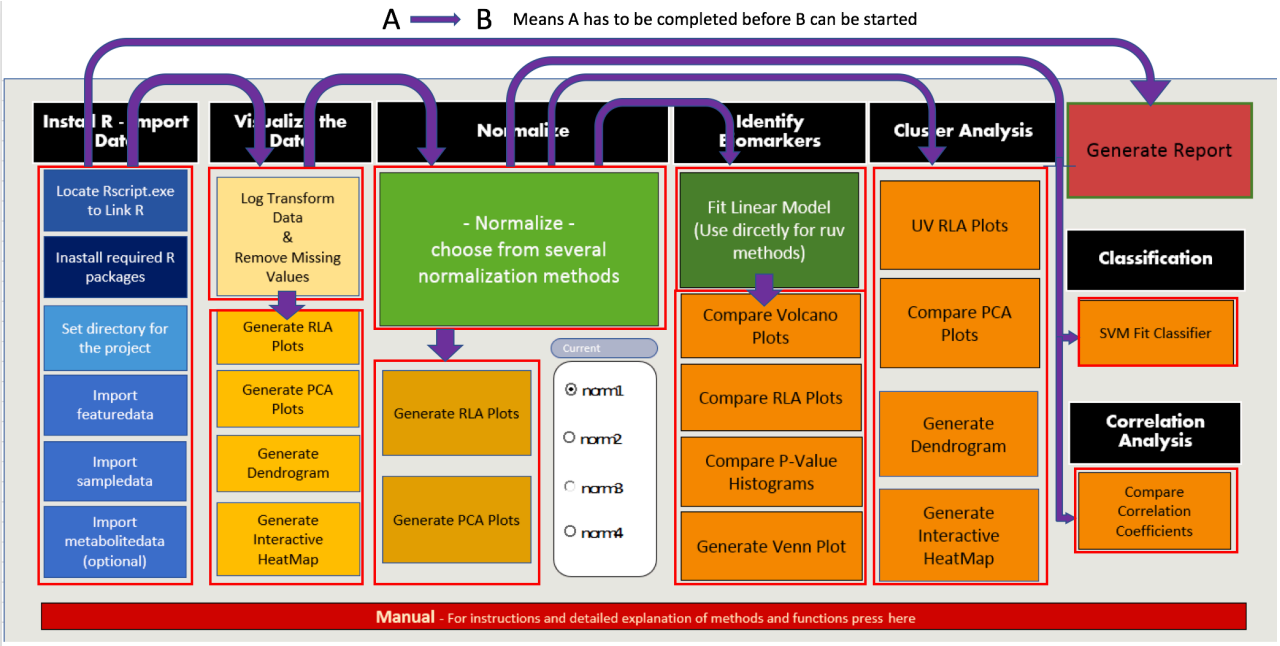
User guide:

Assume a windows machine is being used with no prior software installed other than Microsoft Excel 2016 and a web browser.

Getting the files ready:

To get the required files, go to https://github.com/metabolomicstats/NormalizeMets and download the latest version of the *ExNormalizeMetsSetup.zip*. After downloading unzip the folder, it should contain the following files:
----- Insert Folder Image once final name is decided upon ------
They are all the files needed for installing R and loading NormalizeMets through excel.

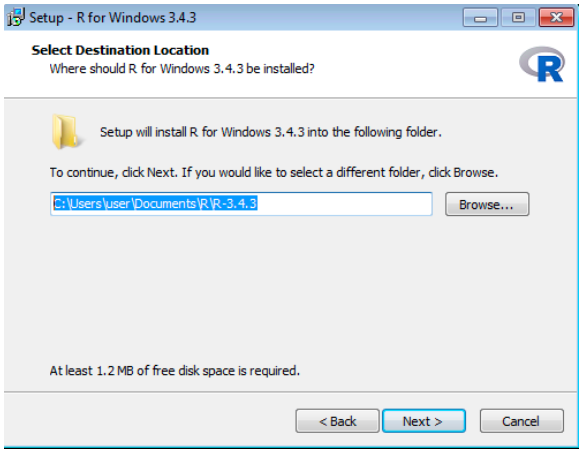
General Workflow:



Installing R:

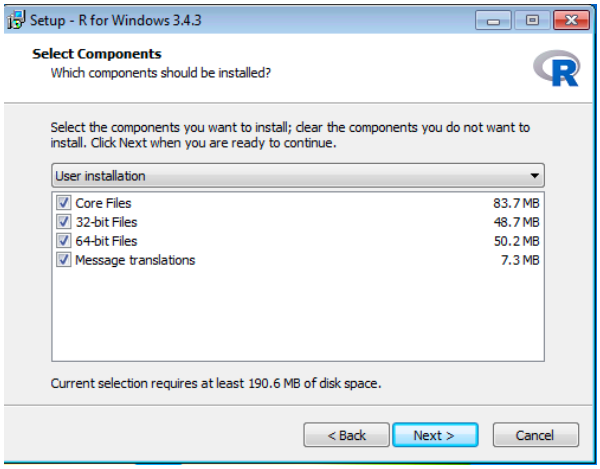
1. Install R from the extracted ExNormalizeMetsSetup folder by running the *R-3.4.3-win* file.

2. Follow the installation instructions and choose the location where R should be installed.

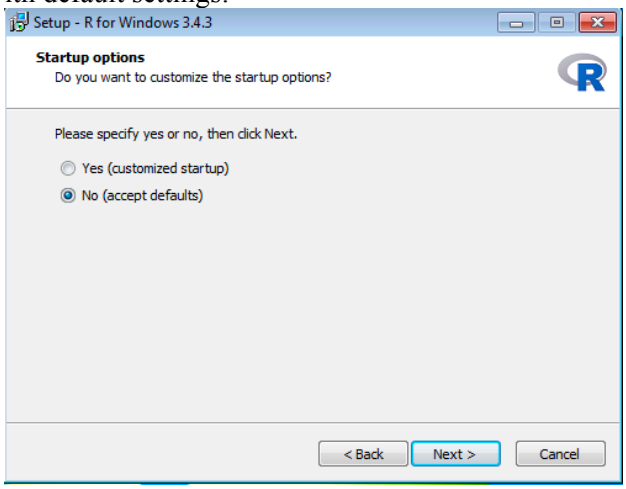


Click *Next* for default location (Recommended) or choose location manually. Make sure you **remember where R is installed** as you will need to locate this folder later to link R to Excel.

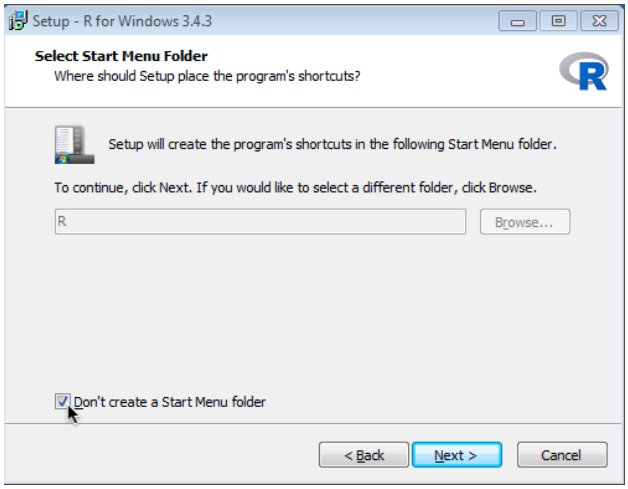
3. Click *Next* to install with the different required settings:

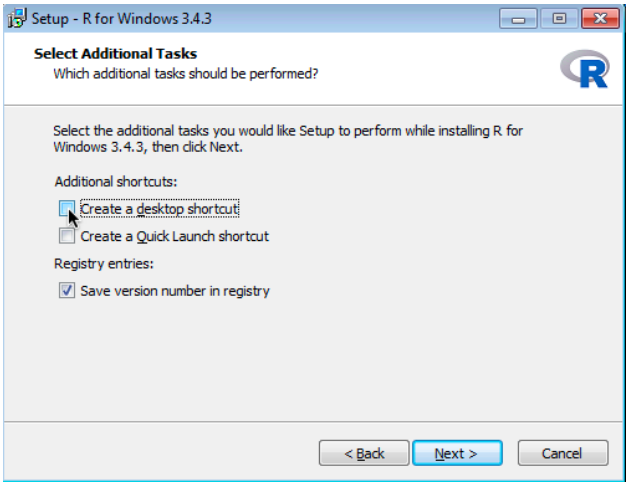


4. Click *Next* to install with default settings:

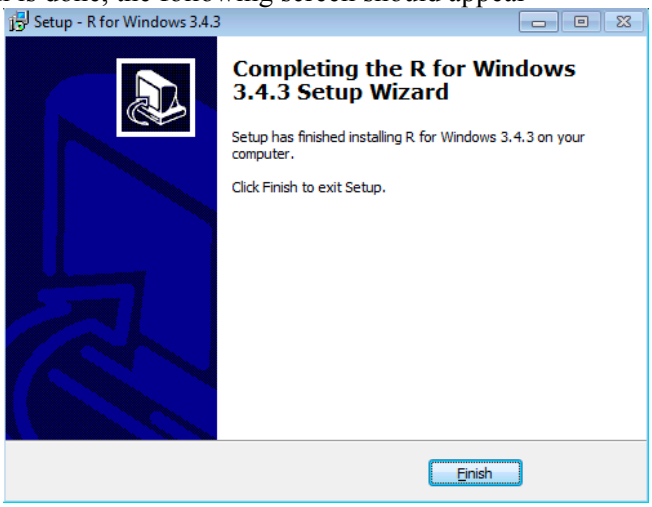


5. In the next screen check the *Don't create a Start Menu folder* icon if you don't intend to use R by itself and click next. Also uncheck the *Create a desktop shortcut* in the next screen:



