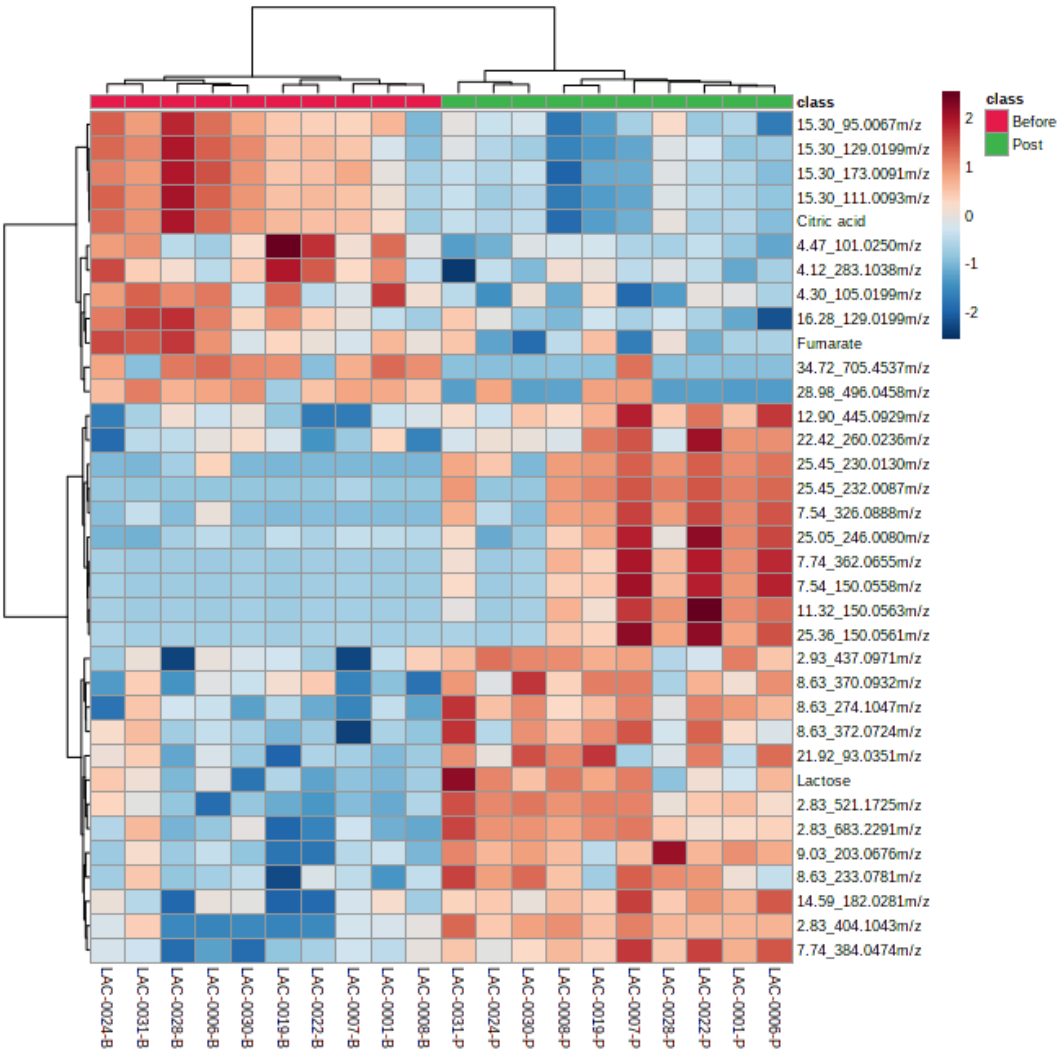
Part 1 (Exercise 1)

Data Processing and Statistics



3rd Oxford Metabolomics Data Processing and Analysis Workshop

Exercise 1: Overview

In this exercise you will be given a dataset from an untargeted metabolomics project and you are asked to process the data to prepare it for statistical analysis and then perform statistical analysis to answer a series of questions. The data table you are given has been compiled from individual LC-MS data files from the analysis of IDH1 mutant and IDH1 wild type glioblastoma cells. Please follow the step by step guide and answer the questions as you progress. We will discuss the results at the end of the exercise.

If you are using MetaboAnalyst for the first time please go through the exercises step by step. This includes screen shots demonstrating how to perform the analysis and display the results. If you have used MetaboAnalyst previously and feel you are sufficiently proficient you can attempt to complete the exercise without the aid of the step by step guide. The questions guiding you in this case start on the penultimate page of this document (page 25) and you can jump to there to complete the exercise and questions.

If you get stuck or have any questions along the way please put up your hand and those running the course will come and help.

To complete this exercise, you will need:

1. Access to MetaboAnalyst online which can be accessed here: https://www.metaboanalyst.ca/MetaboAnalyst/home.xhtml
2. The data table for this exercise: **'Exercise1_DATA_MA.csv'**.
3. These instructions to follow.

You should have already downloaded the data table for Exercise-1: Data Processing and Statistical Analysis if not you can find it on the Microsoft Teams site for the course (under the channel called '4. Data files on Teams). The data file for Exercise 1 is called **'Exercise1_DATA_MA.csv'**. **Please make sure you download a copy of this data file to your local computer before starting the exercise**. You may find it useful to have a hard copy of this exercise sheet or have it open on a separate screen when using MetaboAnalyst to perform the data processing and analysis.

Step by step guide (including screen shots)

Please follow the step by step guide below to analyse the dataset you are given and create a data analysis report in MetaboAnalyst. You will then have some questions to answer about the results (if you are an experienced MetaboAnalyst user you can go straight the list of talks which are given on the last page of this document).

Step by step guide:

1. copy the .csv file called **'Exercise1_DATA_MA.csv'** to your local computer.
2. Open MetaboAnalyst https://www.metaboanalyst.ca/
3. Select: **>>click here to start<<** which will open up the pyramid of modules (see Figure 1 below).

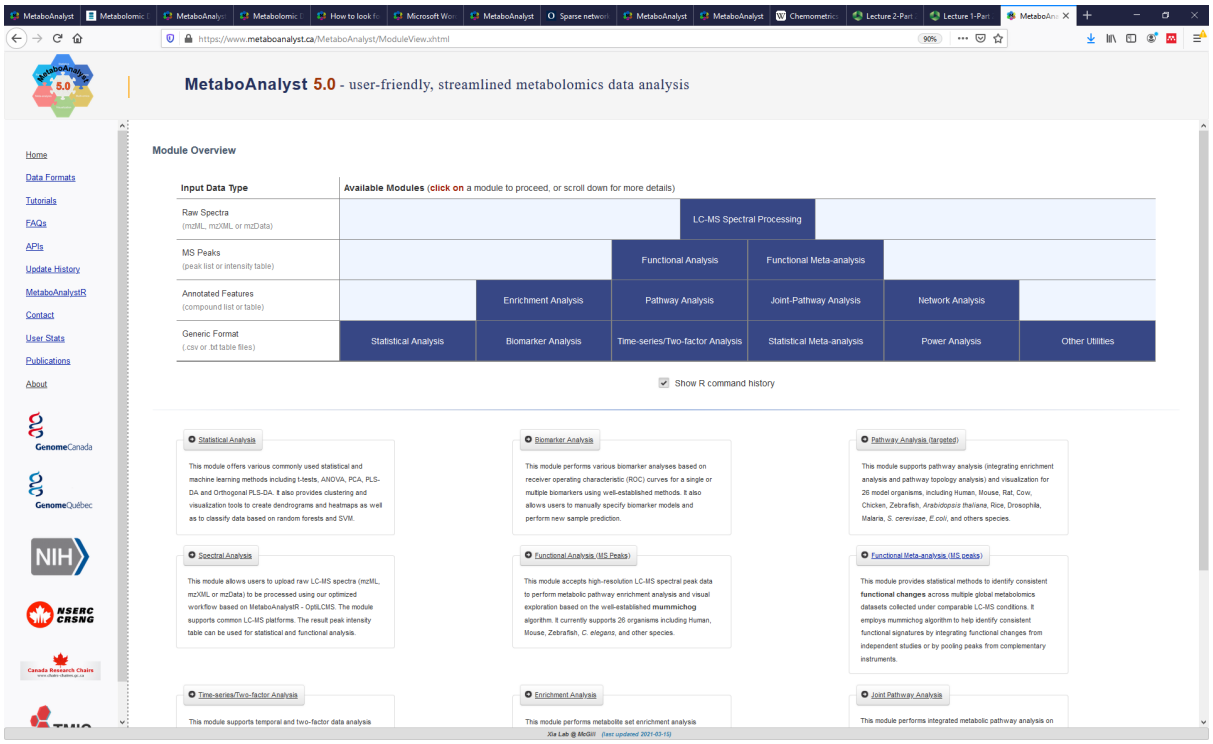


Figure 1: Pyramid of modules in MetaboAnalyst.

4. **Select 'Statistical Analysis'** (bottom left hand side) which will open up a window where you can upload your .csv dataset. Upload your .csv file using the settings as in Figure 2 below. I.e. select 'peak intensity table', 'samples in columns unpaired' and by browsing to the location of your .csv data file called: **'Exercise1_DATA_MA.csv'**

5. Then click on submit to the right hand side (ignore *Try our test data* below but this may be of interest to explore yourself after the workshop).

