exposomeShiny User's Guide

Escribà Montagut, Xavier; González, Juan R.

2020-03-31

Contents

Chapter 1

Overview

ATHLETE
The Exposome from Evidence to Translation
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS
R
Shiny
rexposome
omicRexposome
CTDquerier

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

In order to download and setup the environment to launch exposomeShiny, the latest version of the software has to be downloaded from GitHub. To do so open a new RStudio session. Run the following code on the console once the working directory has been setup.

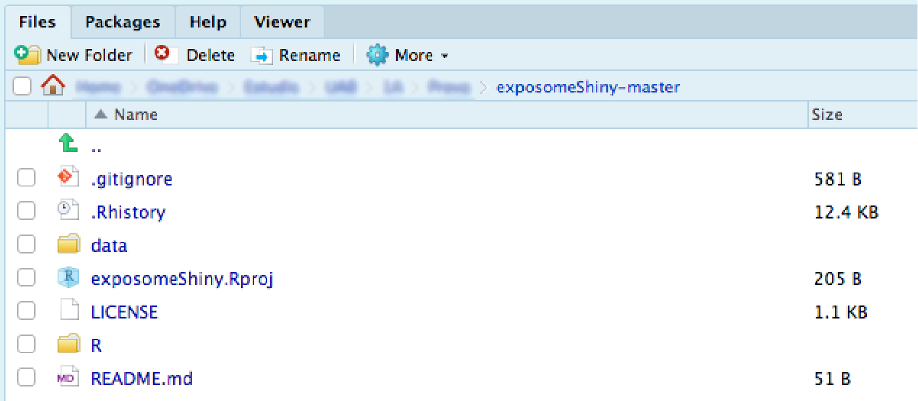
```
  # Set working directory
setwd(dir = "/some/path/")

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

  # 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.



[redacted] exposomeShiny-master

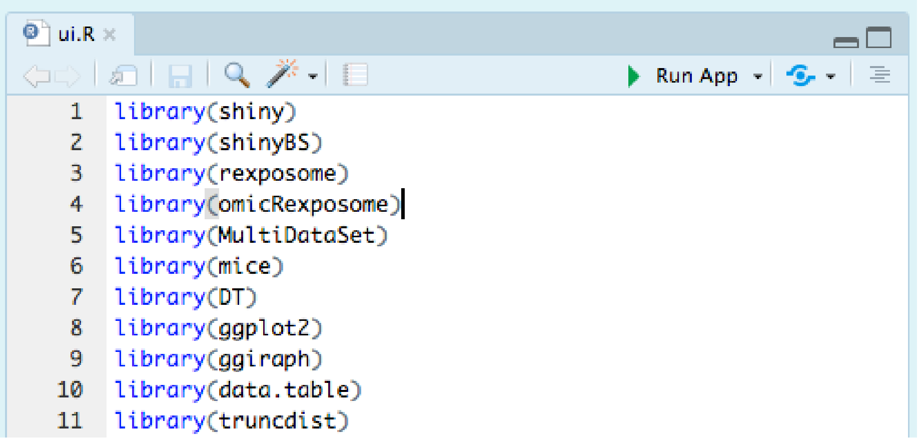
Name	Size
..	
.gitignore	581 B
.Rhistory	12.4 KB
data	
exposomeShiny.Rproj	205 B
LICENSE	1.1 KB
R	
README.md	51 B

Once the project is loaded, run the following code to install `renv` if it's not present on your R packages library and load the dependencies of exposomeShiny on this R session.

```r
install.packages("renv")

renv::restore()
```

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`.