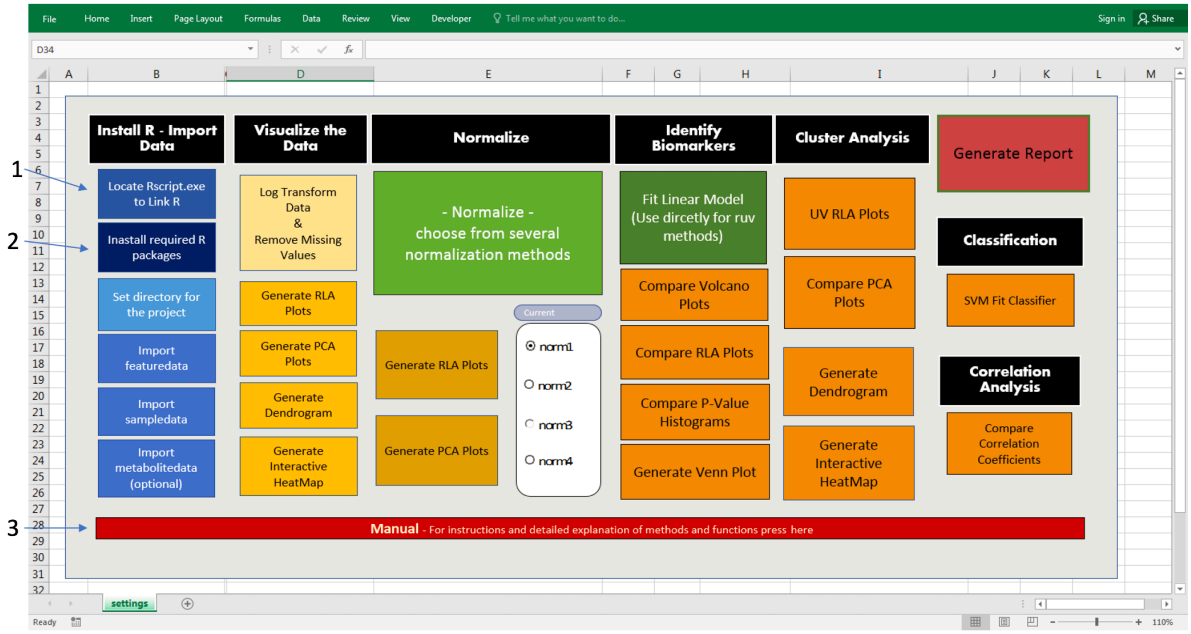


6. Wait until installation is done, the following screen should appear



Linking R to excel and installing the required packages (first use only):

After installing R open the excel file ExNormalizeMets.xlsm, this will open the Excel interface onto the settings sheet:

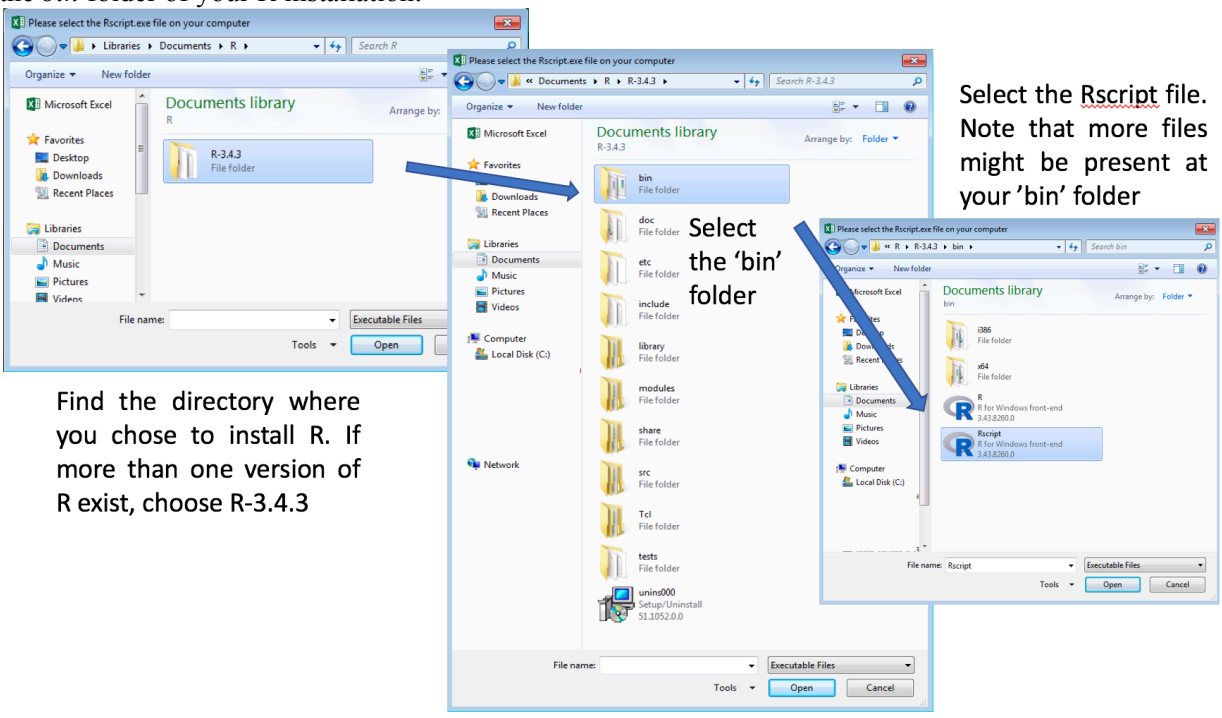


This is your 'Control', any function you want to run, from importing data, Normalizing, viewing results and opening the manual can be done from here.

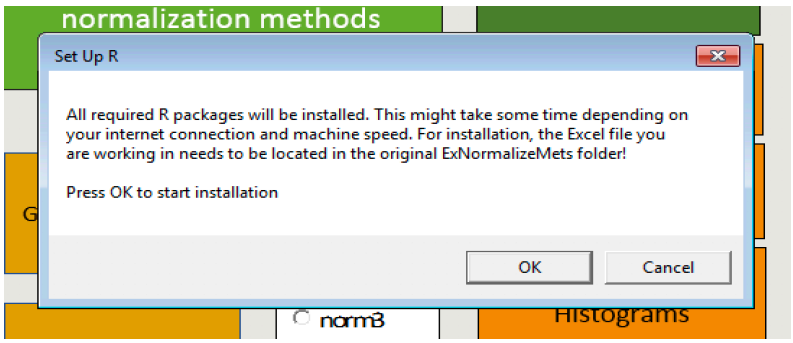
This manual can be accessed at any stage by clicking on (3) but more on this late. For now, first locate the Rscripts.exe file so that excel will know how to run R commands it generates.

Locating Rscript:

Press (1) to locate the Rscript file in the window that opens up. Make sure to select the *Rscript* file in the *bin* folder of your R installation:

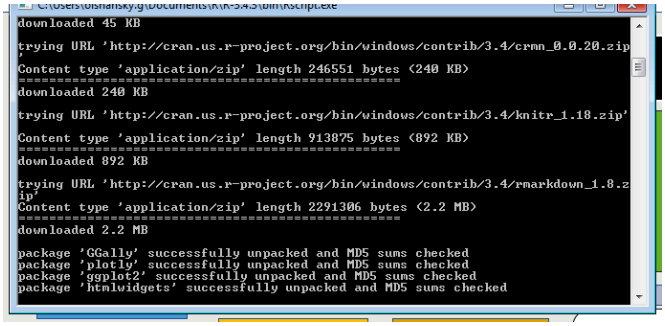


After selecting the file, to install all the required R packages and set up the needed dependencies, press (2).



Pressing ok will start the installation, this might take a few minutes if you are using R for the first time as many of the base packages will need to be installed.

The installation window looks like this:



When the installation is done, a window with the message *Done!* will appear.

Now that the installation is complete, NormalizeMets is ready for use!

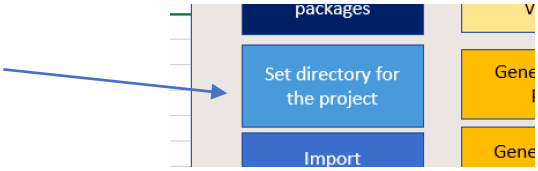
Using NormalizeMets

Data used in the following examples is provided with NormalizeMets (alldata_eg in R), it is located in the *ExampleData* folder in your downloaded *ExNormalizeMetsSetup* file. Future references in this guide refer to this data by default.

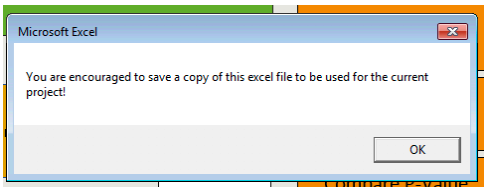
An example excel document containing the data used for the tutorial (MyFirstNormalizeMetsProject_example.xlsm), with all settings identical to those in the tutorial is provided.

Starting a new project:

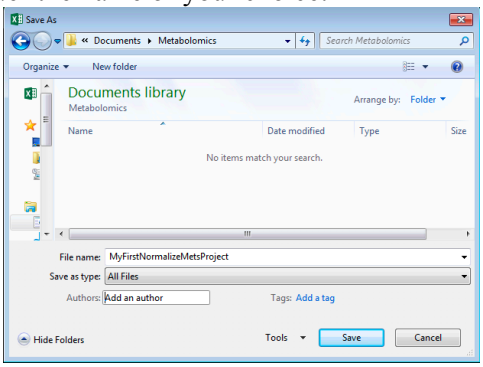
To start a new project, you will need to set up or use an existing directory where the project will 'live', data and plots generated by excel will all be saved to that folder.



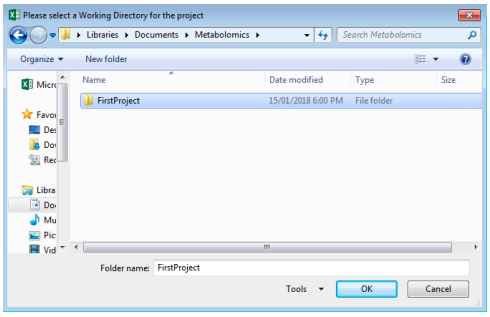
Before selecting the folder, you will be prompted to save a copy of the current version of the excel file, it is recommended to save it with a new name to make sure a 'clean' version always stays in your ExNormalizeMetsSetup folder.