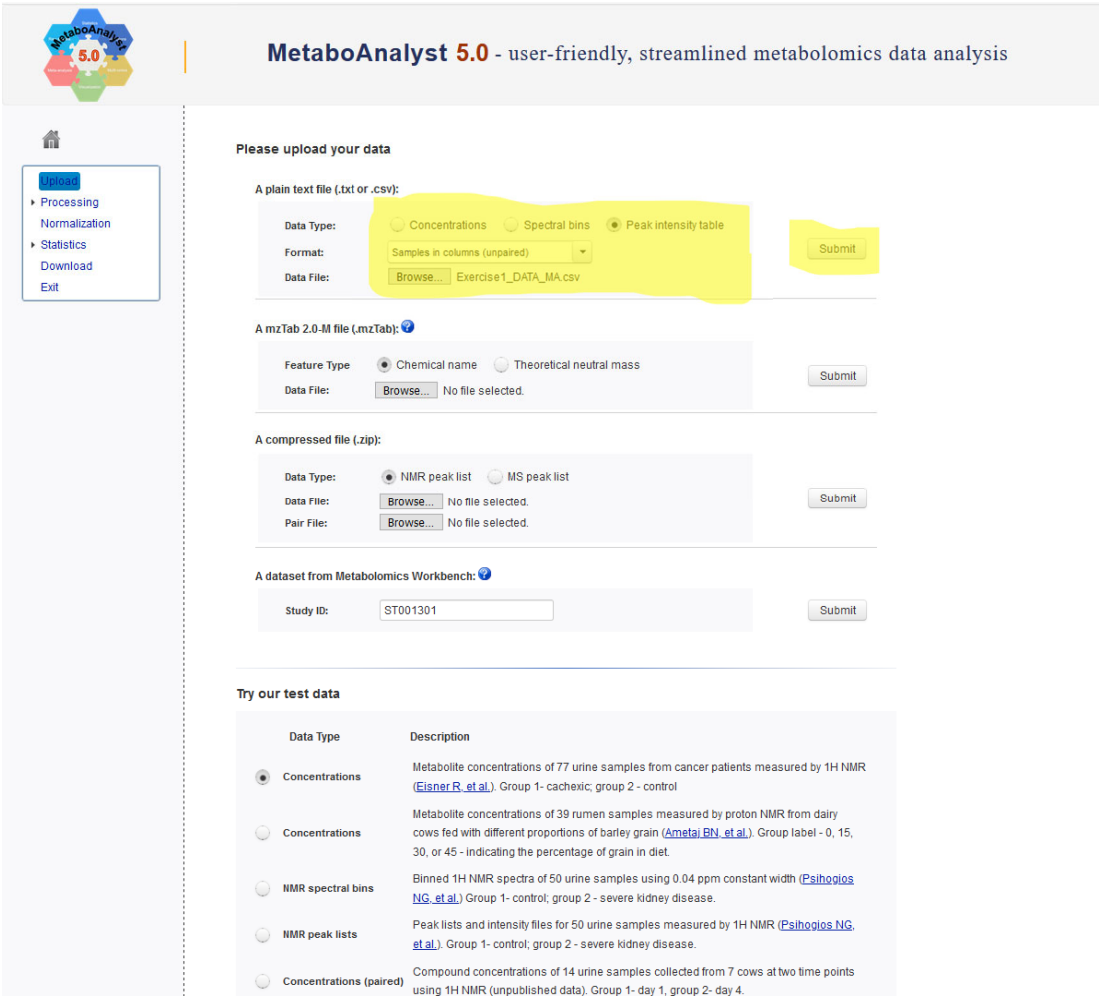


Figure 2: Data upload page under 'Statistical analysis' in MetaboAnalyst

6. If you get an error at this point it is likely that the file extension is not correct (should be .csv) or that there are duplicate entries in the list of compound features/descriptions, or that the headers are not correctly formatted. This should not happen with the dataset provided (unless it has been modified) but is something to watch out for when submitting a new data file for the first time.

7. Upon successful upload **a 'Data Integrity Check'** screen will appear**.** This confirms criteria have been met that enable the data to be analysed effectively. These include data formatting and replacement of zero values with a positive minimum value. Here alternative zero value imputation approaches can be accessed and explored. In this case you don't need to select anything and you can press **'Submit' followed by 'Proceed'.**

8. **The 'Data Filtering' page:** This is used to identify and remove variables that hinder the identification of biologically important signals. For example:
   1) Very small values (close to baseline or detection limit).
   2) Variables that are constant across conditions.
   3) Variables that show low repeatability using %RSD for example.
   4) The dataset is reduced is size for ease of processing and analysis.

   Accept the default (IQR), select 'Submit' followed by 'Proceed' (Figure 3)

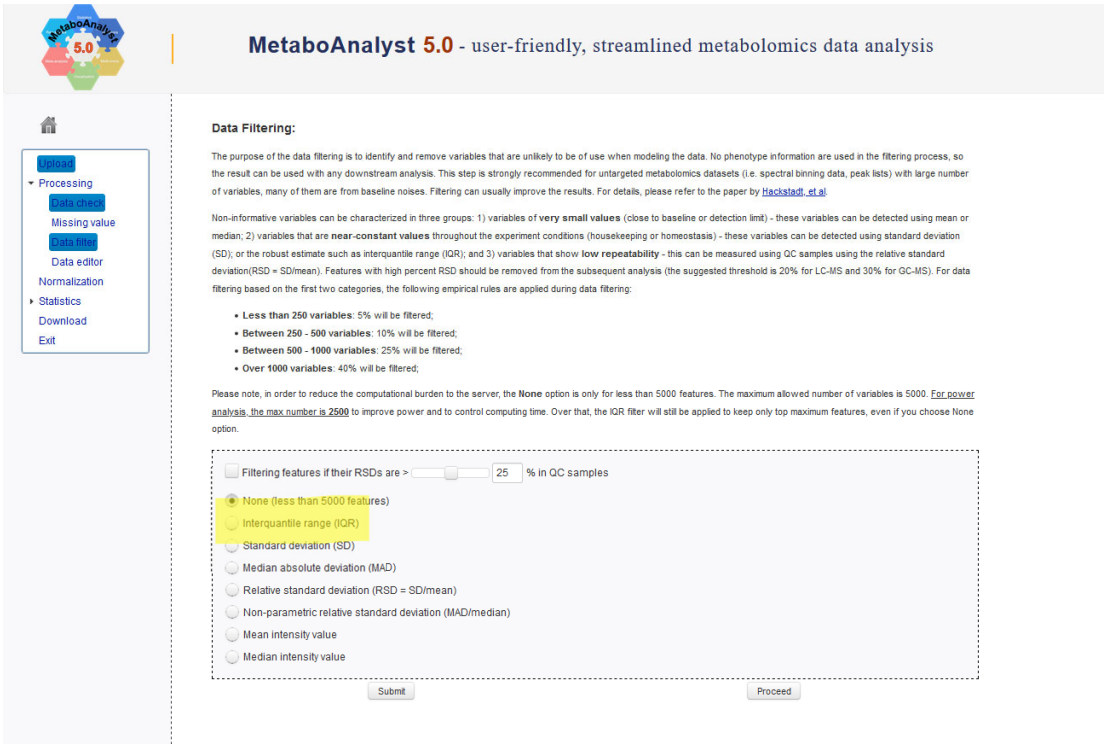


Figure 3: Data Filtering screen

9. **Normalisation Overview page (Figure 4)**: This page enables three data operations**:**
   1) *Normalisation* (ensures sample abundances are comparable across experimental groups).
   2) *Transformation* (ensures the data is not heteroscedastic).
   3) *Scaling* (ensures the data points are on the same scale).

   These are essential operations that ensure effective statistical analysis using subsequent univariate and multivariate statistical tools. The most appropriate combination is dataset dependent and can be compared using the box and density plot output (under 'View Result, see Figure 5).

For this exercise select '**Normalisation by sum'** for normalisation, '**log transformation'** for data transformation and '**Pareto scaling'** under Data scaling as shown in Figure 4.

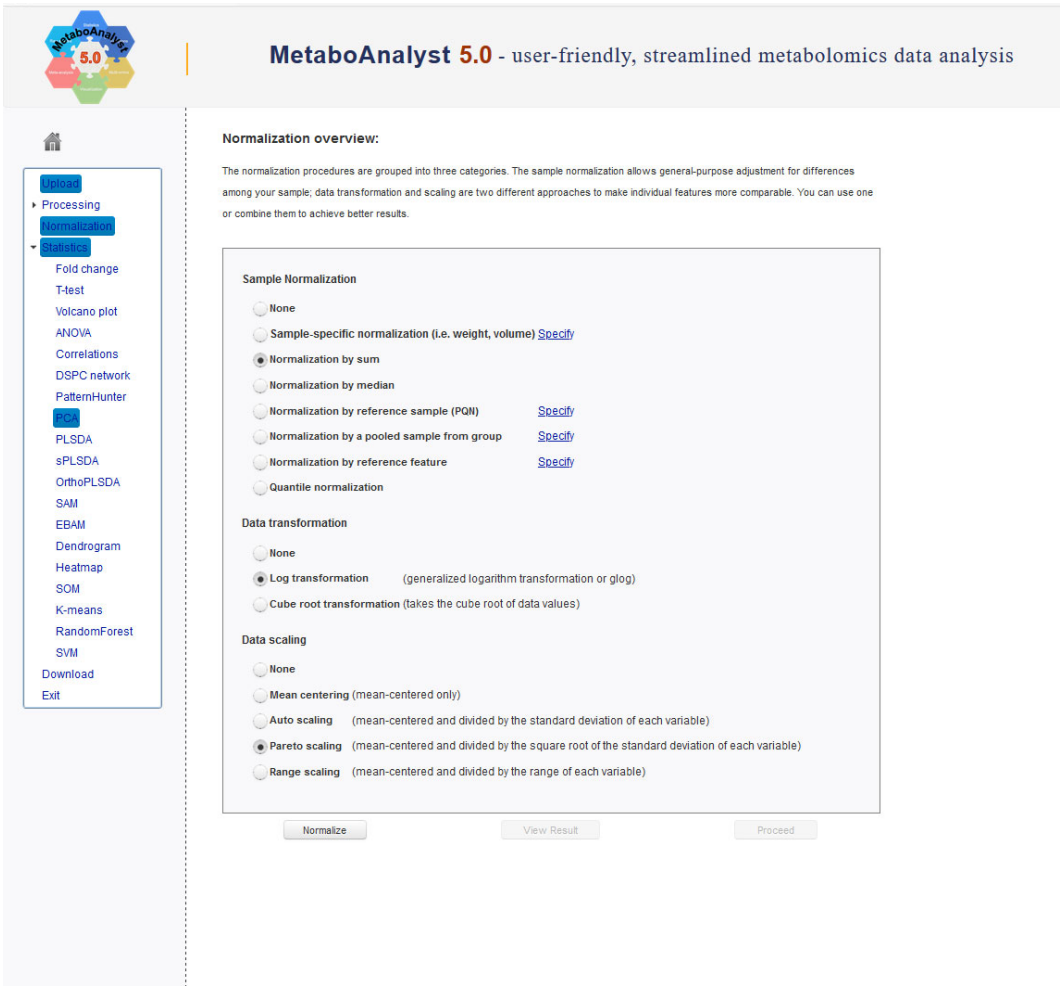


Figure 4: Data normalisation, data transformation and data scaling page.

10. Select '**Normalise'** as the bottom of the screen.
11. Then select '**View result'** to see the effect of the data scaling via box plots and density plots shown in Figure 5. The aim for transformation and scaling is to achieve a centralised Gaussian plot and for the box plots to show that the data is on a similar same scale after as shown in Figure 5. This can be verified visually and the main purpose is to ensure the requirements of a roughly normalised distribution of data are met for subsequent statistical analysis (for example using a student t-test where its assumed by the statistical process).