```r
library(shiny)
library(shinyBS)
library(rexposome)
library(omicRexposome)
library(MultiDataSet)
library(mice)
library(DT)
library(ggplot2)
library(ggiraph)
library(data.table)
library(truncdist)
```

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

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

Chapter 3

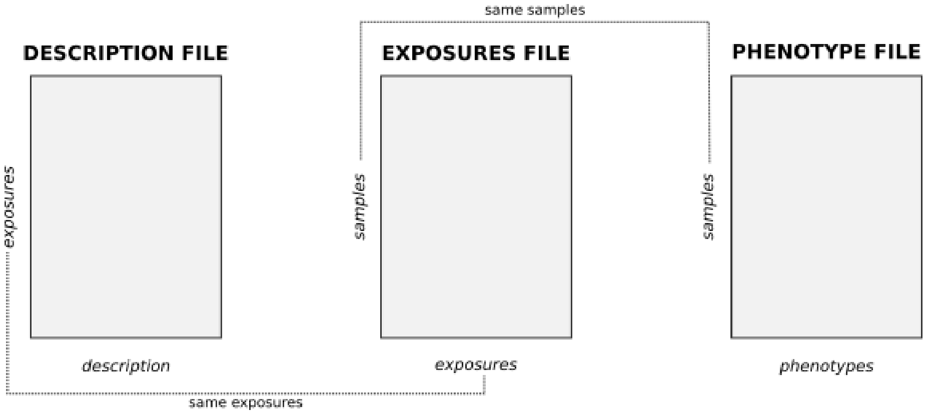
Data sets

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 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.

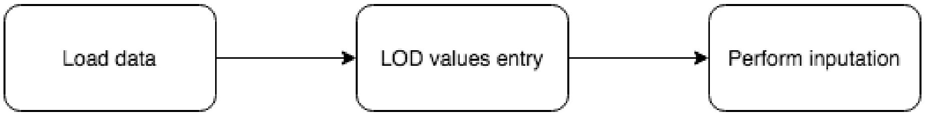
Chapter 5

Analysis flowcharts

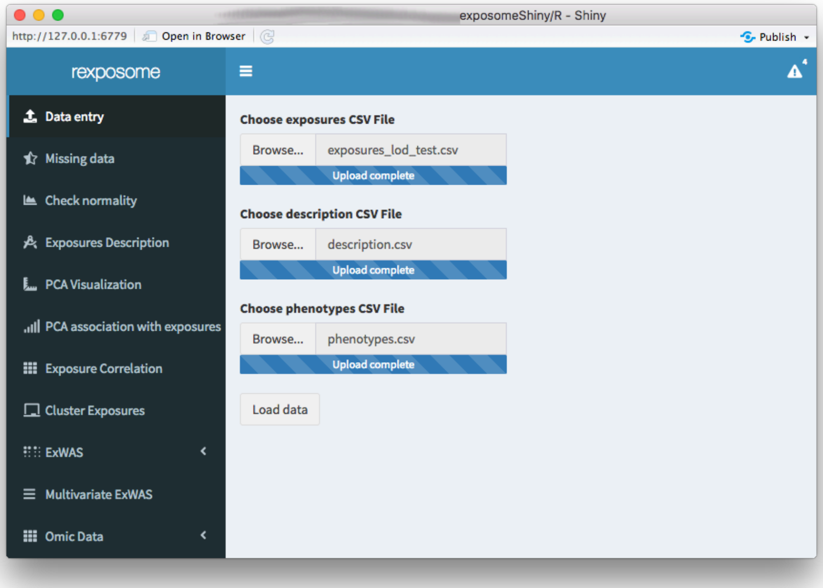
5.1 Exposome analysis

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 rexposomeShiny, 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 LOD imputation



Input the exposures, description and phenotypes files and load them into the application.

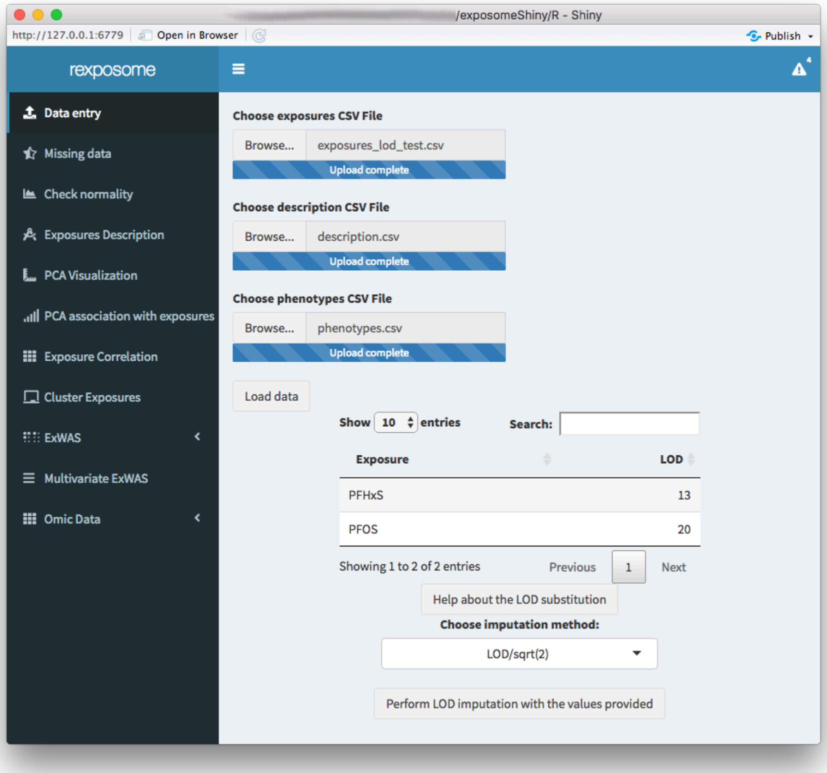


Choose exposures CSV File: exposures_lod_test.csv — Upload complete

Choose description CSV File: description.csv — Upload complete

Choose phenotypes CSV File: phenotypes.csv — Upload complete

If exposomeShiny detects LODs (limit of detection) on the exposures file (exposures with value: -1), 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 instead of LOD/sqrt(2).



Choose exposures CSV File: exposures_lod_test.csv (Upload complete)

Choose description CSV File: description.csv (Upload complete)

Choose phenotypes CSV File: phenotypes.csv (Upload complete)

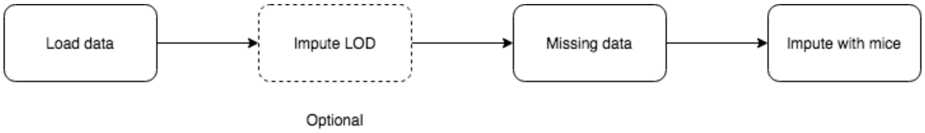
Show 10 entries Search:

Exposure	LOD
PFHxS	13
PFOS	20

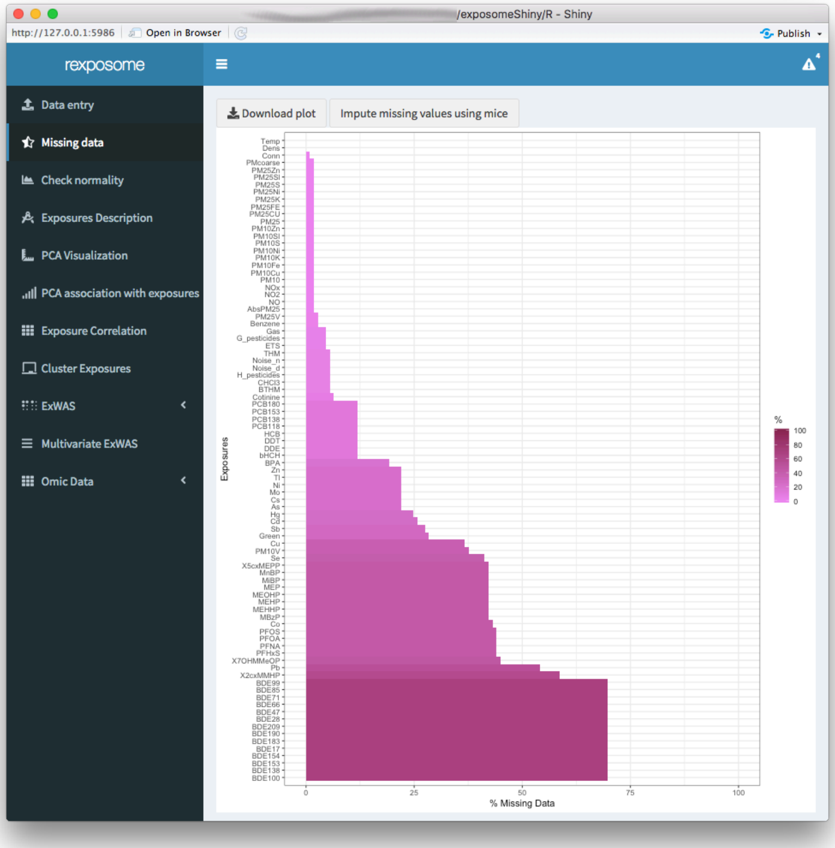
Showing 1 to 2 of 2 entries

Choose imputation method: LOD/sqrt(2)

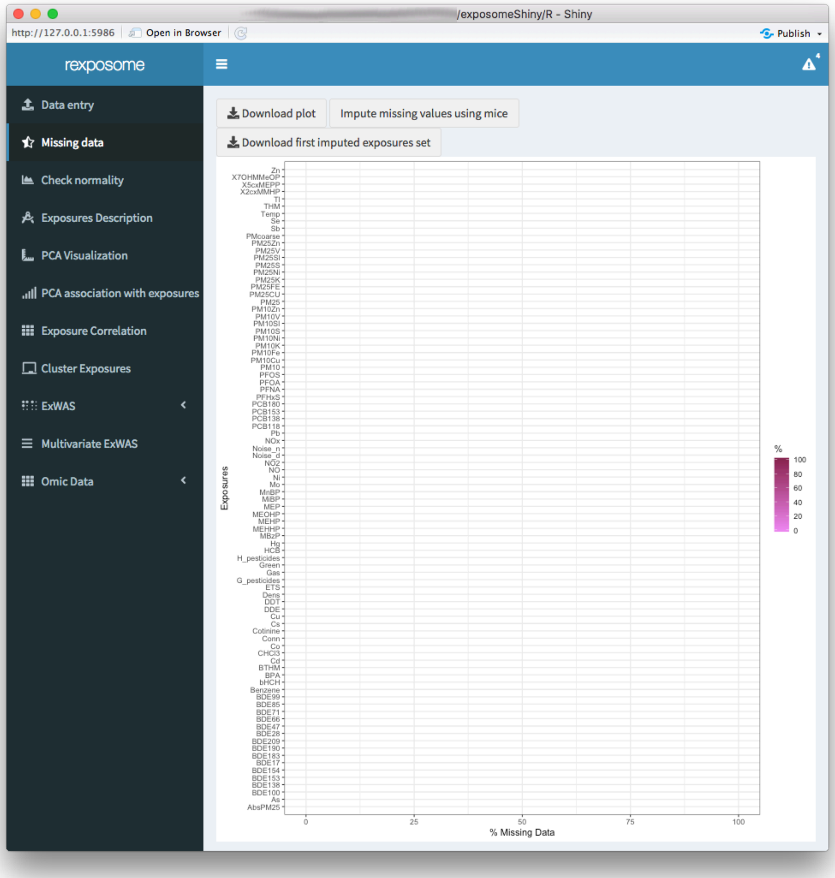
5.1.2 Missing imputation



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.

