

MetaboClust User Guide

Martin Rusilowicz

29th September 2017

1 System requirements

For large datasets, a 64-bit system with 8Gb RAM is recommended. MetaboClust is dependent on the .NET framework. If you are running a recent version of Windows it is more than likely that this is already installed on your computer. If not you will need to download the installer version (see below), or download and install the Microsoft .NET framework from <https://www.microsoft.com/net/download>.

2 Downloading binaries

If the .NET framework is installed, download the most recent MetaboClust zip file from

https://bitbucket.org/mjr129/metabolitelevels_release. After downloading and unzipping, launch *MetaboliteLevels.exe* to start the application.

If the .NET framework is not already installed, a full installation, which includes the .NET framework, desktop and start-menu shortcuts and an un-installer, can be downloaded instead. In this case download the most recent MetaboClust(installer) zip file. After downloading and unzipping, run *Setup.exe* and follow the on-screen instructions. The application will be installed using Microsoft ClickOnce - for troubleshooting and details see <https://msdn.microsoft.com/en-us/library/t71a733d.aspx>. After the install you should be able to run the application from your start menu, or by launching *MetaboliteLevels.exe* from the folder you installed the application to.

Note

If an error message appears when you try to start the application, check that the latest version of the .NET framework is installed and working.

3 Compiling from source

If a full installation is preferred, which (if required) includes the .NET framework, desktop and start-menu shortcuts and an un-installer, the *Installer* version can be downloaded instead. MetaboClust is written in C# using Visual Studio 2015. The source consists of three projects, all of

which must be downloaded:

Project	Relative path	Contents	Download URL
MetaboliteLevels	./MetaboliteLevels/MetaboliteLevels/MetaboliteLevels.csproj	The main application	https://bitbucket.org/mjr129/metabolitelevels
MChart	./MChart/MChart.csproj	Charting library	https://bitbucket.org/mjr129/mchart
MGui	./MGui/MGui.csproj	Helper library	https://bitbucket.org/mjr129/mgui

From the *downloads* page of each of the projects, select *download repository*. Unzip each of the downloads to a new folder on your disk. If any of the above libraries show as missing make sure they are present in the correct folder, or modify your solution to target the correct path.

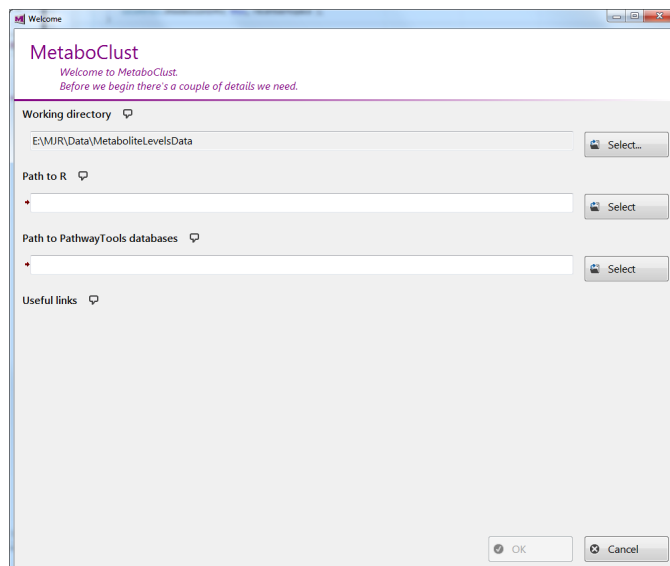
MetaboClust also requires the following libraries. Initially these will show as *missing*, but should be downloaded automatically by *NuGet* during the first build. If you have disabled *NuGet* in VS2015 you will need to add the libraries to the solution manually.

3.1 Running the source

Build and run the *MetaboliteLevels* project to start the application. Note that due to optimisations being skipped, the application will run considerably slower if the build mode is set to *<debug>* and/or a debugger is attached.

