

exposomeShiny User's Guide

Escribà Montagut, Xavier; González, Juan R.

2020-10-13

Contents

1	Overview	5
2	Setup	7
2.1	Downloading the source files, installing the libraries and running the application	7
2.2	Pulling the official Docker image from DockerHub	9
3	Data sets	11
3.1	Exposome dataset	11
3.2	Omics dataset	13
4	Bioconductor packages	15
4.1	rexosome	15
4.2	omicReposome	15
4.3	CTDquerier	15
5	Analysis flowcharts	17
5.1	Exposome analysis	17
5.2	Exposome-Omic analysis	35
6	General application functionalities	53

Chapter 1

Overview



exosomeShiny 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, exosomeShiny relies on previously existent Bioconductor packages (rexosome, omicReXosome and CTDquerier), it uses them in a seamless way so the final user of exosomeShiny can perform the same studies that would conduct using the Bioconductor packages but without writing a single line of code.

Chapter 2

Setup

There are two different ways of setting up and using exposomeShiny

2.1 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.

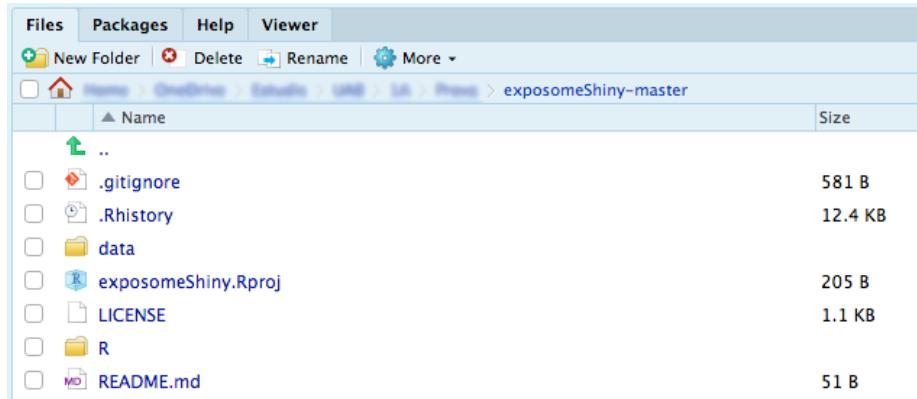
```
# Set working directory
setwd(dir = "/some/path/")

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

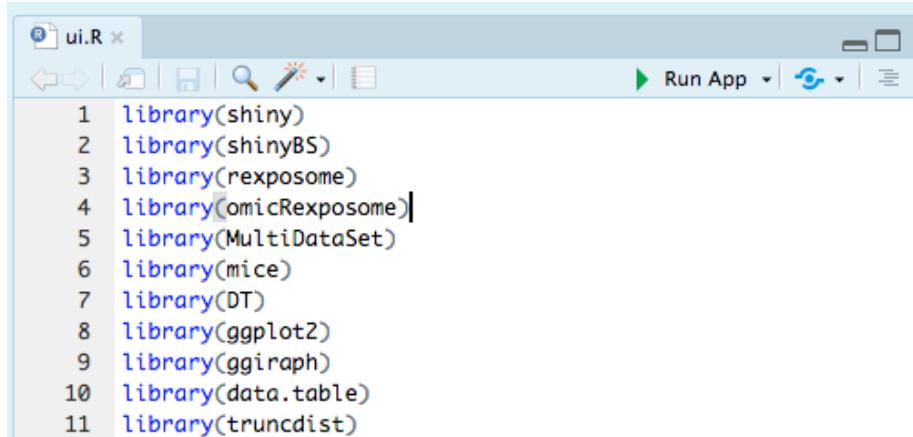


Once the project is loaded, the file found on the source folder called `installer.R` has to be sourced and run. This will install the newest versions of the packages required by Exposome Shiny on this R session. To do so, run the following code on the RStudio console.

```
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 are 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')
```

2.2 Pulling the official Docker image from DockerHub

If there's any trouble 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 your 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 its 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

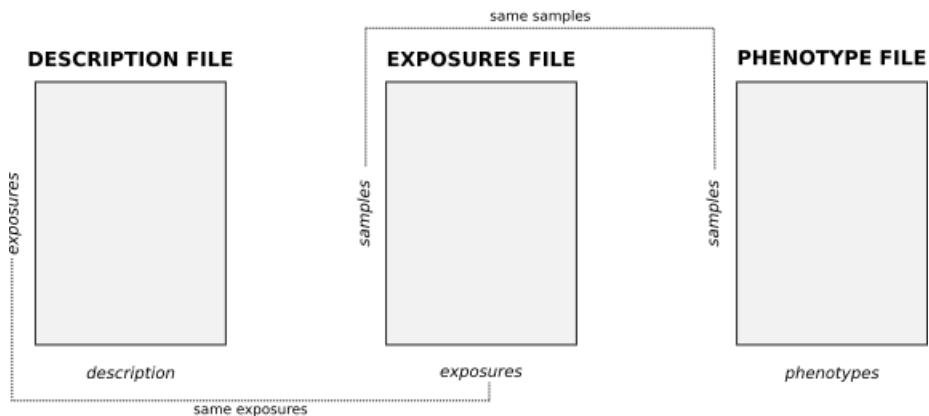
Data sets

3.1 Exosome dataset

The exosome is composed of three different files (in `*.csv` format). Those files are referred 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 include 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.

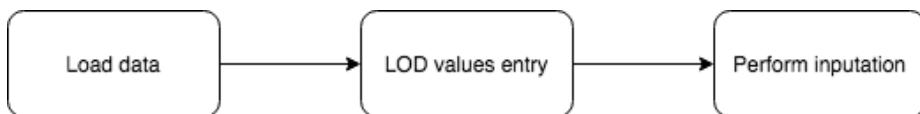
Chapter 5

Analysis flowcharts

5.1 Exposome analysis

As any user would need to do using the Bioconductor packages (rexosome, 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 rexosomeShiny, 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.

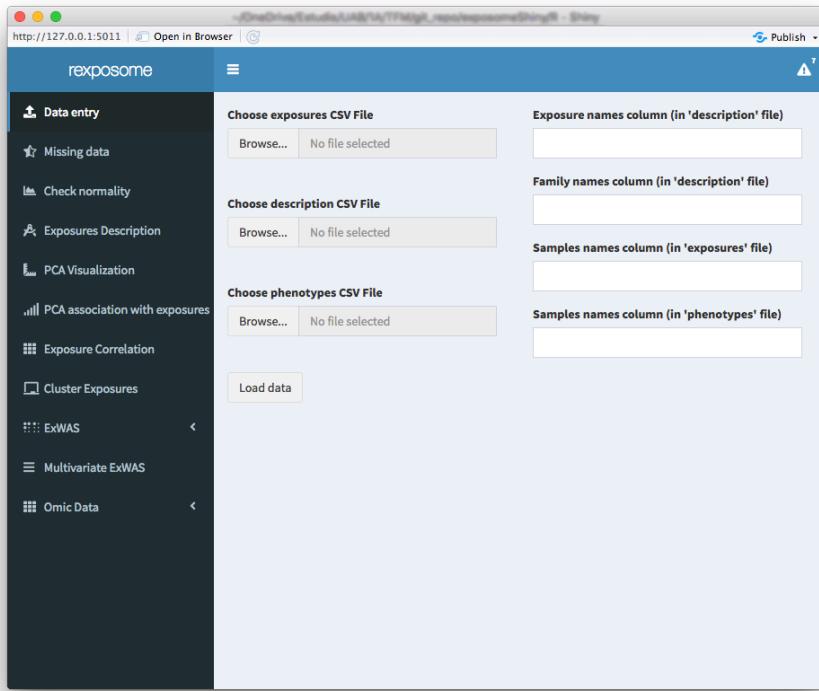
5.1.1 LOD imputation



Input the exposures, description and phenotypes files and load them into the application. On the right of the file browsers, there are four text fields to input the following:

- 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

If the column names do not match the selected data files a warning error will be prompted to the user to check the inputs.



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 impute with random values instead of LOD/sqrt(2).

The screenshot shows the exposome Shiny application interface. The left sidebar lists various analysis options: Missing data, Check normality, Exposures Description, PCA Visualization, PCA association with exposures, Exposure Correlation, Cluster Exposures, ExWAS, Multivariate ExWAS, and Omic Data. The main panel is titled "Data entry" and contains three sections for file upload:

- Choose exposures CSV File:** Browse... exposures_lod_test.csv, Upload complete.
- choose description CSV File:** Browse... description.csv, Upload complete.
- Choose phenotypes CSV File:** Browse... phenotypes.csv, Upload complete.