After saving the workbook under the name of your choice:

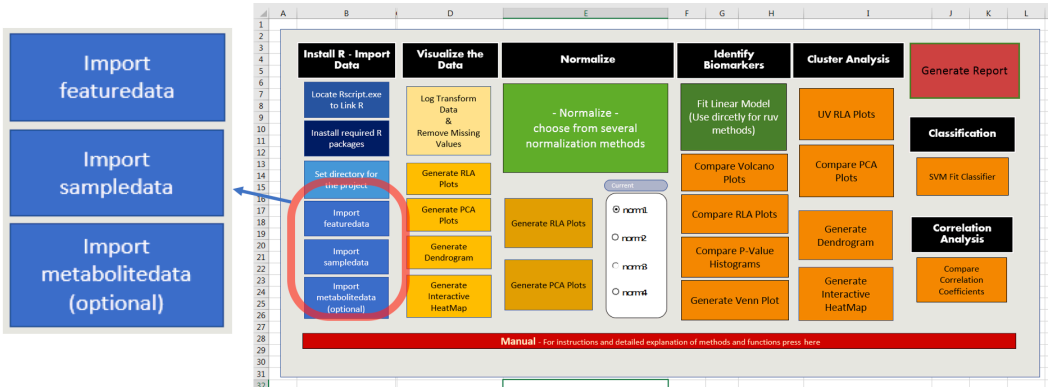


Select the working directory where all files should be generated. We recommend making a new folder for each project.



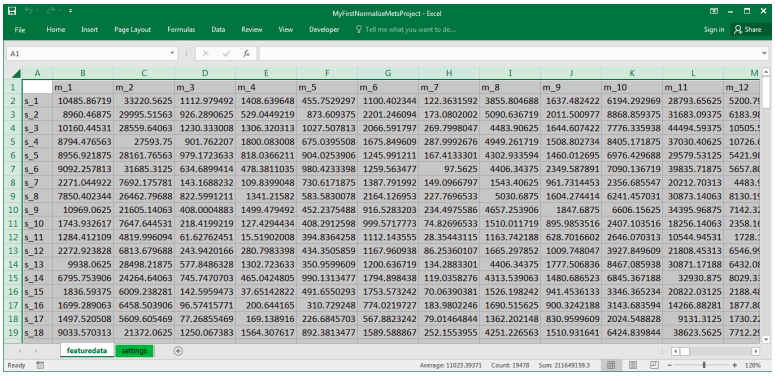
Loading data:

To load data for the project, in turn click on the following to load the relevant data:

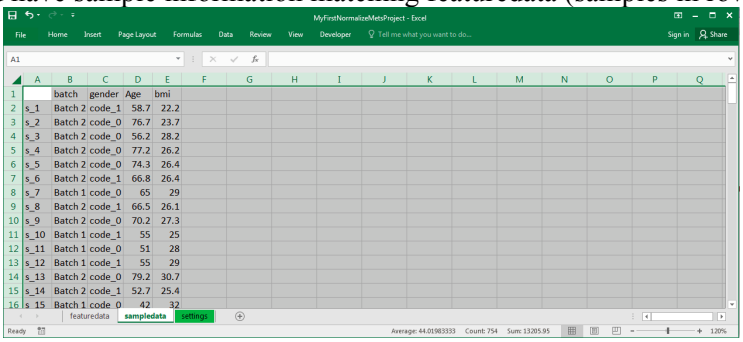


Loaded data needs to be in .csv format. After clicking on the required file, it will open a new sheet, showing the imported data. Select the setting sheet import more data and get back to the options.

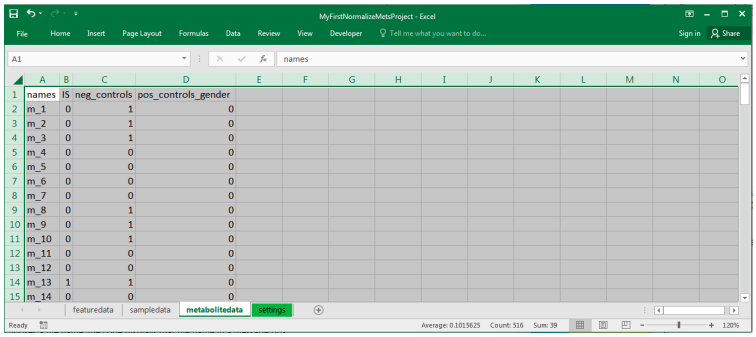
For *featuredata*, set metabolites in columns and samples in rows. Unique sample names should be provided as row names.



sampeldata should have sample information matching featuredata (samples in rows).



Optional *metabolitedata* should have metabolite information matching featuredata with metabolite names in rows.



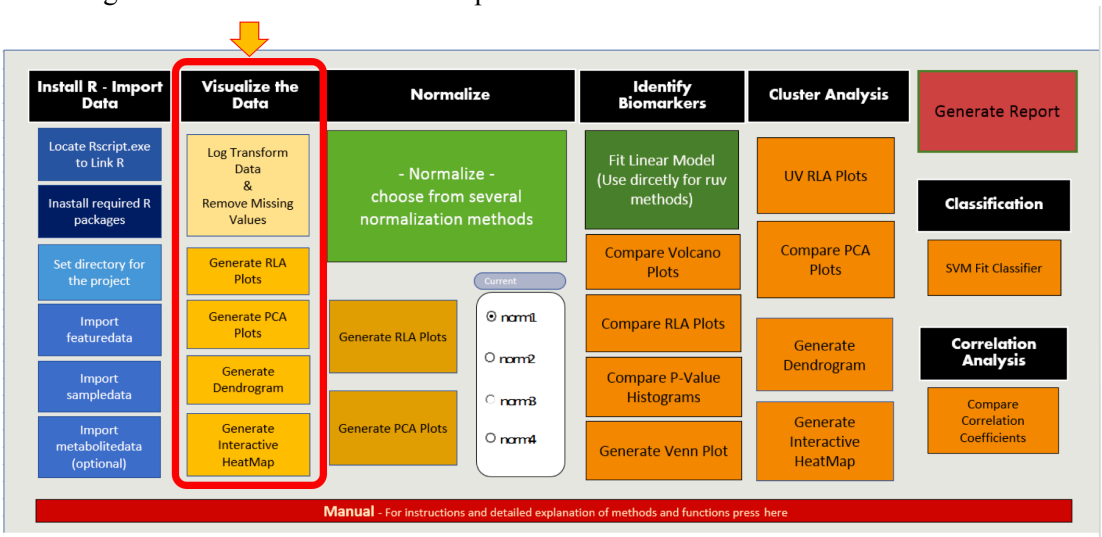
metabolitedata can include any metabolite information such as grouping structures, internal standard metabolites, negative control and positive control metabolites.

After the data is loaded, you are ready to proceed to analyse the data!

NormalizeMets Workflow:

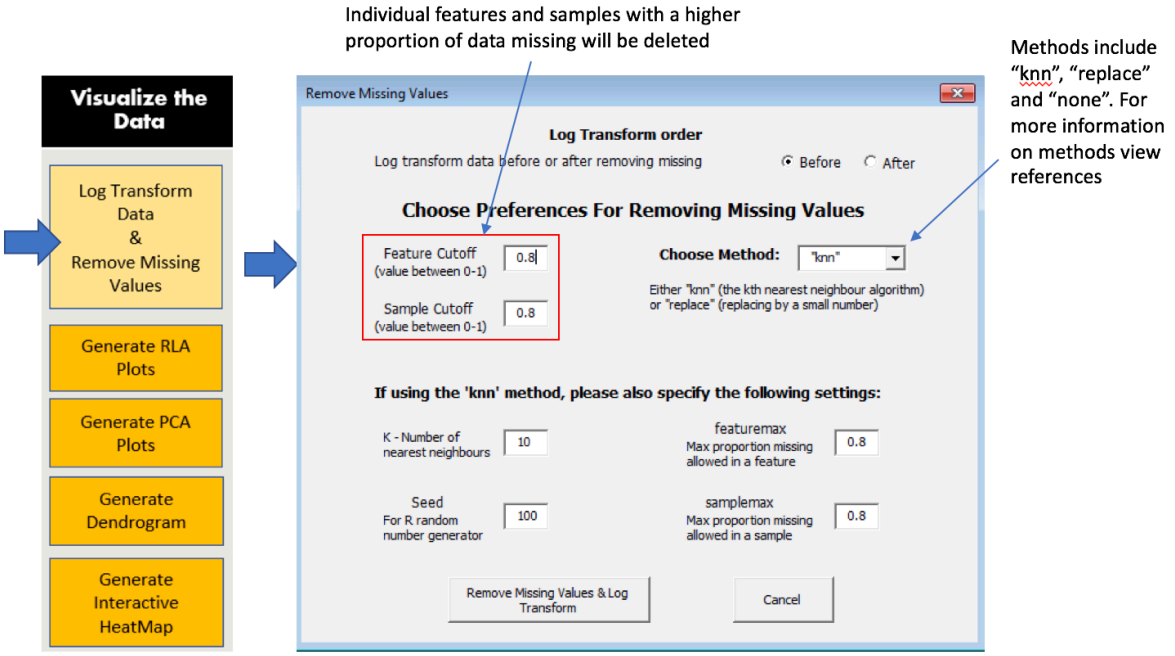
Visualize the Data

The following section refers to the visualize part:



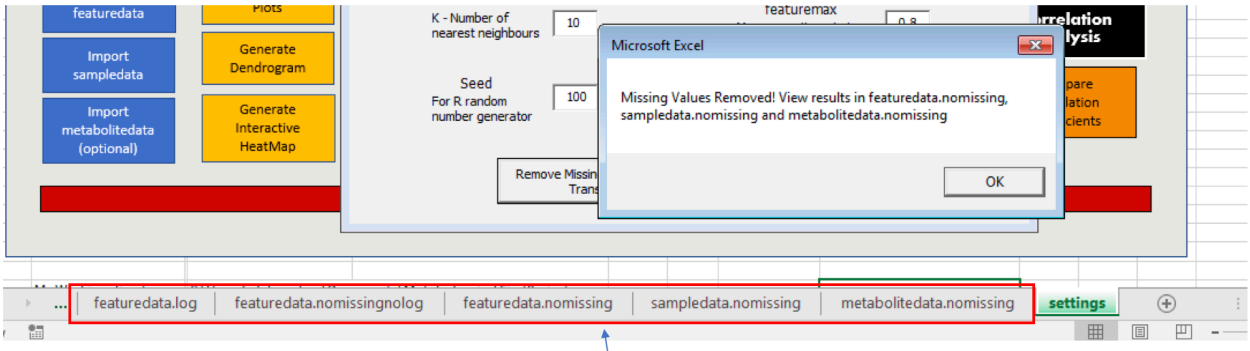
Log Transforming the data and removing missing value (mandatory):

A frequent issue in metabolomics data sets is the occurrence of missing values. It is important to reduce the number of missing values as much as possible by using an effective pre-processing procedure. For example, a secondary peak picking method can be used for LC-MS data to fill in missing peaks which are not detected and aligned.



"knn" – use k nearest neighbours method to replace missing values.
"replace" – replaces missing values by half the minimum value in featuredata.