- MathNet.Numerics
- RDotNet
- JetBrains.Annotations

4 Initial setup



When MetaboClust starts for the first time the initial setup screen shown above is presented and requires the following information to be entered:

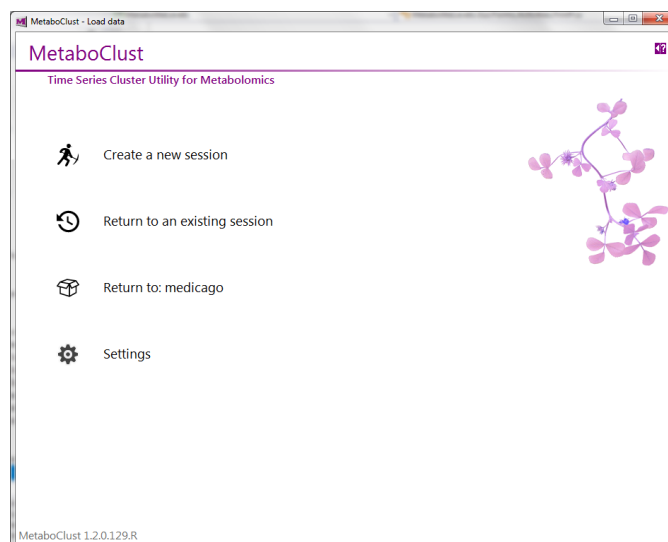
Initial setup options

- **Working directory** – This is where the application stores its data. By default this is the application's home directory. The default value should suffice in most case but can be changed (e.g. if administrator permissions deny read-write access to that folder).
- **Path to R** – MetaboClust uses R to operate and needs to know where R is located. Clicking the *«select»* button to the right of the text box should automatically detect the location of R and present a drop-down list of the versions of R available. If MetaboClust cannot find an R installation, the path to R will need to be specified manually. Pressing the *«select»* button (and then, if required, the *«browse»* option) will prompt you to locate the R installation. On Windows, R is usually located at `C:\Program Files\R\R-x.x.x\bin\x64`, where `x.x.x` is the version. This folder can be identified by the presence of the R library, `R.dll`.
- **Pathway tools databases** – MetaboClust can use Pathway Tools databases to make identifications. If any databases are

already present on the system MetaboClust can be directed to them here. If no databases are available the «*select*» button will offer a default location which can be used to put the databases in when they are available.

When you are done, click the «*OK*» button to commit the selections. MetaboClust detects the presence of errors on most screens. The software will check a connection to R can be established, and check to make sure it has read/write access to the data folders. A greyed out «*OK*» button indicates an error and a small red arrow should point in the direction of anything amiss. Hover the mouse over the arrow for more details.

5 Loading data




Once the initial setup is completed the data-load screen shown above will appear. The ⚙ icon in the bottom right of the window presents a drop down menu and the «*edit paths and libraries*» option here will return you to the *initial setup* screen.

6 Creating a new session

A MetaboClust “session” is a database comprising your data, annotations and analyses. You need to create a session before any analysis is performed. Select «*create a new session*» on the data-load screen to create a new session. The application will walk you through its creation.

Important note

Clicking the «*show help*» button (or in later versions the  icon) will show a **context sensitive** help bar at the side of the screen containing up-to-date details of the input fields. For inputs requesting files the «*show file format details*» button within the help bar describes the expected layout of input files.

Clicking the «*Next*» button progresses to the next stage of input. If this greyed out a small red arrow will point to anything amiss. Hovering the mouse over the arrow should describe the problem.

Loading data

- **Template** – Allows you to start from a previous setup. Normally you will start with the «*blank template*».
- **Session name** – For your reference only
- **Data set**

Data source – If you have LC-MS data MetaboClust needs to know how the adducts are formed. If the data is not sourced from LC-MS, or automated annotations are not required, then select «*Source = Other*», otherwise select the column mode (positive or negative). The «*Source = Mixed mode*» option allows you to mix modes, but your «*peaks*» file must then contain an extra column specifying the mode of each peak («*1*» or «*-1*»).