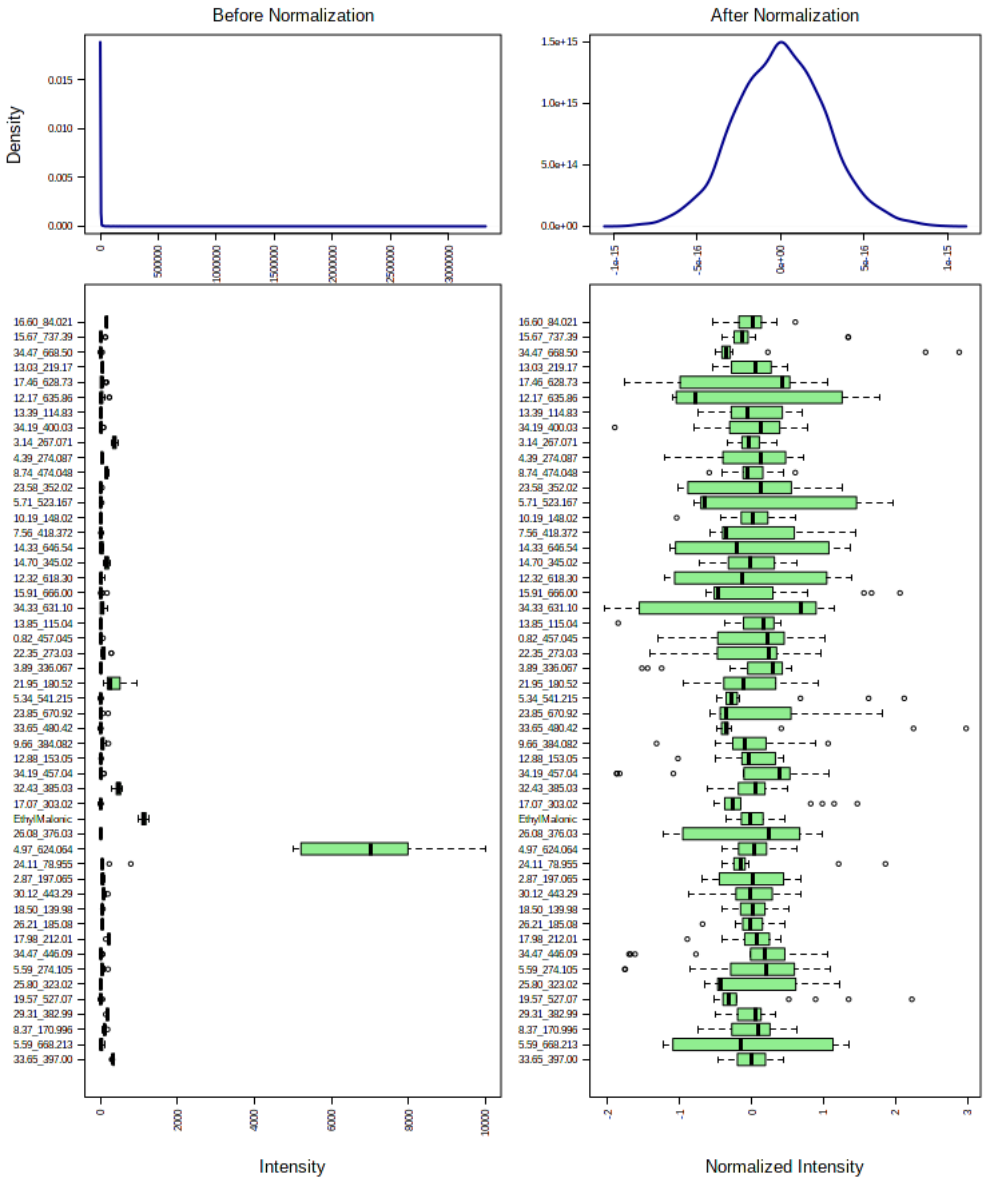


Figure 5: Look for a Gaussian density plot after 'Normalisation' (somewhat confusingly they mean after *transformation and/ scaling'* and not just 'normalisation by sum)

[Note: you can use (non-hierarchically clustered) heat maps (found in the statistical analysis section, see Figure 6 below) to assess the effect of different data-driven normalisation approaches (by sum above). This is usually a good practice as it gives you a visual overview of the entire dataset and allows you to inspect and verify the effect of the data processing parameters (e.g. compare their effects). **Figure 6** shows three examples with Normalisation by Sum the most effective in this case. Comparing the effects of normalisation visually via heat maps is usually sufficient to assess which is most appropriate. It can be seen from the PCA plots in Figure 6 that data normalisation can have a significant effect on the way data clusters using unsupervised multivariate approaches **(note these examples are from a different dataset for illustration only and you don't need to perform these three additional normalisations for comparison for this exercise). You should move on to section 11 when ready to continue).**

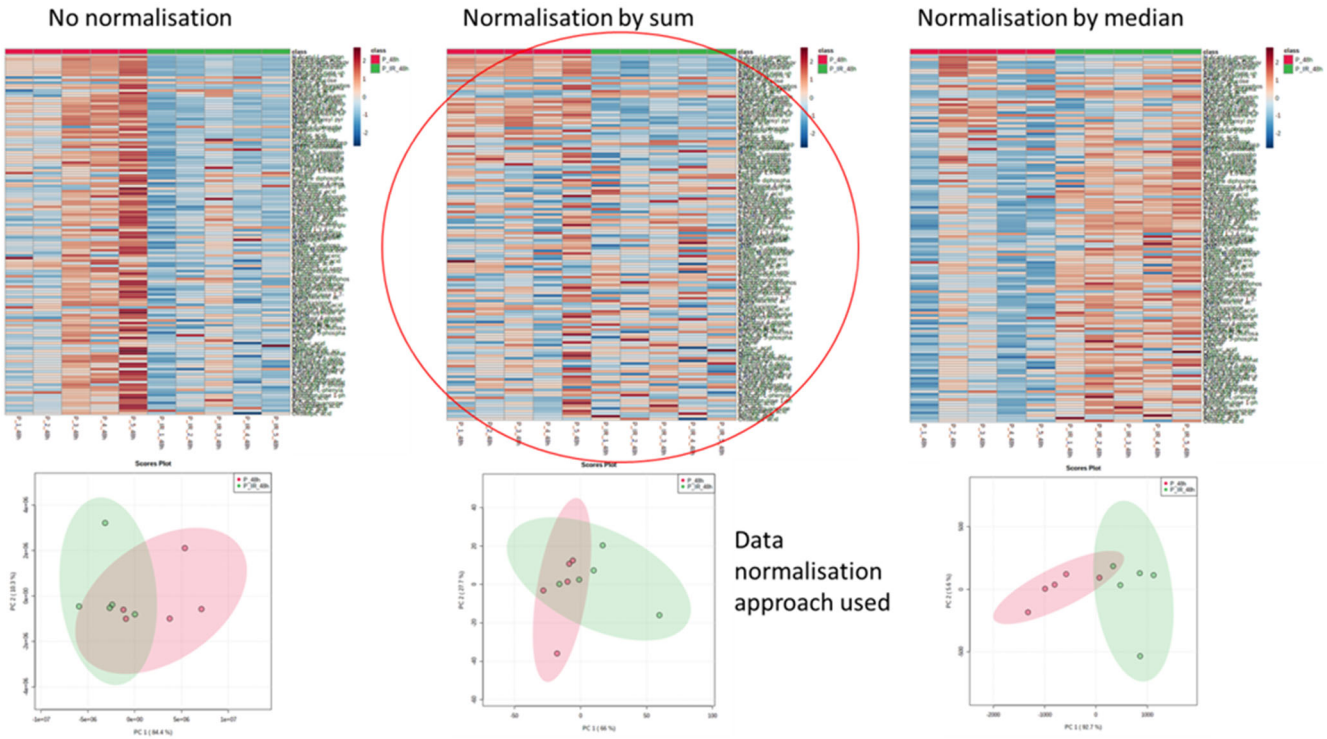


Figure 6: Visually comparing normalisation approaches using non-hierarchically clustered heat maps and PCA plots shows the effect data normalisation is having on your dataset.

12. **Select 'Proceed'.** The **'Select an analysis path to explore'** page opens and various types of data analysis tools including univariate, multivariate, clustering and classification analyses are available. We will start with univariate statistical analysis of the dataset and **'Fold-change Analysis.**

13. **Fold-change Analysis:** Select this and now the left-hand pane lists all the statistical analysis tools available starting with 'Fold Change Analysis' highlighted in blue. The main page shows a graph with the default fold-change threshold of 2. Each circle is a compound feature; those in grey have a fold-change <2 and those in red/blue >2 (both increases and decreases in abundance between the two experimental groups, Figure 7). Note you can alter the fold-change threshold and doing so followed by submit will update the graph you can repeat this as many times as you like but whatever the final choice made will be the result saved in a list and incorporated automatically in the MetaboAnalyst report at the end of the data analysis process (this exercise). You can also click on a circle and a box plot will be provide for that particular compound-feature. Select a fold change of threshold of 2 and submit. **Question: How many compound-features with a fold-change >2 are elevated in mutant samples and how many have a fold-change<2?**

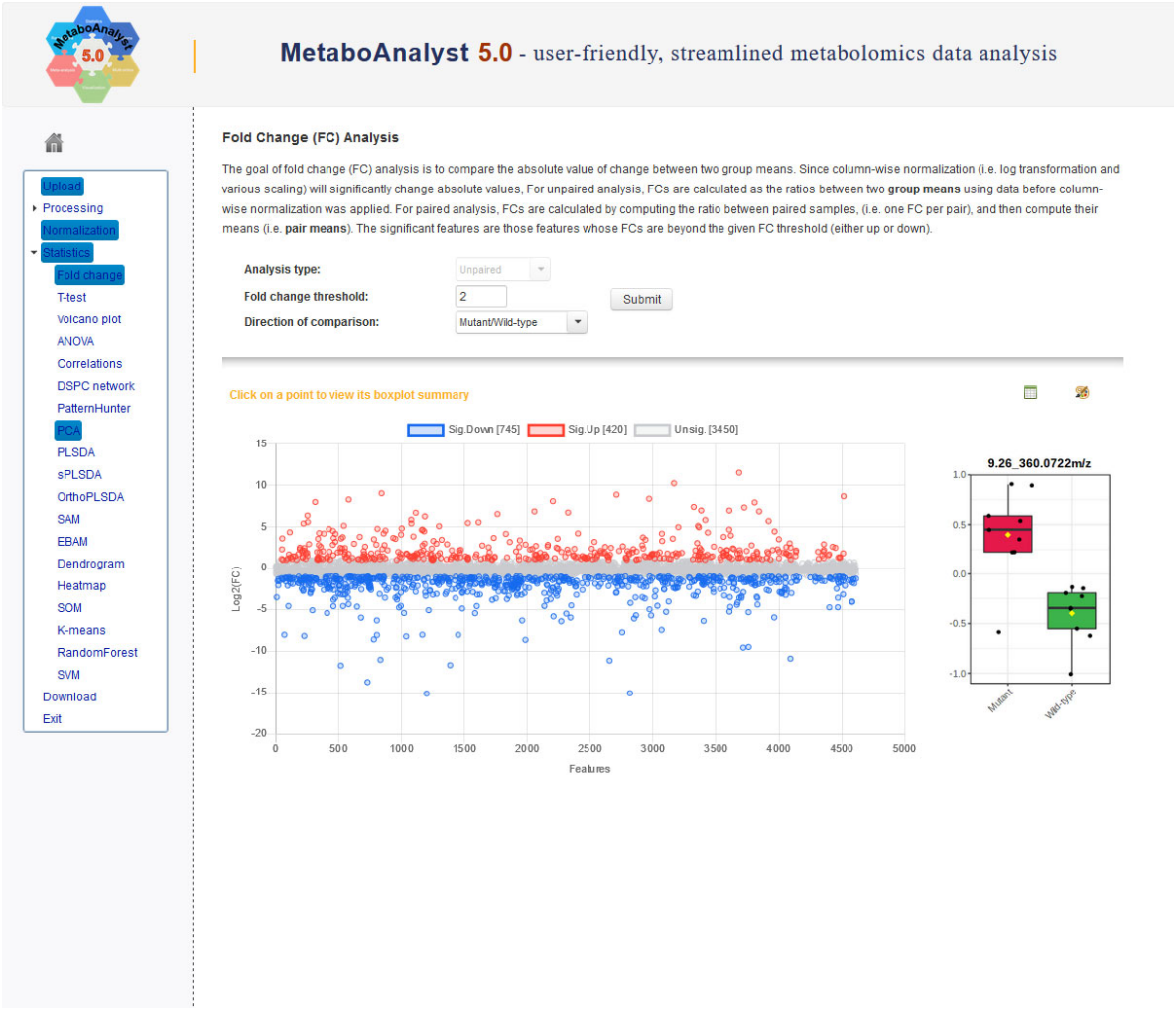


Figure 7: Fold-change analysis: blue and red circles show a fold-change >2.

