# Missing Value Imputation Summary:

Missing data in untargeted MS-based metabolomics data occur for various reasons. First, it is possible that the molecules are truly absent from the sample, a situation that may occur e.g. for drug metabolites that only appear in a subset of people taking that medication. On the other hand, there are several technical reasons that could result in missing values, including: (i) instrument sensitivity thresholds, below which concentrations of a specific metabolite might not be detectable in a sample (i.e., below the limit of detection, LOD); (ii) matrix effects that impede the quantification of a metabolite in a sample through other co-eluting compounds and ion suppression; (iii) declining separation ability of the chromatographic column and increasing contamination of the MS instrument; and (iv) limitations in computational processing of spectra, such as poor selection and alignment of the spectral peaks across samples. A widely used and flexible class of missing data strategies is imputation, which involves the replacement of missing values by reasonable substitute values. The most commonly used imputation approaches for metabolomics data assume that missing data occur because they are below the limit of detection (left-censoring, a variant of MNAR).[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6153696/].

### Input Summary:

Input Dataset: test7 -> e.csv

Output Datasets and Files: test7 -> Missing Value Imputation -> result summary.csv -- e.csv

**- Defination of Missing Values:** . What in your dataset can be treated as missing values. "Empty cells" is where there is no value in the excel sheet. "Characters" is where the value is a character but not a number. "Zeros" is when the value is 0. "Negative Values" is where the value is less than 0. Or you can also cusomize a threshold to determine the missing value by selecting the "Values Less Than...". Any value in the excel sheet less than the VALUES LESS THAN will be treated as missing values.

**- Values Less Than:** . The the "values less than..." is selected, any value less than 500 will be treated as missing value.

**- Remove Compounds with Too Many Missing Values:** When the Remove compounds... is checked. The compounds with more than 50% of missing values will be eliminated before imputation.

**- Missing Value Imputation Method (with a disturb):**

replace by half minimum: replace the missing values with random samples from a normal distribution with half-minimum of non-missing value as mean and 1/10 of half-minimum of non-missing value as standard deviation for each compound.

replace by 1/10 minimum: replace the missing values with random samples from a normal distribution with 1/10-minimum of non-missing value as mean and 1/10 of 1/10-minimum of non-missing value as standard deviation for each compound.

replace by minimum: replace the missing values with random samples from a normal distribution with minimum of non-missing value as mean and 1/10 of minimum of non-missing value as standard deviation for each compound.

replace by mean: replace the missing values with random samples from a normal distribution with mean of non-missing value as mean and 1/10 of mean of non-missing value as standard deviation for each compound.

replace by median: replace the missing values with random samples from a normal distribution with median of non-missing value as mean and 1/10 of median of non-missing value as standard deviation for each compound.

imputed by kNN: replace the missing values by k Nearest Neighbors Imputation method. The original kNN imputation was developed for high-dimensional microarray gene expression data (n<p, n is the number of samples, and p is the number of variables). For each gene with missing values, this method finds the k nearest genes using Euclidean metric and imputes missing elements by averaging those non-missing elements of its neighbors. In metabolomics studies, we applied kNN to find k nearest samples instead and imputed the missing elements.

imputed by kNN: replace the missing values by k Nearest Neighbors Imputation method. The original kNN imputation was developed for high-dimensional microarray gene expression data (n<p, n is the number of samples, and p is the number of variables). For each gene with missing values, this method finds the k nearest genes using Euclidean metric and imputes missing elements by averaging those non-missing elements of its neighbors. In metabolomics studies, we applied kNN to find k nearest samples instead and imputed the missing elements.

Bayesian PCA missing value estimation (BPCA): see [https://rdrr.io/bioc/pcaMethods/man/bpca.html]

probabilistic PCA (PPCA): see [https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5169998]

Single Value Decomposition imputation (SVD): SVD imputation will initialize all missing elements with zero then estimate them as a linear combination of the k most significant eigen-variables iteratively until reaches certain convergence threshold. In metabolomics data, we scaled and centralized the data matrix first and then applied this imputation approach with the number of PCs setting to five.

missForest (Imputation with Random Forest): This imputation method applies random forest, a powerful machine learning algorithm, to build a prediction model by setting particular target variable with non-missing values as the outcome and other variables as predictors, then to predict the target variable with missing values iteratively.

QRILC (Quantile Regression Imputation of Left-Censored data): QRILC imputation was specifically designed for left-censored data, data missing caused by lower than LOQ. This method imputes missing elements with randomly drawing from a truncated distribution estimated by a quantile regression. A beforehand log-transformation was conducted to improve the imputation accuracy.