On the right, there are four input fields:

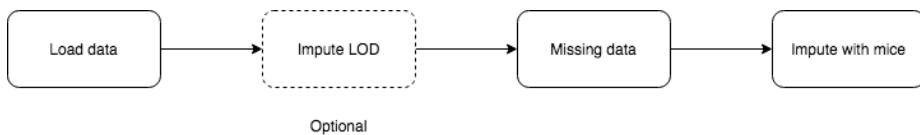
- Exposure names column (in 'description' file):** Exposure
- Family names column (in 'description' file):** Family
- Samples names column (in 'exposures' file):** idnum
- Samples names column (in 'phenotypes' file):** idnum

Below these is a table with two entries:

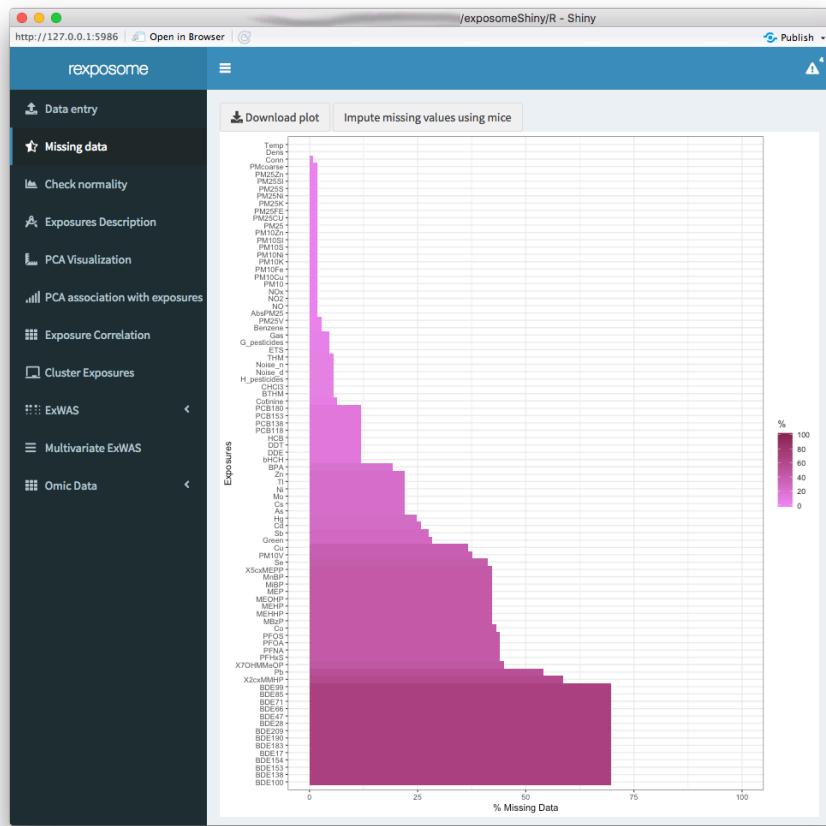
Exposure	LOD
PFHxS	1
PFOS	2

Buttons include "Load data", "Show 10 entries", "Search:", "Help about the LOD substitution", "Choose imputation method:" (set to LOD/sqrt(2)), and "Perform LOD imputation with the values provided".

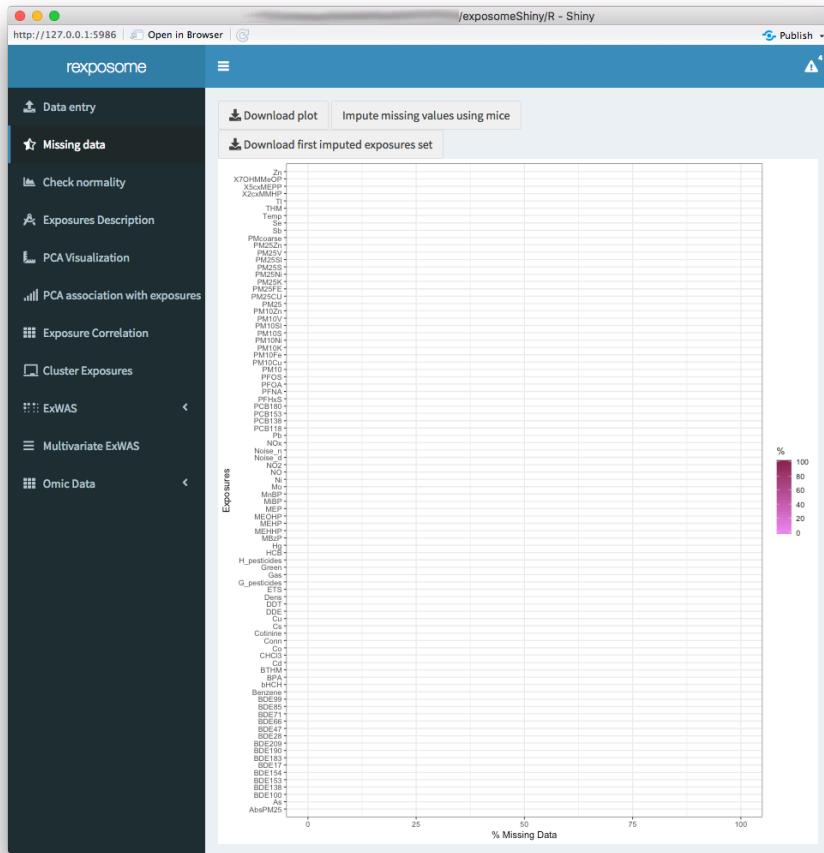
5.1.2 Missing imputation



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.

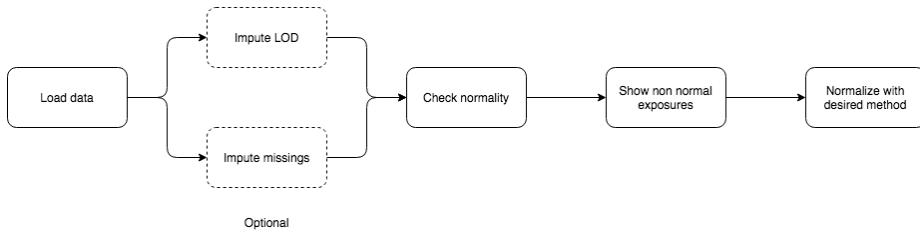


To impute the missing values select “Impute missing values using mice”. After the process finishes, the expect output should be a new missing data graph where there's no missing of any exposure.



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.

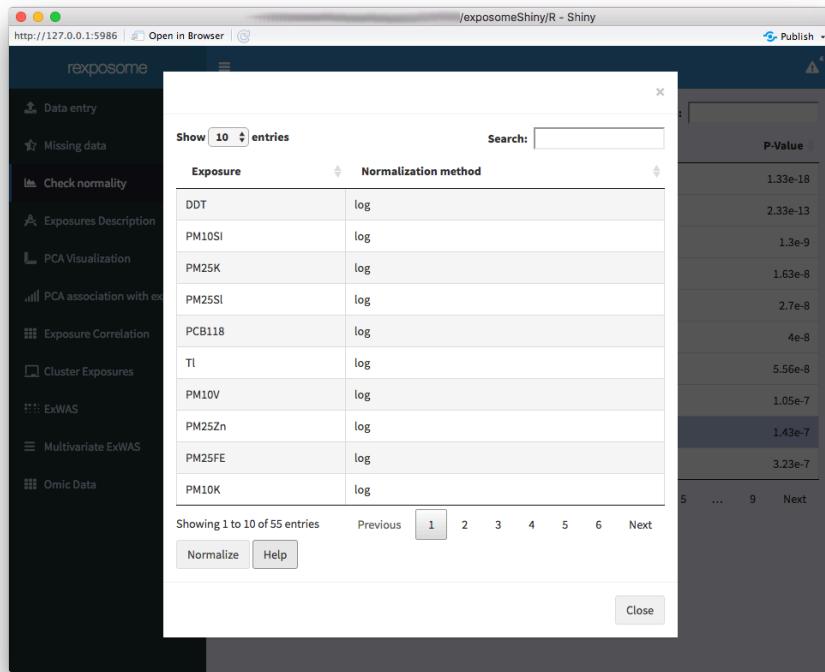
5.1.3 Normality correction



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.

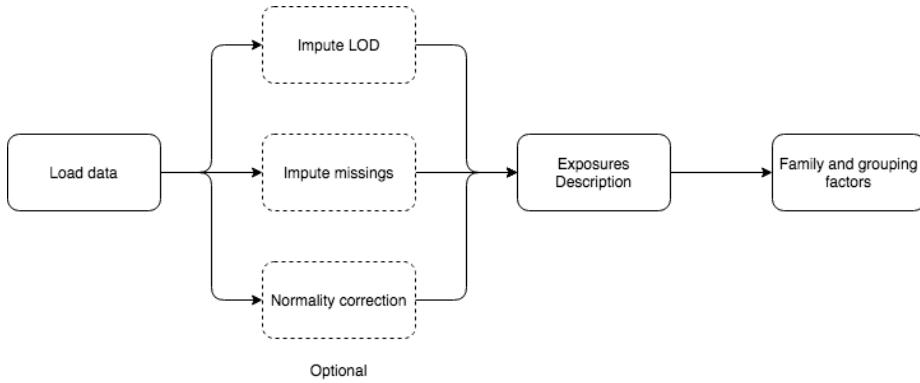


By clicking the “Show false” button, all the non normal exposures are listed with the method that will be applied to normalize, this table can be edited (the “Normalization method” column) by double clicking on the desired row. There are three possible methods to use, “log” (default), “ $\sqrt[3]{\cdot}$ ” and “sqrt”. If no method is desired to be applied to an exposure input “none”.



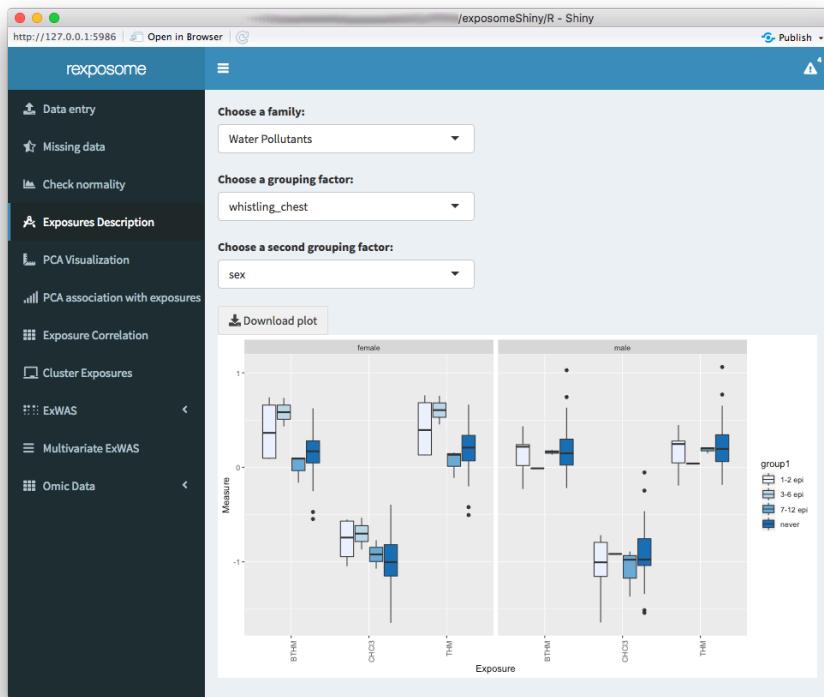
Click “Normalize” and the normalization method selected will be applied, the table on the “Check Normality” tab will be updated with the results of the normalization.