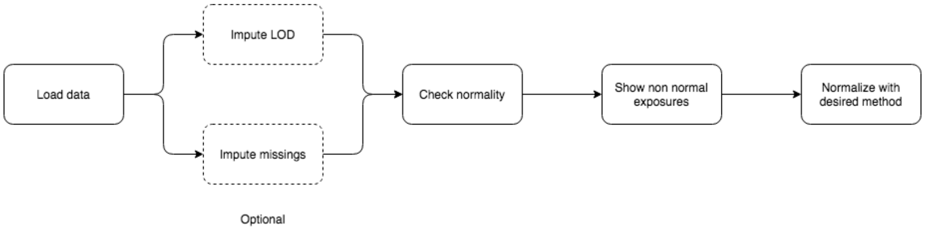


To impute the missing values select "Impute missing values using mice". After the process finishes, the expect output should be a new missing data graph where there's no missing of any exposure.

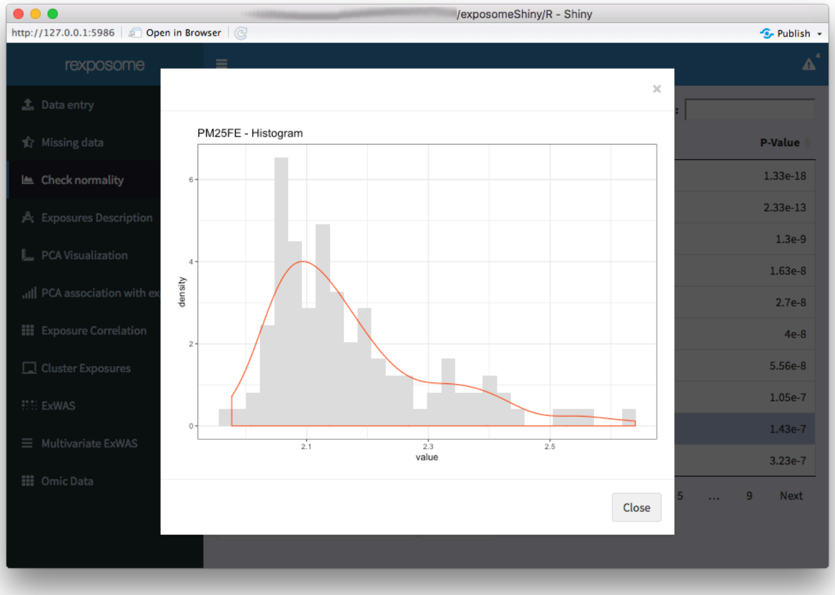


The new imputed exposures set can be downloaded as a `*.csv` file, please note that the downloaded file just assigns numbers to the `idnum` column, if the data you are using has different `idnum` format it's needed to format it properly so that it matches the `idnum` on the phenotypes input file when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

5.1.3 Normality correction

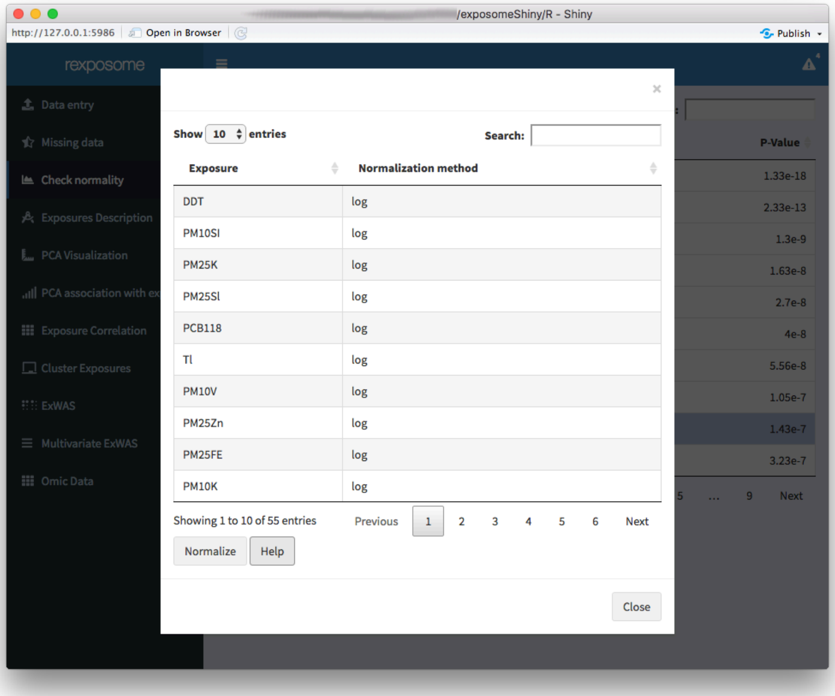


Once the dataset is loaded into the Shiny, look at the “Check Normality” tab to check which exposures are not normal (Normality = false). By selection from the table the desired exposure and clicking the “Plot histogram of selected exposure”, as the label of the button implies, a histogram of the selected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



P-Value
1.33e-18
2.33e-13
1.3e-9
1.63e-8
2.7e-8
4e-8
5.56e-8
1.05e-7
1.43e-7
3.23e-7

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), “^1/3” and “sqrt”. If no method is desired to be applied to an exposure input “none”.



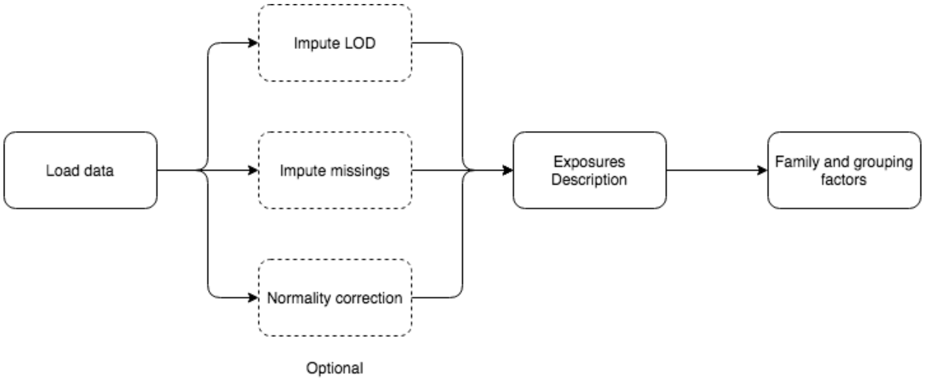
Show 10 entries Search:

Exposure	Normalization method
DDT	log
PM10SI	log
PM25K	log
PM25SI	log
PCB118	log
Tl	log
PM10V	log
PM25Zn	log
PM25FE	log
PM10K	log

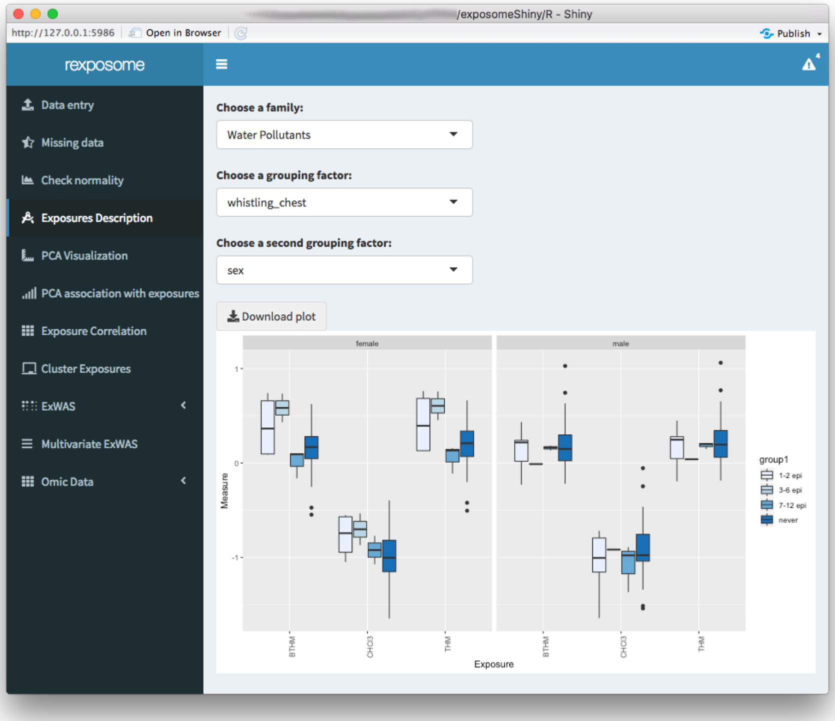
Showing 1 to 10 of 55 entries

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.

5.1.4 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).

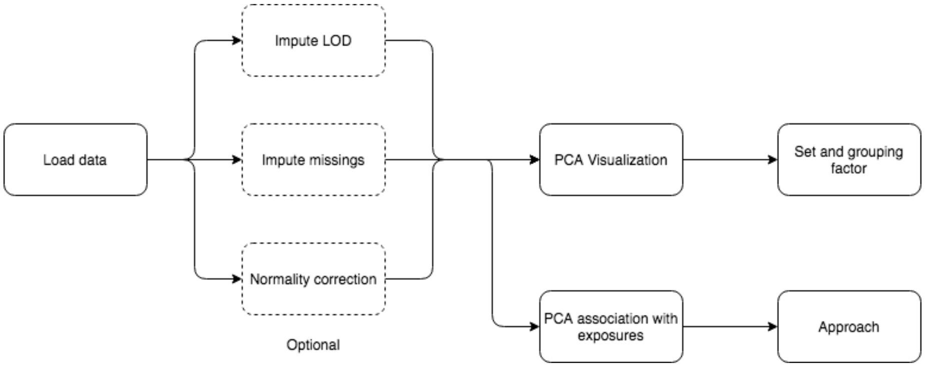


Choose a family:
Water Pollutants

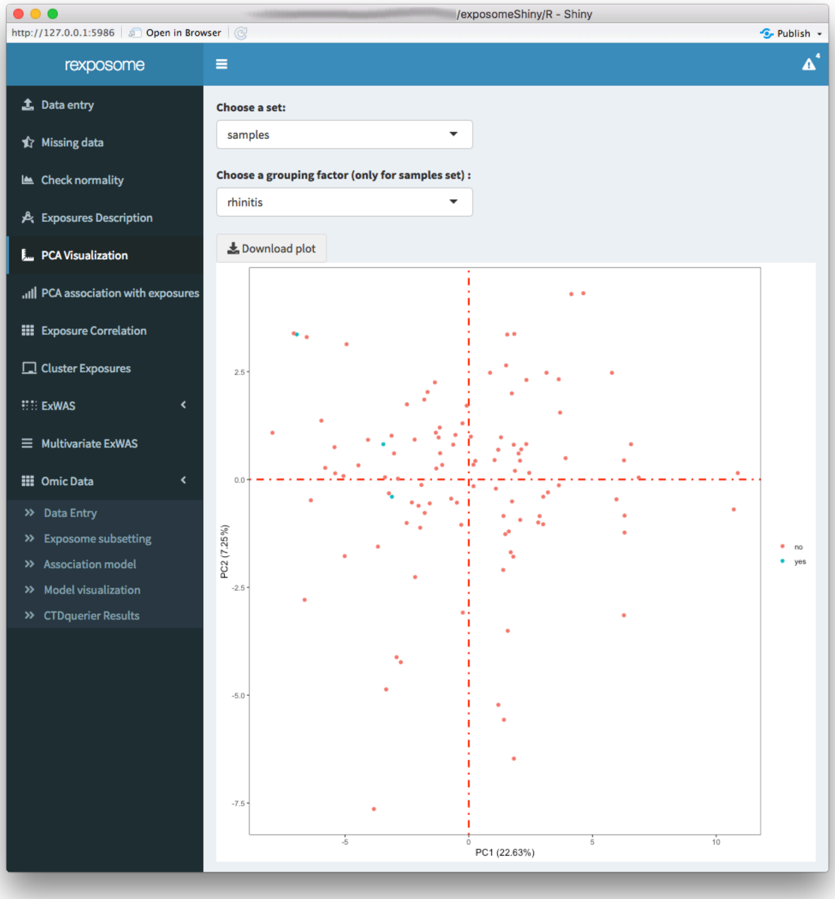
Choose a grouping factor:
whistling_chest

Choose a second grouping factor:
sex

5.1.5 PCA Analysis



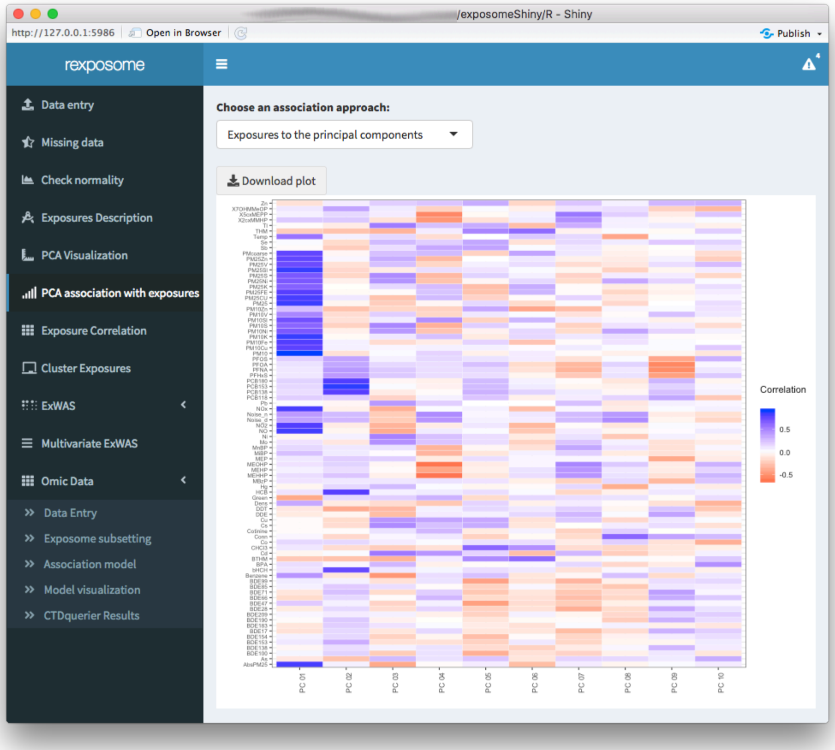
To see the results of a PCA (principal component analysis) study, load the data and check the PCA Visualization tab, there a set and grouping factor can be choose, it's important noting (as it's already stated on the Shiny) that the grouping parameter only works when the set is selected to "samples".



Choose a set: samples

Choose a grouping factor (only for samples set) : rhinitis

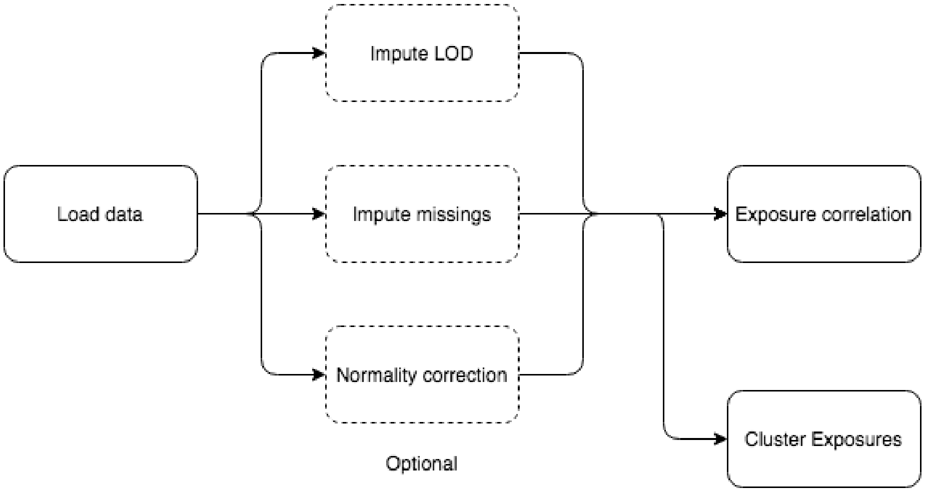
If the association of the PCA analysis with the exposures is desired to visualize, check the "PCA association with exposures" tab, there are two grouping methods to visualize, the phenotypes to principal components and the exposures to principal components.



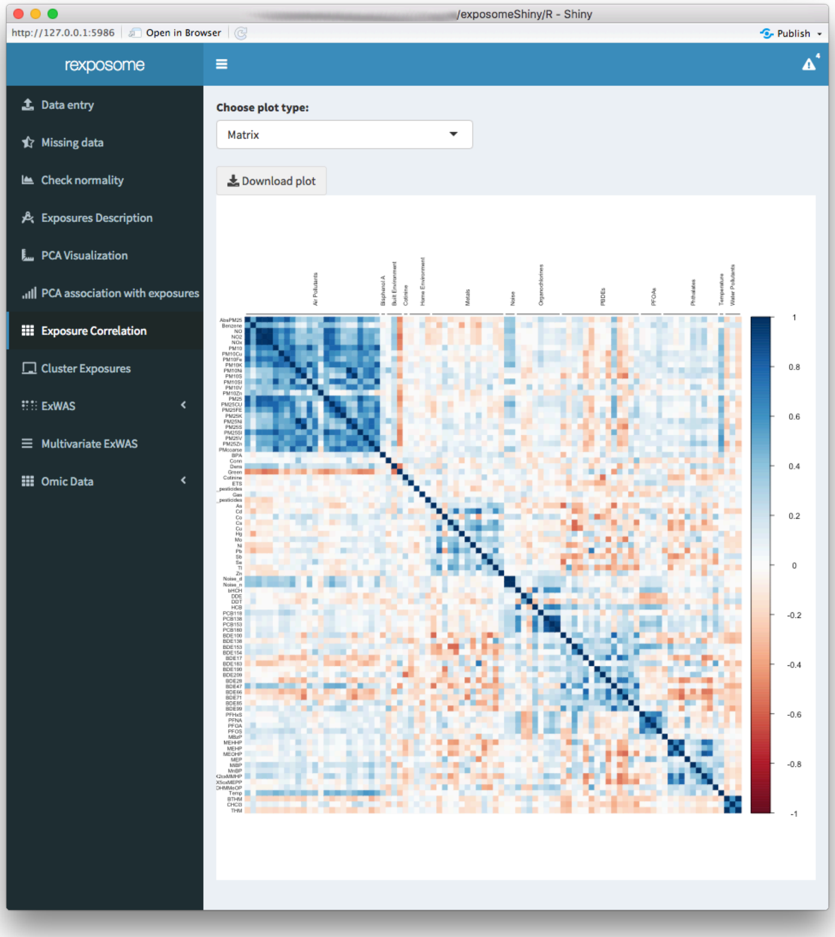
Choose an association approach:

Exposures to the principal components

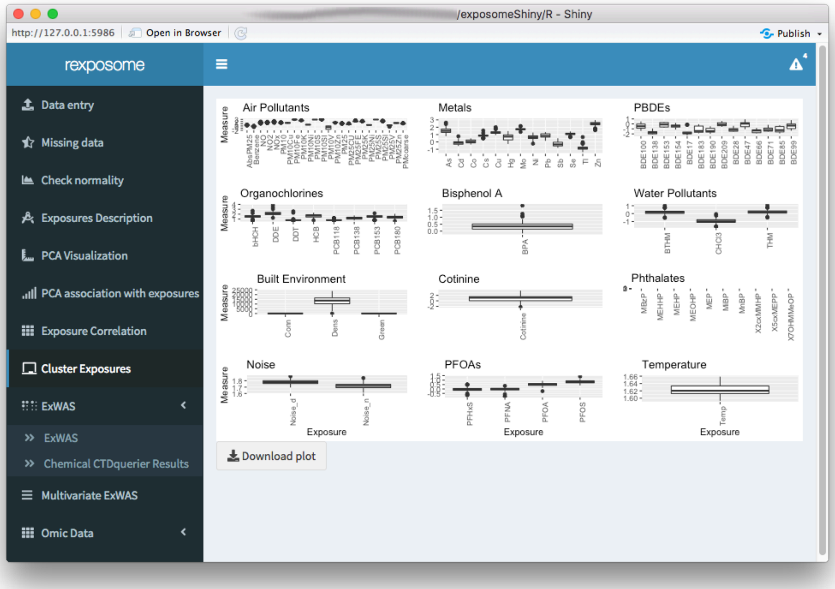
5.1.6 Clusterization and correlation of exposures



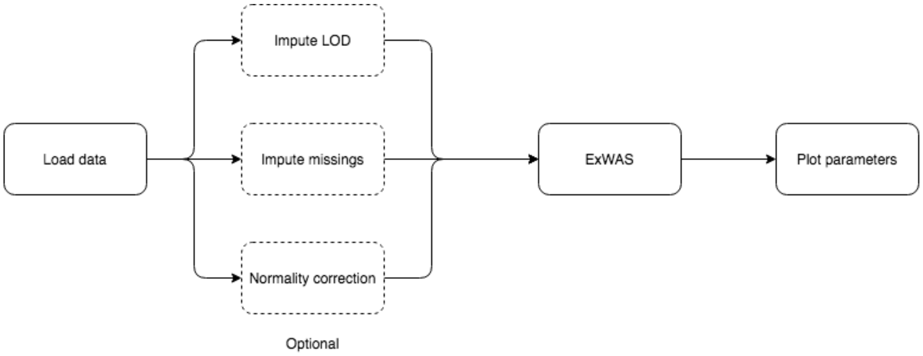
To see the results of the exposure correlation and clustering, select the corresponding tab to each analysis. For the exposure correlation analysis there are two visualizations, the matrix representation and the circos.



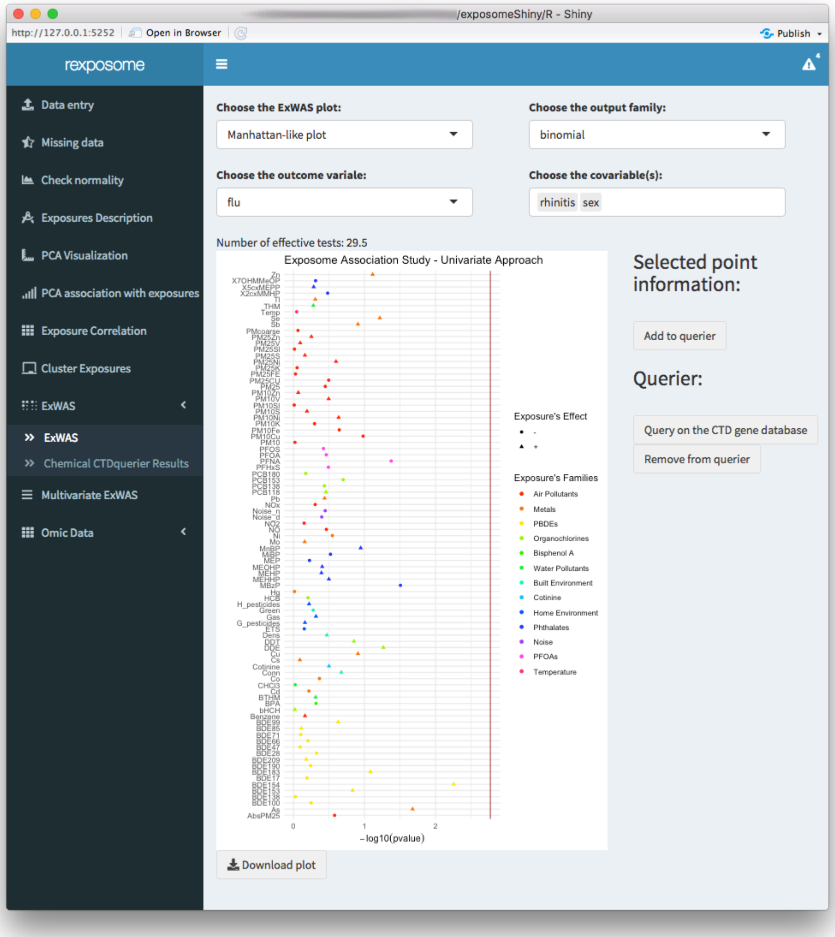
Choose plot type: Matrix



5.1.7 ExWAS



To perform an ExWAS (exposome-wide association) 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.



Choose the ExWAS plot: Manhattan-like plot

Choose the output family: binomial

Choose the outcome variale: flu

Choose the covariable(s): rhinitis sex

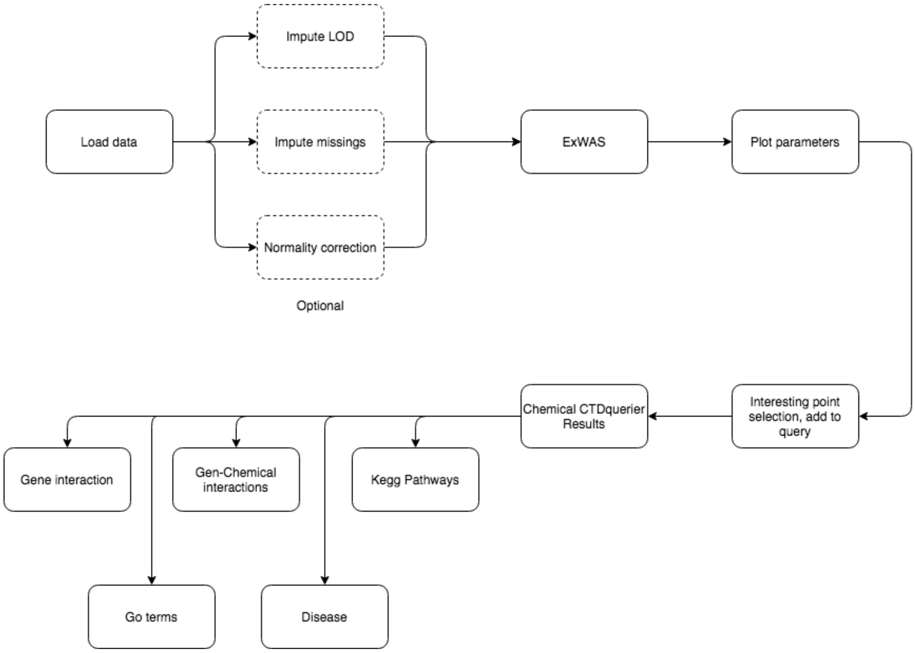
Number of effective tests: 29.5

Selected point information:

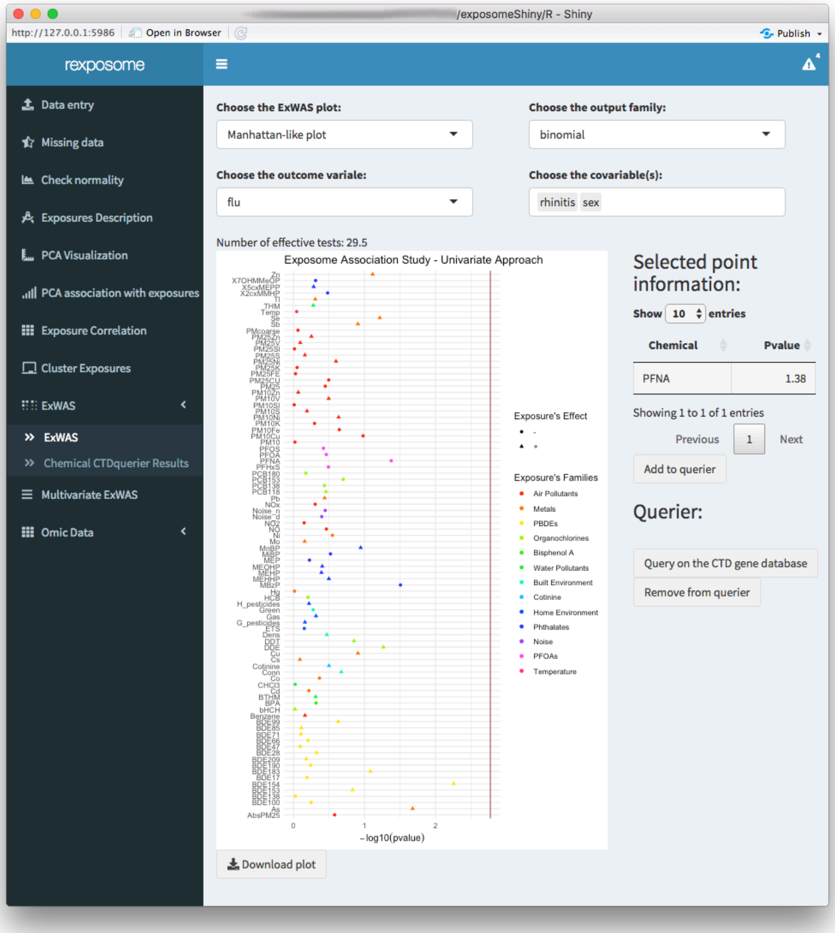
Querier:

5.1.8 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".



