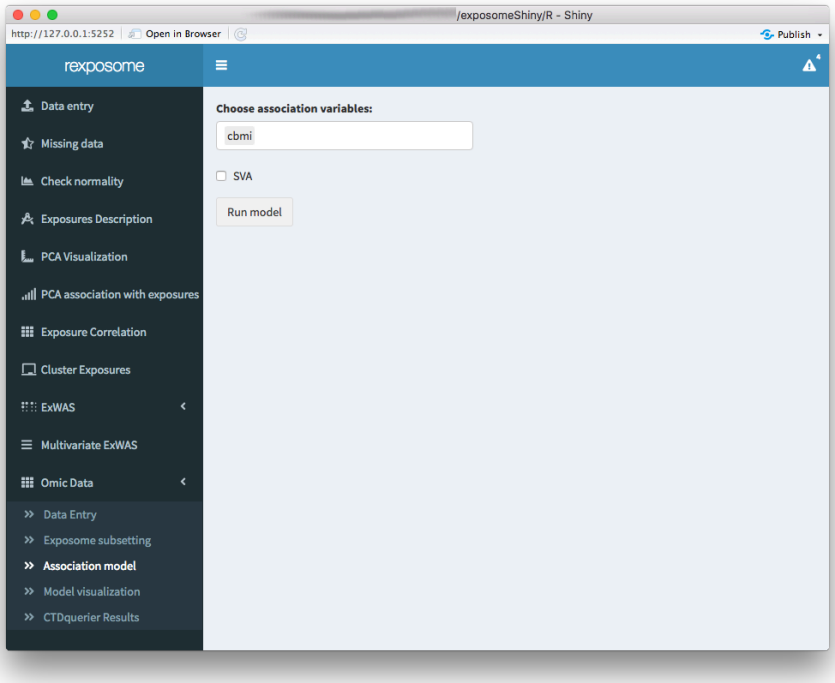

Do first the proceeding of exposome data load and corresponding treatment if desired, then proceed to load the omic dataset on the "Data Entry" subtab. The omic data should be provided as a `*.RData` file.



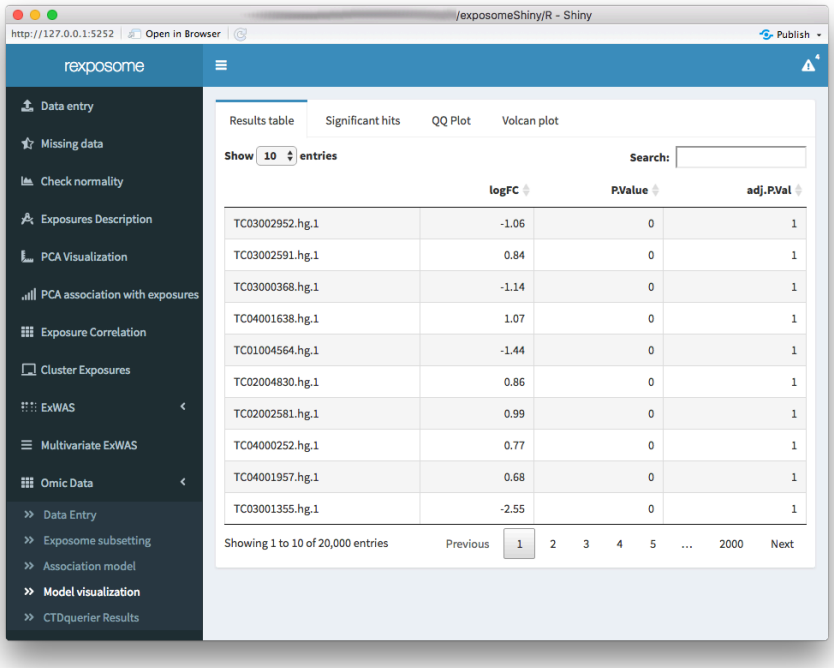

The exposome dataset can be subseted by families for this analysis, on the "Exposome subsetting" subtab select the families that are desired to be included in this new set to study, if all the families are desired just don't input any and proceed to click the "Subset and add", which will trigger the action to combine the subsetted (or not) exposome dataset with the provided omic dataset.