### Result Summary:

The defination of missing values are empty cells, characters, zeros, negative values, values less than 500. In total, there are 754 missing values in the dataset. Compounds with more than 50% missing values were excluded. In total, there were 61 compounds excluded. Then the missing values in the dataset were imputed by the method of replace by half minimum. For more details, please see Table 1.

Table Explanation.

- index: the index of compounds, mainly for sorting the table.

- label: compound labels.

- empty.cells: the number of empty cell for each compound.

- characters: the number of character values for each compound.

- zeros: the number of zeros for each compound.

- negative.values: the number of negative values for each compound.

- values.less.than.500: the number of values less than 500 for each compound.

- total: the total number of missing values for each compound.

- excluded: if Yes, the compound is excluded.

| label | empty.cells | zeros | negative.values | characters | values.less.than.500 | total | excluded |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Thromboxane B2 | 0 | 0 | 0 | 0 | 12 | 12 | Yes |
| Prostaglandin D2 | 0 | 0 | 0 | 0 | 12 | 12 | Yes |
| 9,12,13-trihydroxyoctadec-10-enoic acid | 0 | 0 | 0 | 0 | 12 | 12 | Yes |
| 9S,10R-dihydroxy-stearic acid | 0 | 0 | 0 | 0 | 12 | 12 | Yes |
| 12,13-dihydroxyoctadec-9-enoic acid | 0 | 0 | 0 | 0 | 12 | 12 | Yes |
| 9,10-dihydroxyoctadec-12-enoic acid | 0 | 0 | 0 | 0 | 12 | 12 | Yes |
| 15,16-dihydroxyoctadeca-9,12-dienoic acid | 0 | 0 | 0 | 0 | 12 | 12 | Yes |
| 12,13-dihydroxyoctadeca-9,15-dienoic acid | 0 | 0 | 0 | 0 | 12 | 12 | Yes |
| 9,10-dihydroxyoctadeca-12,15-dienoic acid | 0 | 0 | 0 | 0 | 12 | 12 | Yes |
| 14,15-dihydroxyeicosa-5,8,11-trienoic acid | 0 | 0 | 0 | 0 | 12 | 12 | Yes |

Table 1: several compounds with most missing values.

# Student t-test Summary:

Two Student t-tests show which compounds have the power to differentiate the different two-groups in the data set. The two sample t-test is applied to one metabolite at a time (i.e., a univariate analysis) to determine whether the mean values of the two groups are different. The null hypothesis for the test is H0 : mu\_group1 = mu\_group2, and the alternative hypothesis is Ha : mu\_group1 != mu\_group2. If the p-value for the test is smaller than a cutoff value, typically 0.05, the null hypothesis is rejected. If the p-value is large, there is no significant difference between the mean values for the two groups, indicating the metabolite has little power to separate them.

The 0.05 cutoff value is often used when the t-test for a metabolite is examined individually, without considering the tests for other metabolites. A multiple comparison procedure can be employed, in which a smaller cutoff value is used, to control the overall error caused by using all the t-tests together.

### Input Summary:

Input Dataset: test7 -> Missing Value Imputation -> e.csv

Output Datasets and Files: test7 -> Missing Value Imputation -> Student t-test -> student\_t\_test.csv

**- Treatment Group:** Genotype. The student t-test will be performed on each compound to detect those significantly altered by Genotype.

**- Variance Equality Assumption:** FALSE. If TRUE, the equality of variance is assumed, and the tests are Student's t-test, otherwise Welch t-test.

**- Correct the False Discovery Rate:** Benjamini -- Hochberg (1995). When conducting multiple tests, the rate of incorrectly reject a null hypothesis will increase. FDR-controlling procedures are designed to control the expected proportion of "discoveries" (rejected null hypotheses) that are false (incorrect rejections). The suggested method for metabolomics is the Benjamini -- Hochberg procedure. For more information, please visit https://en.wikipedia.org/wiki/False\_discovery\_rate.

### Result Summary:

Welch t-test (not assuming equal variance) was performed on each compound to test if the mean average of Lmbrd1 different from Null. Out of 679 compounds, 69 are significant with p-value less than 0.05. To control the false discovery rate (FDR), the Benjamini -- Hochberg (1995) procedure was used. After FDR correction, 1 compounds are still significant. See Table 2 for more detail.

Table Explanation.

- index: the index of compounds, mainly for sorting the table.

- label: compound labels.

- p\_values: p-values from t-tests.