Data matrix – The data matrix is a grid containing the recorded intensities, with 1 row per observation and 1 column per variable (peak). Row and column names must be provided and must specify unique names for all observations (which can be O1, O2, O3,... for example) and variables (V1, V2, V3, ... for example). See the help bar as described above for exact details.

Observation information – The observation information matrix gives details for each observation, with one observation on each row and one field of information in each column. Row names should contain the observation IDs as specified for the «*Data*» matrix and column headers should contain the field

names. Most fields are optional, but some features do require specific fields (for instance batch correction requires the *«batch»* and/or *«acquisition order»* fields). Since the exact file format may change with a new release, please see the help bar in the software itself for the list of fields (column headers) available.

Peak/variable information – The peak/variable information matrix gives details for each covariate. The software refers to these variable as peaks to avoid ambiguity with other variables, such as algorithm parameters. Please see the help bar for the list of fields available.

Alternate intensities – Sometimes another version of your data may be available, e.g. scaled or unscaled, normalised or raw. The alternate intensities option allows this to be loaded in for quick reference later. This data can be viewed, but it will not be used in the actual analysis. This feature is not present from version 1.2 onwards as an unlimited number of intensity matrices can be loaded from the file menu.

Condition names – If your experimental groups do not have intuitive names, e.g. just “1”, “2” and “3”, then this allows them to be mapped to more informative names.

- **Conditions**

Specify conditions – Details of the experimental groups can be provided here. The conditions should be the same as in the *observation information* file or, if present, the *condition names* file. This information is not mandatory, but if specified will allow the software to generate default statistics and filters (described later). Conditions can also be added manually later.

- **Statistics**

Auto-create statistics – If conditions have been specified, *t*-tests between experimental groups and controls, as well as Pearson correlations of the intensities for each group against time, can be generated. These options are not available if conditions were not specified, but can be calculated later.

Perform corrections – The UV-scale and centre data correction can be added to your pipeline here. This and other corrections can also be added or modified later.

- **Compound libraries** – These are the compound and pathway libraries used for annotations and pathway analysis. One or more

of these must be selected to enable automated annotations. If you don't have any libraries on your system then the list will be empty.

Adduct libraries – Adduct libraries are used for automated annotations. There are two adduct libraries built into MetaboClust, *All* and *Refined*. All adducts listed on <http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/MS-Adduct-Calculator> (Kind, 2016) are incorporated in the *All* library with a subset in the *Refined* library.

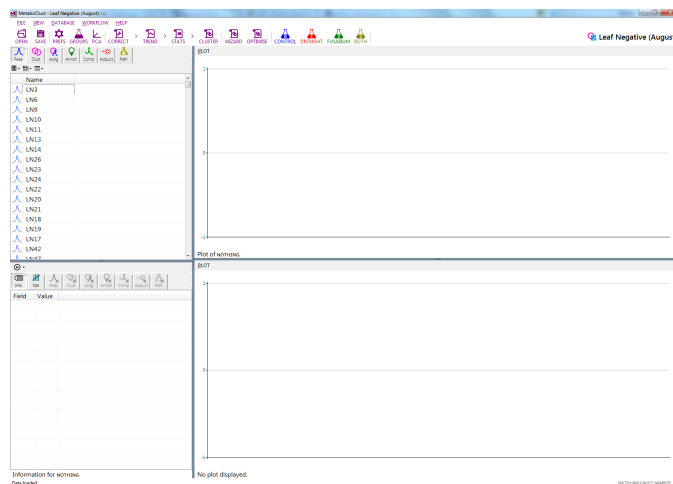
- **Automated identification** – Peaks can be annotated with potential metabolite identifications. For this option to be selected the `<tolerance>` and `<annotation status>` must be specified. Possible annotation statuses are `<tentative>` (unconfirmed), `<affirmed>` (computationally confirmed) or `<confirmed>` (experimentally confirmed).