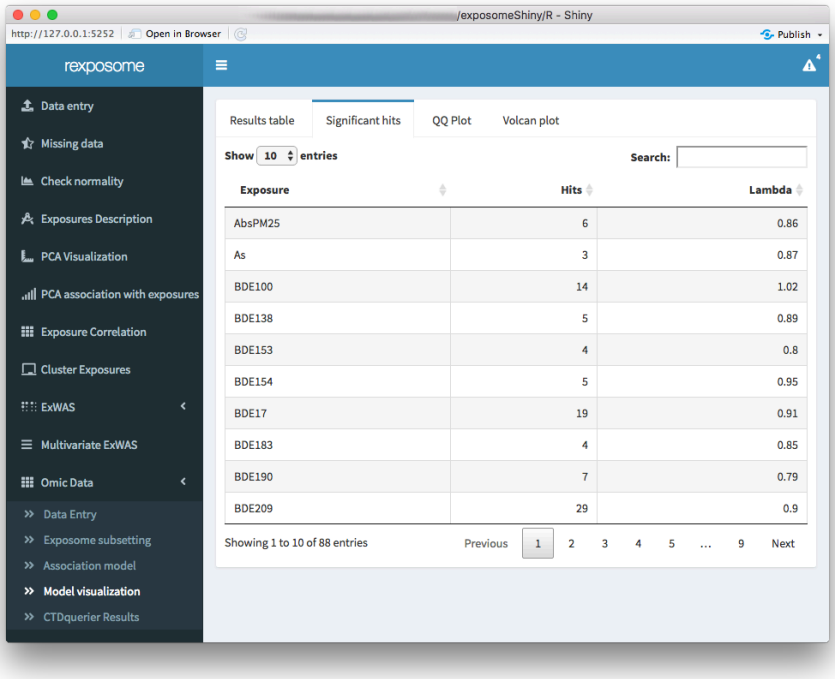

Select the variables for the association analysis (linear models) and if SVA (surrogate variable analysis) is wanted on the "Association model" subtab.



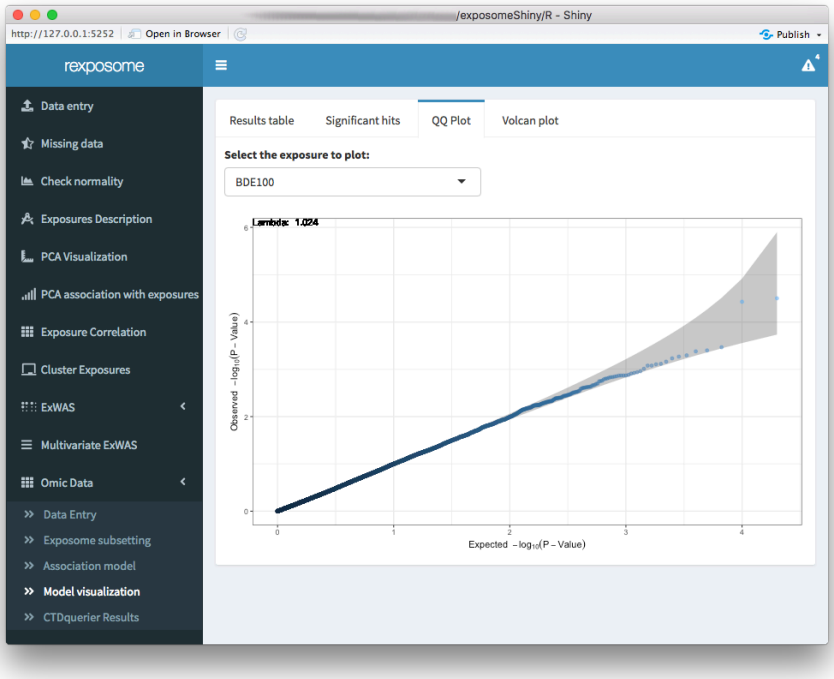

There are tabs to visualize the results of running the association model, all of the are on the "Model visualization" subtab. The "Results table" shows the gene, log of the fold change, p-value and adjusted p-value.