Choose the ExWAS plot: Manhattan-like plot

Choose the output family: binomial

Choose the outcome variale: flu

Choose the covariable(s): rhinitis sex

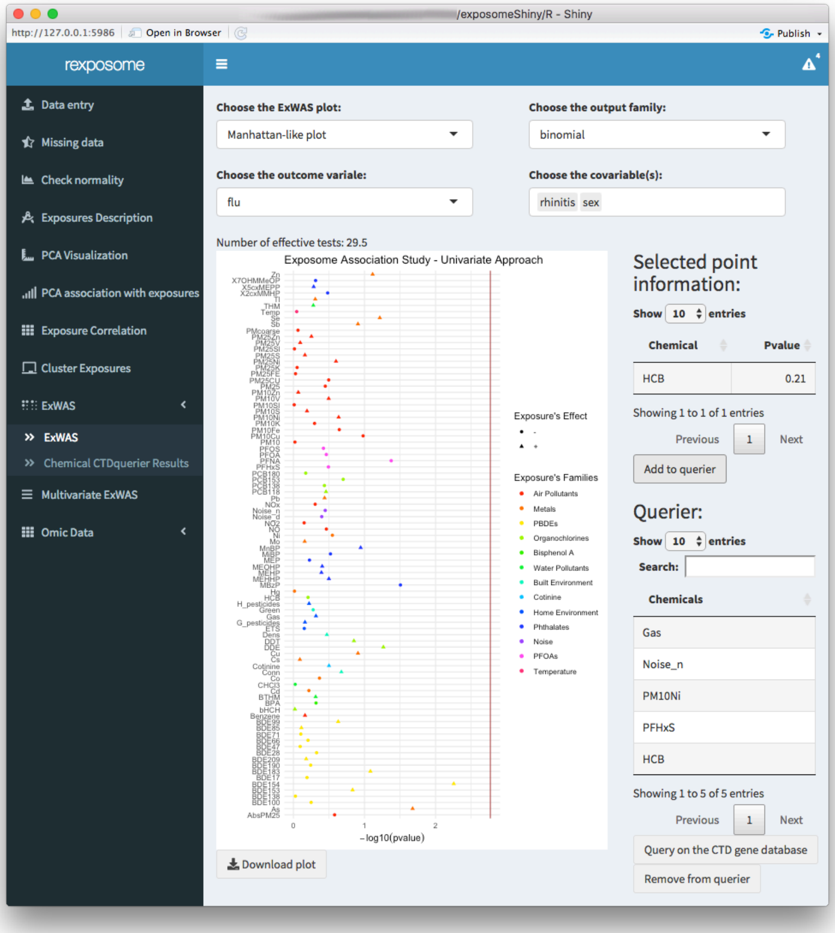
Number of effective tests: 29.5

Selected point information:

Show 10 entries

Chemical	Pvalue
PFNA	1.38

Showing 1 to 1 of 1 entries



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

Selected point information:

Show 10 entries

Chemical	Pvalue
HCB	0.21

Showing 1 to 1 of 1 entries

Querier:

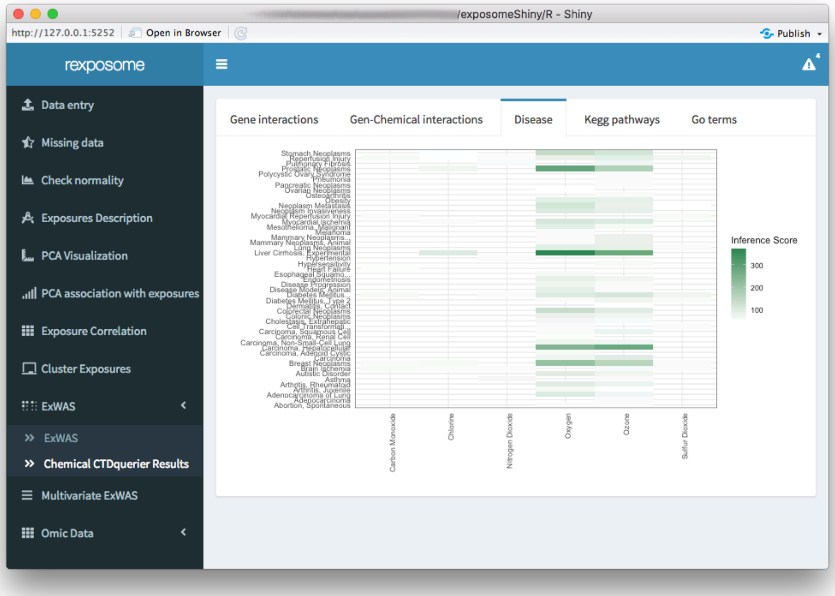
Show 10 entries

Search:

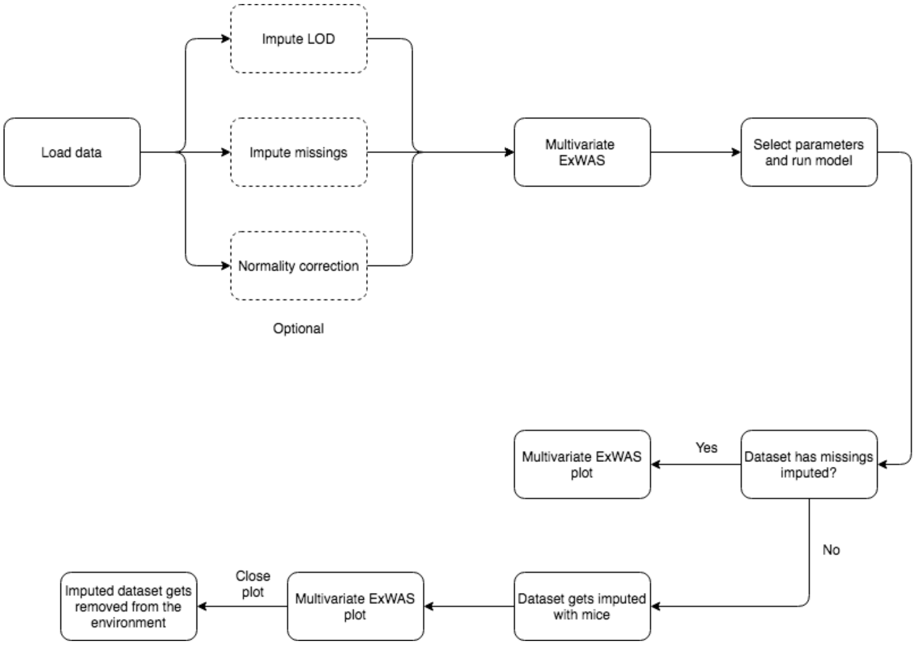
Chemicals
Gas
Noise_n
PM10Ni
PFHxS
HCB

Showing 1 to 5 of 5 entries

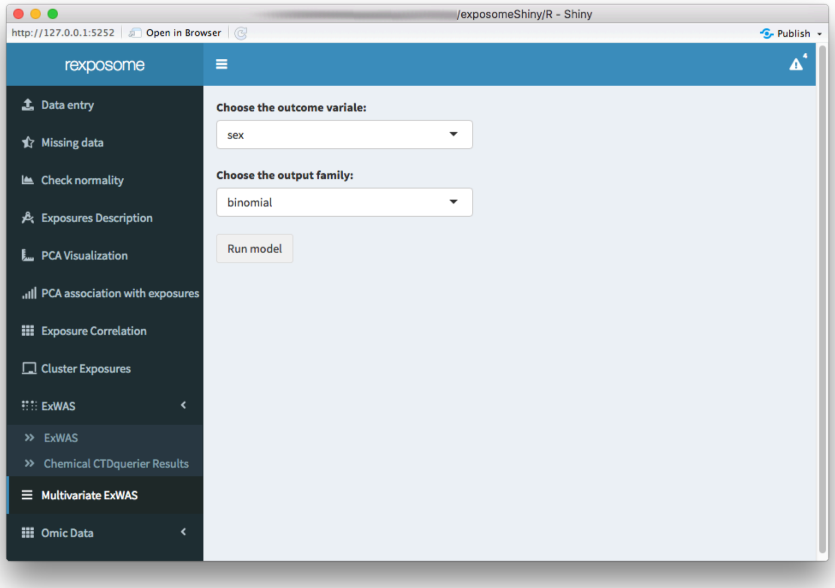
To do the query of the chemicals to de CTD database click on “Query on the CTD gene database” and see the results on the “Chemical CTDquerier Results” subtab. It’s important noting that on the “Kegg pathways” and “Go terms” the input field corresponds to the negative exponent of the filter.



5.1.9 Multivariate ExWAS

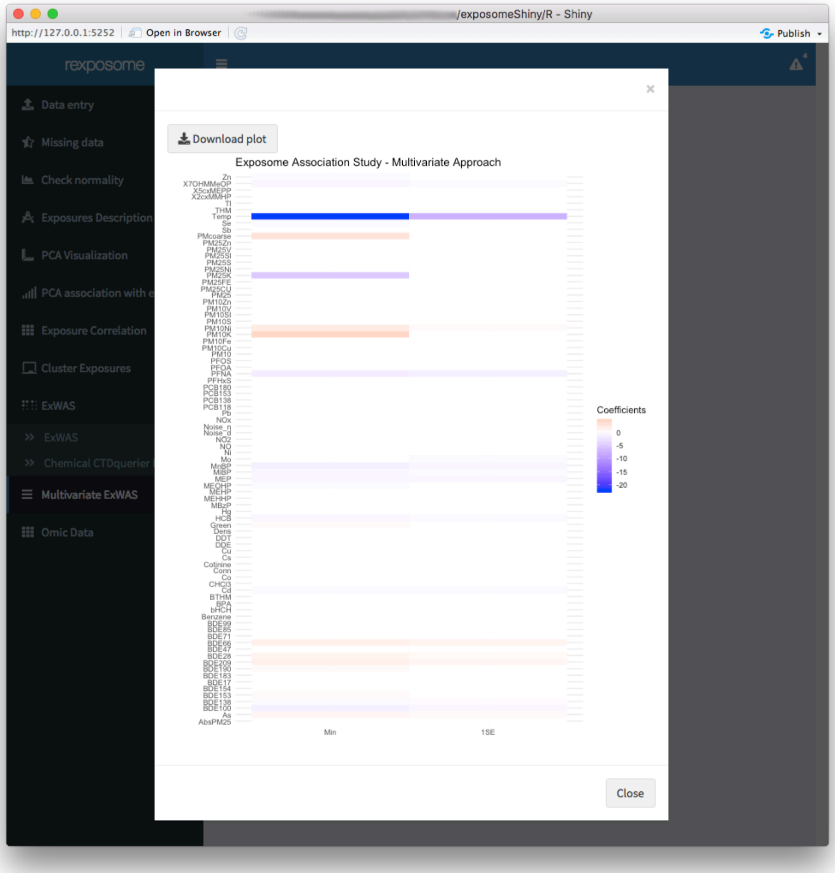


To perform a multivariate ExWAS study, check the Multivariate ExWAS tab and select the desired output parameter, click on run model to generate the plot. As on the ExWAS plot options there's implemented an internal check to advise the user on which parameters to select depending if the selected outcome 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.



Choose the outcome variale: sex

Choose the output family: binomial

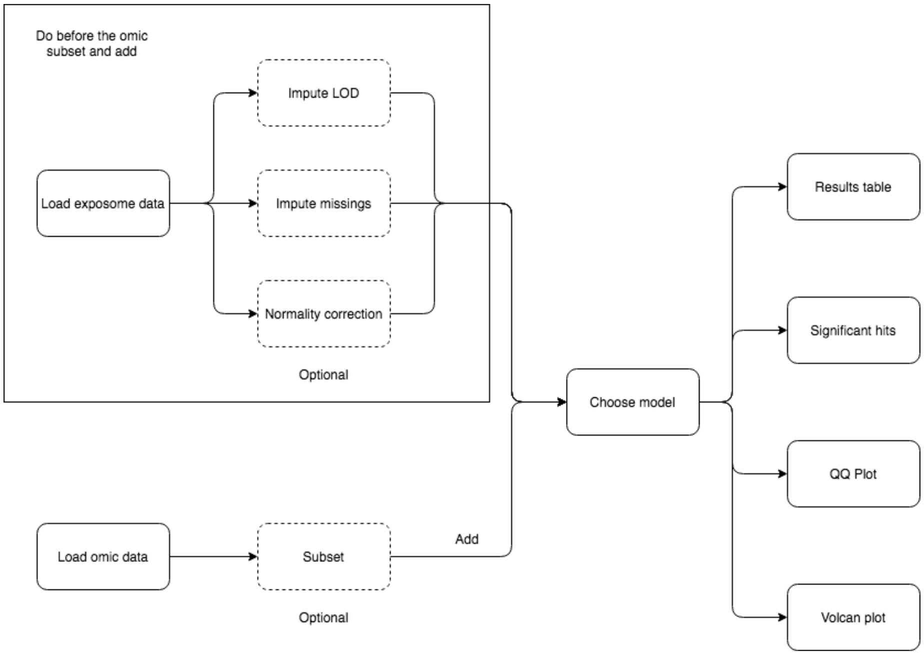


5.2 Exposome-Omic analysis

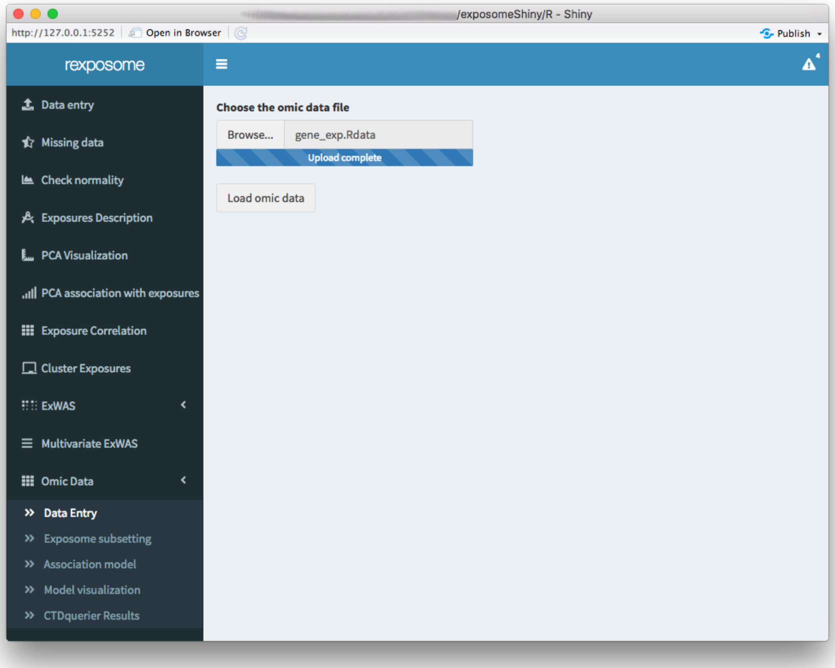
It's important noting that the maximum size of the omics data is 30 MB, if the omics file to be analyzed is bigger, change the line number 2 of the `server.R` file.