5.1.4 Exposures description

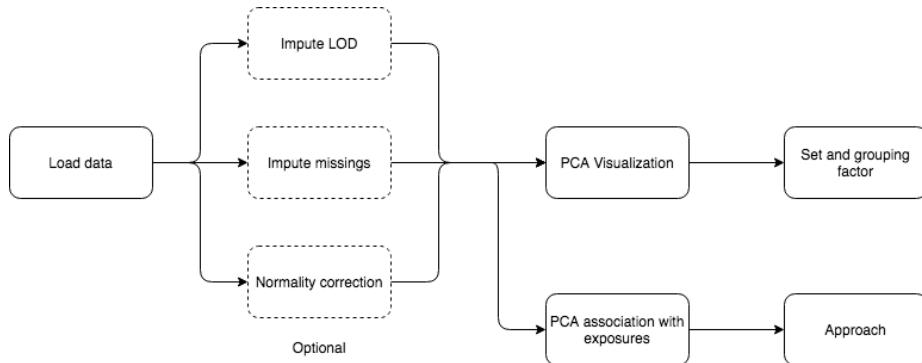


To see all the insights of the exposures dataset loaded into the Shiny, once loaded it check the exposures description tab, there are three options to dig into

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

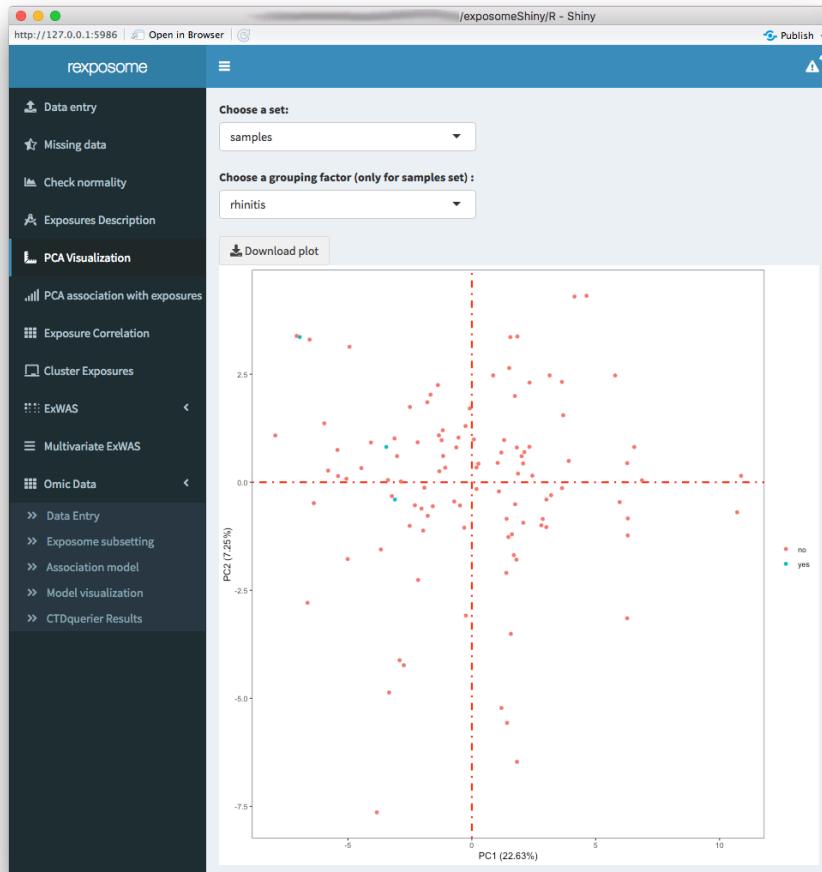


5.1.5 PCA Analysis

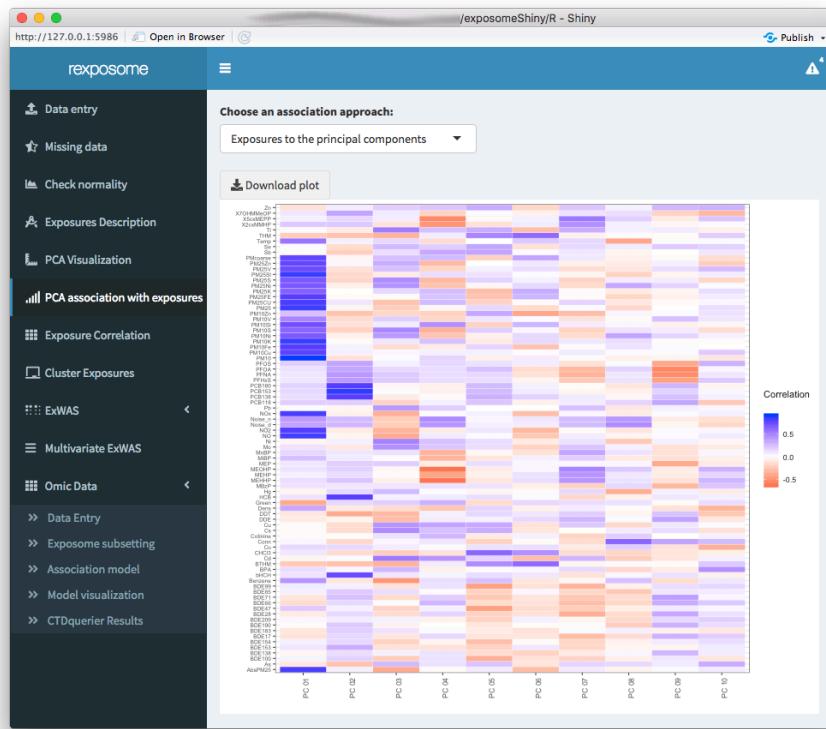


To see the results of a PCA (principal component analysis) study, load the data and check the PCA Visualization tab, there a set and grouping factor can be choose, it's important noting (as it's already stated on the Shiny) that the

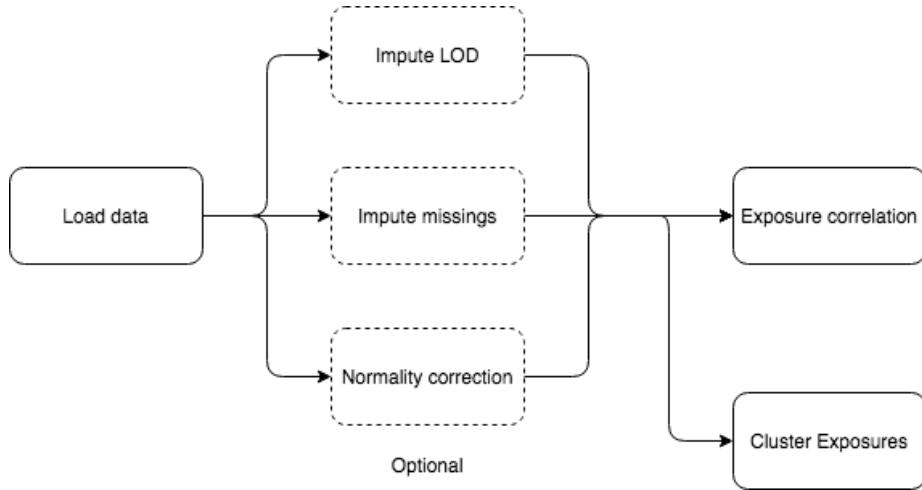
grouping parameter only works when the set is selected to “samples”.



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.