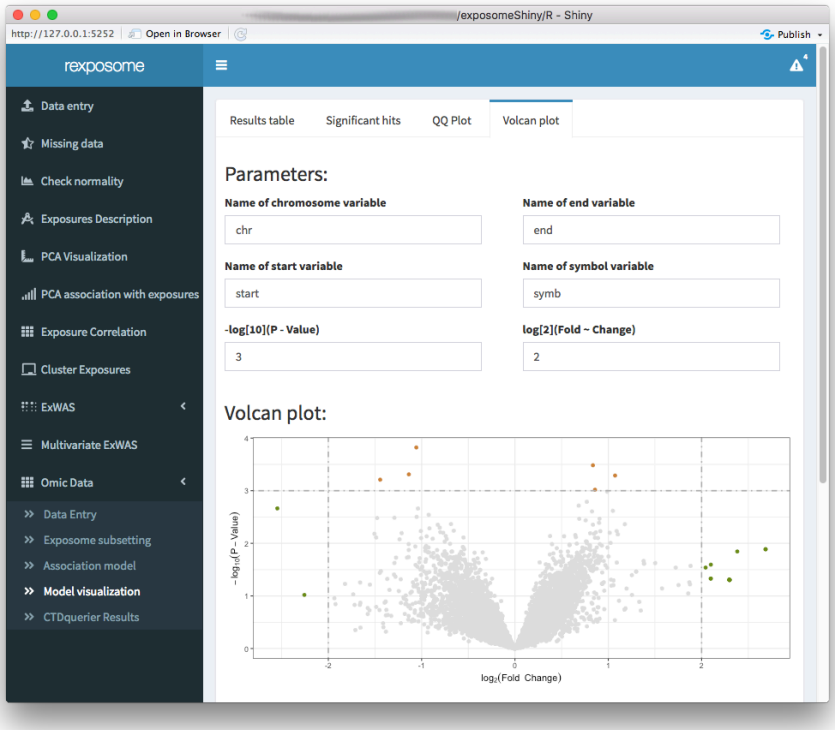

The “Significant hits” shows the exposure, hits and lambda.



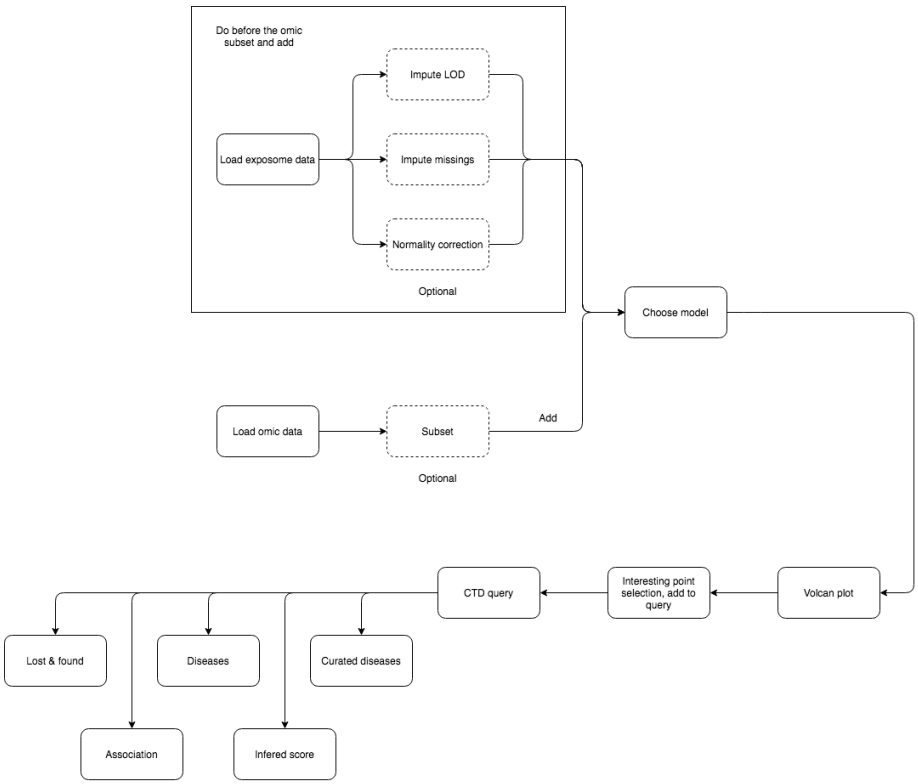

The QQ Plot shows a QQ plot (expected vs. observed -lo10(p-value)) for the selected exposure.