```
  # 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 of the “Omic Data” tab. The omic data should be provided as a `*.RData` file.

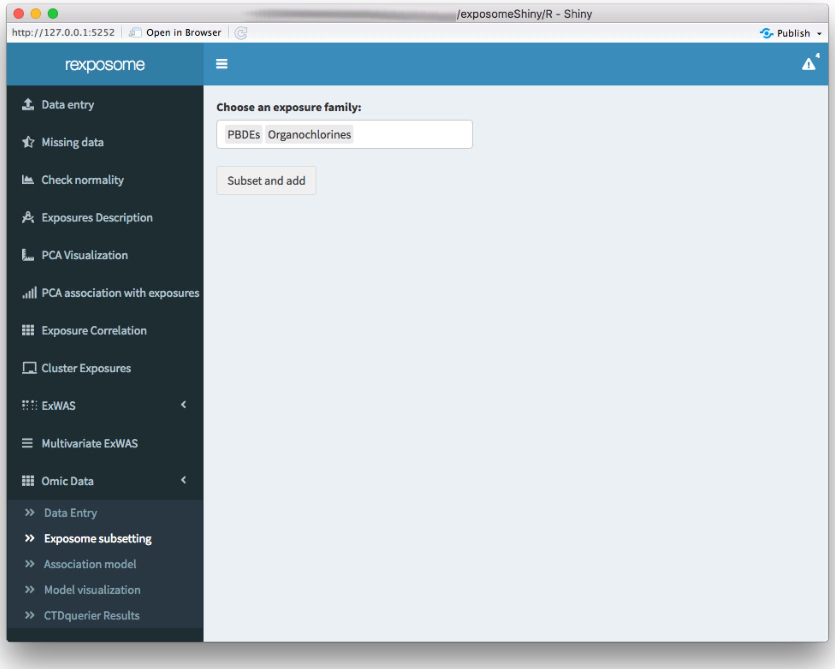


Choose the omic data file

gene_exp.Rdata

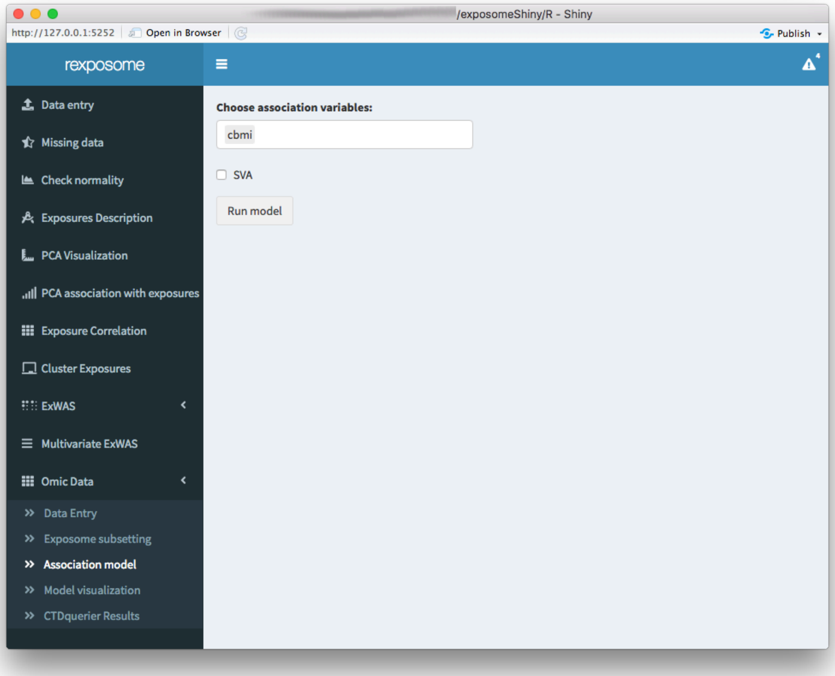
Upload complete

The exposome dataset can be subseted by families, 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.



Choose an exposure family: PBDEs Organochlorines

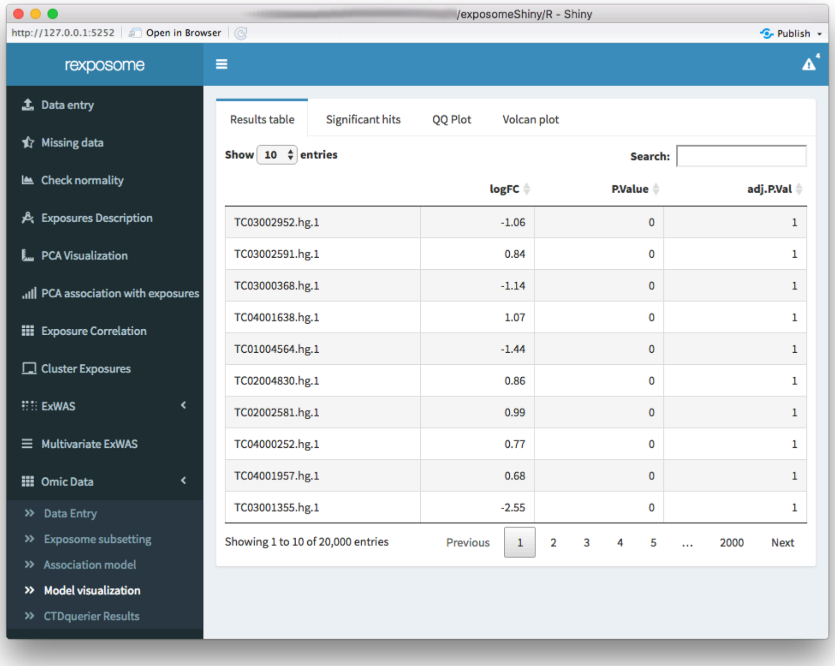
Select the variables for the association analysis and if SVA is wanted on the "Association model" subtab.



Choose association variables: cbmi

□ SVA

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.

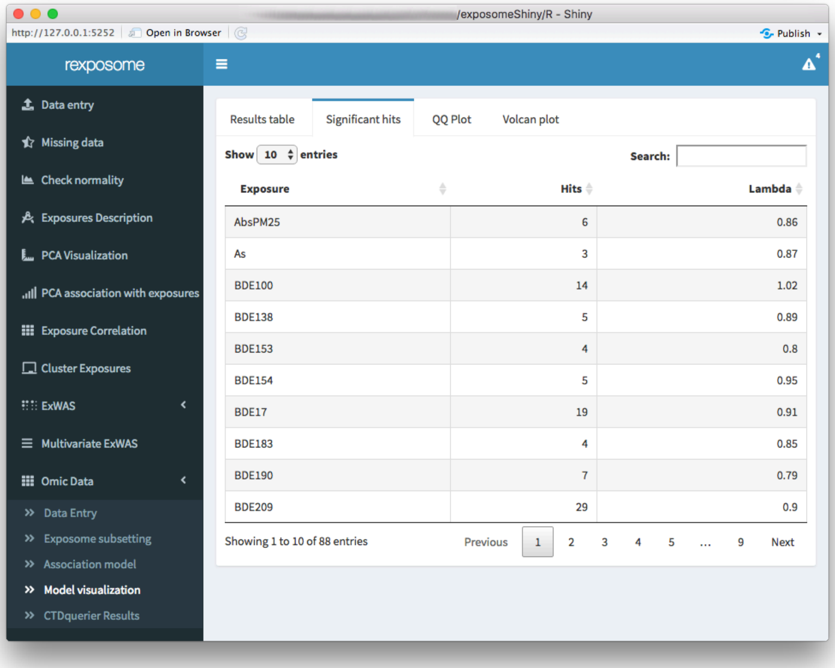


Show 10 entries Search:

	logFC	P.Value	adj.P.Val
TC03002952.hg.1	-1.06	0	1
TC03002591.hg.1	0.84	0	1
TC03000368.hg.1	-1.14	0	1
TC04001638.hg.1	1.07	0	1
TC01004564.hg.1	-1.44	0	1
TC02004830.hg.1	0.86	0	1
TC02002581.hg.1	0.99	0	1
TC04000252.hg.1	0.77	0	1
TC04001957.hg.1	0.68	0	1
TC03001355.hg.1	-2.55	0	1

Showing 1 to 10 of 20,000 entries Previous 1 2 3 4 5 … 2000 Next

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

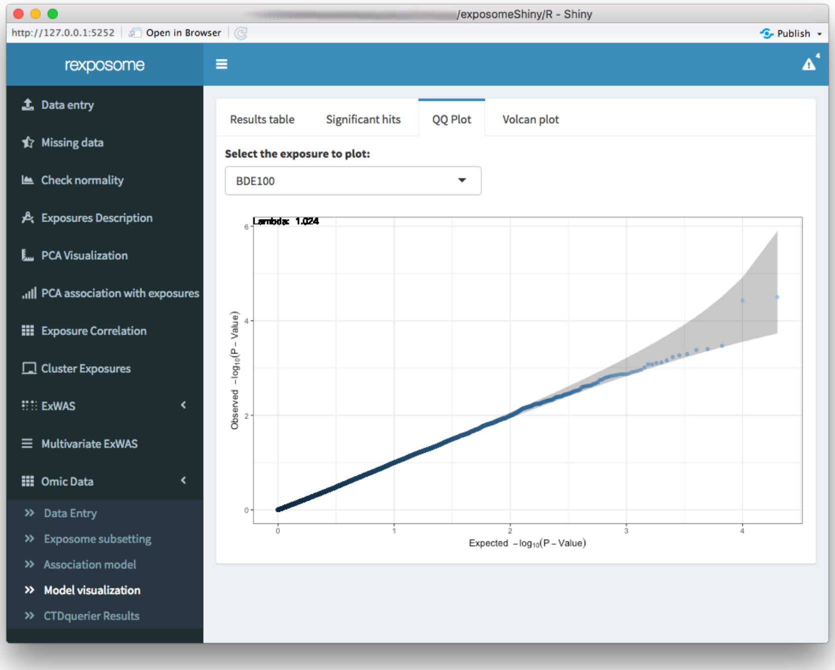


Show 10 entries

Exposure	Hits	Lambda
AbsPM25	6	0.86
As	3	0.87
BDE100	14	1.02
BDE138	5	0.89
BDE153	4	0.8
BDE154	5	0.95
BDE17	19	0.91
BDE183	4	0.85
BDE190	7	0.79
BDE209	29	0.9