Clicking '*Remove Missing Values & Log Transforming'* the following appears:

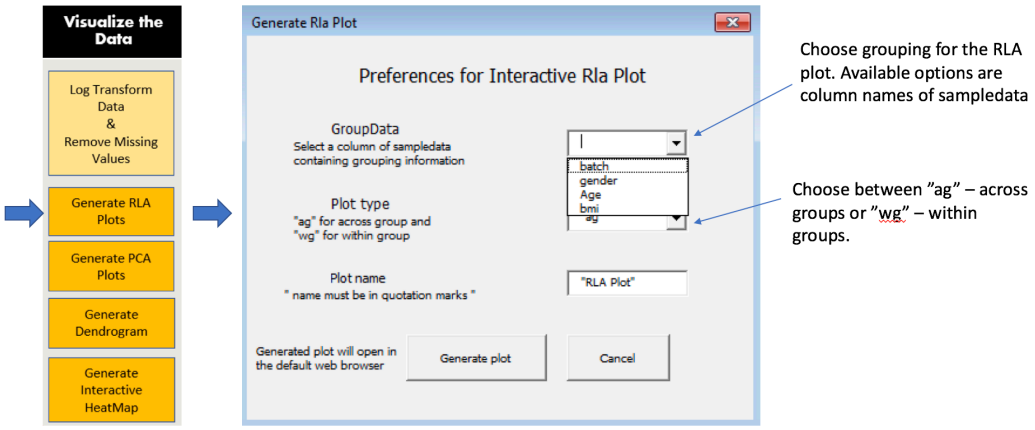


Note the new sheets that appeared, they have respectively the regular log transformed featuredata, the log transformed featuredata with missing values removed, featuredata with missing values removed and without log transformation, sampledata with rows removed corresponding featuredata.nomissing, metabolitedata with rows removed corresponding to featuredata.nomissing.
Unless you are interested to view or copy any of this data, those sheets are only going to be used for further internal functions.

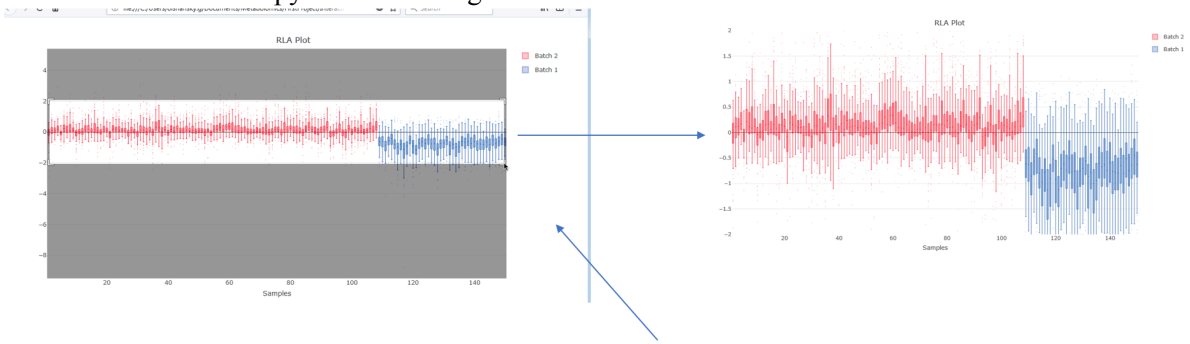
Now the plots in *Visualize the Data* can be generated! The data generated is also going to be used for the Normalization section.

RLA plots

One way of visualising the log transformed metabolomics data is the use of *across group* or *within group* relative log abundance (RLA) plots (De Livera et al. 2012 De Livera et al. (2015)).

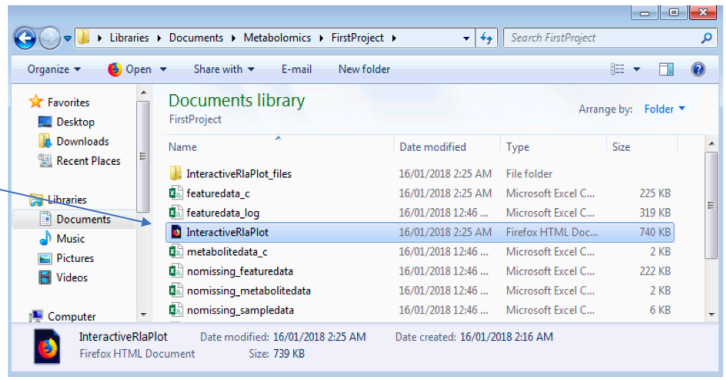


Setting groupdata to *batch* and selecting *Generate plot* opens the interactive plot in the default web browser and saves a copy in the working folder.



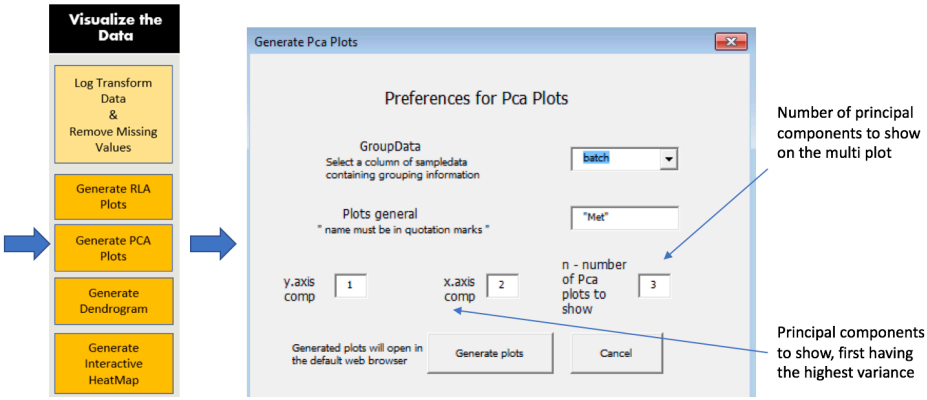
Interactive plot opens in the default browser, to zoom in, simple select the required part

Copy of all plots together with all other generated files are stored in the working directory



PCA plots

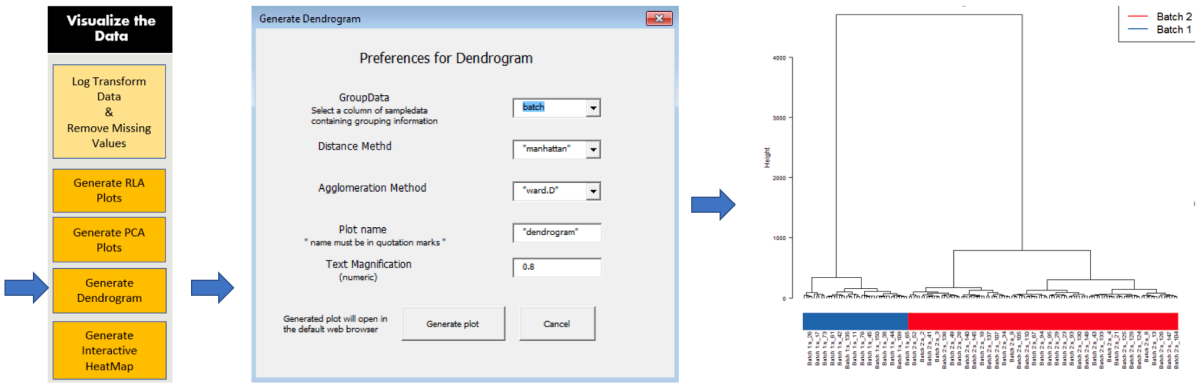
The following function can be used to obtain multiple plots for exploration of the principal components of the *featuredata* matrix: a bar plot indicating the variance explained by each principal component, scores and loading plots with specified axes (interactive and non-interactive), and a pairs plot of the first *n* principal components. These plots are useful in identifying any outlying samples and getting a preliminary understanding of the structure of the data.



All produced plots are stored in the working directory, with interactive plots opened in the browser and static plots located in the new *plots* sheet.

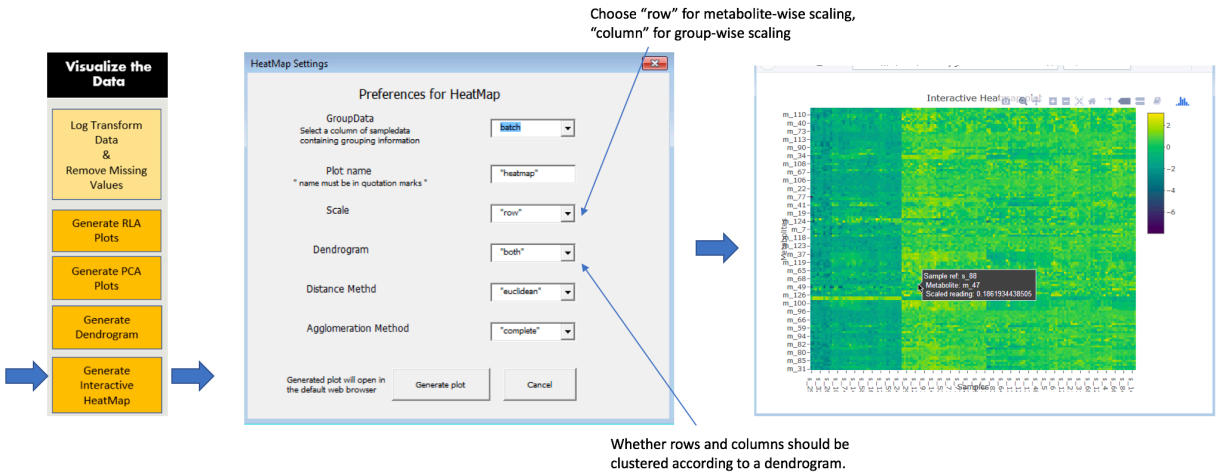
Dendrogram

Generates a dendrogram to visualise clustering structures in the data, many different methods ae available.