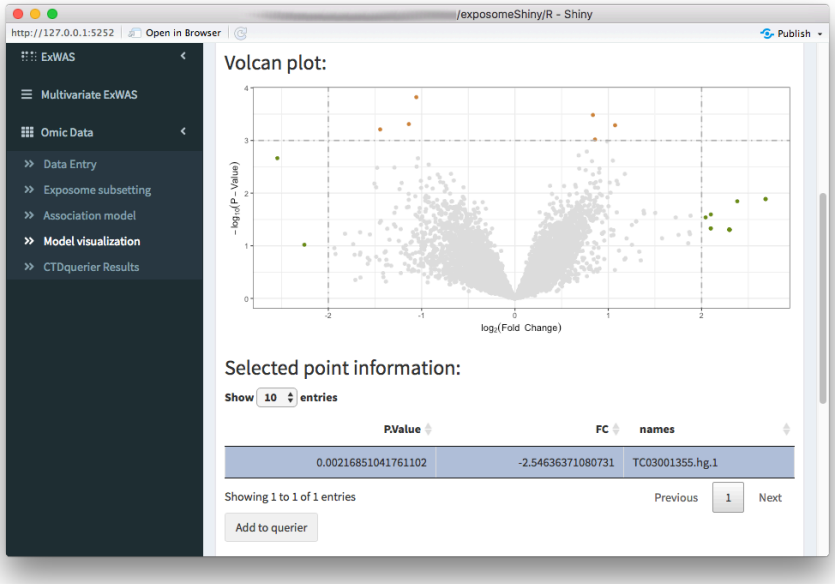

The Volcan plot shows a volcan plot (log2(fold change) vs -log10(p-value)). For this plot there are two input cells to adjust the horizontal and verital limit lines to filter out the results.


/exposomeShiny/R - Shiny
http://127.0.0.1:5252
Open in Browser
Publish
rexposome
Data entry
Missing data
Check normality
Exposures Description
PCA Visualization
PCA association with exposures
Exposure Correlation
Cluster Exposures
ExWAS
Multivariate ExWAS
Omic Data
Data Entry
Exposome subsetting
Association model
Model visualization
CTDquerier Results
Results table
Significant hits
QQ Plot
Volcan plot
Parameters:
Name of chromosome variable
chr
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
Volcan plot:
log2(Fold Change)