14. Next, we will investigate the **statistical significance** of compound-features by selecting '**T-test'** and determining p-values. A new graph will appear with additional options at the top (see Figure 8).

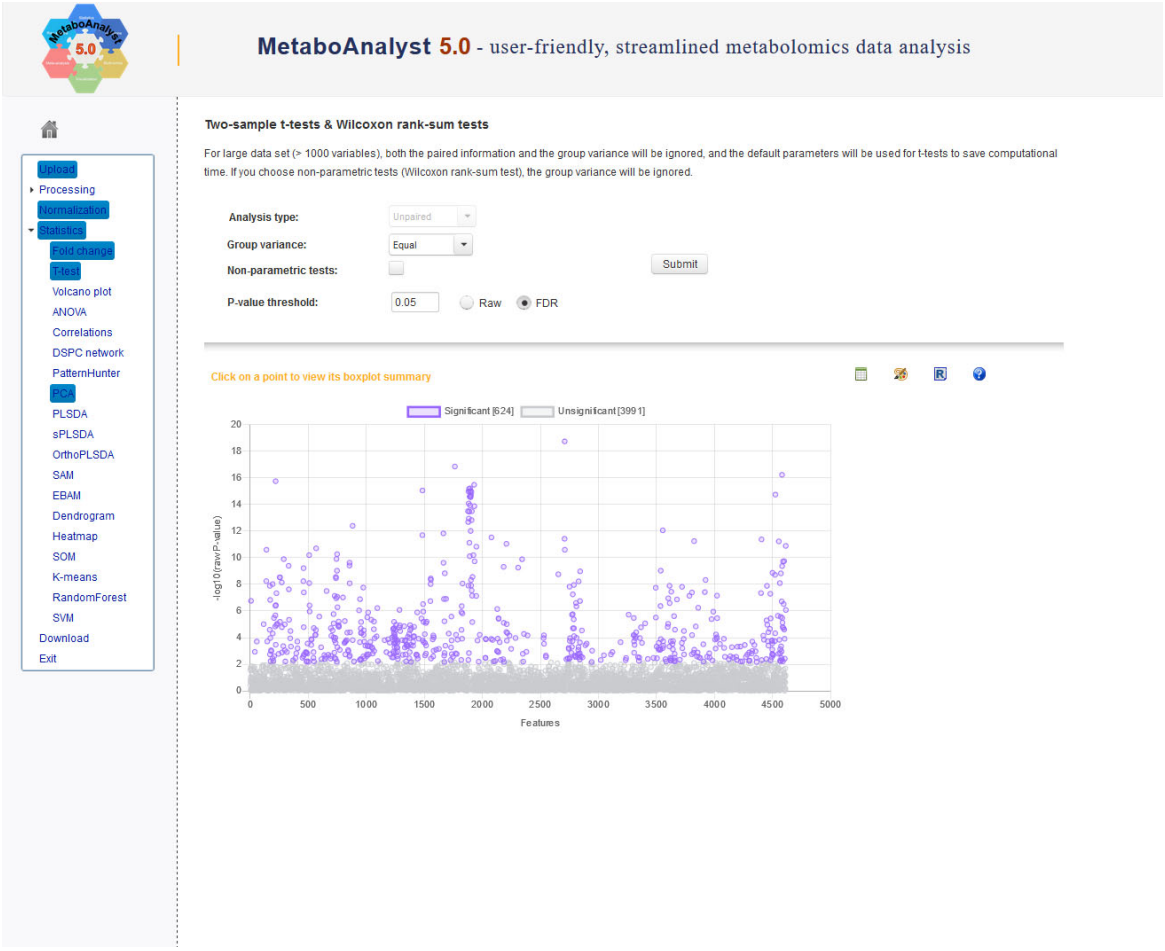


Figure 8: T-tests (and multiple testing by false-discovery rate (FDR))

The FDR-adjusted p-value threshold is set to 0.05 by default. It is also possible to select parametric on non-parametric t-tests here depending on your dataset. For the purposes of this exercise keep the default values. You will see there are a lot of significant compound-features. See Box 1 for further notes on univariate and multivariate analysis of the data. The output report at the end will pick up the processed data from whichever tabs have been selected in the list on the LHS.

Note for univariate statistical analysis using data compiled from multiple methods: (T-test, volcano plot etc) the ranking of metabolites is based solely on p-value and fold-change. I.e. where compound-features are treated as independent variables. Hence it does not matter what kind of scaling you use for this output or how many different LC-MS methods you have combined. This is worth noting for some studies where a range of LC-MS methods have been used to gain greater metabolite coverage. It is important to correct the p-values for multiple testing by using the FDR-adjusted p-value and not simply the p-value. This is because the t-test will not be testing a single hypothesis when applied to the entire metabolomics dataset but rather many hypotheses, as many as there are compound-features.

Note for multivariate statistical analysis using data compiled from multiple methods: For the ranking produced from this type of analysis (VIP scores for example) the scaling makes a significant difference to the ranking of the metabolites. This is because multivariate methods link patterns between variables as well as uses p-value, fold change etc. This means that differences in absolute abundance between metabolites can be a factor in their ranking using VIP scores unless the data is transformed and scaled appropriately). Different Transformation and Scaling approaches can help to highlight different characteristics of the data and are therefore useful for different applications. This can be explored in more advanced analysis settings but it is important to remember that the scaling and transformation largely affects the multivariate statistical analysis and not the univariate analysis.

Box 1: Notes on univariate and multivariate analysis of the data

15. **Next select 'Volcano plot'** and adjust 'Fold-change' to 2.0 and 'p-value threshold' to 0.05 and select FDR-adjusted (see Figure 9). Then **click on 'Submit'.** Note that the most important compounds-features in a volcano plot of FC and p-value will be at the top right and top left of the graph. In the example there is compound-feature at the top right-hand side (circled in red on Figure 9). Left click on this data-point to bring up a box plot for that particular compound-feature (this can be done for any feature visible). This is all you need to do to add the data in this section to your report for this exercise, however, if you want to output figures for presentations or publications take note of the following**:**

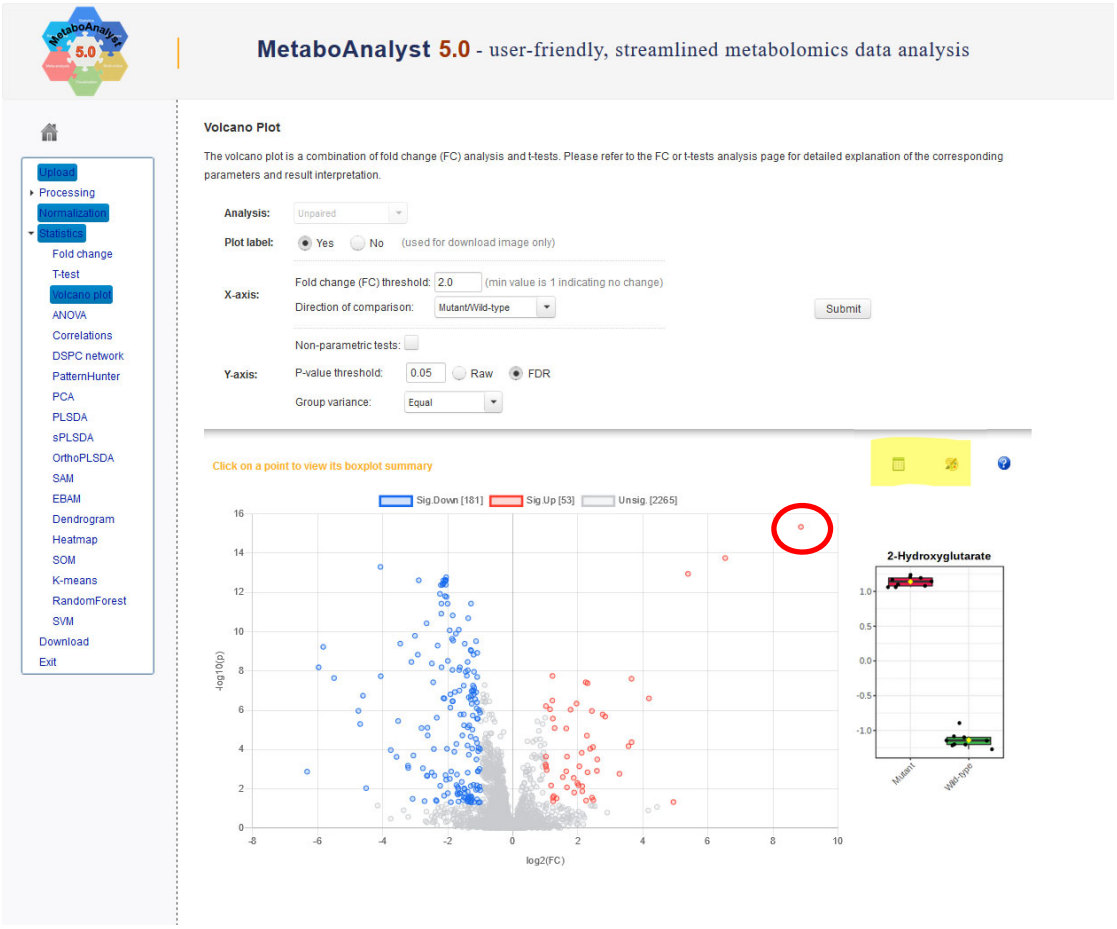


Figure 9: A volcano plot of p-value and fold-change values enables selection of only those compound features that have statistically significance differences in abundance above a selected fold-change value. These can be determined based on parametric or non-parametric tests as well as equal or unequal group variance.