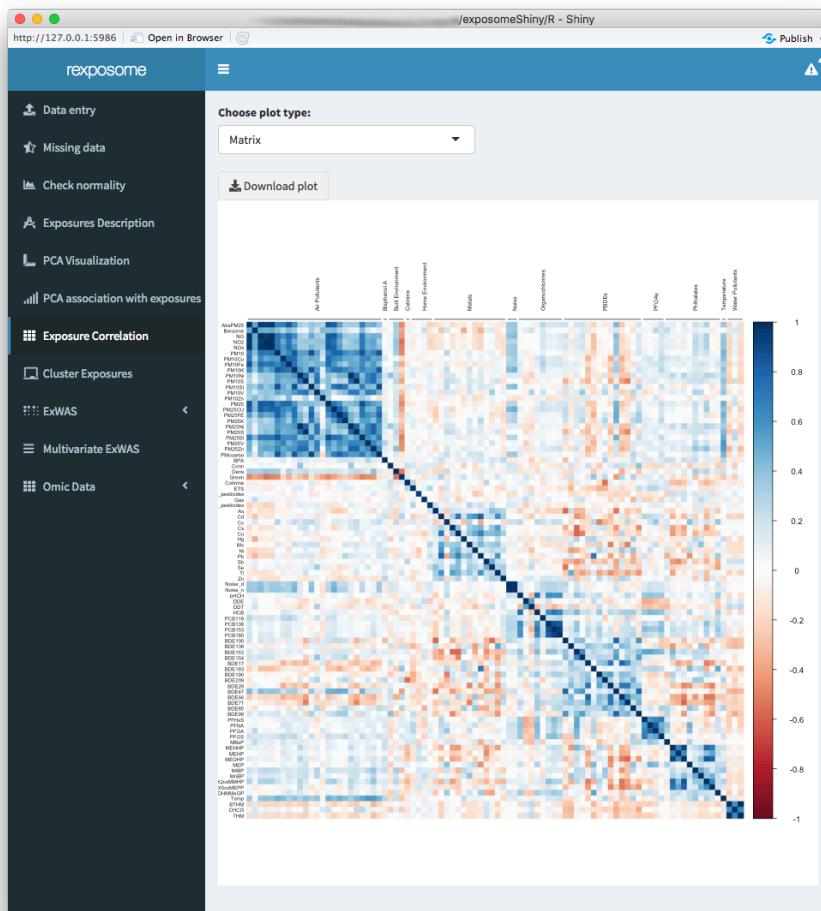


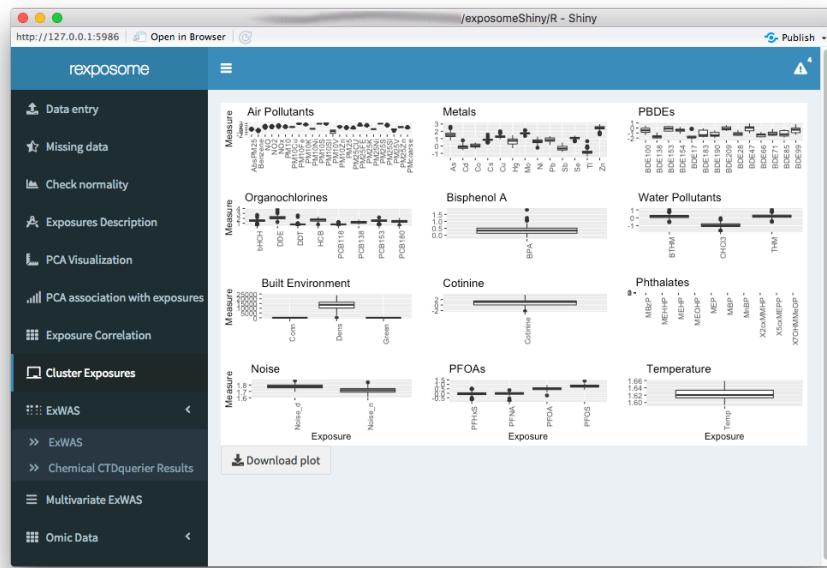
5.1.6 Clusterization and correlation of exposures



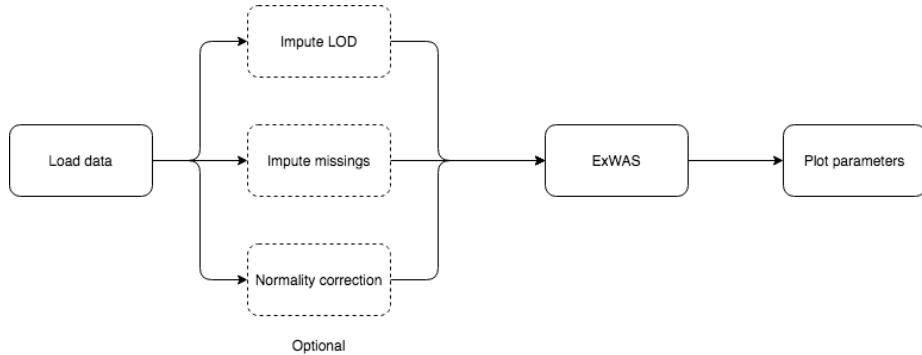
To see the results of the exposure correlation and clustering, select the corre-

sponding tab to each analysis. For the exposure correlation analysis there are two visualizations, the matrix representation and the circos.

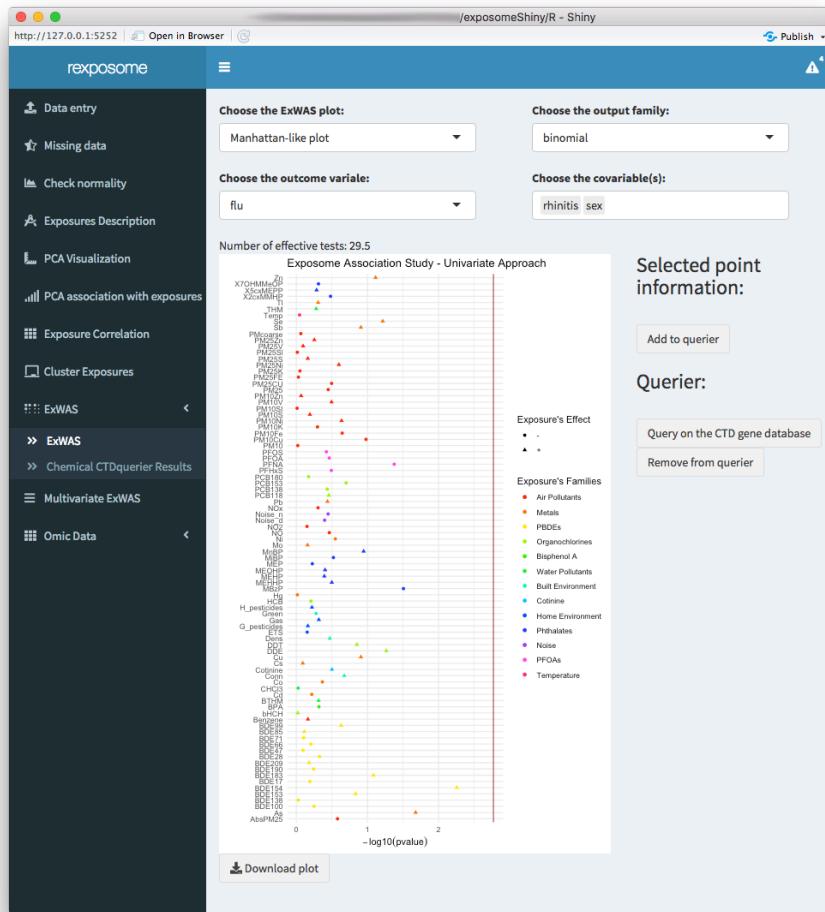




5.1.7 ExWAS

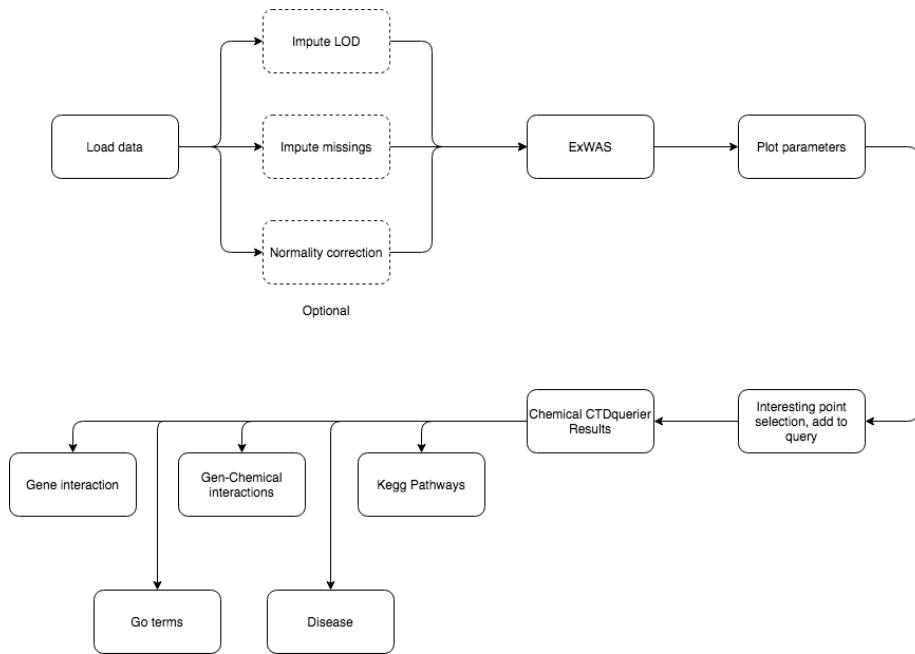


To perform an ExWAS (exposome-wide association) study, check the ExWAS tab and select the addequate parameters for the ExWAS plot, there are two different plot representations, the output variable to choose (phenotype), the output family and as many covariables (phenotypes) as the user wants. There are internal checks to advise the user on which parameters to select depending if the selected outcome is numerical or bionomial.