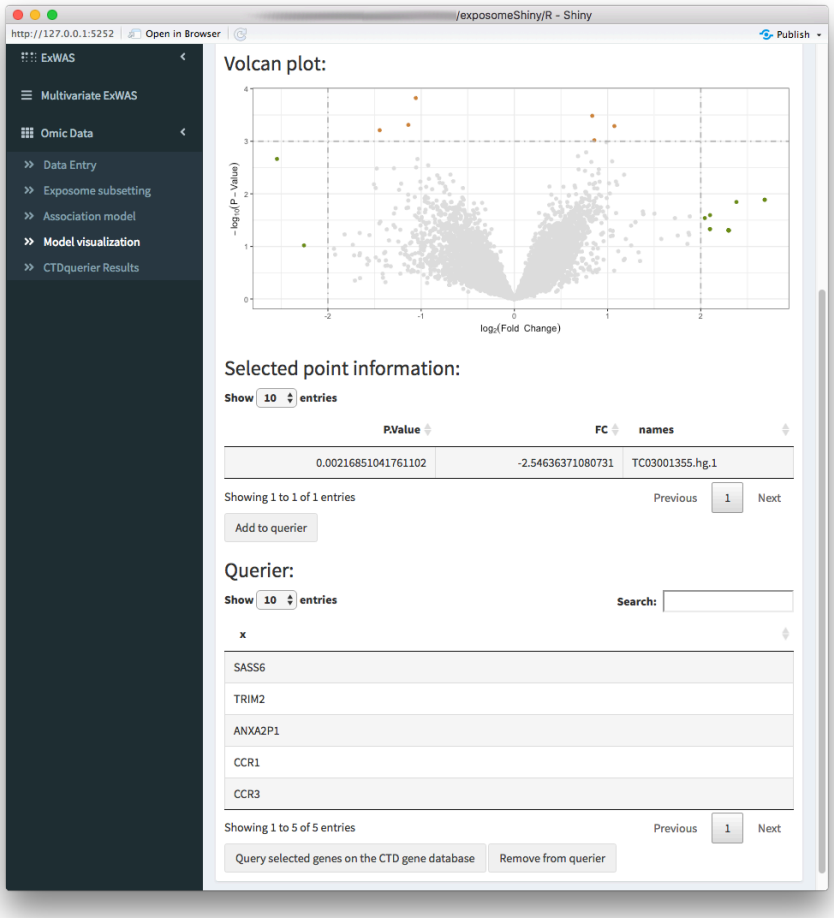
### 5.2.2 CTD querier



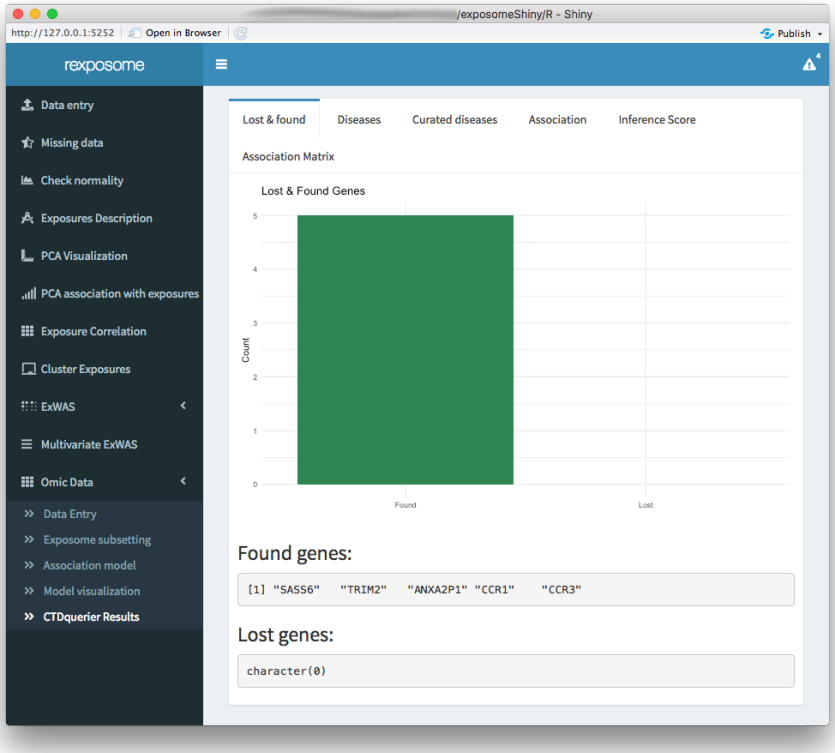

To perform a CTD querier study of the exposome-omic analysis, as before, load both datasets and run the desired model with them, check the Volcan plot and click on the desired point on the Volcan plot, the information of the selected plot will appear on the table below the plot (sometimes there are many points close so more than one rows can appear on the table), select from the table the desired point to add to the query and click "Add to querier". It's important noting that when trying to add to the querier the Shiny will find on the fields of the omic dataset that the user specifies on top of the plot. If the search does not return any symbol a prompt will appear, however if it's found it will be added to the lower table corresponding to the genes to query.



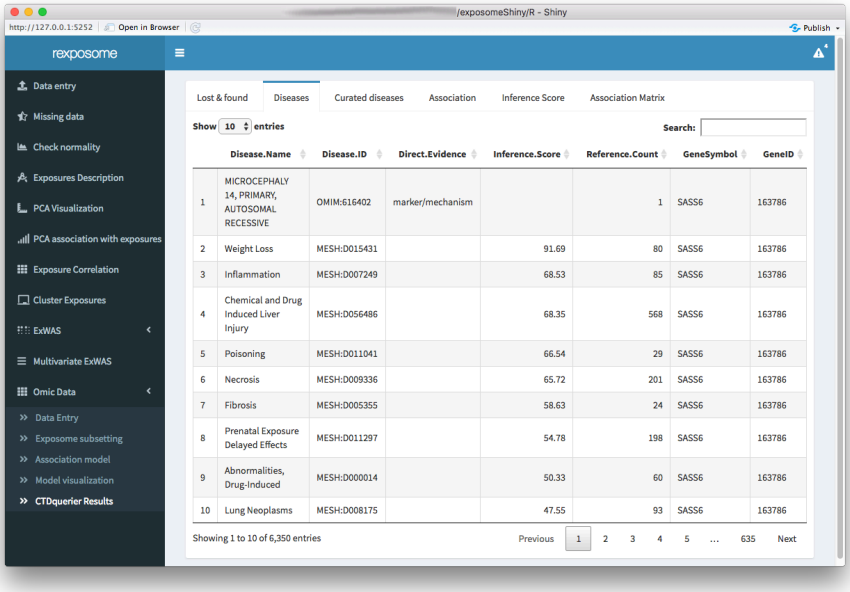

If by mistake some gene (or genes) were introduced to the querier, select them by clicking on the table row and click “Remove from querier”. Click on “Query selected genes on the CTD gene database” to perform the query of all the symbols of the querier list.