- p\_values\_adjusted: p-values adjusted by the FDR correction procedure.

| label | p\_values | p\_values\_adjusted |
| --- | --- | --- |
| 3,6-anhydro-D-galactose | 1.245943e-08 | 8.459951e-06 |
| 9,12-Octadecadiynoic Acid | 6.707096e-04 | 2.277059e-01 |
| gamma-Glutamylmethionine | 2.858643e-03 | 4.356493e-01 |
| .gamma.-L-Glu-.epsilon.-L-Lys | 3.590251e-03 | 4.356493e-01 |
| Thiazolidine-4-carboxylic acid | 3.614101e-03 | 4.356493e-01 |
| Bicyclo[2.2.1]heptane-2-methanol | 3.849626e-03 | 4.356493e-01 |
| Palmitamide | 5.212657e-03 | 4.702353e-01 |
| glycerol | 6.045576e-03 | 4.702353e-01 |
| 3-Ureidopropionic acid | 7.647428e-03 | 4.702353e-01 |
| TG (56:7) | 8.109736e-03 | 4.702353e-01 |

Table 2: most significant compounds (i.e. small p-values)

# Fold Change Summary:

Fold change is often used to reveal the direction as well as the magnitude of changing of two groups for a compound. A fold change is defined as the average value of one group divided by another (i.e. the baseline group). There are two metrics for the average, mean and median. The mean fold change can be deemed as parametric, while the median as the non-parametric. A fold change greater than 1 means the compounds increased compared to the baseline and vise versa.

### Input Summary:

Input Dataset: test7 -> Missing Value Imputation -> e.csv

Output Datasets and Files: test7 -> Missing Value Imputation -> Fold Change -> fold\_change.csv

**- Treatment Group:** Genotype. The fold change, defined as the average of Lmbrd1 divided by Null, will be performed on each of the compounds.

**- Mean or Median:** median. Mean is the Arithmetic average, while the median is the 'middle value' (i.e. 50% quantile) average. Mean can be deemed as parametric, while median as non-parametric.

### Result Summary:

Fold change, defined as the median average ratio of Lmbrd1 to the Null were calculated for each of the compound. Out of 679 compounds, there are **336** (49%) increased from Null to Lmbrd1 with fold\_change greater than 1, while **343** (51%) decreased.

Table Explanation.

- index: the index of compounds, mainly for sorting the table.

- label: compound labels.

- fold\_changes: the median average of Lmbrd1 divided by the median average of Null. A fold change greater than 1 indicates a increasing from Null to Lmbrd1.

See Table 2, Table 3 and Figure 1 for more detail.

| label | fold\_changes |
| --- | --- |
| Tetradecylamine | 77.068123 |
| 1-Hexadecylamine | 30.540887 |
| 1-Octadecanamine | 20.494183 |
| Cotinine | 16.199981 |
| Bicyclo[2.2.1]heptane-2-methanol | 13.896599 |
| Palmitamide | 7.593477 |
| n-Pentadecylamine | 6.861397 |
| Methenamine | 6.421594 |
| 9,12-Octadecadiynoic Acid | 5.291904 |
| hippuric acid | 4.270760 |

Table 2: most increased compounds (from Null to Lmbrd1).

| label | fold\_changes |
| --- | --- |
| ribose | 0.1113365 |
| Urocanic acid | 0.1744019 |
| 3-(3-hydroxyphenyl)propionic acid | 0.2147682 |
| Serotonin | 0.2149211 |
| pyruvic acid | 0.2293504 |
| TG (46:1); TG(12:0/16:0/18:1); | 0.2833557 |
| TG (48:1); TG(14:0/16:0/18:1); | 0.2875649 |
| TG (48:3); TG(14:0/16:1/18:2); | 0.2899897 |
| SM (d32:2) | 0.2932567 |
| uric acid | 0.2973056 |

Table 3: most decreased compounds (from Null to Lmbrd1).