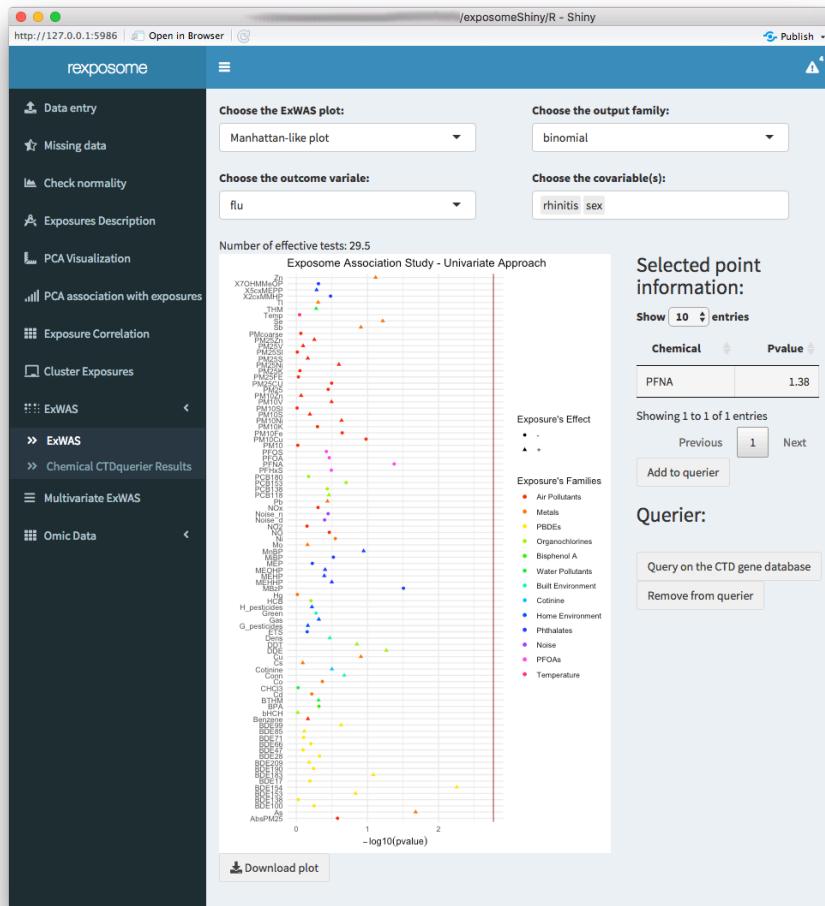


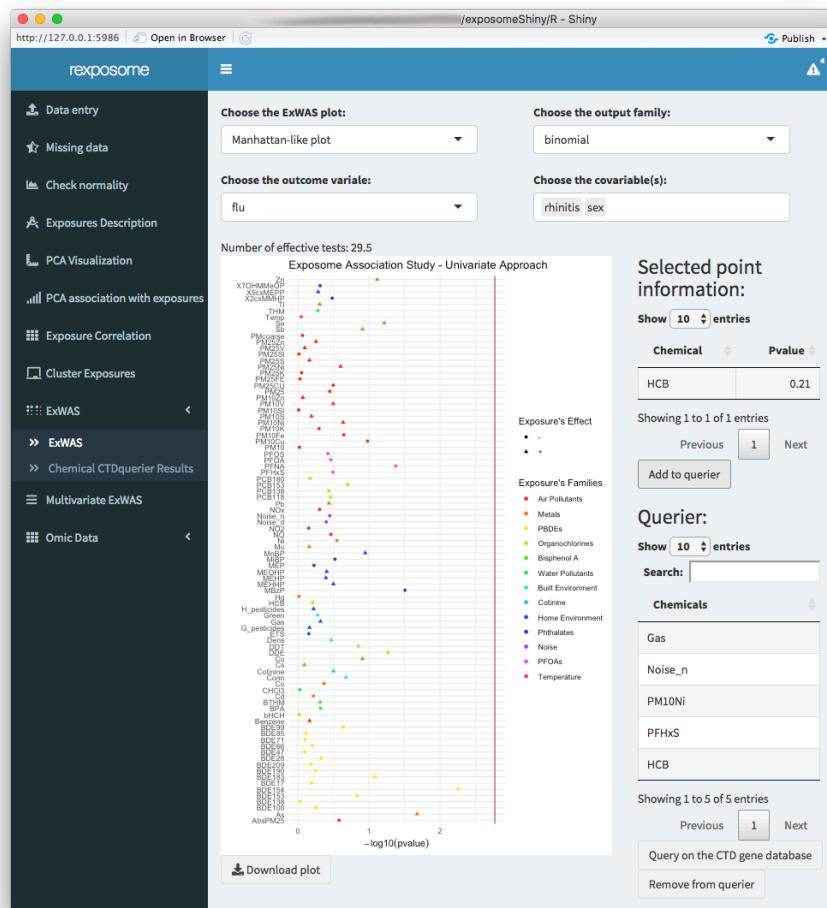
5.1.8 ExWAS - CTDquerier

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

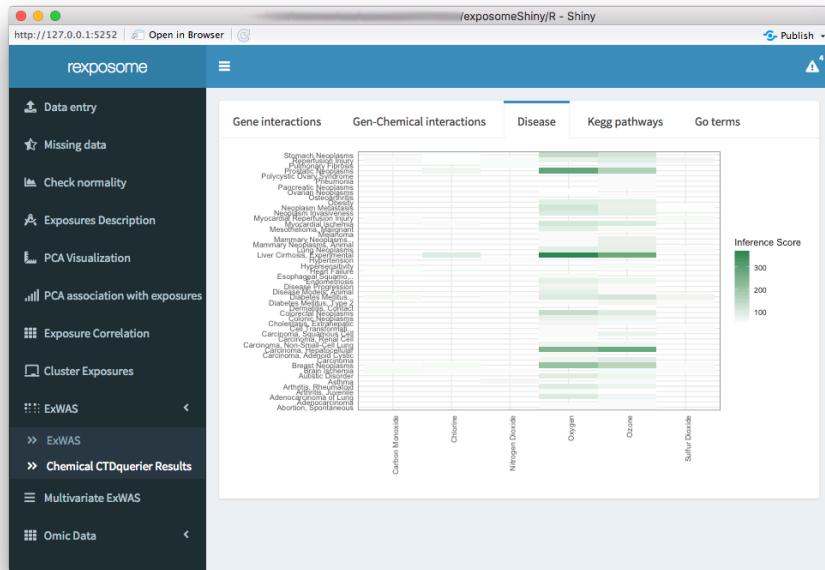


To perform a CTDquerier of chemicals with the results of the ExWAS, click on the desired exposure to preload it into the query, when clicked, a chemical name with it's associated P-Value will appear on the table on the right, if that's the desired chemical to add to the query list click "Add to querier". 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".

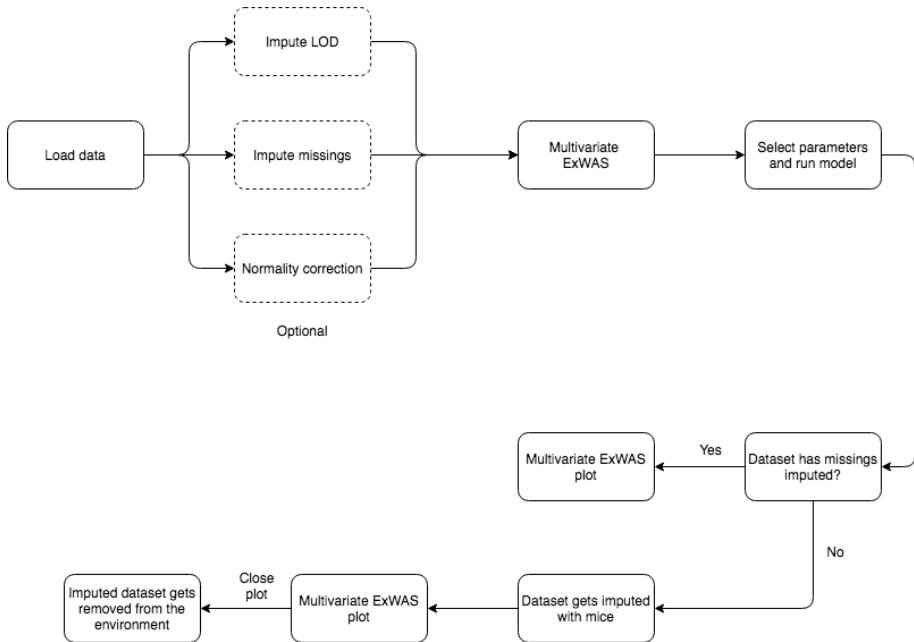




To do the query of the chemicals to de CTD database click on “Query on the CTD gene database” and see the results on the “Chemical CTDquerier Results” subtab. It’s important noting that on the “Kegg pathways” and “Go terms” the input field corresponds to the negative exponent of the filter.

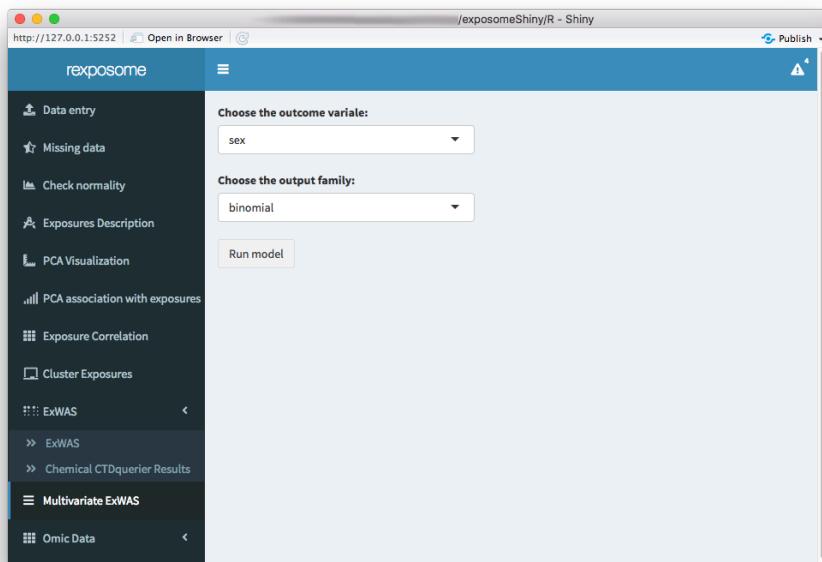


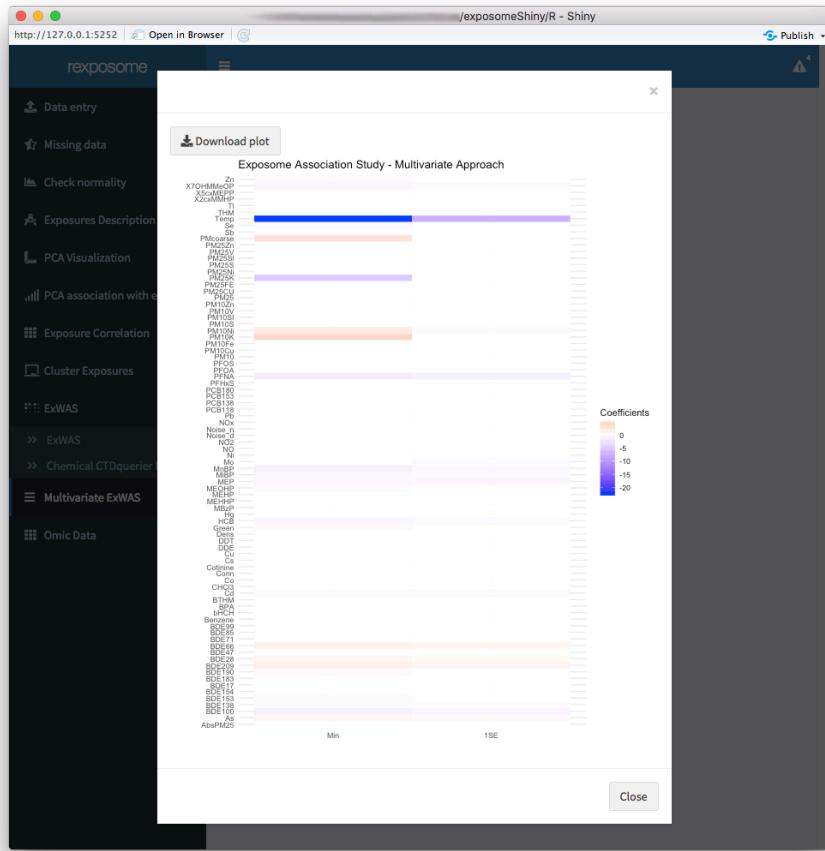
5.1.9 Multivariate ExWAS



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 binomial, 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.