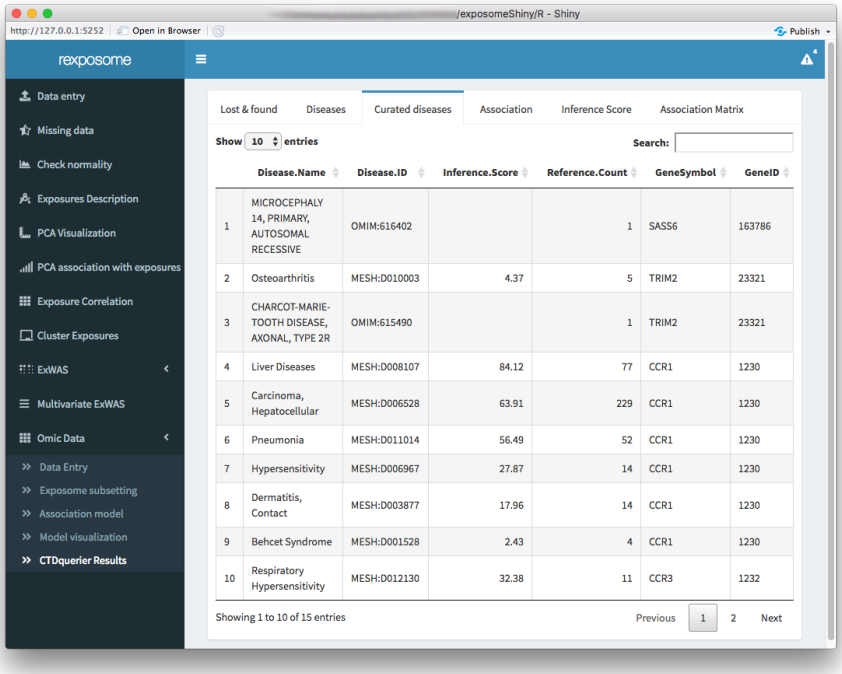

To visualize the results of the query, go to the "CTDquerier results" subtab. There are six tabs showing different results interpretations. First there's the "Lost & found" tab which a plot to see the amount of genes found on the CTD database and the ones that were not found them, ther's also two lists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



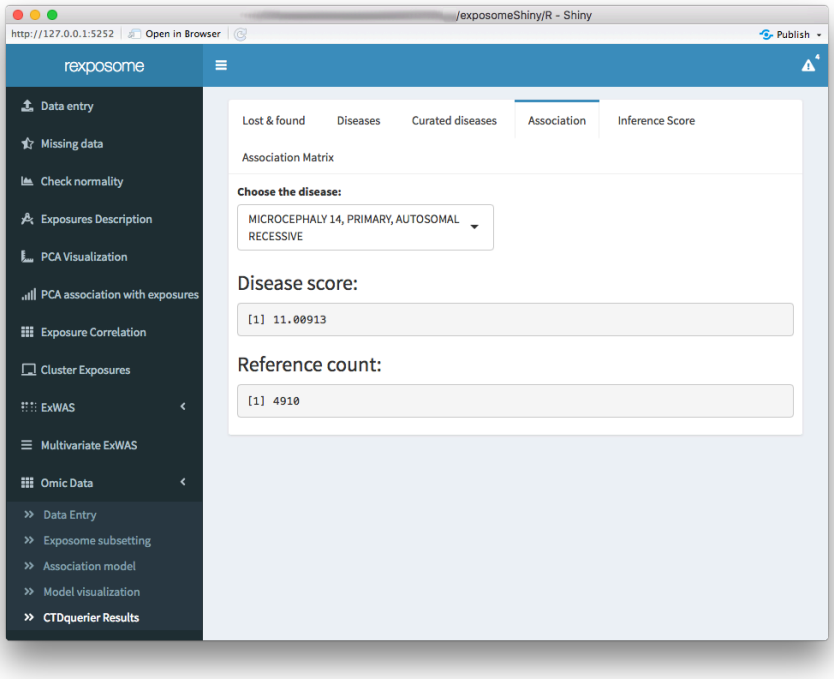

The diseases tab shows a table of all the associated diseases found on the CTD database.