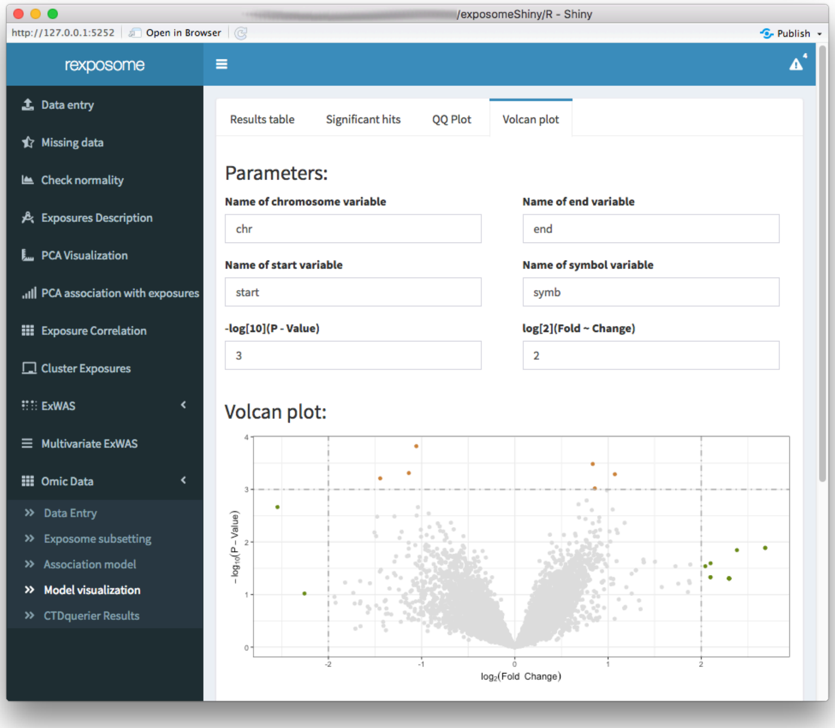
Showing 1 to 10 of 88 entries

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



Select the exposure to plot: BDE100

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.



Parameters:

Name of chromosome variable: chr

Name of end variable: end

Name of start variable: start

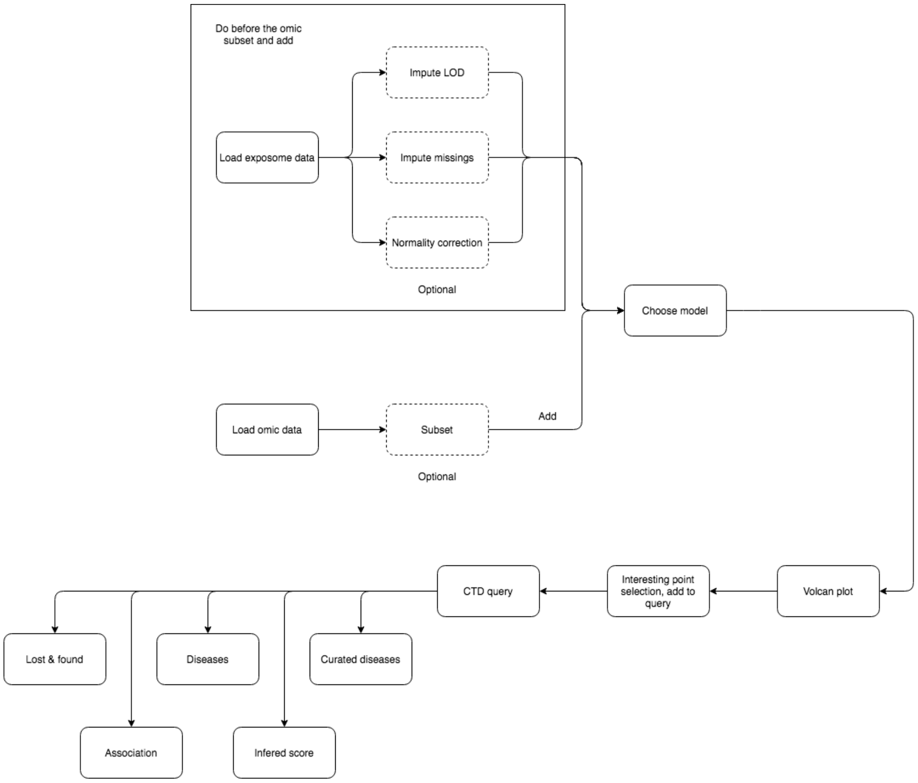
Name of symbol variable: symb

-log[10](P - Value): 3

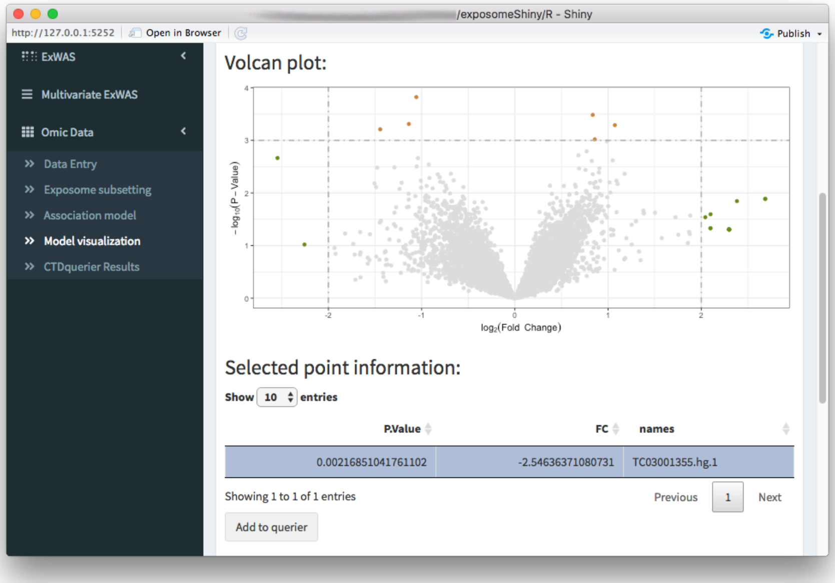
log[2](Fold ~ Change): 2

Volcan plot:

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.



Volcan plot:

Selected point information:

Show 10 entries

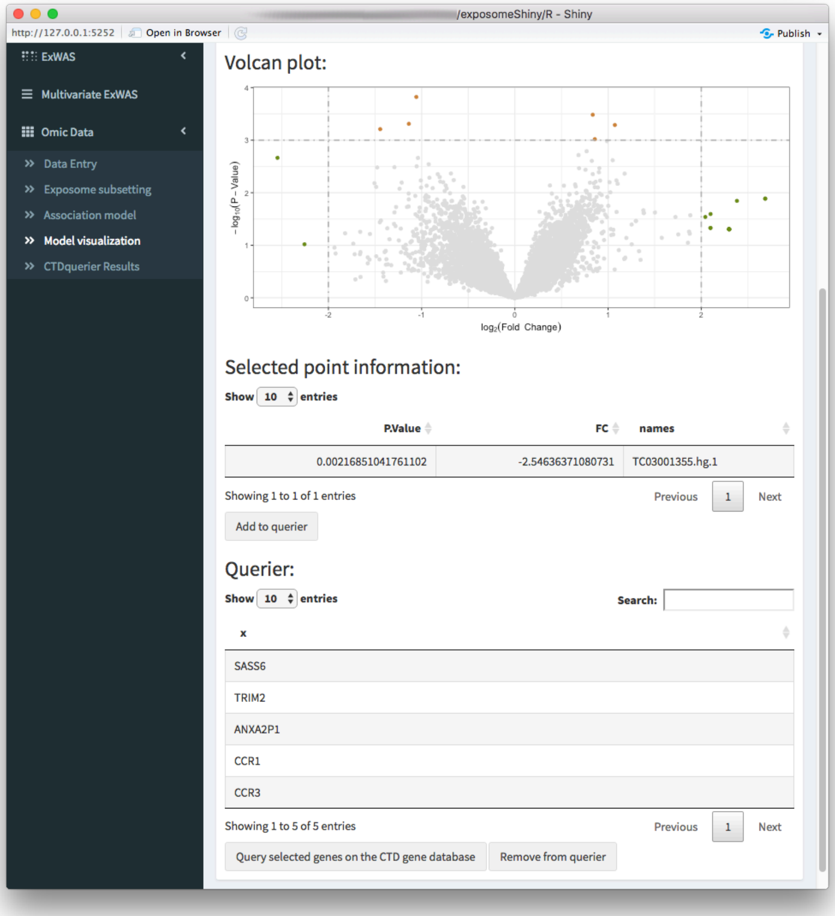
P.Value	FC	names
0.00216851041761102	-2.54636371080731	TC03001355.hg.1

Showing 1 to 1 of 1 entries

Previous 1 Next

Add to querier

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.



Volcan plot:

Selected point information:

Show 10 entries

P.Value	FC	names
0.00216851041761102	-2.54636371080731	TC03001355.hg.1

Showing 1 to 1 of 1 entries

Querier:

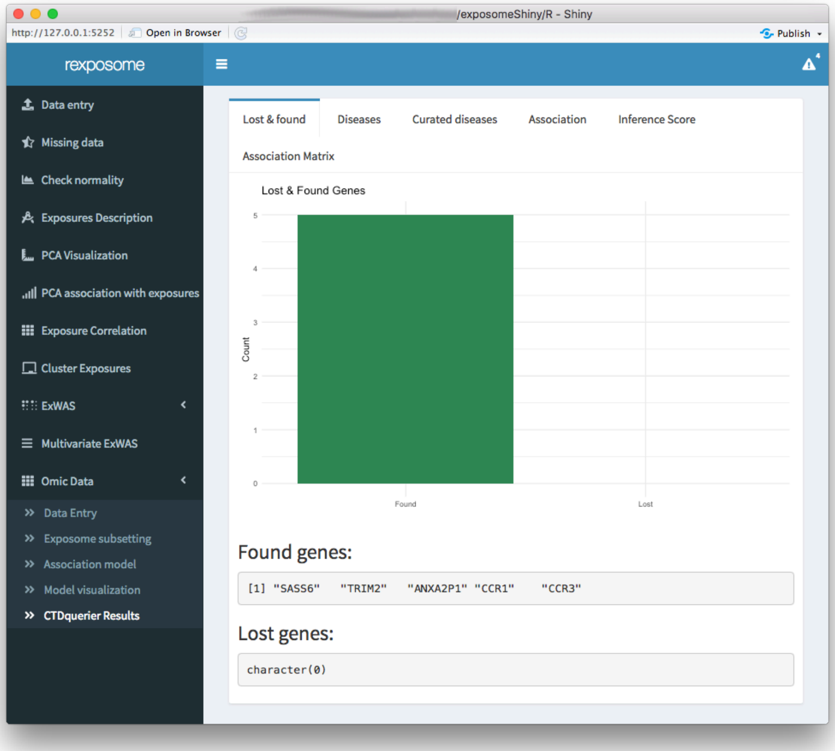
Show 10 entries

Search:

x
SASS6
TRIM2
ANXA2P1
CCR1
CCR3

Showing 1 to 5 of 5 entries

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.