Peak-peak matching – This annotates peaks with other peak IDs based on m/z similarity and is primarily used to search for related compounds.

Manual identifications – Manual identifications can be uploaded. See the help bar for the exact file format. The `<annotation status>` specified here will only be used if that information is missing from the file itself.

When all the fields you wish to select are complete click the `<OK>` button to load the data. This may take a few minutes, especially if automated peak-compound annotations are being performed. Saving the session will avoid this delay in future.

7 Data exploration



Once you have created or loaded a session you will be presented with the main screen, shown above. There is no fixed series of steps in data analysis, but a brief overview of possibilities will be presented here. The images shown are taken from the analysis of the *Medicago* leaf data, provided as a sample data set. This LC-MS dataset (negative mode) is provided as a sample data set and comprises 184 observations and 2920 peaks, with four experimental groups, C (Control), D (Droughted), F (Fusarium treated), and B (Both Fusarium treated and Droughted), as well as QC samples. Peaks were automatically annotated using the MedicCyc database (Urbanczyk-Wochniak et al., 2007, *Bioinformatics* 23(11), pp 1418-1423).

8 Univariate statistics


Double clicking a peak in the list to the top-left of the window will present a plot of the chosen peak should be displayed to the right. This is a useful first step to ensure the data has loaded correctly.

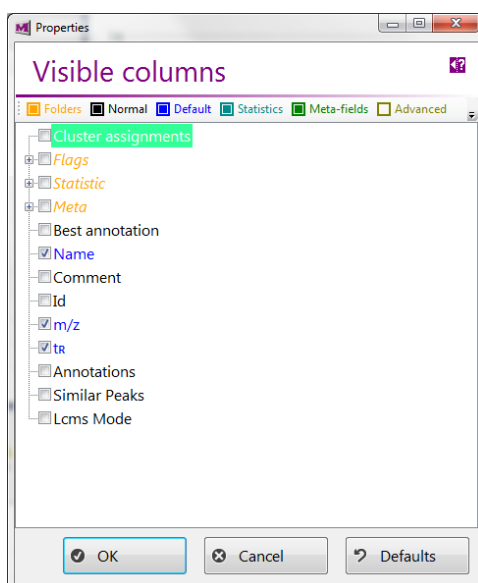
Graph controls

- **Left or right click** – Select point or series. Details on that point will be displayed above the graph. Repeatedly clicking will cycle through the display options.
- **Left click and drag** – Box-zoom

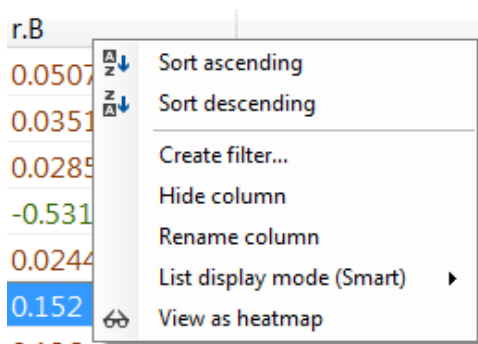
- **Mouse wheel** – Zoom in or out
- **Middle click** – Restore zoom and cancel selection.

The «plot» button above the graph provides plotting options, including exporting the plot to a file and toggling display of the legend.

To show more information about each peak, click the  icon above the peak list. In the example below, a particular statistic is chosen for display. The p-value for the t-test between experimental group G and the control group is denoted p.G and the Pearson correlation between the intensities for group G with time is denoted r.G.