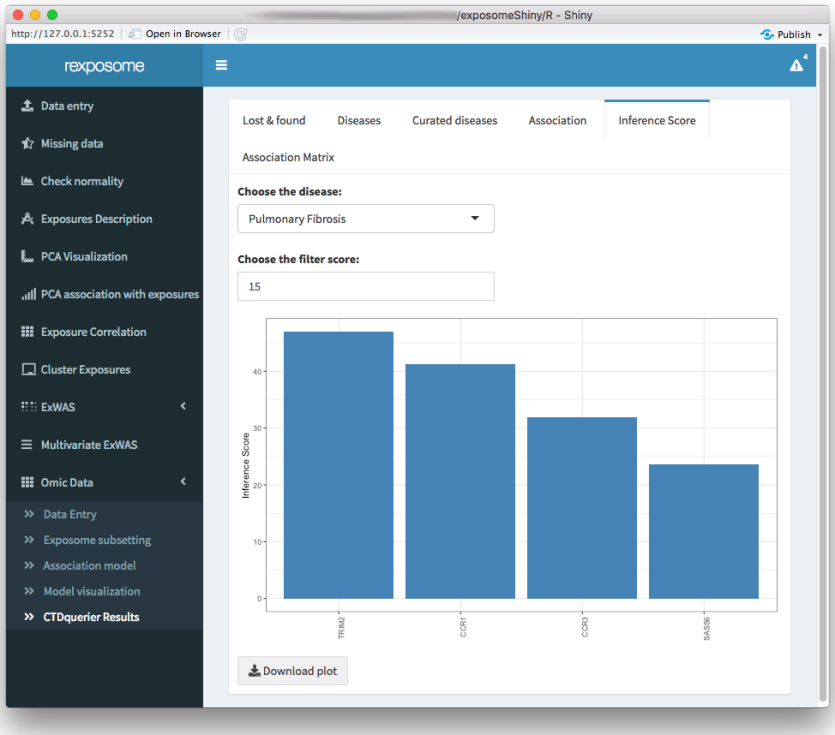

The curated diseases tab shows the table of associated diseases but only shows the ones with direct evidence.



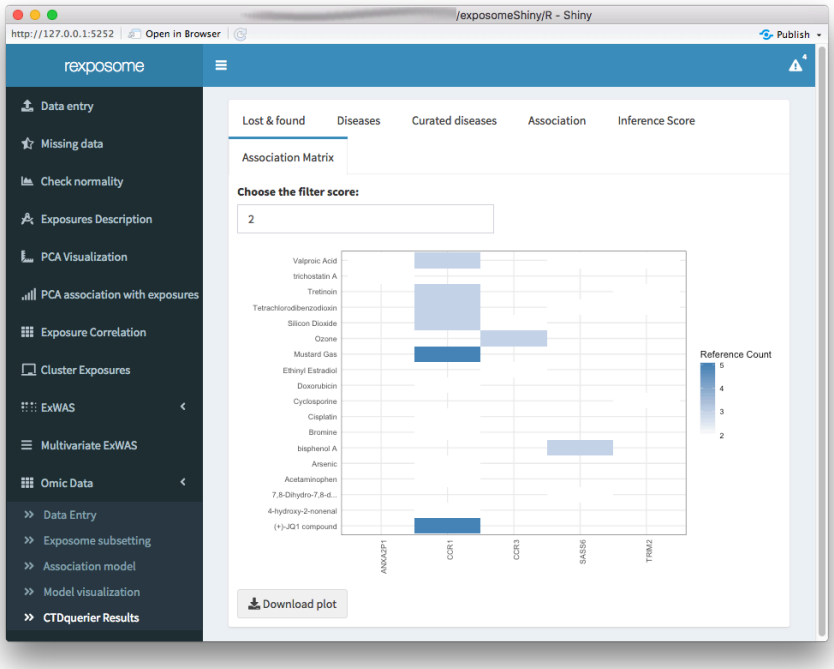

The association tab shows information about all the direct evidence associated diseases. Select the disease of interest to see the score and reference count of it.




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.