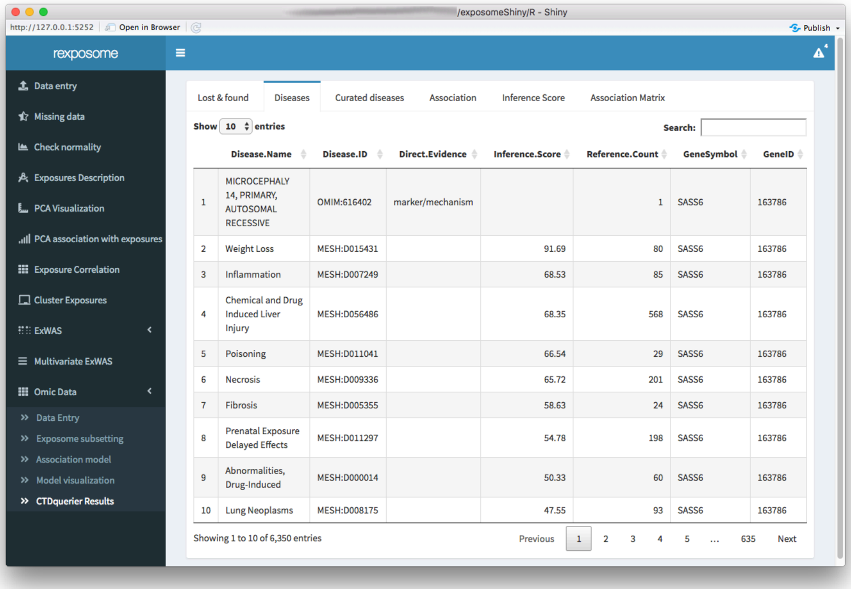
Found genes:

```
[1] "SASS6"   "TRIM2"   "ANXA2P1" "CCR1"    "CCR3"
```

Lost genes:

```
character(0)
```

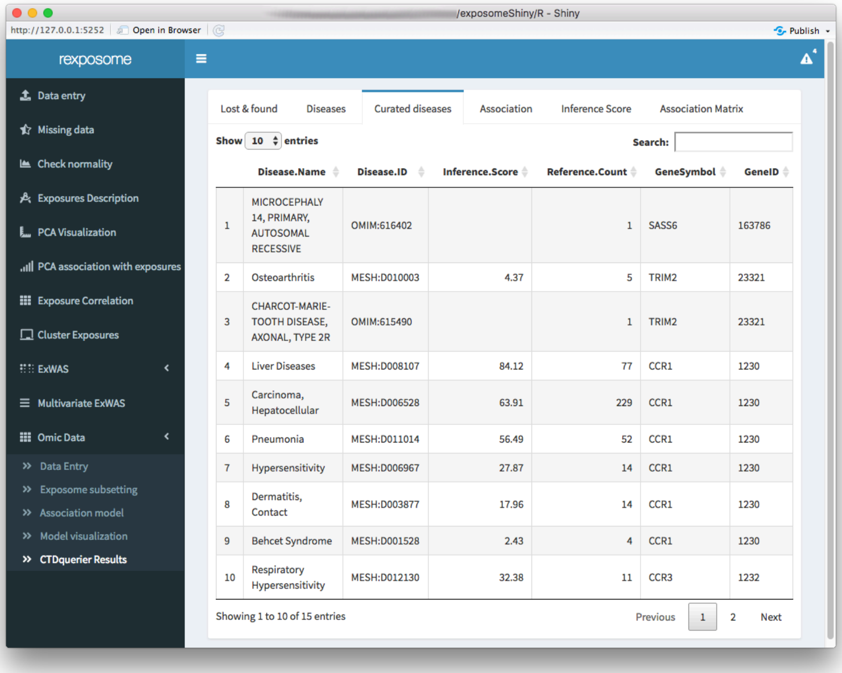
The diseases tab shows a table of all the associated diseases found on the CTD database.



	Disease.Name	Disease.ID	Direct.Evidence	Inference.Score	Reference.Count	GeneSymbol	GeneID
1	MICROCEPHALY 14, PRIMARY, AUTOSOMAL RECESSIVE	OMIM:616402	marker/mechanism		1	SASS6	163786
2	Weight Loss	MESH:D015431		91.69	80	SASS6	163786
3	Inflammation	MESH:D007249		68.53	85	SASS6	163786
4	Chemical and Drug Induced Liver Injury	MESH:D056486		68.35	568	SASS6	163786
5	Poisoning	MESH:D011041		66.54	29	SASS6	163786
6	Necrosis	MESH:D009336		65.72	201	SASS6	163786
7	Fibrosis	MESH:D005355		58.63	24	SASS6	163786
8	Prenatal Exposure Delayed Effects	MESH:D011297		54.78	198	SASS6	163786
9	Abnormalities, Drug-Induced	MESH:D000014		50.33	60	SASS6	163786
10	Lung Neoplasms	MESH:D008175		47.55	93	SASS6	163786

Showing 1 to 10 of 6,350 entries

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

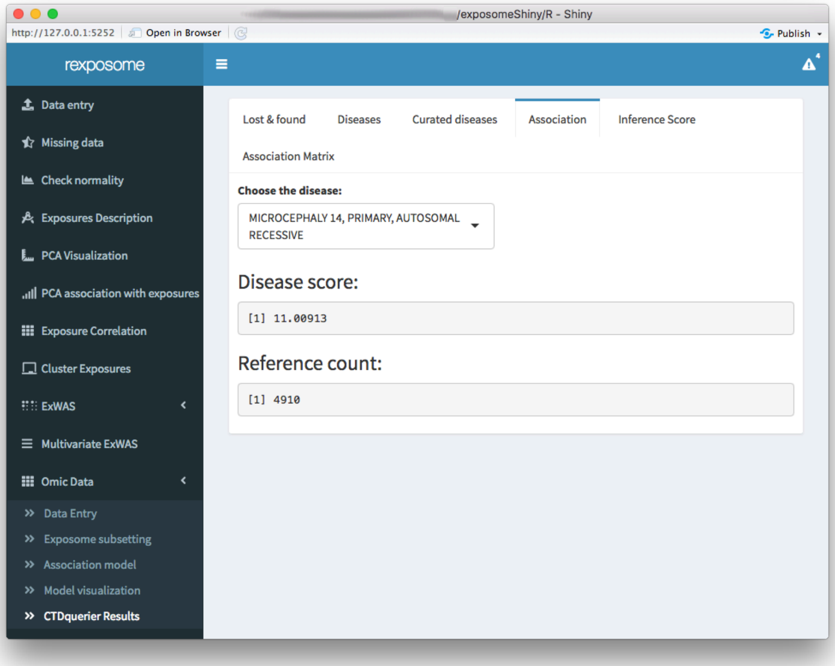


Show 10 entries Search:

	Disease.Name	Disease.ID	Inference.Score	Reference.Count	GeneSymbol	GeneID
1	MICROCEPHALY 14, PRIMARY, AUTOSOMAL RECESSIVE	OMIM:616402		1	SASS6	163786
2	Osteoarthritis	MESH:D010003	4.37	5	TRIM2	23321
3	CHARCOT-MARIE-TOOTH DISEASE, AXONAL, TYPE 2R	OMIM:615490		1	TRIM2	23321
4	Liver Diseases	MESH:D008107	84.12	77	CCR1	1230
5	Carcinoma, Hepatocellular	MESH:D006528	63.91	229	CCR1	1230
6	Pneumonia	MESH:D011014	56.49	52	CCR1	1230
7	Hypersensitivity	MESH:D006967	27.87	14	CCR1	1230
8	Dermatitis, Contact	MESH:D003877	17.96	14	CCR1	1230
9	Behcet Syndrome	MESH:D001528	2.43	4	CCR1	1230
10	Respiratory Hypersensitivity	MESH:D012130	32.38	11	CCR3	1232

Showing 1 to 10 of 15 entries

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.



Choose the disease:

MICROCEPHALY 14, PRIMARY, AUTOSOMAL RECESSIVE

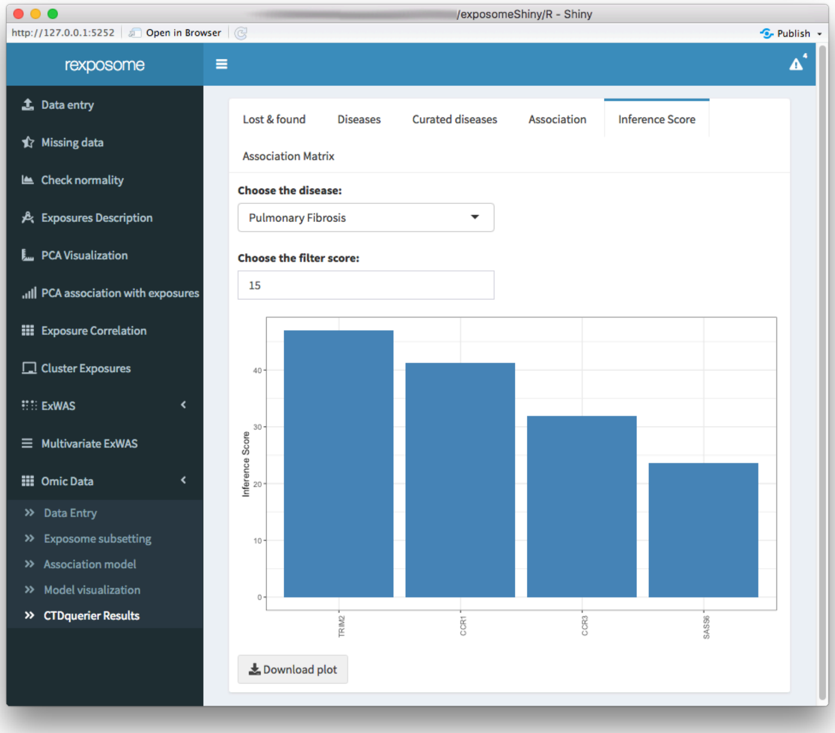
Disease score:

```
[1] 11.00913
```

Reference count:

```
[1] 4910
```

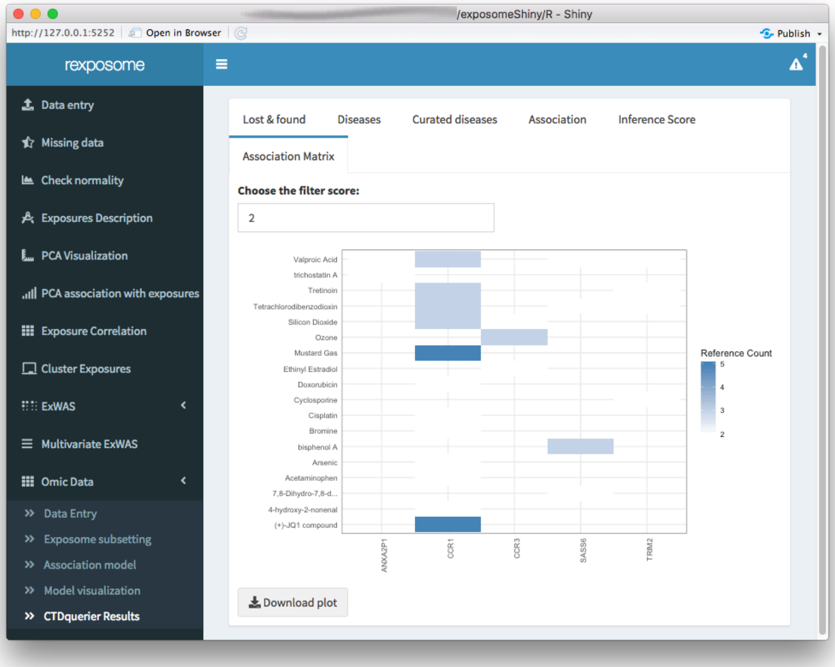
The inference score tab shows the inference score for each gene for a selected disease, the filter parameters puts out the genes with an inference score lower than the selected filter.



Choose the disease: Pulmonary Fibrosis

Choose the filter score: 15

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.



Choose the filter score:

2