By clicking on the artists palette at the top right of the plot this will take you to details for downloading high quality images for presentations and publications (this can be done for most of the statistical outputs which is useful for presentation and creating figures). For a quick screen grab you can right click directly on the image and copy/paste into an appropriate software such as PowerPoint for example.

16. Click on the Excel icon next to the pallet (top right of volcano plot) and note the highest ranked metabolite. **What is the name of the two highest ranked identified metabolites? Select 'view' next to each to see a box plot for that metabolite. Which feature is elevated in 'Mutant' cells and which is depleted**?

17. Ignore 'ANOVA' and 'Correlations' for the purposes of this exercise and next select **'PatternHunter'**. Type '2-Hydroxyglutarate' into 'feature of interest' box (see Figure 10). Keep all the other defaults and then **click on 'Submit'.** A list of compound-

features that are positively and negatively corrected with 2-Hydroxyglutarate will be displayed. You can do this of any compounds-feature you like (need to make sure the name is exactly as it appears in the data table submitted).

18. **What are the accurate mass values of the two compound-features that are positively correlated with 2-hydroxyglutarate**?

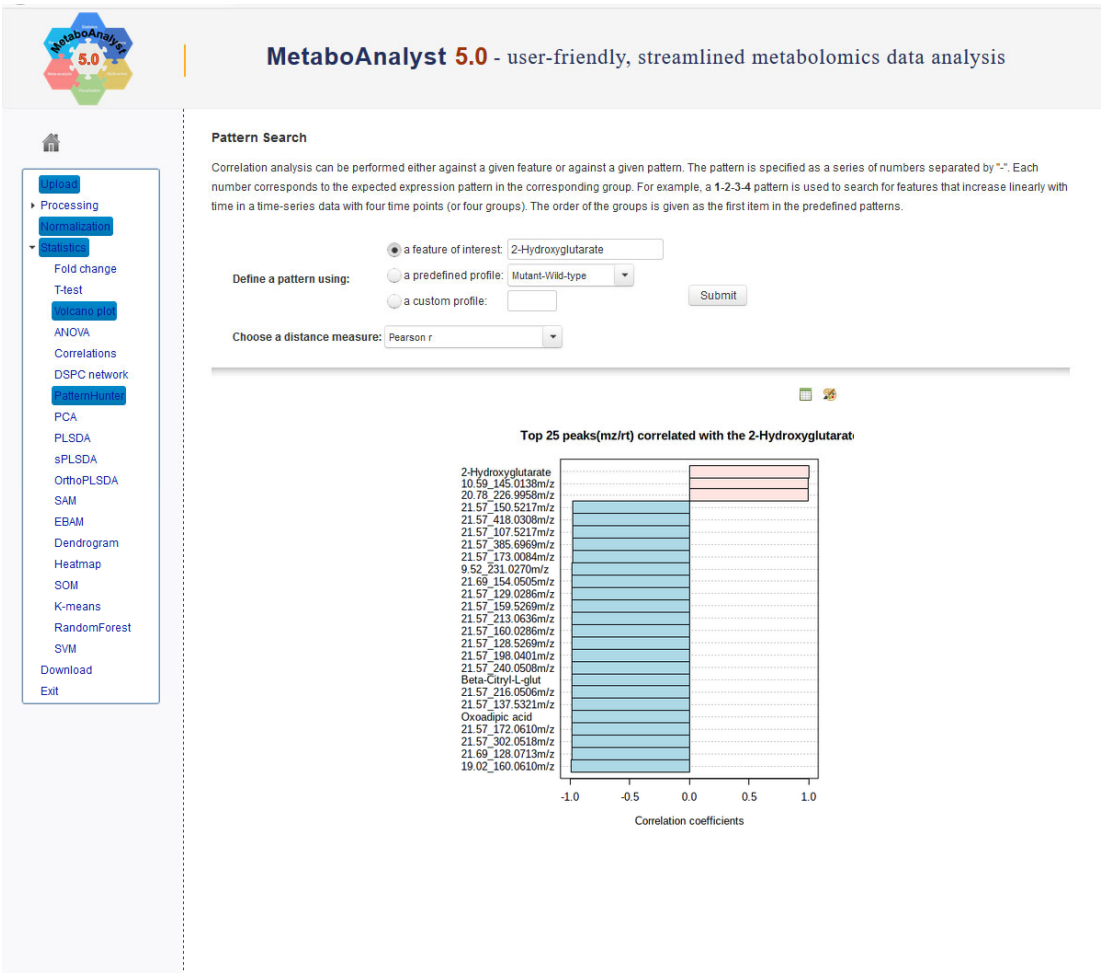


Figure 10: PatternHunter uses correlation analysis (Pearson r coefficient) to identify compounds-features that are highly correlated and anti-correlated with a selected feature.

19. Explore possible identities for these two compound-features: Open a new webpage and navigate to **https://hmdb.ca/.** The Human Metabolome Database is a very useful repository of all compounds found in the human metabolome (includes endogenous and exogenous metabolites.

a. Click on 'Search' and then 'LC-MS Search in the dropdown menu **(see figure 11)**

b. Complete the information required: put the accurate mass (e.g. to all 4 decimal places) in the 'Query masses' box on the left-hand side and then update the information on the right-hand side appropriate to the experiment (see figure 12). Select 'Search'

c. **What are the suggested putative identifications for both compound-features? (how many for each).**

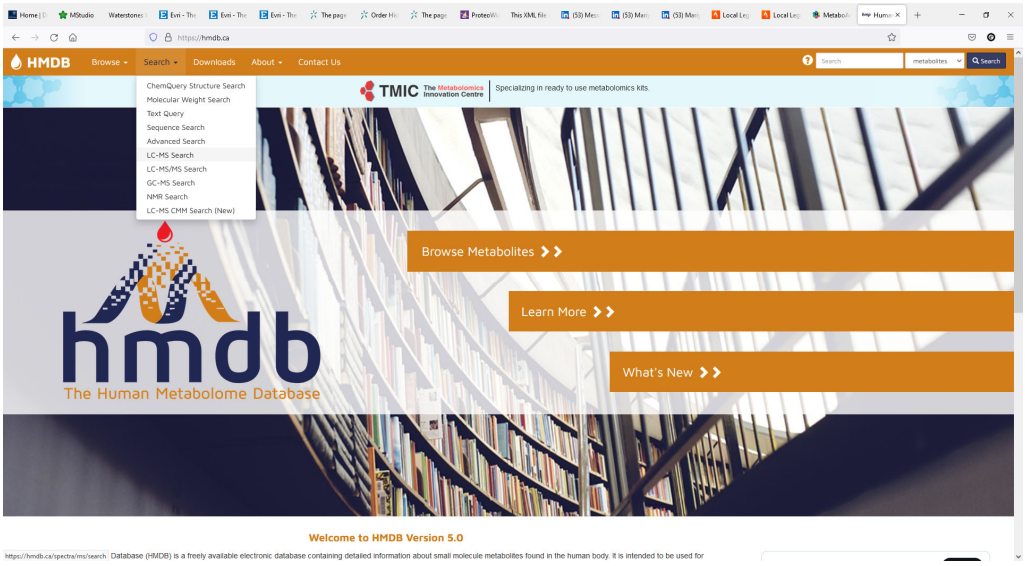


Figure 11: The Human Metabolome Database (https://hmdb.ca/) is a very useful online repository of all compounds linked to the human metabolome and can be used to search for putative identification based on accurate mass data.

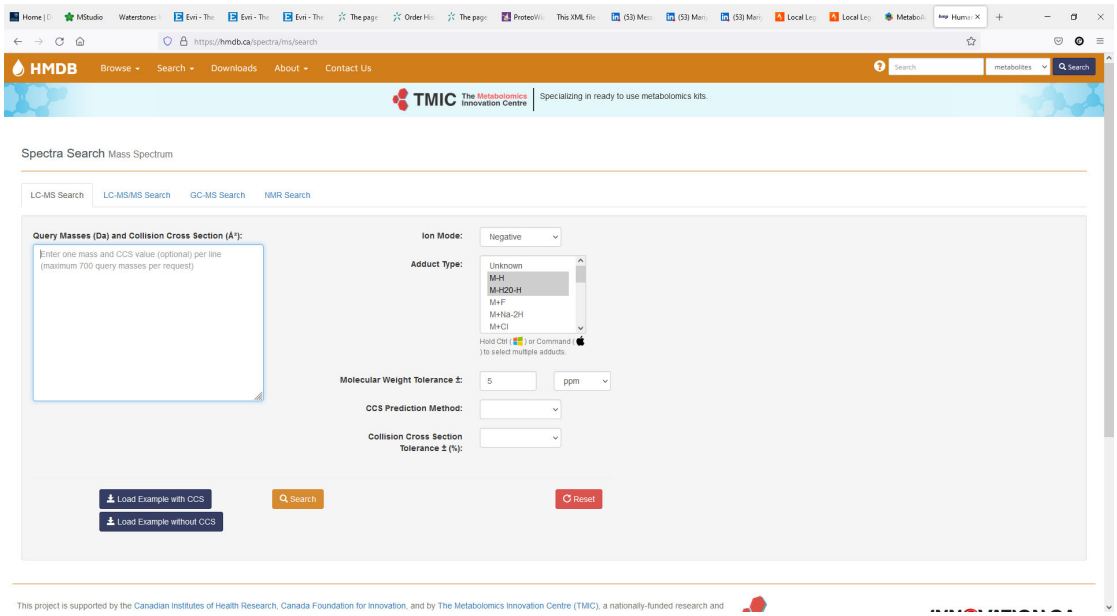


Figure 12: complete the information requested and add the accurate mass query in the box on the left-hand side.

20. Next select 'PCA'. The overview slide provides small graphs of up to the first 8 principle components. Select the '2D scores plot' output (along the top) (see Figure 13). Note there is a radio button to add labelling of variables. Select it and click on 'update'. This should provide you with an update plot where the sample names are visible (this can be useful to identify outliers).

21. Loadings plot: select the 'Loadings Plot' (next tab along). the Excel icon, next to the pallet for downloading images, enables the list of features to be downloaded as well as box plots.

PCA plots are often used to identify outliers and any problems associated experimental design rather than provide direct insights into biological context itself although, as in this case, when significant differences in metabolome composition are present this is often reflected in the PCA plot. This however, is not always the case as PCA is an unsupervised multivariate statistical tool which simply provides (a very useful) projection of the full dataset. You will often see it presented in the scientific literature as a way to show that two group can be differentiated. There is nothing wrong with this but need to be clearly demonstrated that this is due to biologically-derived differences and that it should be kept in mind that biologically relevant changes may not be picked up in a PCA plot if less important differences predominant (e.g. sampling bias or a large amount of biological variability or heterogeneity). PLS-DA and OPLS-DA are on the other hand similar vector-based approaches but supervised and hence are able to reveal biomarkers using multivariate statistics often in a more powerful and effective way than PCA.