HeatMap

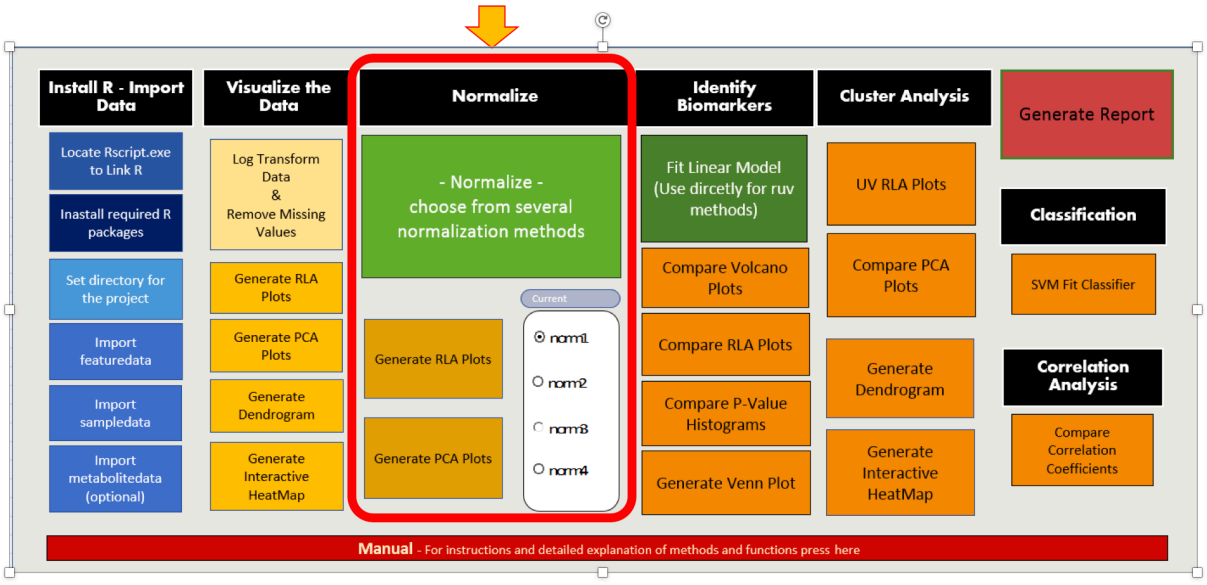
The HeatMap produced can reveal interesting structures in the data.



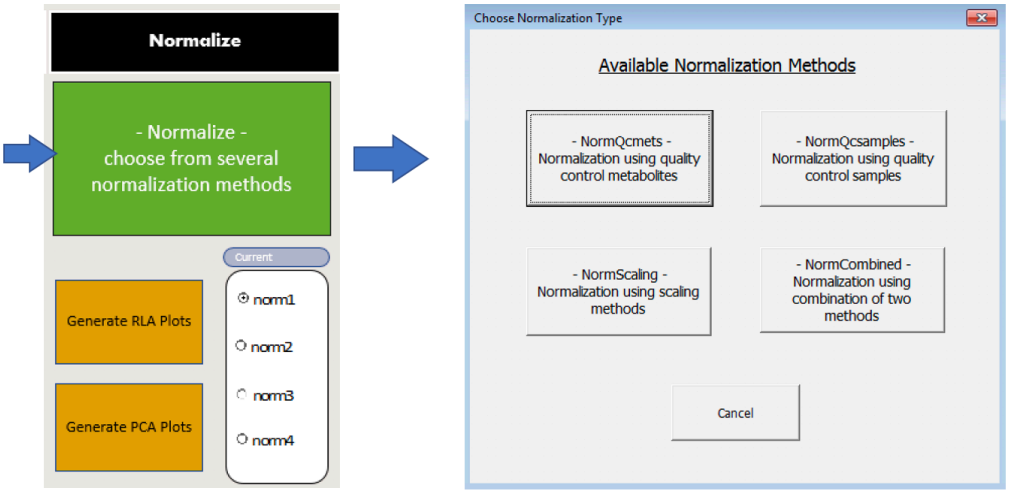
Normalization

Normalization methods presented in this package are divided into four categories, as those which use (i) internal, external standards and other quality control metabolites (*NormQcmets*) (Sysi-Aho et al. 2007, Redestig et al. (2009), De Livera et al. (2012), De Livera et al. (2015), Gullberg et al. (2004)) (ii) quality control samples (*NormQcsamples*) (Dunn et al. 2011), (iii) scaling methods (*NormScaling*) (Scholz et al. 2004, Wang et al. (2003)), and (iv) combined methods (*NormCombined*) (Kirwan and Broadhurst (2013)).

The normalization methods are accessible in the following section:



Clicking on the Normalize button opens the following menu enabling the choice of different normalization methods.

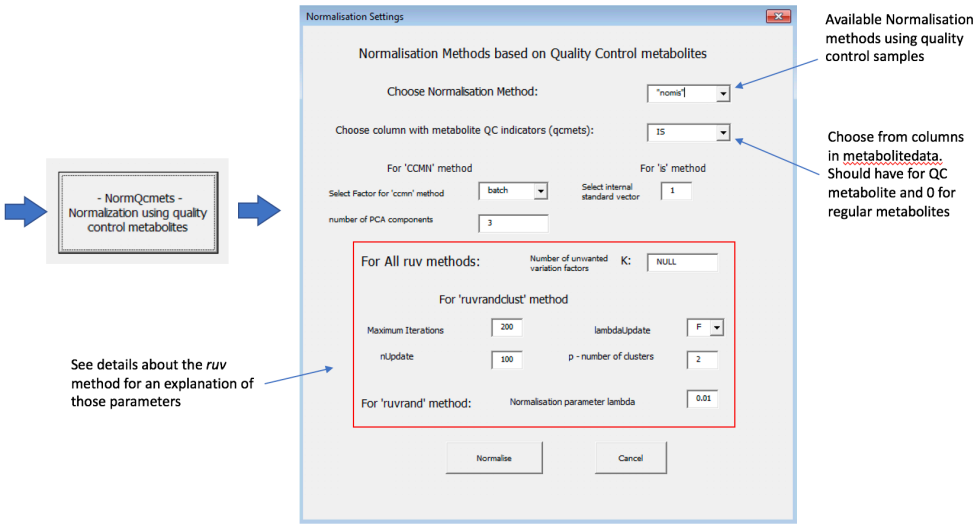


NormQcmets

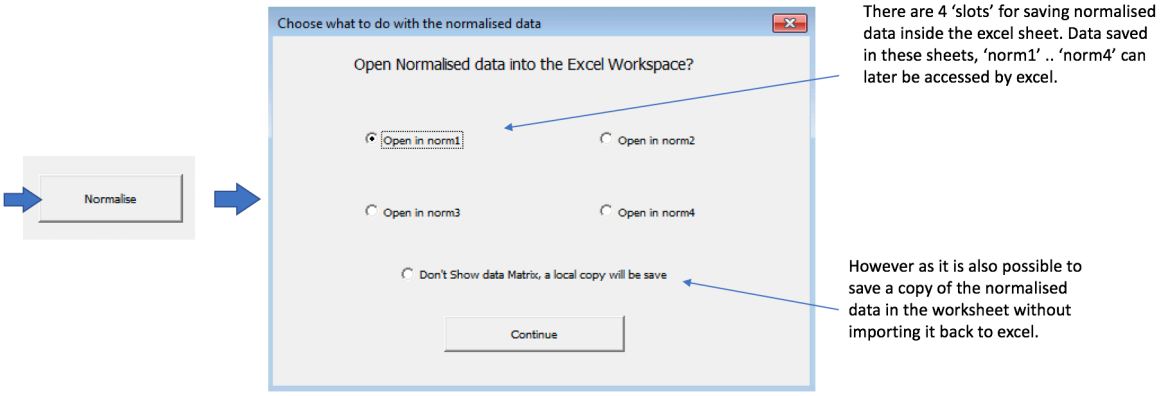
The approaches in *NormQcmets* use internal, external standards and other quality control metabolites. These include the *is* method which uses a single standard (Gullberg et al. 2004), the *ccmn* (cross contribution compensating multiple internal standard) method (Redestig et al. 2009), the *nomis* (normalization using optimal selection of multiple internal standards) method (Sysi-Aho et al. 2007), and the remove unwanted variation methods (J. A. Gagnon-Bartsch, Jacob, and Speed 2014) as applied to metabolomics using “ruv2” (De Livera et al. 2012), “ruvrand” and

"ruvrandclust" (De Livera et al. 2015). Note that *ruv2* is an application specific method designed for identifying biomarkers using a linear model that adjusts for the unwanted variation component.

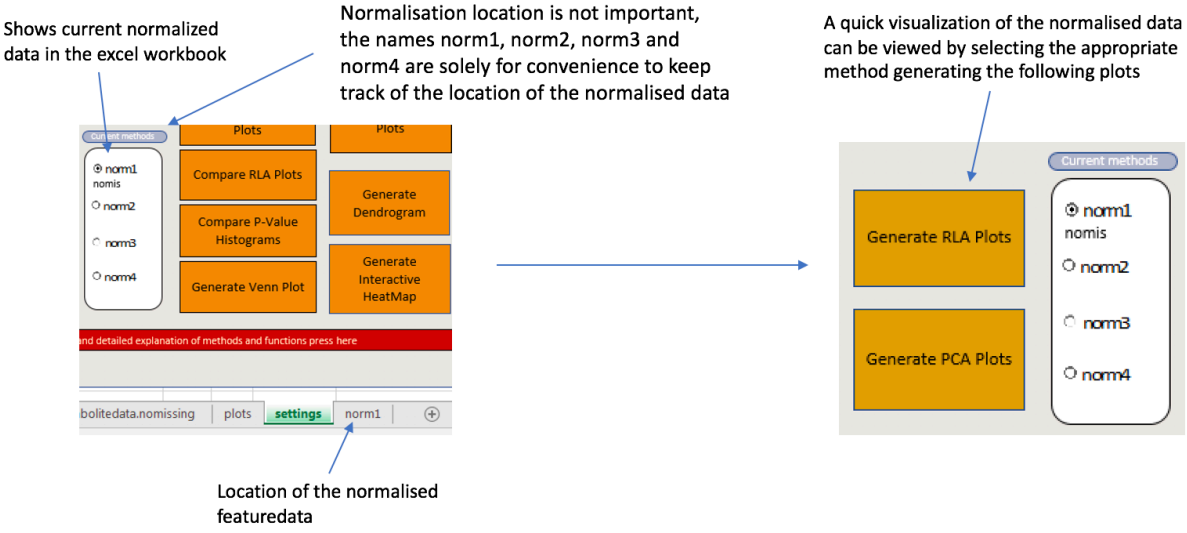
To Normalize:



After Clicking the Normalise button the screen asking you where the normalized data is to be saved appears.