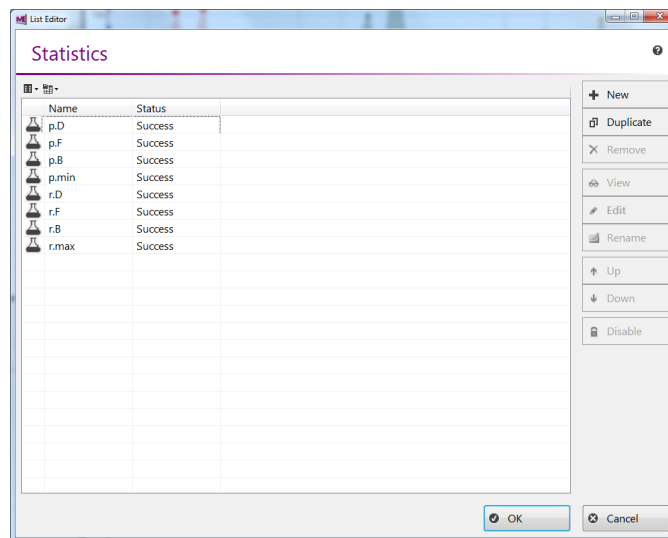
The values of the chosen statistic for each peak will be displayed in a new column and the peaks can then be sorted on these values by clicking on the column header to allow the most “significant” peaks to be located.



Clicking *«view as heatmap»* will display the column as a heatmap with brighter colours indicated higher values. Note that heat-map is ordered the same as the column, so if the peaks have been sorted by this column, the heat-map will appear as a gradient.

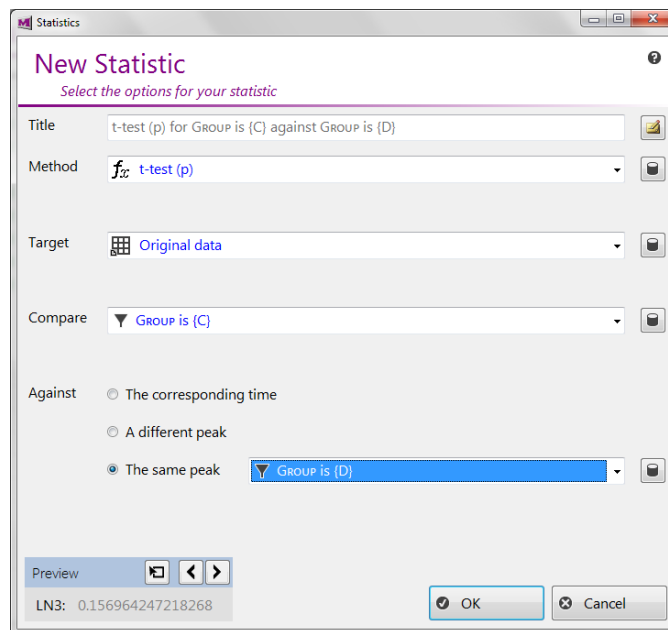


To add univariate statistics at this point, click the  *«stats»* option from the tool-bar, or select *«Database/Workflow/Statistics»* from the menu.






Name	Status
p.D	Success
p.F	Success
p.B	Success
p.min	Success
r.D	Success
r.F	Success
r.B	Success
r.max	Success

Select *«New»* to create a new statistic and the window below will appear.



Statistic fields

- **Title** – A title for the statistic. A name will be provided for you if you don't specify one. Clicking the  icon provides space to add detailed comments.
- **Method** – The statistic to calculate, click the  button to the right of the method to define your own methods.
- **Parameters** – If the method takes any parameters, enter them here. Multiple parameters are separated by commas. Clicking the button next to the text-box displays the parameters as individual inputs rather than a single-line text-box.
- **Target** – The intensity matrix for which the statistics are to be calculated. This may be from various stages of the analysis: “origin” indicates the original intensity matrix that was loaded and will be the only option until data-correction has been performed. Items marked with an asterisk designate dynamic sources. Pre-version 1.2 only two intensity matrices are available, being the latest set of observations (*<<*final correction>>*) and latest trend (*<<*final trend>>*).
- **For** or **Compare** – Selects the filter to define the set of

observations to be used in the test/comparison. Click the  button to define new filters.