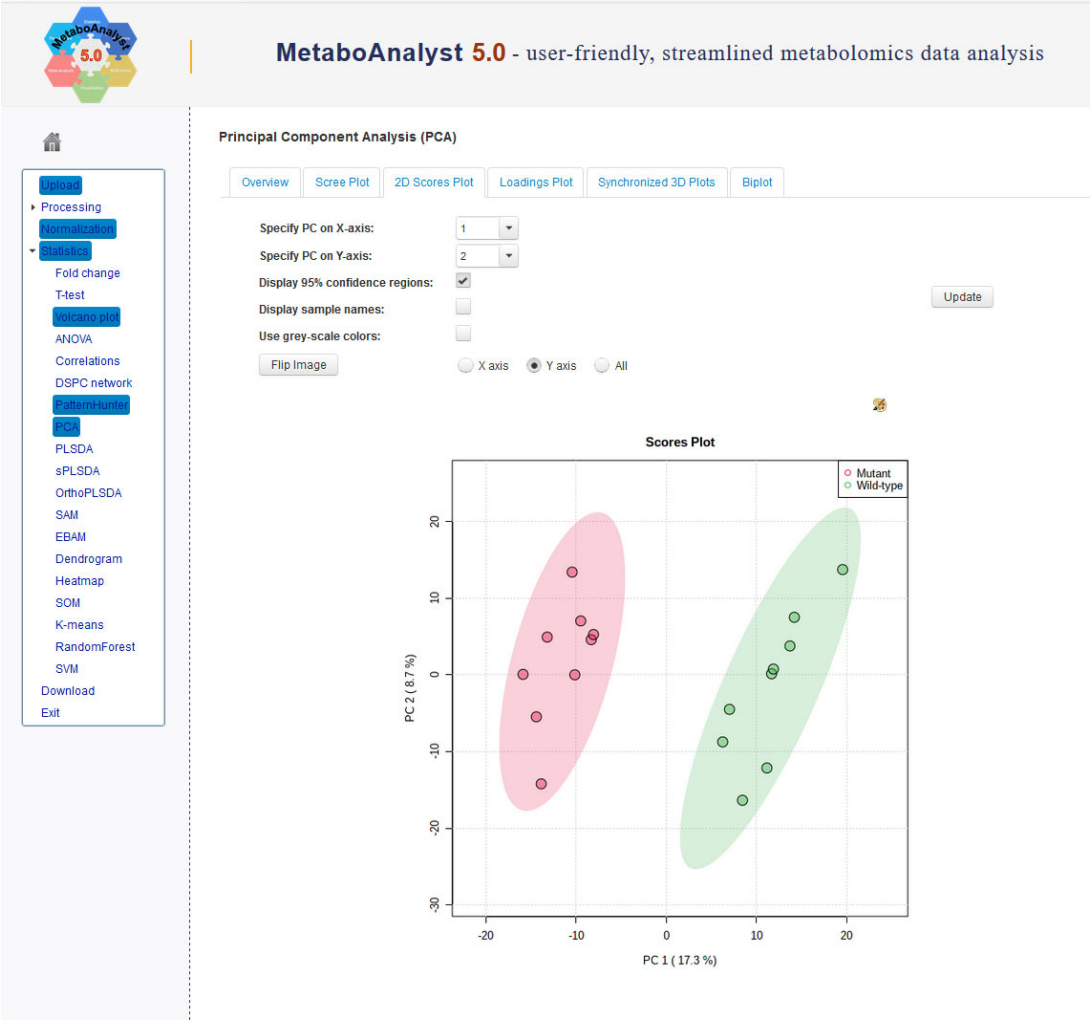


Figure 13: PCA Scores plot showing unsupervised separation of experimental groups.

Questions: What are the names of the two samples which represent the greatest outliers in each experimental group according to the PCA plot? Are they within or outside the 95% confidence boundary?

22. Next **Select 'PLS-DA'** from the left-hand panel. The PLS-DA looks very similar to the PCA scores plot but is a *supervised* multivariate statistical modelling tool which enables ranking of compound-features using Variable Importance in the Projection (VIP) scores. These help to select compound-features which make the biggest difference between the selected experimental groups. This is useful for biomarker discovery but care must be taken as the modelling can tend to over-fit the data and it is important to validate the model before accepting the results.

    Select '2D Scores Plot' and note the increased separation achieved compared to the PCA. Next select **the 'Imp. Features'** tab. In the latter the top 15 VIP scores are

displayed by default (**Figure 14**). These can be downloaded to Excel or outputted to high-resolution images as required.

Question: Suggest the identity and or chemical formula of the two most important metabolites in the dataset for differentiating the two experimental groups?

Select 'Cross Validation' to assess the Accuracy, R2 and Q2 values which provide information about how good the model is (click on Excel download icon to see values). As a rule of thumb the model should have a high accuracy with R2 and Q2 values above 0.5, the Q2 should normally be less than the R2 values but not by more than 0.2 units (**Figure 15**). Permutation testing (final tab) provides an alternative (significance-based) approach to PLS-DA model validation. Note that in this dataset the Q value is just below the 0.2 difference from R2 and that the permutation testing does not return a pvalue <0.05. Its often hard to achieve well validated models when the number of samples versus the number of experimental variables is low such as is often the case for tissue culture experiments as here. These modelling approaches can nevertheless be useful indicators of what the important features are in the dataset. *sPLSDA* and *OrthoPLS-DA* (tabs on the RHS below PLS-DA) are alternative supervised, vector-based multivariate modelling tools. We will not explore these as part of this exercise, but both provide a variation on the PLS-DA modelling approach.

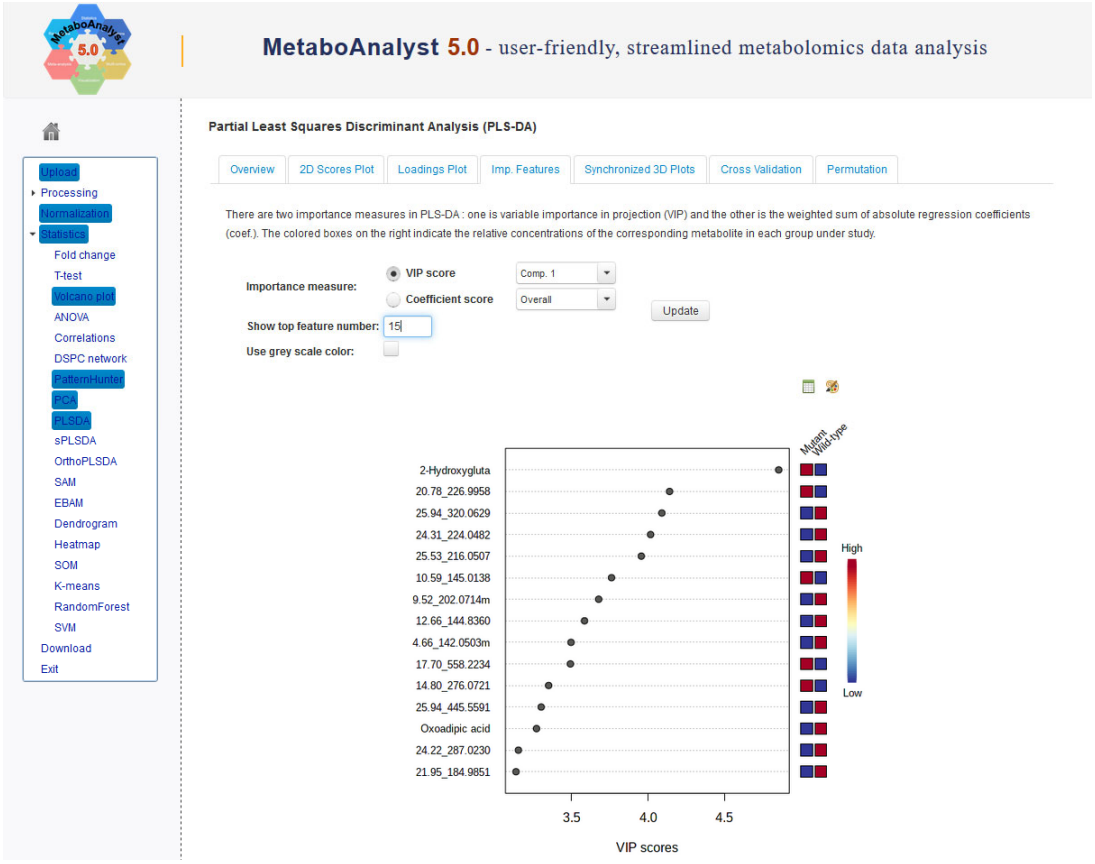


Figure 14: PLS-DA plot showing the top 15 VIP scores

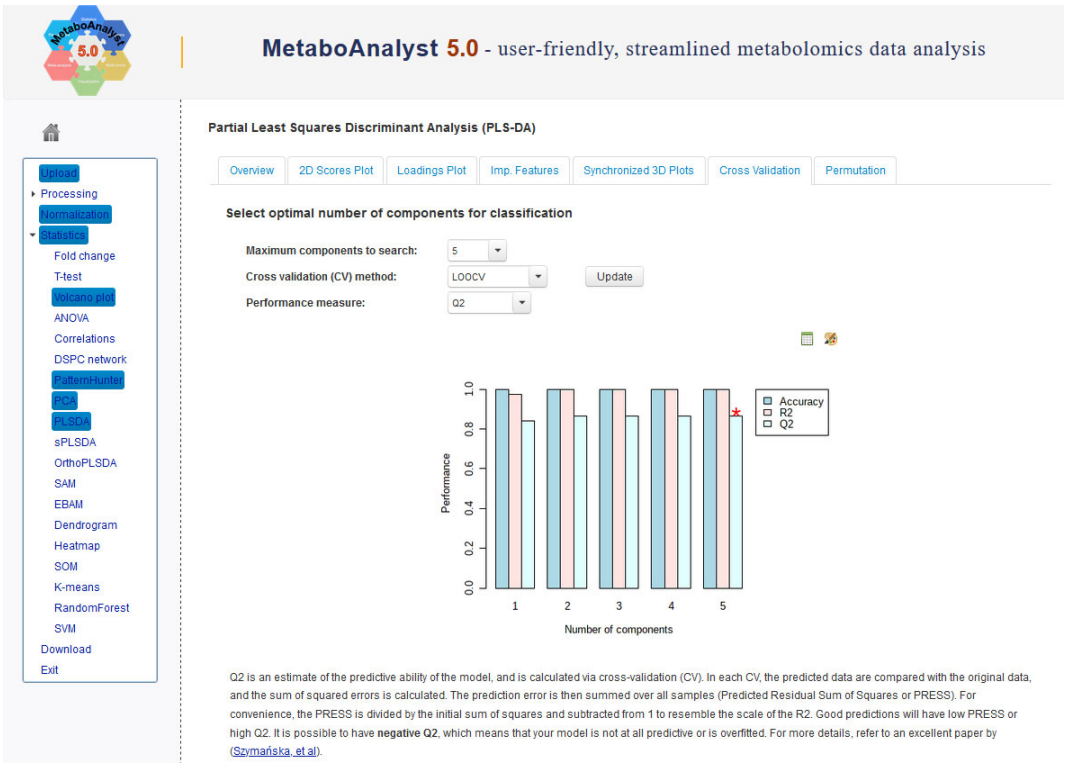


Figure 15: PLS-DA cross validation analysis

23. Dendrogram: Hierarchical-clustering is an unsupervised multivariate classification approach which uses a 'distance measure' (here we can choose from Euclidean, Spearman or Pearson) along with a clustering algorithm to link samples together in terms of how close they are to each other in metabolite composition. **Figure 16** shows that the primary division between all the samples is whether they contain the mutation or not (e.g. they are primarily divided into the two experimental groups Mutant and Wild type).