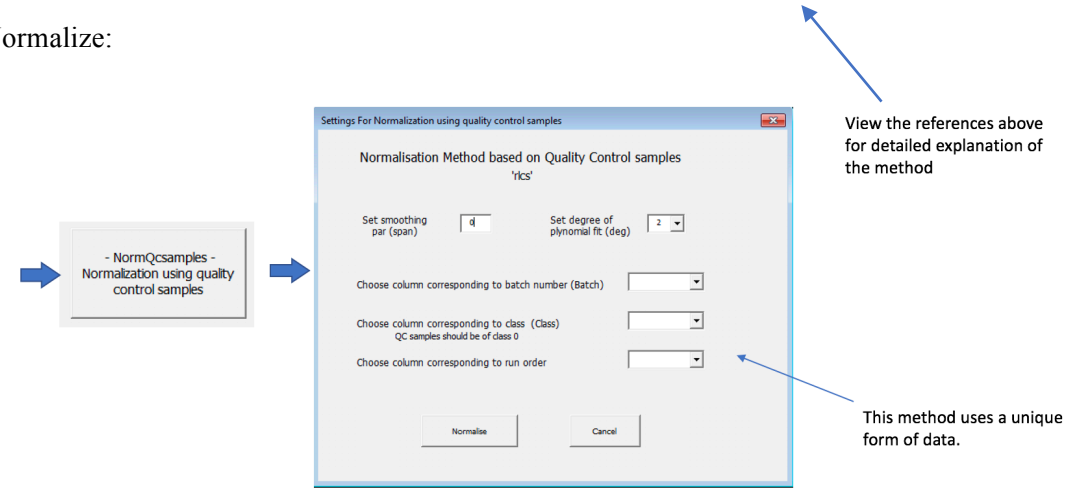
Upon clicking continue, you will return back to the settings sheet but you can notice some changes:



NormQcsamples

This function is based on the quality control sample based robust LOESS (locally estimated scatterplot smoothing) signal correction (QC-RLSC) method as described by Dunn et al. (2011) and impletemented statTarget (Luan 2017)

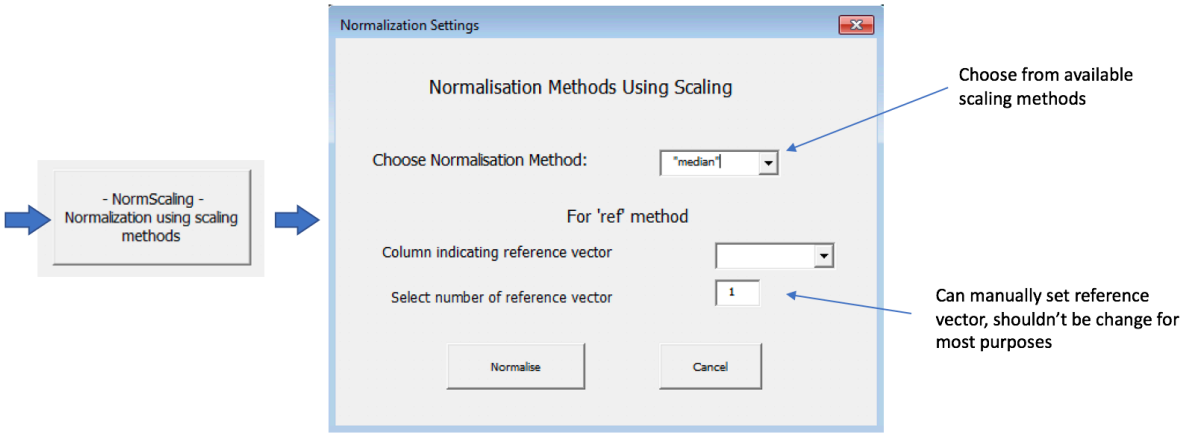
To Normalize:



NormScaling

The scaling normalization methods (Scholz et al. 2004, Wang et al. (2003)) included in the package are normalization to a total sum, normalisation by the median or mean of each sample, and are denoted by *sum*, *median*, and *mean* respectively. The method *ref* normalises the metabolite abundances to a specific reference vector such as the sample weight or volume.

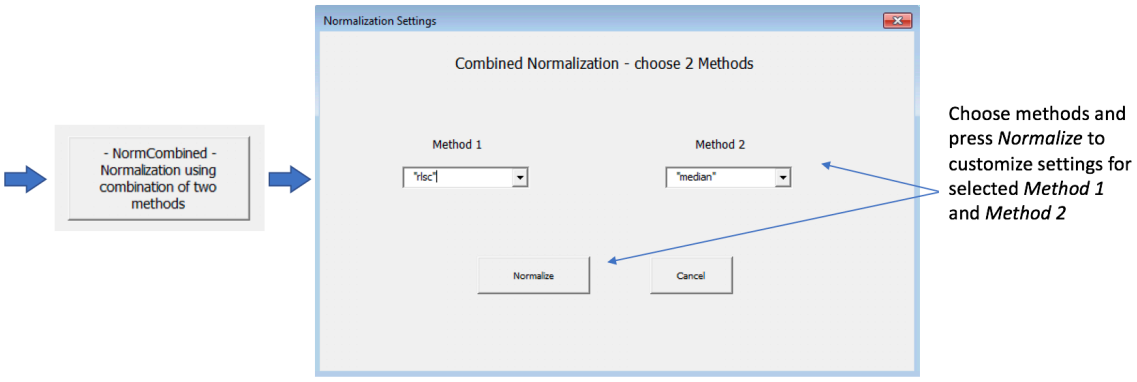
To Normalize:



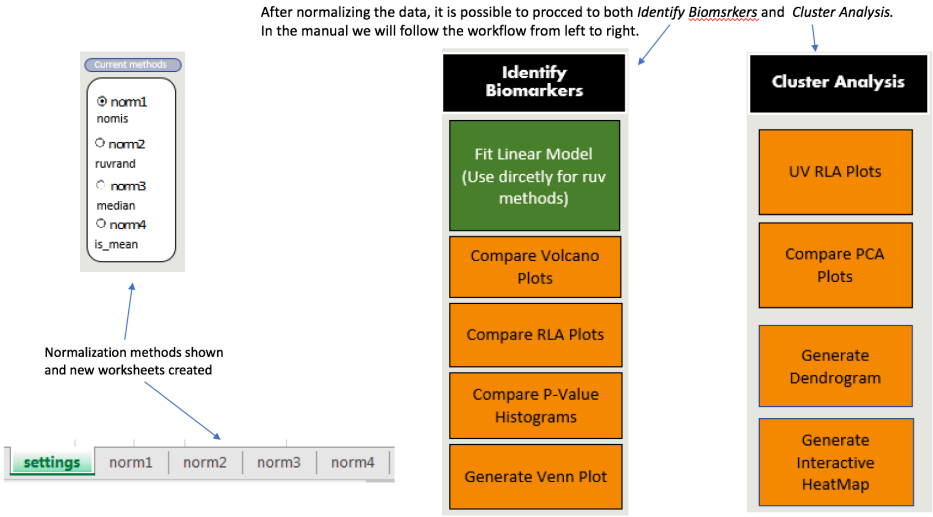
NormCombined

In some circumstances, researchers use a combination of the above normalizations (i.e., one method followed by another). This can be achieved using the *NormCombined* function. The function defaults to employing 'rlsc' approach followed by the `median'.

To Normalize:



Note that normalizing the data is not necessary to proceed to fitting a linear model although it is highly recommended to try a few normalization methods when analysis data.



Assessing and choosing normalization methods

The criteria for assessing and choosing a normalization method implemented NormlizeMets have been described in detail by De Livera et al. (2012), De Livera et al. (2015) and J. A. Gagnon-Bartsch, Jacob, and Speed (2014).

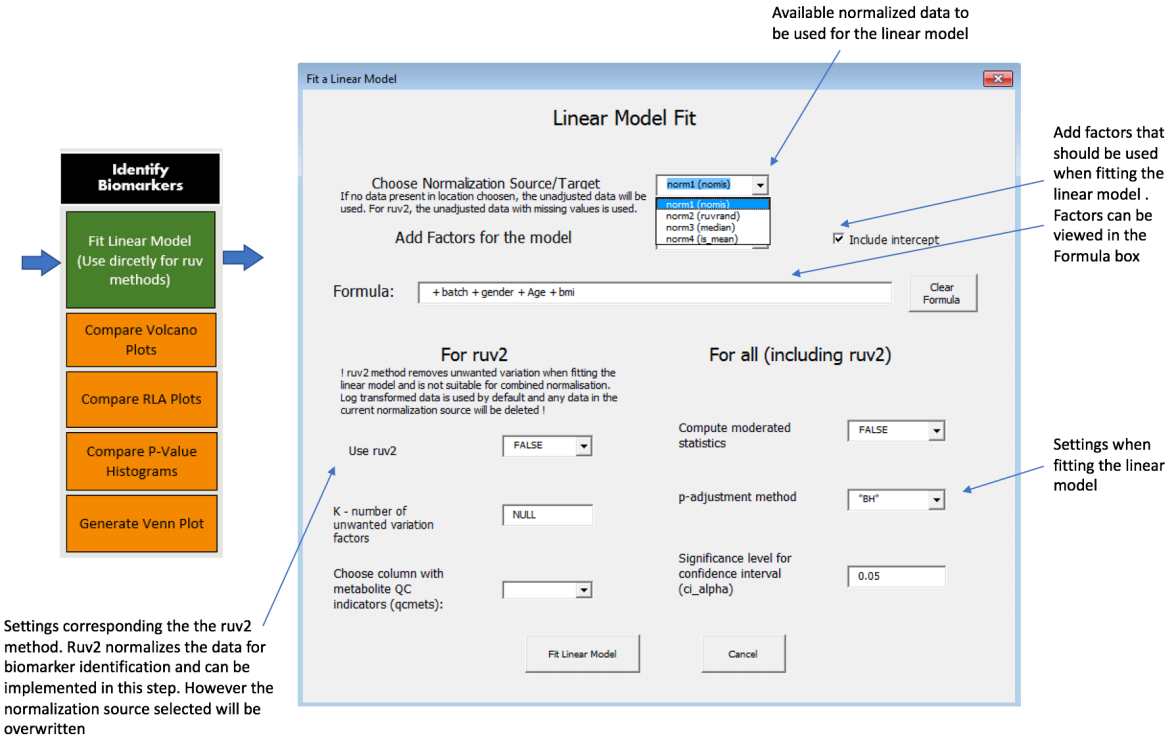
Identifying Biomarkers

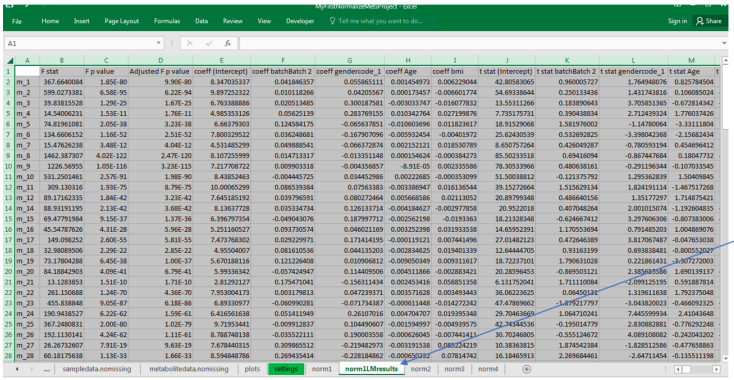
To view and compare the biomarkers identified, first a linear model has to be fitted to the data.