- **Against** – Only available for bivariate statistics, specifies the second input vector:


The corresponding time – The times corresponding to the first input vector (e.g. to correlate intensity against time)

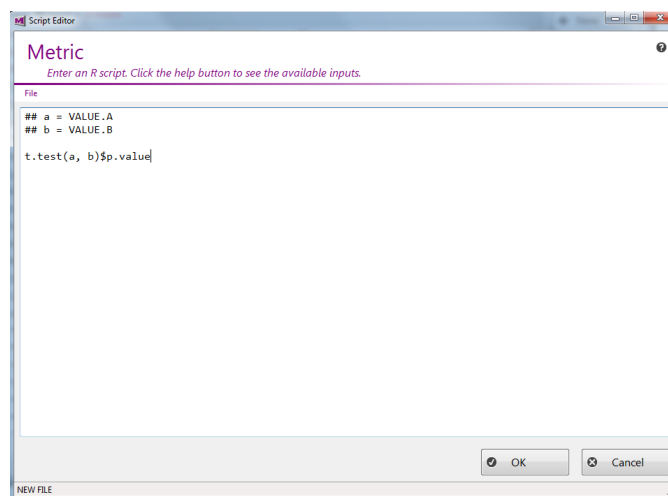
A different peak – The set of intensities corresponding to the first input vector for a different peak (e.g. to find similar peaks)

The same peak – The set of intensities sourced from different observations on the same peak (e.g. to contrast experimental and control observations).

For example, to calculate the mean of the QC samples select *«Method = Mean»* and *«For = Group is Q»*. The *«Preview»* box allows you to preview the result of your calculation on individual peaks. Select *«OK»*. Click *«OK»* again to leave the *«List editor»*. Any new or modified statistics will be recalculated. Edit the columns for the peaks list to show your new statistic.

9 Using additional R packages and user-defined scripts


Users own scripts for statistics, metrics or trends can be added via the new statistics window. To do this, click the  next to the method box, and then, rather than choosing from the list of available options, click *«new»* which will open the script editor as in the example below.



The script for any available R package could be added this way.

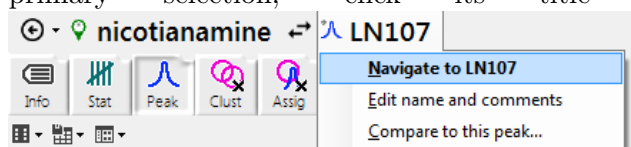
10 Exploring annotations




The set of coloured icons above the list allows paging between the database contents. If automated annotation was chosen when loading the data try selecting the  tab and viewing the annotations. Double click an annotation to view it. As there are no graphs associated with annotations, nothing will be displayed in the top-right, but the secondary list in the bottom-left should update to reflect the selected annotation. Above the secondary list select the *peak* tab to display the peaks associated with the selected annotation. Double-click the peak which appears in the list to plot the peak associated with the annotation.

Data exploration

Almost all of the data in MetaboClust can be explored in this way. The primary list (top) selects items within the dataset, whilst the secondary list (bottom) allows you to explore items within the context of the primary selection. To select an item in the secondary list as the primary selection, click its title above the list:



11 Multivariate statistics

An overview of your data can be obtained using PCA. Click the  button in the menu strip to launch the PCA window.

12 MVA

The MVA window presents a PCA plot of the dataset. The options to the left control the method of PCA and the display of the scores.

MVA controls

- **Method** – Switch between PCA and PLSR plots
- **Source** – Perform the analysis on observations or variables (peaks)
- **View** – Toggle between scores and loadings plots.
- **Legend** – Select what the colours on the graph represent.
- **Corrections** – View your data with various corrections.^a
- **Input** – Choose between performing PCA of all observations, or just your trend line (useful for noisy datasets).
- **Observations** – Select a filter to determine the set of observations to explore.
- **Peaks** – Select a filter to determine the set of peaks to explore.
- **View on main** – Displays the selected peak or observation on the main screen.^b
- **Mark as outlier** – Applies a filter to exclude the selected observation.
- **Next component** – Views the next principal component.
- **Previous component** – Views the previous principal components
- **Plot options** – Displays the set of plot options, including toggling display of the legend.