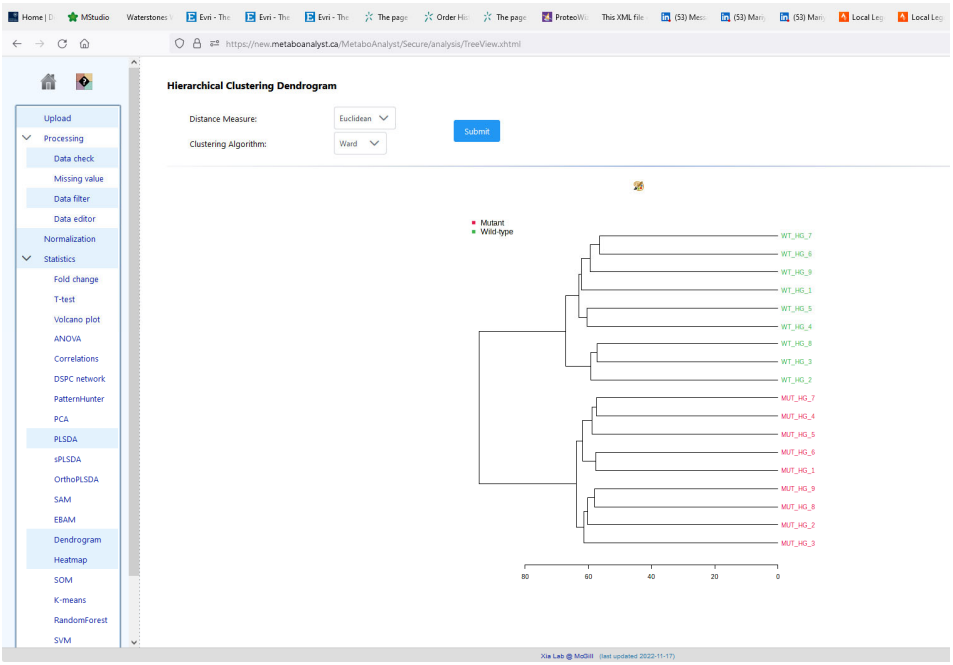


Figure 16: Hierarchical Clustering Dendrogram

24. Select 'Heatmap' next. Hierarchical clustering can also be applied to the metabolites as well and visualised in a heat-map output. This enables both samples and compound-features to be arranged according to how close they are in their composition **(Figure 17**). It is useful for:

1) identifying whether samples naturally cluster into their experimental groups.

2) Identifying individual metabolites that separate the experimental groups most strongly and visualising these (an alternative approach to PLS-DA).

3) In a modified output it can be useful to assess the effects of normalisation methods (see point 18 below). The default is to provide clustering of both samples and variables (metabolites) but either can be addressed separately or both can be

removed. Selecting the top 25 compounds features is useful for identifying potentially the most well correlated compound-features (**Figure 18**).

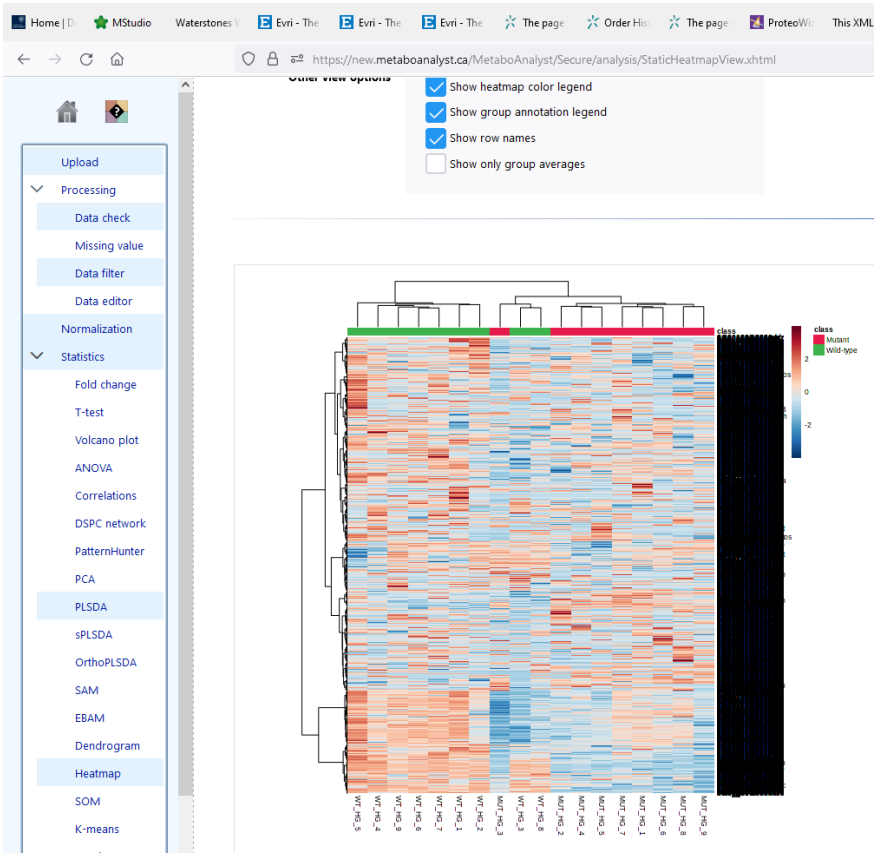


Figure 17: Hierarchical clustering of all samples with heatmap output

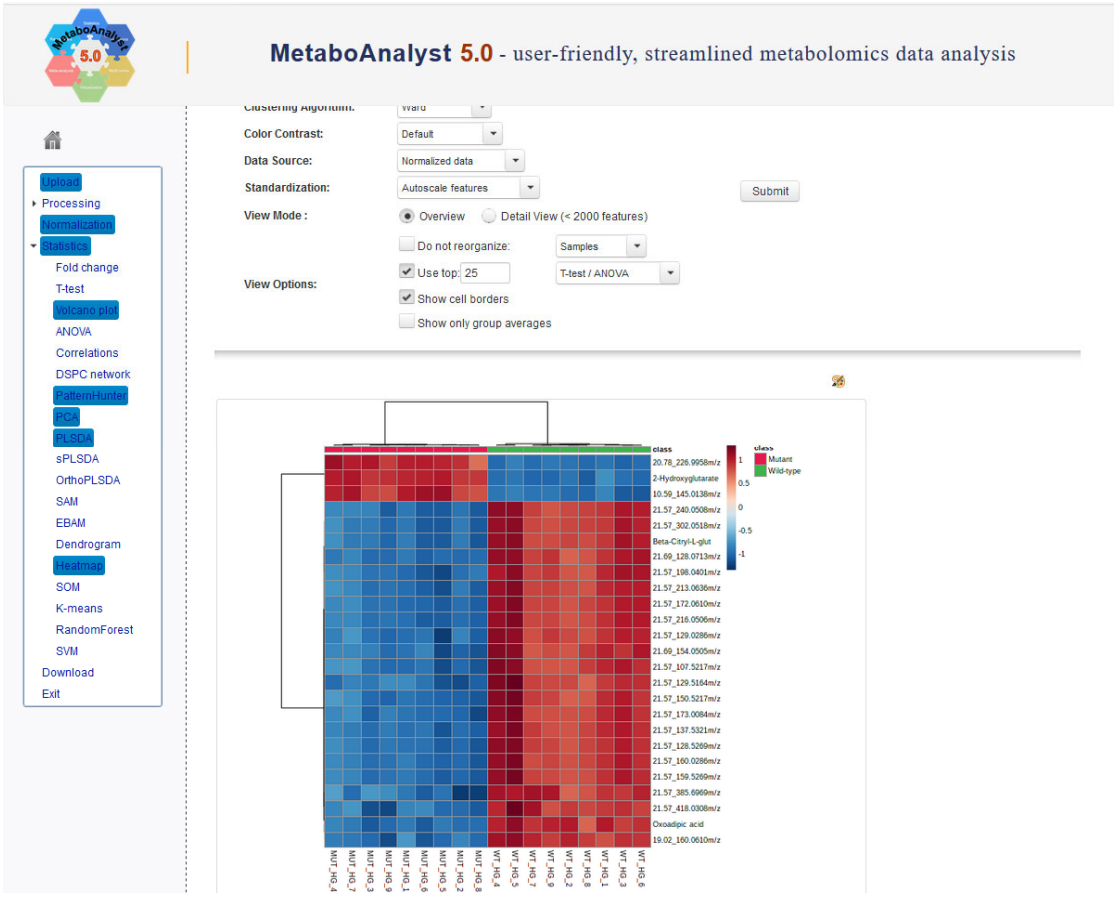


Figure 18: Top 25 compound features identified by Heat-map dendrogram visualisation

Tick the 'Top 25' box and then submit to show the heatmap output for the top 25 compound-features only.

Question: Which sample does not quite cluster within its experimental group when clustering the full dataset?

Question: What are the names of the top three *identified metabolites* in the hierarchically-clustered data?

Note on heatmaps - assessing the effect of normalisation: A modified output of the same Heatmap approach can be useful for assessing the effects of normalisation. Example heatmaps with and without normalisation are compared in Figure 19 below from which it can be seen that in the example why data normalisation can be an important data processing.