Fit Linear Model

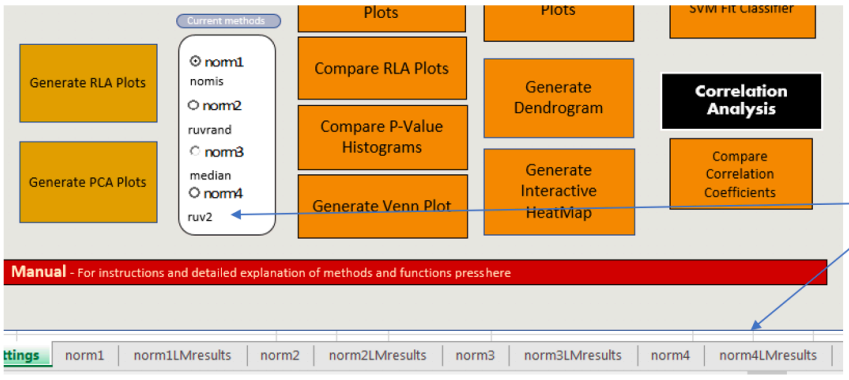
A linear model has to be fitted for every Normalization method that is to be used down the line for Biomarker identification. Setting from one ‘run’ of the Fit Linear model will be saved for the next.

To Fit Linear Model:





A new sheet is created, it stores the linear model output. This data should not be altered manually as other functions will need to access it.

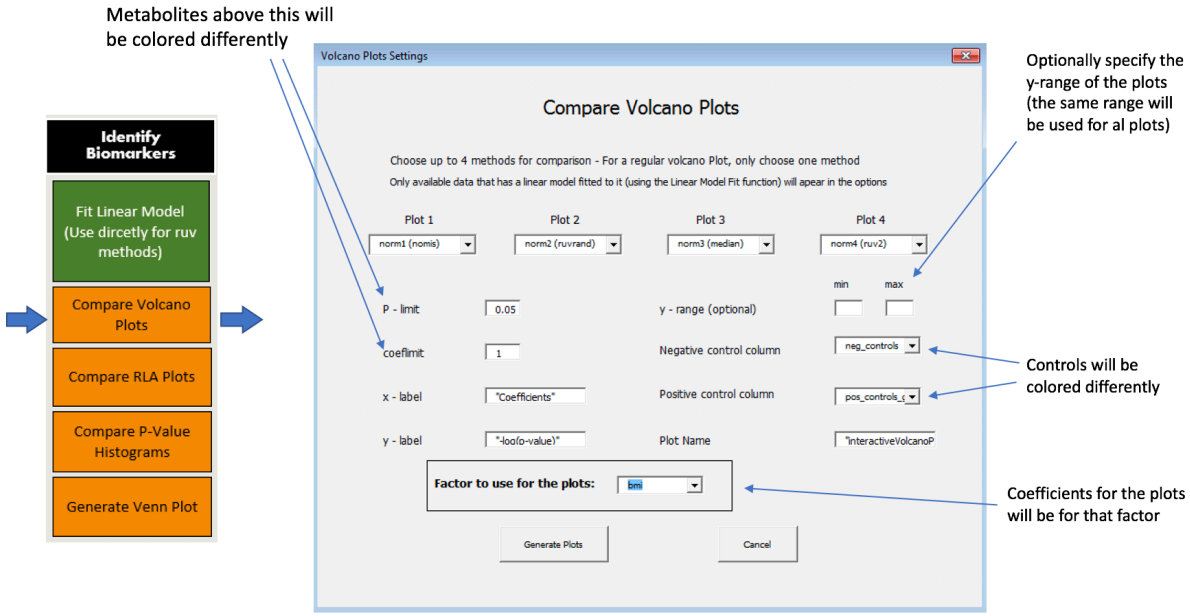


Note that choosing ruv2 overrides the normalization target

Volcano Plots

Volcano plots are useful in identifying biomarkers and generally assessing the normalization.

Compare Volcano Plots:

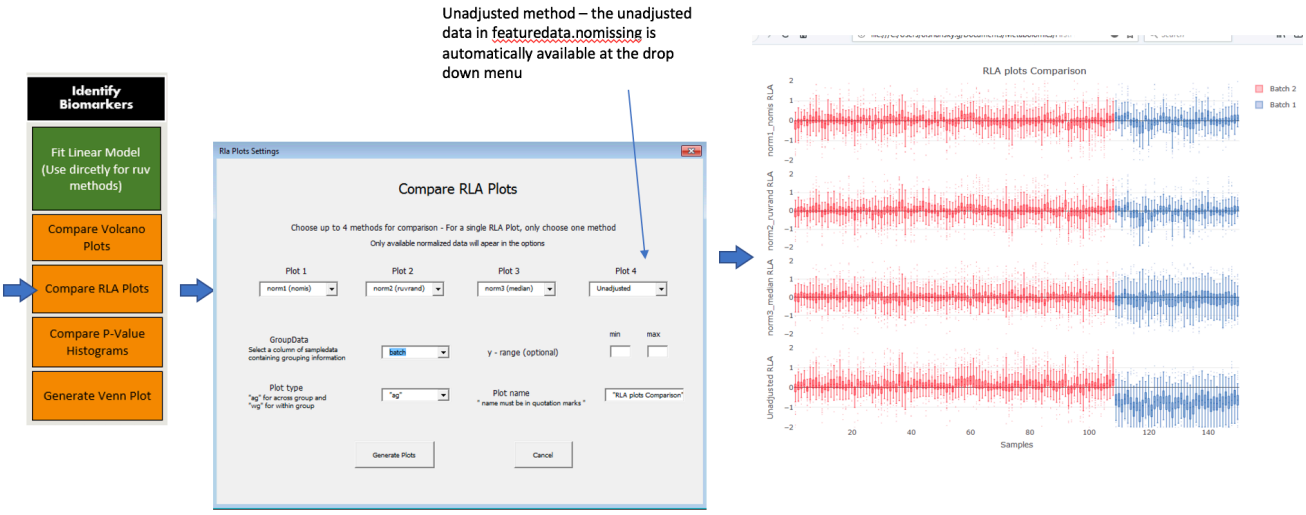


The plot is saved in the working directory and opens in the default browser.

Compare RLA plots

Used to assess normalization by comparing relative log abundance plots, similar input to the Generate RLA plots function

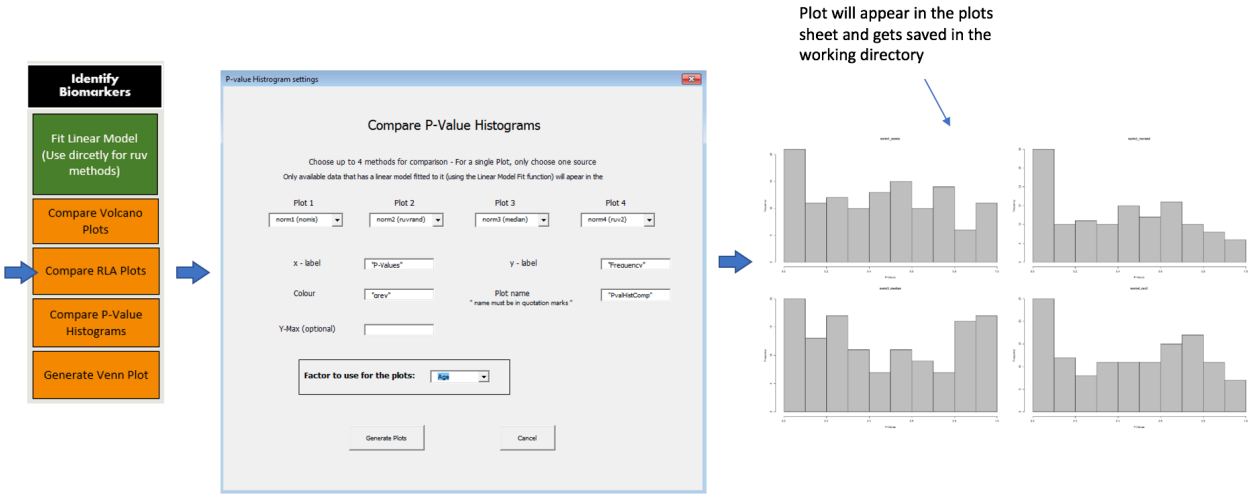
Compare Rla Plots:



Compare P-Value Histograms

Compare histograms of the coefficient's p-values. The distribution of the p-values should be used to assess the success of the normalization.

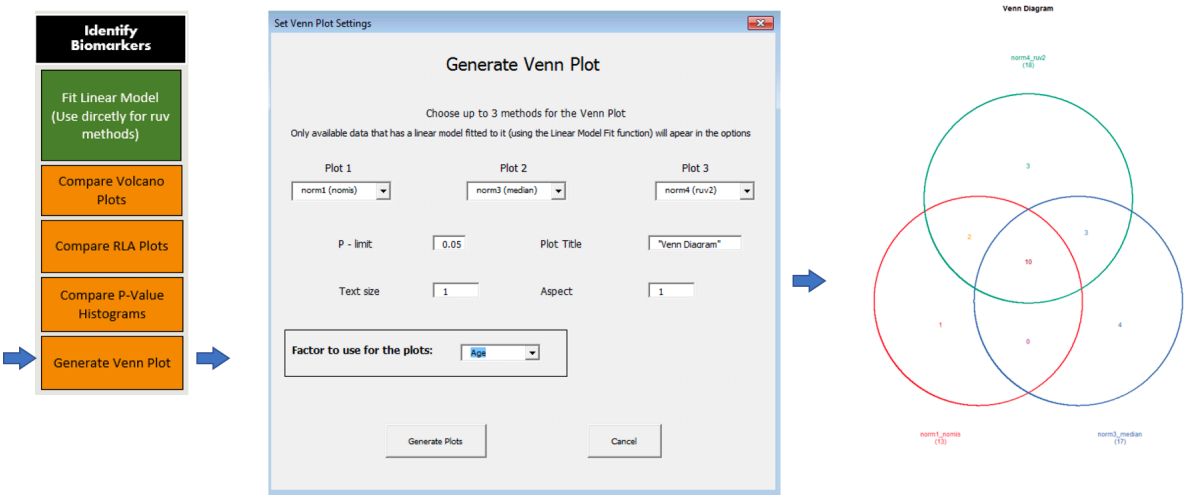
Compare P-Value histograms:



Generate Venn Plot

Generates a Venn plot that compares the biomarkers identified by the different normalization methods.