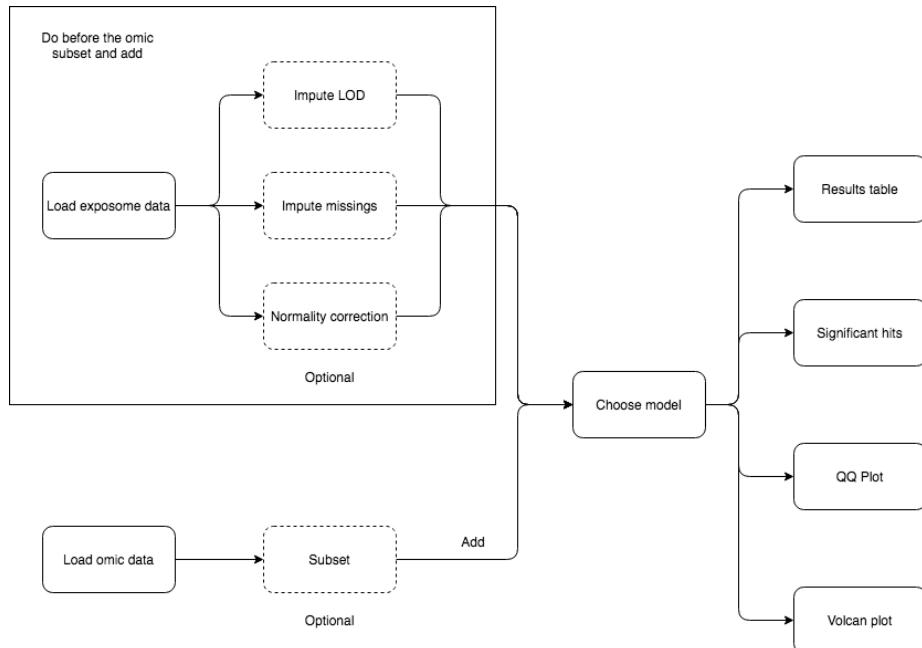


5.2 Exposome-Omic analysis

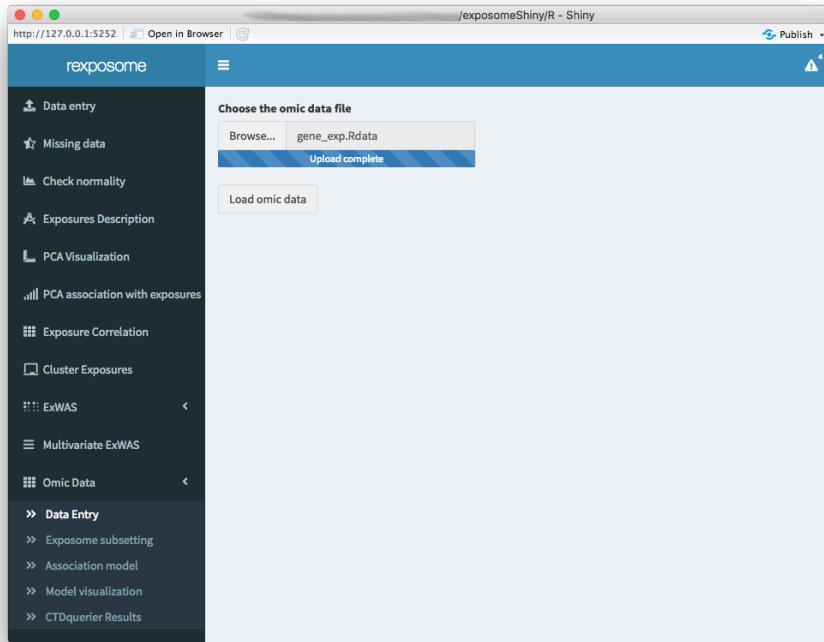
It's important noting that the maximum size of the omics data is 30 MB, if the omics file to be analyzed is bigger, change the line number 2 of the `server.R` file.

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

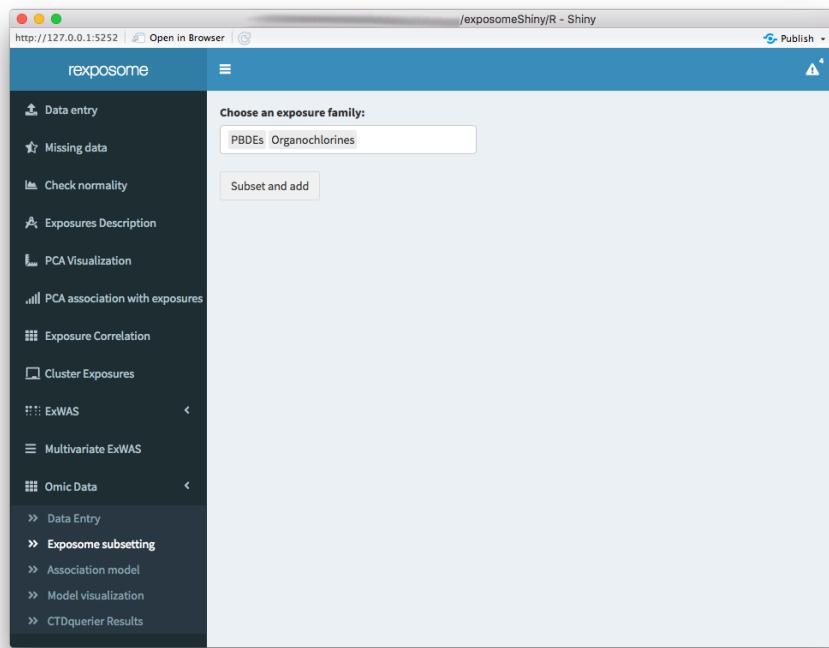
5.2.1 Association analysis



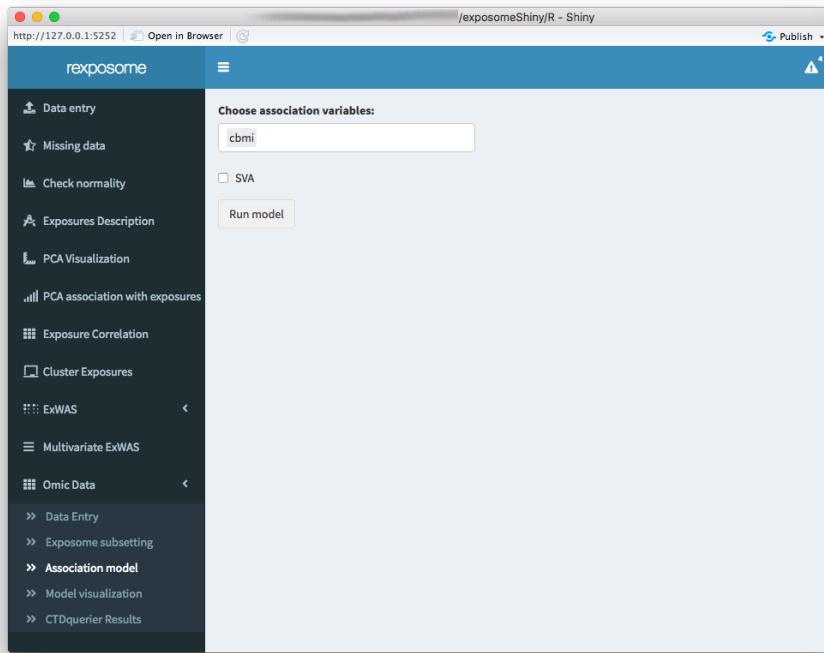
Do first the proceeding of exposome data load and corresponding treatment if desired, then proceed to load the omic dataset on the “Data Entry” subtab of the “Omic Data” tab. The omic data should be provided as a `*.RData` file.



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.



Select the variables for the association analysis and if SVA 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.

	logFC	PValue	adj.PVal
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

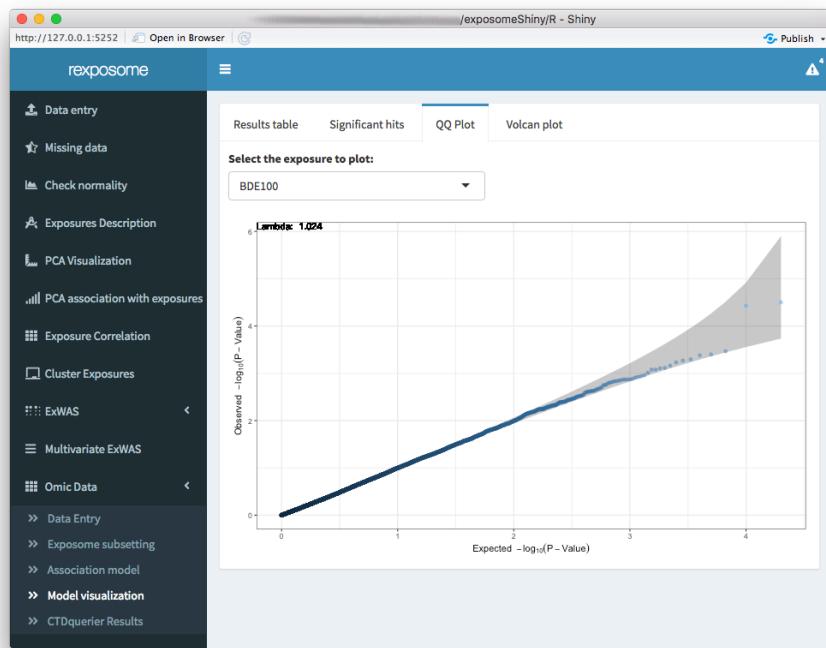
Showing 1 to 10 of 20,000 entries Previous 2 3 4 5 ... 2000 Next

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

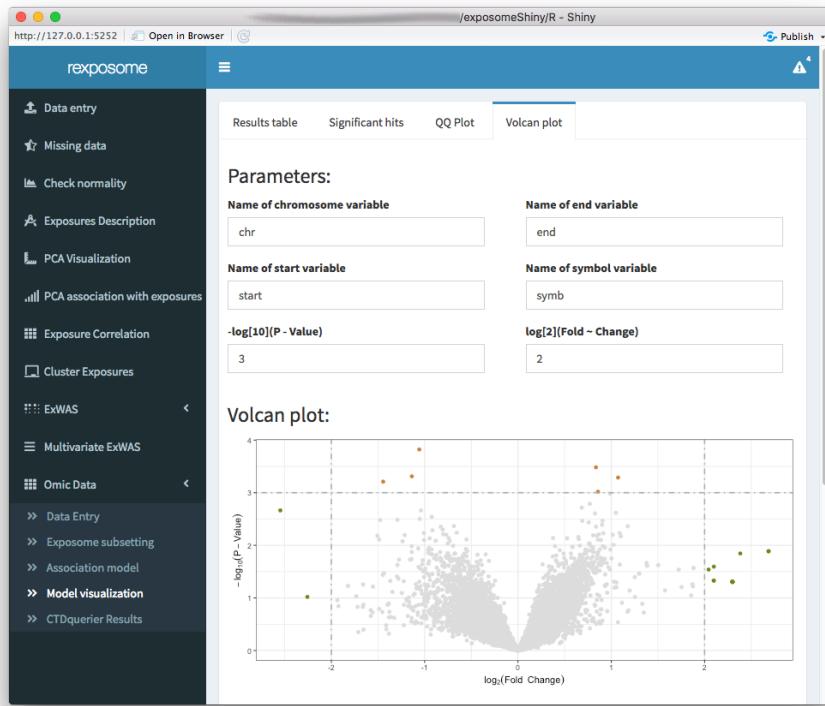
The screenshot shows a web-based application titled "exposomeShiny/R - Shiny" running at <http://127.0.0.1:5252>. The interface has a dark sidebar on the left containing various menu items under the "reposeosome" header. The main area is a table titled "Results table" with three tabs above it: "Significant hits" (selected), "QQ Plot", and "Volcan plot". The table displays 10 entries of significant hits, each with columns for "Exposure", "Hits", and "Lambda". The table footer indicates there are 88 entries in total, with page navigation from 1 to 9.