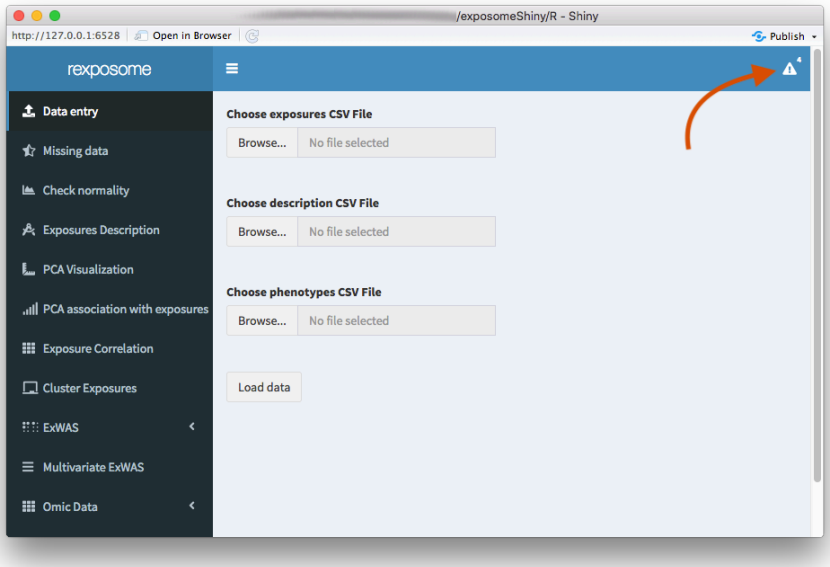

The association matrix tab shows a matrix of genes vs. chemicals with a heatmap representing the existing papers (references) providing evidence about the association between chemicals and genes.


/exposomeShiny/R - Shiny
http://127.0.0.1:5252
Open in Browser
Publish
rexposome
Data entry
Missing data
Check normality
Exposures Description
PCA Visualization
PCA association with exposures
Exposure Correlation
Cluster Exposures
ExWAS
Multivariate ExWAS
Omic Data
Data Entry
Exposome subsetting
Association model
Model visualization
CTDquerier Results
Lost & found
Diseases
Curated diseases
Association
Inference Score
Association Matrix
Choose the filter score:
2
Valproic Acid
trichostatin A
Tretinoin
Tetrachlorodibenzodioxin
Silicon Dioxide
Ozone
Mustard Gas
Ethinyl Estradiol
Doxorubicin
Cyclosporine
Cisplatin
Bromine
bisphenol A
Arsenic
Acetaminophen
7,8-Dihydro-7,8-d...
4-hydroxy-2-nonenal
(+)-JQ1 compound
ANXA2P1
CCR1
CCR3
SAS56
TRIM2
Reference Count
5
4
3
2
Download plot

# Chapter 6

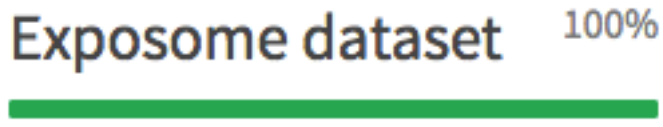
# General application functionalities

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:

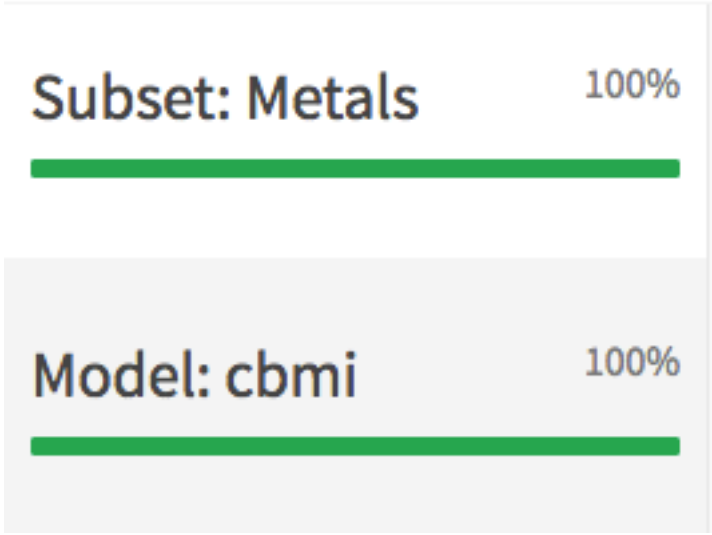
- Exposome dataset: Turns to 100% when an exposome dataset is loaded in the environment. Here's a graphical example.


- 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 Multivariate Association Analysis

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.

# 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.

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

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.