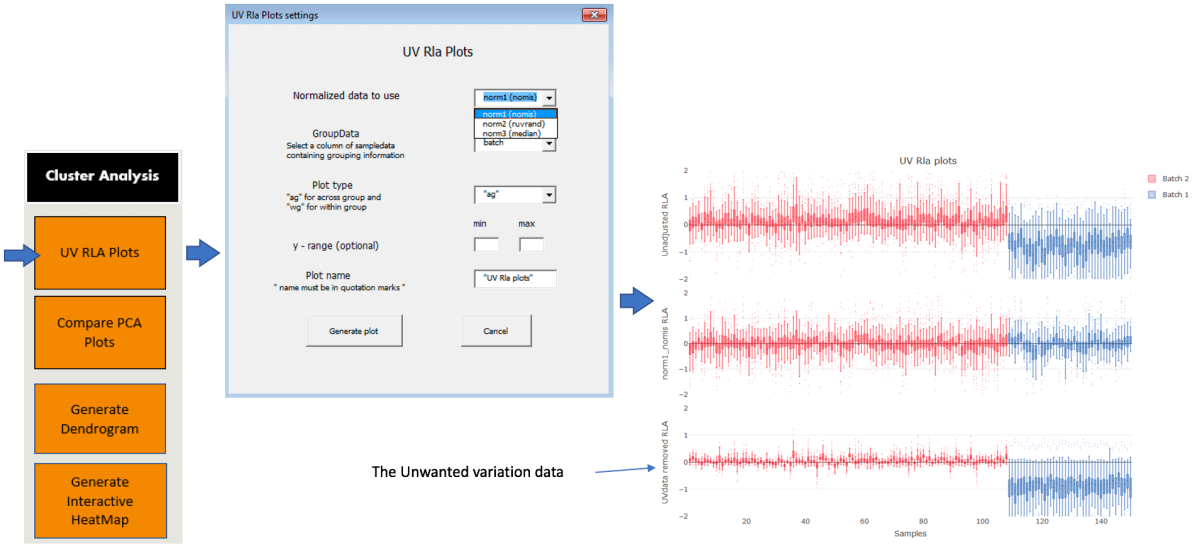
Generate Venn Plot:



Cluster Analysis

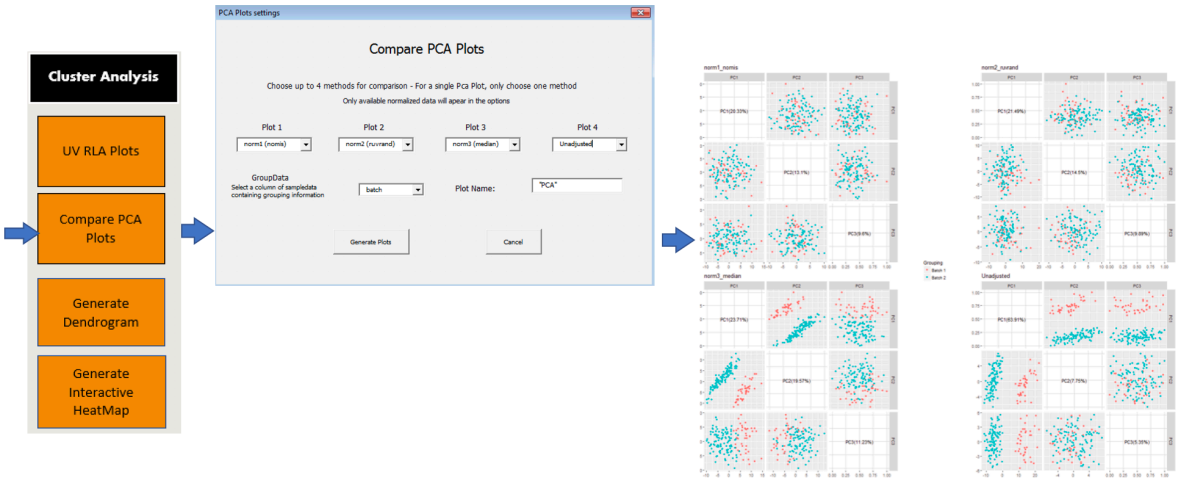
UV RLA plots

Unwanted Variation relative log abundance plots enable visualisation of the unwanted variation removed by each normalization method.



Compare PCA Plots

Compare principal component multi-plots for differed normalization methods.



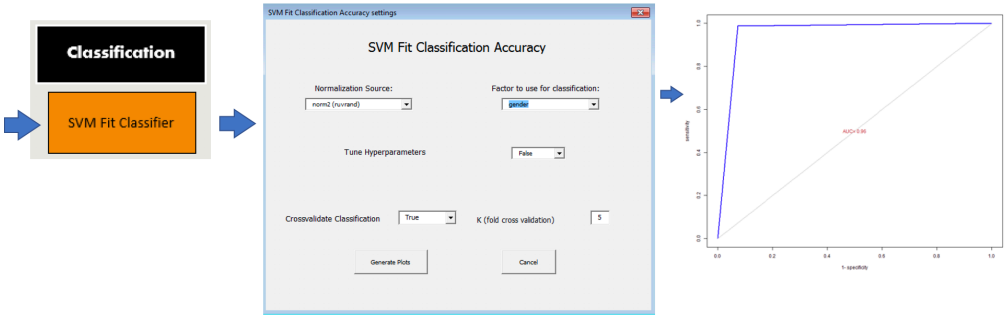
Generate Dendrogram and Generate Interactive HeatMap are identical to those discussed in the *Visualize the Data* section. The user has to choose the normalized data to be used.

Classification

Classification accuracy is a good way to assess the success of a given normalization method

SVM Fit

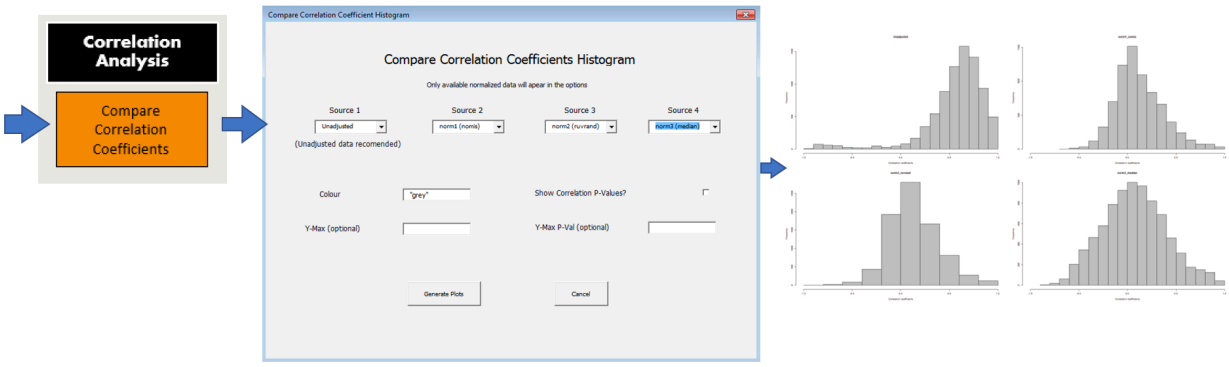
The Support Vector Machines method is used to classify the data, the classification accuracy is then assessed based on the factor specified.



Correlation Analysis

It is important to look at correlation coefficients when normalizing data.

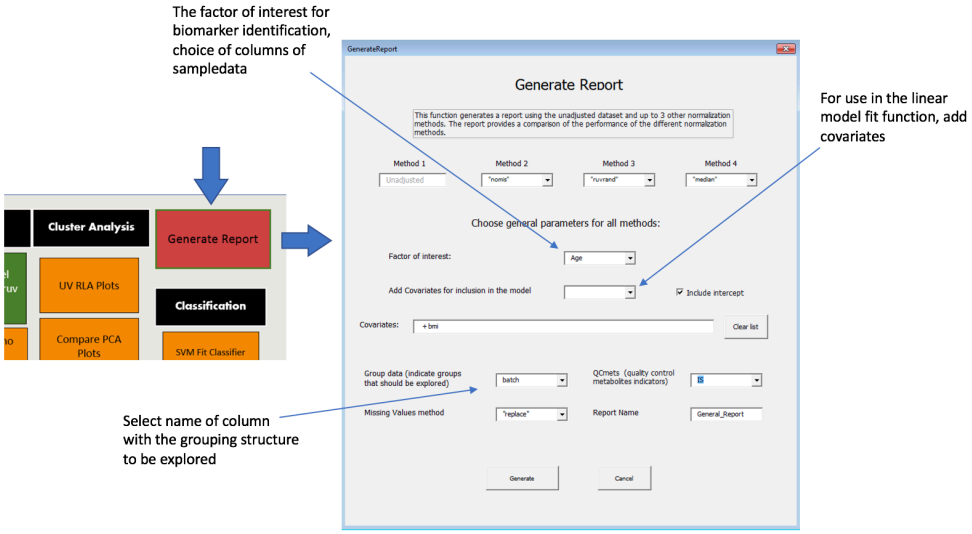
Compare Correlation coefficients



Generate Report

The *Generate Report* function generates an interactive report based on basic user input. There is a choice of up to 3 normalization methods to be included together with the unadjusted data. The report includes various plots and diagnostic to assess the normalization. Guidance on interpretation of the various plots, together with notes of what the user should look for when assessing the results is provided in the generated document.

Generate Report:



An example report is available in your downloaded ExNormalizeMetsSetup.