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

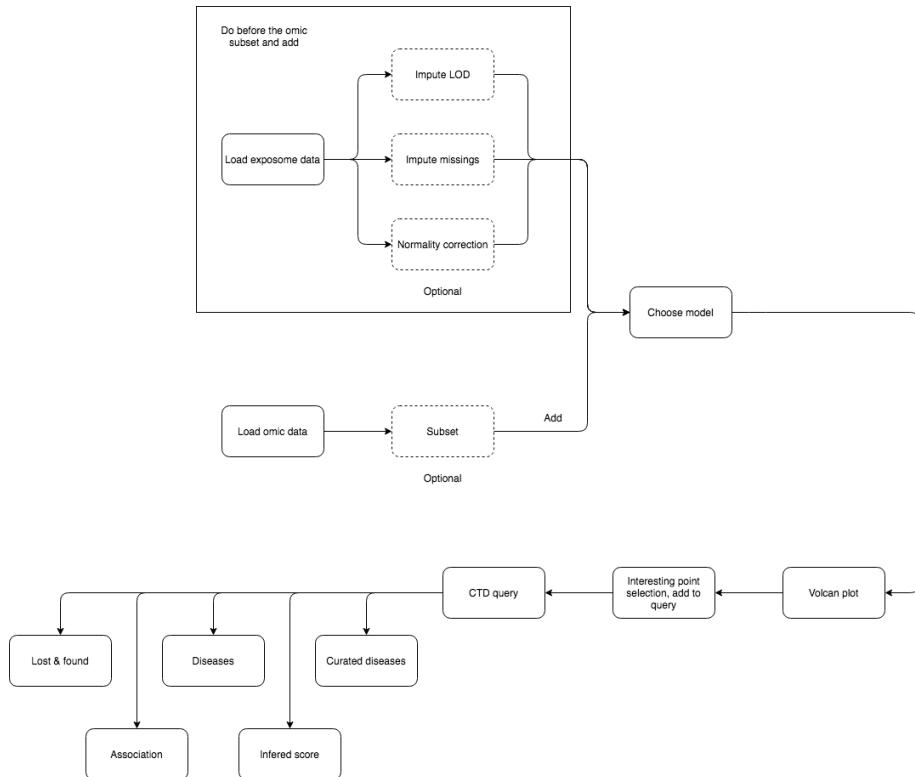
The QQ Plot shows a QQ plot (expected vs. observed -lo10(p-value)) for the selected exposure.



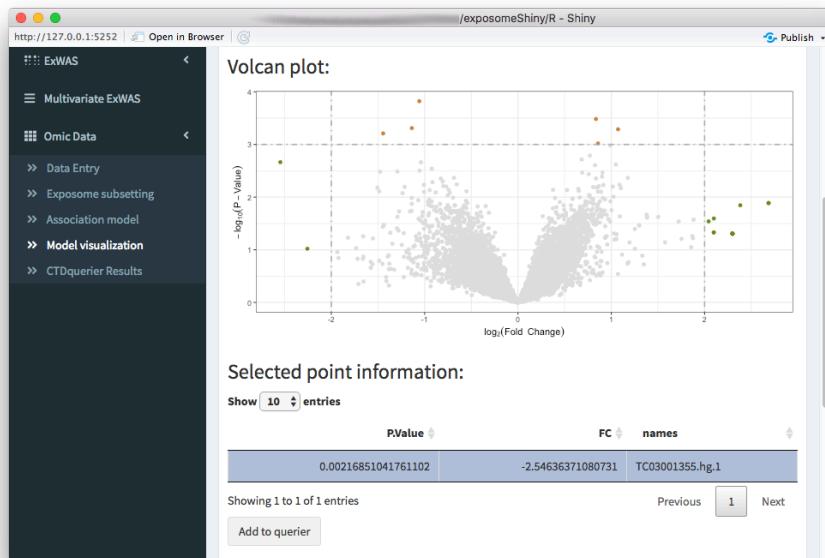
The Volcan plot shows a volcano plot ($\log_2(\text{fold change})$ vs $-\log_{10}(\text{p-value})$). For this plot there are two input cells to adjust the horizontal and vertical limit lines to filter out the results.



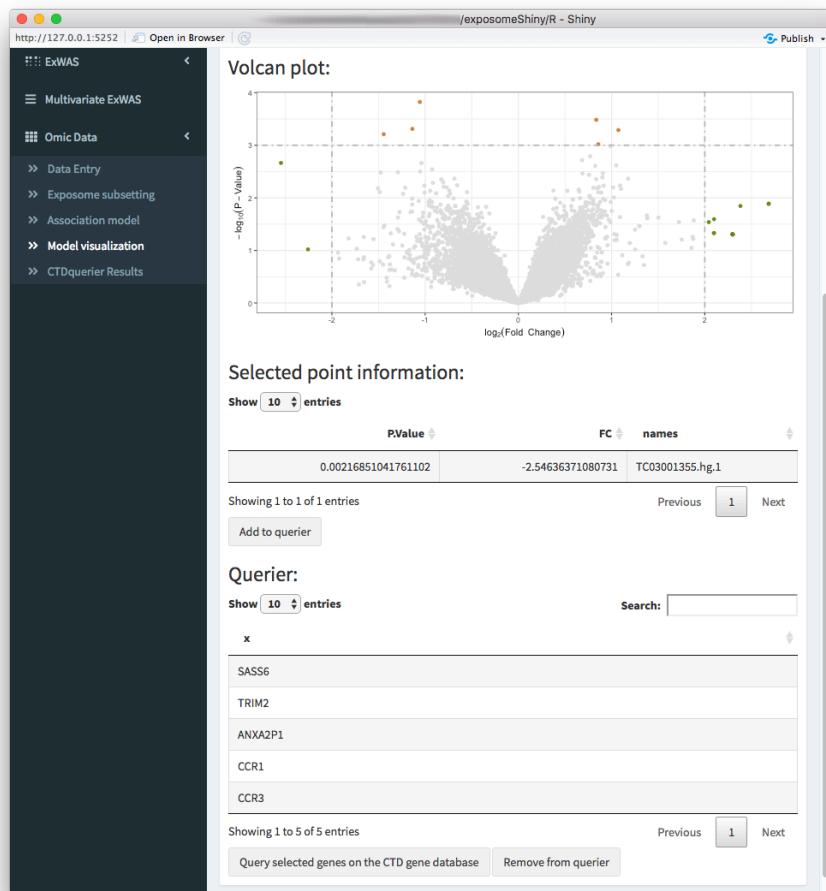
5.2.2 CTD querier



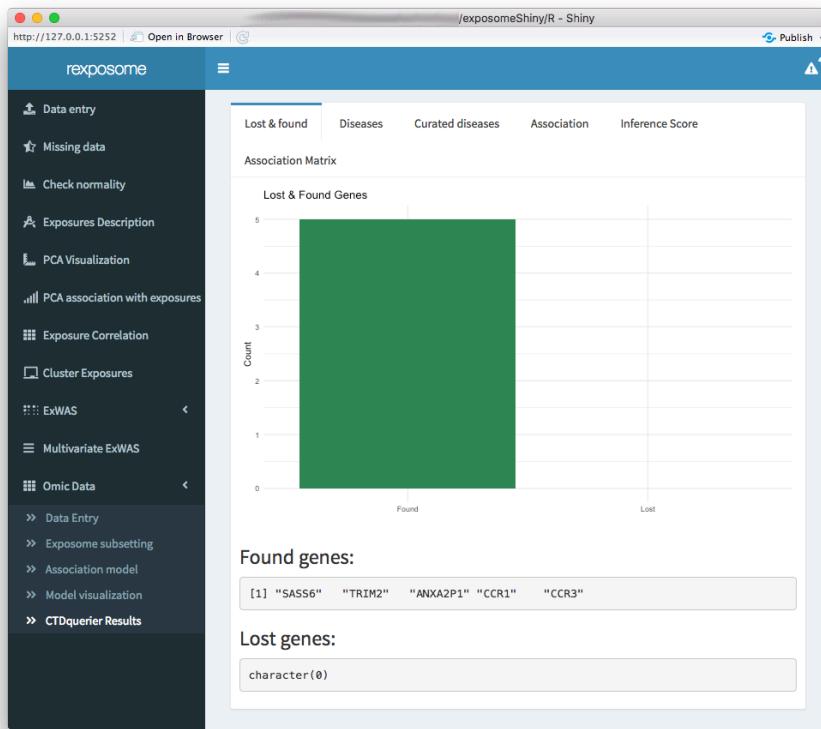
To perform a CTD querier study of the exposome-omic analysis, as before, load both datasets and run the desired model with them, check the Volcan plot and click on the desired point on the Volcan plot, the information of the selected plot will appear on the table below the plot (sometimes there are many 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 screenshot shows a web-based application titled "exposome" running in a browser. The left sidebar contains a navigation menu with various options: 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, and CT querter Results. The main content area is titled "Diseases" and displays a table of associated diseases. The table has columns for Disease.Name, Disease.ID, Direct.Evidence, Inference.Score, Reference.Count, GeneSymbol, and GeneID. The table shows 10 entries from a total of 6,350. The first entry is MICROCEPHALY 14, PRIMARY, AUTOSOMAL RECESSIVE, with OMIM:616402 as the Disease.ID and marker/mechanism as the Direct.Evidence. The last entry is Lung Neoplasms, with MESH:D008175 as the Disease.ID and no evidence listed. The table includes a search bar at the top and a navigation bar at the bottom showing pages 1 through 635.