^a Corrections and trends are defined from the main screen.

^b Requires an object in the plot to have been selected first

For example, if the data were collected in batches, click the **LEGEND** – the **BATCH** to colour the plot by batch. Subsets of the data can be selected using filters. Click the **OBSERVATIONS** menu to show a list of observation filters by experimental group. As when creating a new statistic, if no filters are available you can click *«Observations/Edit observation filters...»* to create a new filter. The same can be done with peak filters by selecting *«Peaks/Edit peak filters...»*.

PCA can also be used for outlier removal. Click an observation in the plot



and select the MARK AS OUTLIER button. A new filter will be created, excluding that observation (or peak) from the dataset.

13 Data correction

Select «Correct» from the menu-bar of the main screen to open the corrections list. Click «new» to create a new one.

The data correction window presents a list of data-correction methods, as well as trend generation methods. Data-correction methods, such as "UV scale and centre" act alone, whilst the trend-generators can be used to perform batch correction and control correction.

Data correction options

- **Title** – A title for the correction method. A name will be provided for you if you don't specify one. Clicking the  icon provides space to add detailed comments.
- **Source** – Data on which to perform the correction.
- **Method** – The correction to be performed. Click the  button to the right of the method to define your own methods.
- **Parameters** – If the method takes any parameters, enter them here. Multiple parameters are separated by commas. Clicking the button next to the text-box displays the parameters as individual inputs rather than a single-line text-box.
- **Operator** – *Only available for trend-based corrections.* The correction takes the form $x' = f(x,t)$, where f is defined as divide or subtract. Generally a batch correction will use «divide» and control correction «subtract».
- **Filter** – *Only available for trend-based corrections.* Selects the set of points used to generate the trend.

The preview window allows you to preview the correction on an individual peak. For trend-based corrections the trend used will be highlighted to the left.

13.1 Examples

13.1.1 QC correction

Dividing by the mean of the QC samples in the batch is a fairly standard method of correcting for batch-differences in LC-MS, to use this select: «*Method = straight line across mean*», «*Corrector = Batch*», «*Operator = Division*» and «*Filter = Group is Q*».

13.1.2 Background correction


To perform background correction, as described in Rusilowicz et al., (2016) *Metabolomics* 12(3), pp 1-11, select «*Method = moving median*», «*Corrector = Batch*» and «*Filter = All*». You will need to enter the window width «*w*» parameter in this case. Experiment with values to find one that looks good in the plot.

13.1.3 Scale and centre


Select «*Method = UV scale and centre*». This correction should generally be performed *after* batch correction and therefore the «*Source*» parameter should point to the intensity matrix generated by your batch correction – QC or Background correction as described above.

13.2 Viewing corrections



Back on the main screen you will need to select your corrected dataset before your changes can be viewed. Click  «*dataset*» or the drop-down list next to it to select your modified data. You can use the «**Final correction*» meta-option to always keep your display up-to-date with the latest correction.

14 Trend line generation

A trend can be generated across observations to be used for background correction or clustering observations, or a trend can be obtained for each variable by averaging over the replicates. Select  «*trend*» from the tool-bar or «*Database/workflow/trends*» to define a trend. You will be presented with a list much like the «*correction*» window. (Pre-version 1.2 this will contain a «*no-trend*» entry, click «*remove*» to get rid of it). Click «*new*» to create a new trend.

Trend options

The «*New trend*» options are largely the same as those described in section 13.

14.1 Example

The simplest trend is the mean for each time-point. Select METHOD = MOVING MEAN with value $w = 1$. (w is the window width for the moving mean – a window width of 1 effectively the mean of the replicates).




14.2 Viewing trends

Back on the main screen you will need to select your trend before your changes can be viewed. Click «*trend*» or the drop-down list next to it to select your modified data. You can use the «**Final trend*» meta-option to keep your display up-to-date with the latest correction. After selecting your trend any peaks you plot will show the specified trend. You may see bold lines through 0 on peak plots pre-version 1.2 before a trend line has been defined.