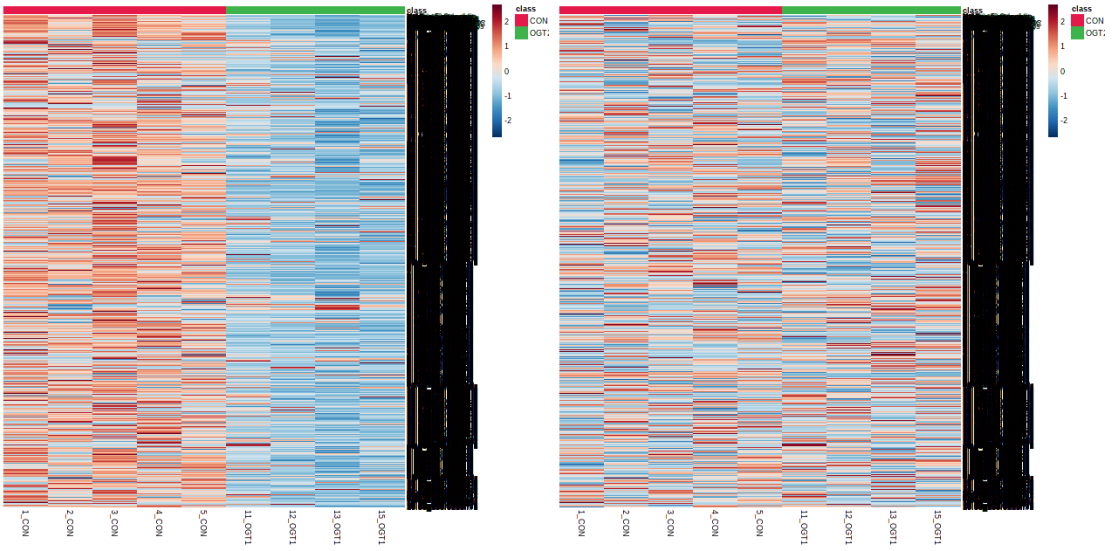


Figure 19: Assessing the effect of normalisation: The two non-hierarchically clustered heat-maps above use the same dataset. On the right hand side **no data normalisation** has been applied and on the left hand side **'quantile normalisation'** is shown. The bias inherent in the un-normalised dataset is largely removed by quantile normalisation.

This concludes the main statistical approaches we most often use in the statistical analysis toolkit. There are other tabs we have not looked at and these can be explored in more detail yourselves. We only used these when we have a specific application they may unusually require it. The tools we have looked at today are what we used 90% of the time in our untargeted metabolomics data analysis workflow.

25. To conclude the data analysis **select 'Download'** at the bottom of the list on the LHS. This provides a list of the figures and tables of data created so far. **Select the 'Generate results' icon** above this list (**Figure 20**) and this will then create a single PDF file of the results. When the 'Analysis Report' has been created, select it to view the completed report of your data analysis and save this to your computer.

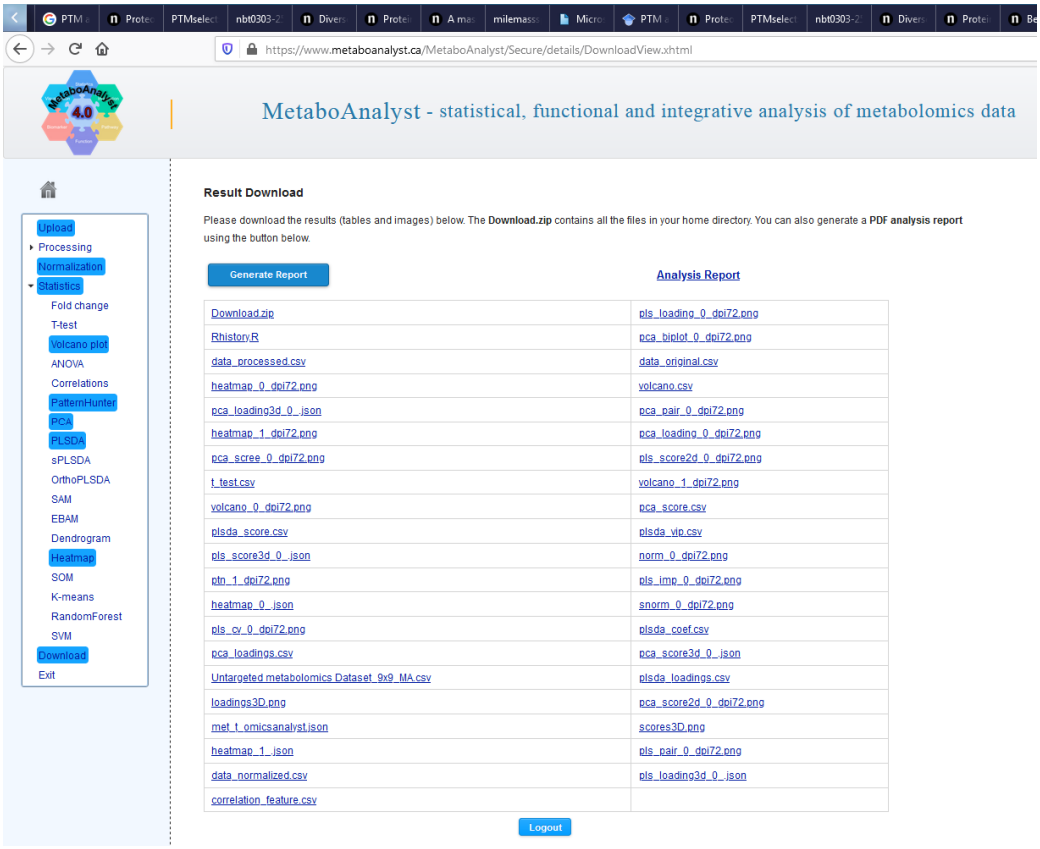


Figure 20: Select 'Generate Report' and then 'Analysis Report' once it has been created.

The PDF report contains the data analysis and commentary and explanations (see example in Figure 21).

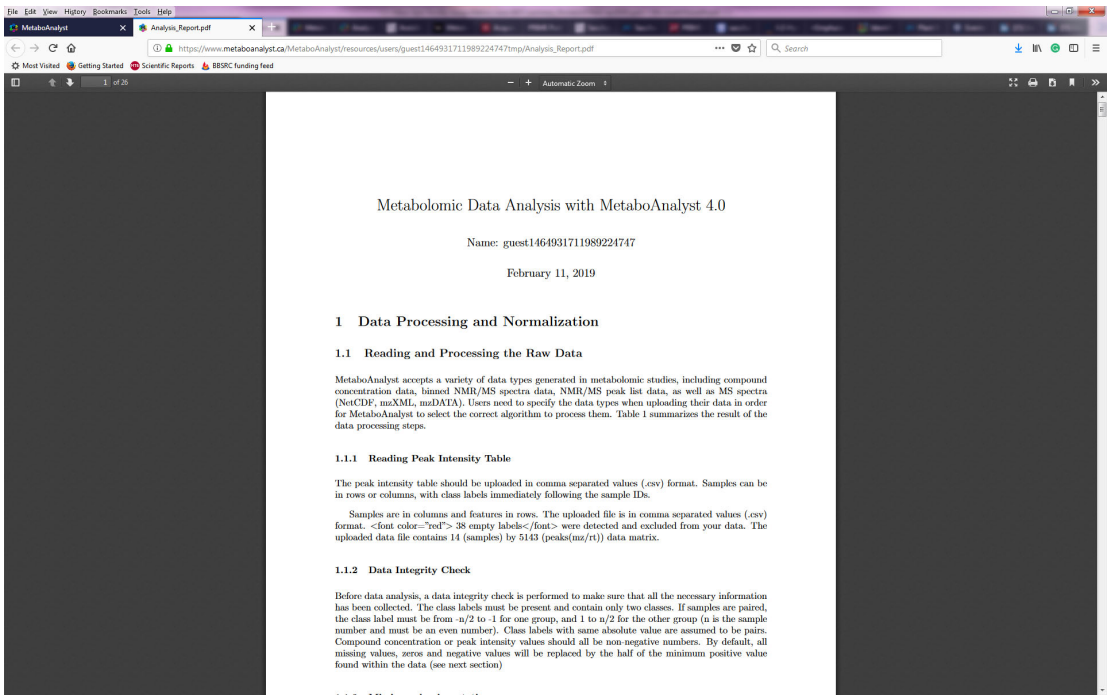


Figure 21: PDF report in MetaboAnalyst

Open the report. This now should provide a useful summary of the data processing and statistical analysis of your data. **Save the PDF report to your computer for future reference**. It may also be worth copying figures at each stage into PowerPoint for presentational purposes at a later date. There are options for each statistic to output high-resolution images and to download the data from which the figure is created, usually as an Excel file.

Note not all the statistical tools available in MetaboAnalyst have been explored here but a range of informative approaches for identifying a list of metabolites and compound-features that are altered in mutant IDH cells compared to Wild-type have been highlighted, along with ways to visualise these differences. Others can be explored yourself, there are a set of tutorials available for download on the homepage of MetaboAnalyst.

The dataset used in this exercise is from the following study published in Communications Biology by our group in 2020:

Walsby-Tickle, J., Gannon, J., Hvinden, I. *et al.* Anion-exchange chromatography mass spectrometry provides extensive coverage of primary metabolic pathways revealing altered metabolism in IDH1 mutant cells. *Commun Biol* **3,** 247 (2020). https://doi.org/10.1038/s42003-020-0957-6

List of Data processing and analysis tasks

If you are familiar with using MetaboAnalyst follow steps 1-7 directly below and create a report using MetaboAnalyst.

1. Perform data-driven normalisation, transformation and scaling of the dataset. Explore the effects of the different parameters using a heatmap output. Demonstrate the effectiveness of your data processing.

2. Create an overview of the data using a PCA scores plot – are there any sample outliers?

3. Create a volcano plot with a fold-change cut-off set to 2 and FDR-corrected p-value threshold of 0.05. How many compound-features are significantly i) elevated in abundance and ii) depleted?

4. Create a bar graph using a Pearson coefficient to determine the top 25 positively and negatively correlated compounds features with the metabolite **'2-Hydroxyglutarate'.**

5. Produce a PLS-DA score plot, followed by a VIP graph and demonstrate that the accuracy of the model is > 95% using cross validation and report the p-value using a permutation test withy 100 permutations.

6. Produce a hierarchical clustering heatmap of the data showing the top 25 compounds features. Which identified metabolites are present?

7. Download the data and generate a report in MetaboAnalysat. Select the Analysis Report and save it.

Please answer the following questions from Exercise 1: Data Processing and Analysis

1. **Fold-change:** How many compound-features with a fold-change >2 are elevated in mutant samples and how many have a fold-change<2?

2. **Statistical Significance: (volcano plot)**

- What is the name of the two highest ranked identified metabolites?
- Which feature is elevated in 'Mutant' cells and which is depleted?

3. **Hunting for patterns (correlations with 2-HG):** What are the accurate mass values of the two compound-features that are positively correlated with 2-hydroxyglutarate?

4. **PCA Plot:**
    - What are the names of the two samples which represent the greatest outliers in each experimental group according to the PCA plot?
    - Are they within or outside the 95% confidence boundary?

5. **PLS-DA:** Suggest the identity and or chemical formula of the two most important metabolites in the dataset for differentiating the two experimental groups?

6. **Hierarchical clustering:**
    - Which sample does not quite cluster within its experimental group when clustering the full dataset?
    - What are the names of the top three *identified metabolites* in the hierarchically-clustered data?