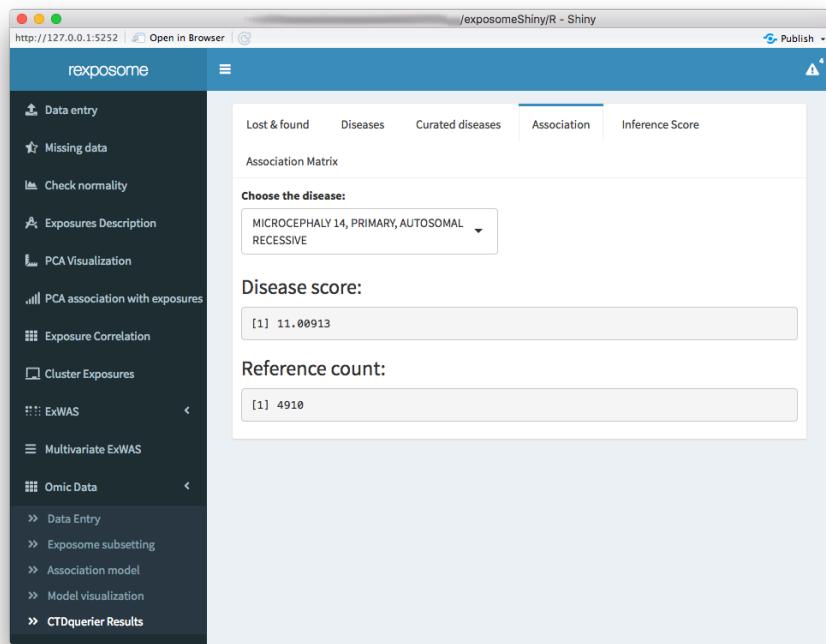
	Disease.Name	Disease.ID	Direct.Evidence	Inference.Score	Reference.Count	GeneSymbol	GeneID
1	MICROCEPHALY 14, PRIMARY, AUTOSOMAL RECESSIVE	OMIM:616402	marker/mechanism		1	SASS6	163786
2	Weight Loss	MESH:D015431		91.69	80	SASS6	163786
3	Inflammation	MESH:D007249		68.53	85	SASS6	163786
4	Chemical and Drug 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

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

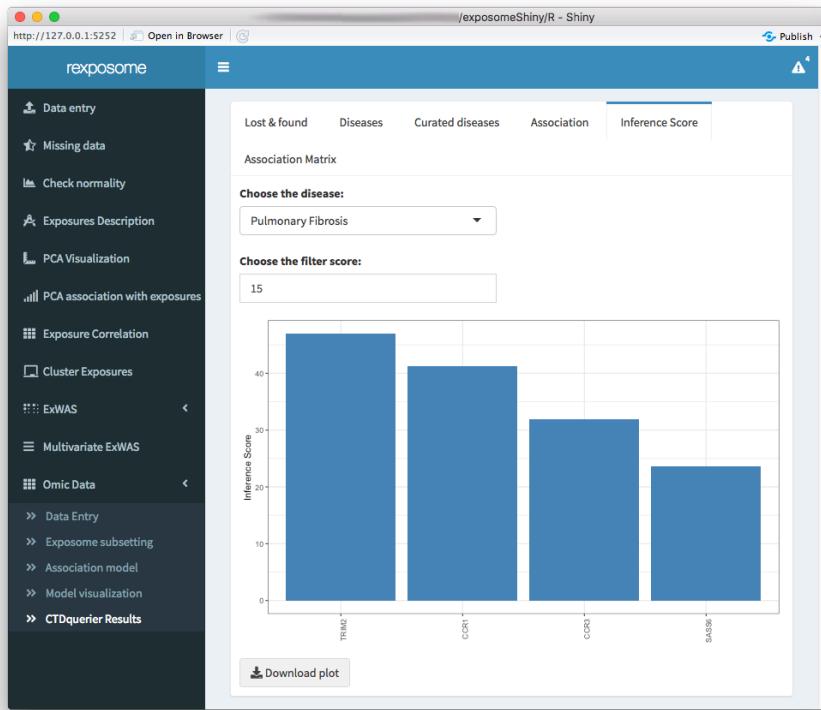
The screenshot shows the rexposome Shiny application interface. The left sidebar contains a navigation menu with various options like Data entry, Missing data, Check normality, Exposures Description, PCA Visualization, PCA association with exposures, Exposure Correlation, Cluster Exposures, ExWAS, Multivariate ExWAS, Omic Data, Data Entry, Exposome subsetting, Association model, Model visualization, and CTDquerier Results. The main content area has tabs for Lost & found, Diseases, Curated diseases, Association, Inference Score, and Association Matrix. The Diseases tab is active, showing a table with 10 entries. The columns are Disease.Name, Disease.ID, Inference.Score, Reference.Count, GeneSymbol, and GeneID. The data includes entries for MICROCEPHALY, Osteoarthritis, CHARCOT-MARIE-TOTH DISEASE, Liver Diseases, Carcinoma, Hepatocellular, Pneumonia, Hypersensitivity, Dermatitis, Contact, Behcet Syndrome, and Respiratory Hypersensitivity.

	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:D0006967	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

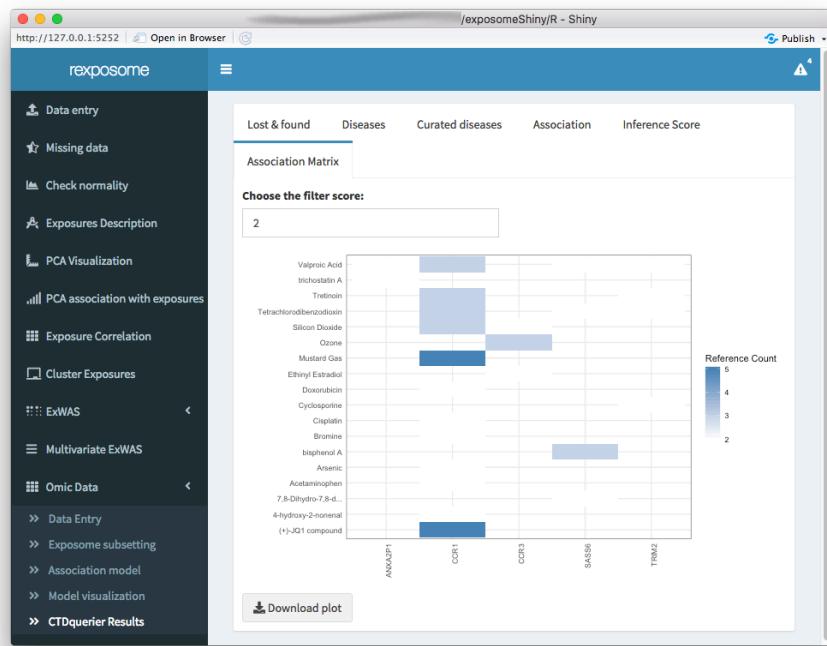
The association tab shows information about all the direct evidence associated diseases. Select the disease of interest to see the score and reference count of it.



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.



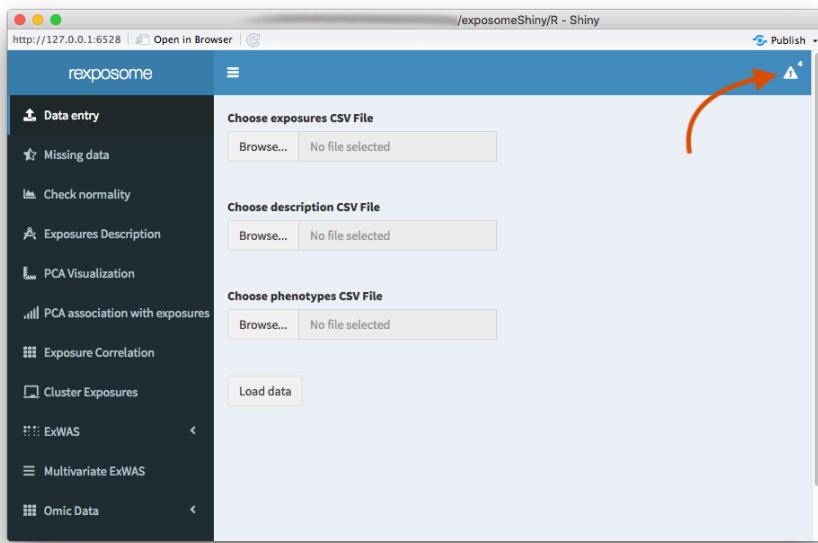
The association matrix tab shows a matrix of genes vs. chemicals with a heatmap representing the existing papers (references) providing evidence about the association between chemicals and genes.



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 be performed on the inputted 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.

Exposome dataset 0%

Exposome dataset 100%

- 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.