15 Clustering

Clusters are created in the same way as corrections, statistics or trends. Select the «*Cluster*» option from the tool-bar of the main window.

Clustering options

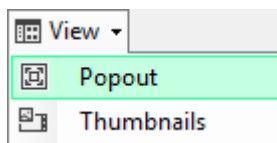
- **Title** – A title for the cluster analysis. A name will be provided for you if you don't specify one. Clicking the  icon provides space to add detailed comments.
- **Method** – The clustering algorithm to be performed. Click the  button to the right of the method to define your own methods.
- **Parameters** – If the method takes any parameters, enter them here. Multiple parameters are separated by commas.
- **Peaks** – Which peaks to cluster. A filter can be defined by selecting the  icon next to the list.

- **Distance** – The distance metric to be used to calculate statistics on clustering performance(next option).
- **Parameters** – Parameters required for the distance metric.
- **Statistics** – Statistics to calculate on clustering performance.
- **Source** – Data on which to perform the correction.
- **Observations** – The set of observations to use in the clustering vectors.
- **One vector per experimental group** – Normally one vector is created per-peak, select this option to “split” the peaks into one vector for each experimental group.
- **Parameter optimiser** – If the clustering algorithm takes parameters this option can be used to optimise them using statistics such as *silhouette width* or *BIC*.

15.1 Viewing clusters


On the main screen click the *«cluster»* icon above the primary list to view the clusters. Double click a cluster to plot it in the cluster plot area. Note that clusters are always plotted using the vectors with which they were created, so the *«trend»* and *«dataset»* visual options will have no effect on the cluster plot.

Clicking a vector within the cluster plot will select the peak associated with that vector as the secondary selection. Alternatively, select the *«peaks»* tab from the secondary list to show a list of peaks assigned to the selected cluster. Double clicking a peak in this list will plot the peak and highlight it in the cluster plot.






If you want to see a quick overview of all clusters, then click the *«View»* and *«Popout»* options above the list of clusters. By default each plot is scaled to fit the plot area, so flat clusters may appear as noisy. To change this and scale all clusters to the same Y-axis, change the plot options by going to the *«Prefs»* window and setting the *«Cluster»* – *«Y-axis range»* to *«Scale to matrix»*.

15.2 Metabolite and pathway exploration

With a cluster selected, clicking the «*compounds*» or «*pathways*» options will show compounds and pathways potentially highlighted by that cluster. Double-clicking these compounds or pathways will highlight the overlap between them and the cluster in the cluster plot. You can show or hide the degree of overlap, or sort clusters by overlap, by selecting the  icon in the secondary list.

A reverse exploration can also be performed, selecting a «*pathway*» or «*compound*» in the primary list will plot the trends of the peaks associated with the pathway or compound in the cluster plot. (You can change the «*dataset*» or «*trend*» in this case.) As for the clusters, selecting individual trends will plot the actual peak. Selecting «*clusters*» in the secondary list will show the clusters affected by peaks annotated with the pathway or compound.

16 General options

- Show or hide observations from experimental groups – Click the group icon in the main tool-bar to toggle group visibility, or select the  «*groups*» icon.
- Rename groups, peaks, etc. – Select the «*Database*» menu to show the database, then edit the group or peak. The groups database can be accessed quickly from the  «*groups*» icon in the tool-bar. Clicking the name of the session in the top-right of the main screen allows you to rename the session.
- Change display options – Select the  icon from the tool-bar.
- Find out which files were used to create a session – Select «*Help/Session information*» from the main menu
- Find an individual item – Click the «*name*» column of the peaks list and select «*filter*» from the menu to search for individual items.
- Get an overview of the session, including peak and observation counts – Select «*View/Miscellaneous functions*» and then «*View statistics*» from the window that appears.

17 Known bugs

MetaboClust is beta software. A list of known bugs is maintained on the download page. Please submit any bugs you find to this list.