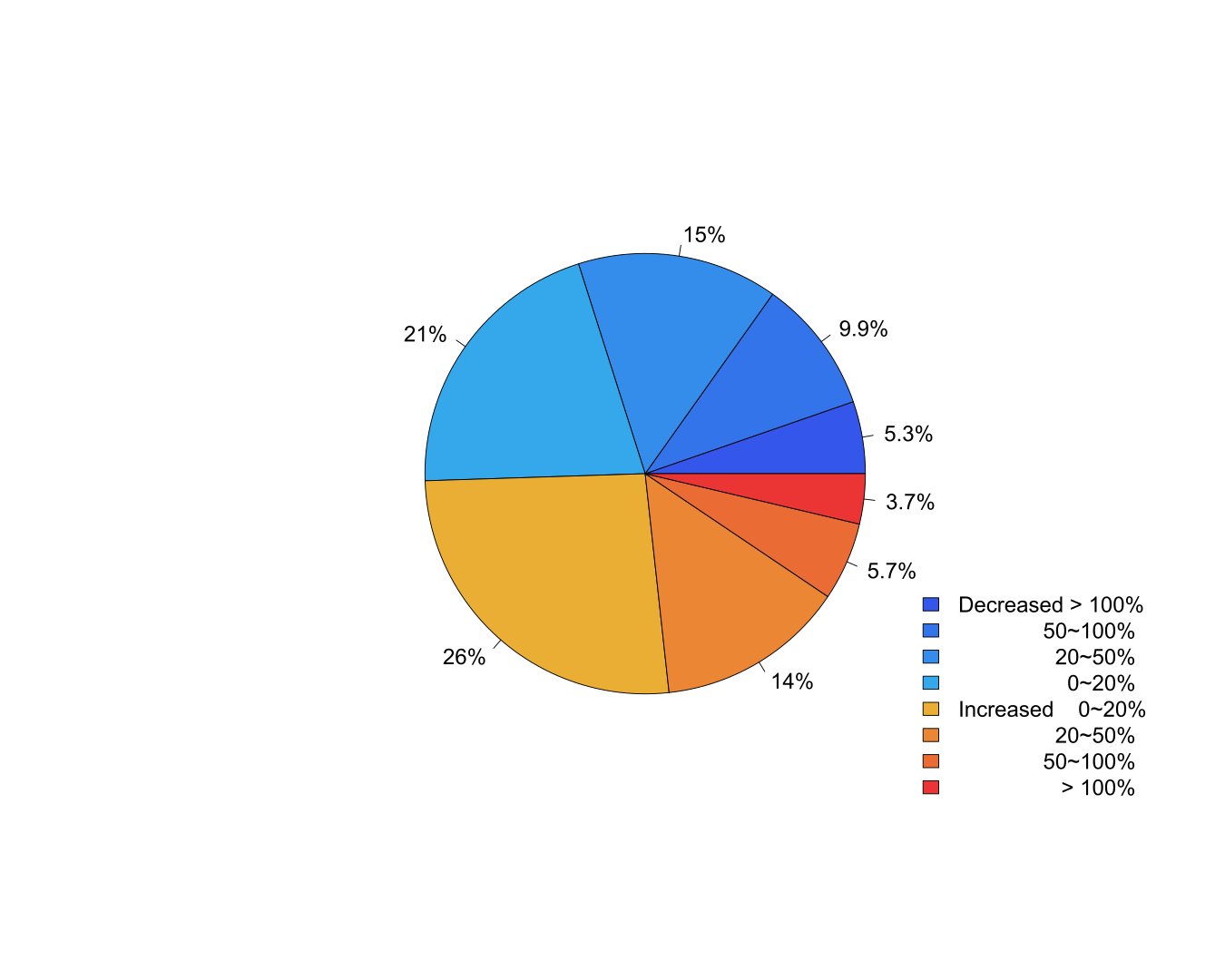


Figure 1: Detailed summary of Fold Changes.

Figure 1 shows a more detailed summary of fold changes, where the percentage of compounds with 0~20%, 20~50% and more than 100% increasing and decreasing are shown.

# Volcano Plot Summary:

Volcano plots are sometimes used for visualization of statistical results of omics data such as differential expression of genes measured through microarrays. The interactive volcano plot has the power to show at a click of a mouse button which metabolites show a stronger combination of fold change and statistical significance. They represent significance from a statistical test (such as a p-value) on the y-axis and fold-change on the x-axis. They can also compare metabolite levels with different experimental conditions. As a consequence, metabolites in the volcano plot that have a relatively low fold-change between the two samples appear near the center and metabolites that have significant p-values are found in the upper-right or upper-left. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3629923/]

### Input Summary:

Input Dataset: test7 -> Missing Value Imputation -> Fold Change -> Student t-test -> fold\_change.csv -- student\_t\_test.csv

Output Datasets and Files: test7 -> Missing Value Imputation -> Volcano Plot -> volcano\_plot.svg

**- P-value Cut-off:** 0.05.

**- Fold Change Cut-off:** 1.5 (Fold change greater than 1.5 is considered to be large increase, whereas fold change less than 1/1.5=0.67 is considered to be large decrease).

### Result Summary:

Volcano plot shows the relationship between the p-values and the fold changes. See Figure 2

Figure 2 is the volcano plot. There are 23 compounds significantly increased more than the fold change cut-off, while 24 compounds significantly decreased more than the fold change cut-off. The y axis is the -log10 p-values. The higher the more significant. The x axis is the log2 Fold Change. The further from the origin the larger the fold change.

Figure 2

Visualize the p-values and fold changes for each compound. The x-axis shows the log2 scale fold change. The further from the origin the larger the fold change. The y-axis is the -log10 p-values. The higher the more significant. Only p-values less than the criterion and fold change greater than the criterion, the dots are colored. See TRACES for more options.

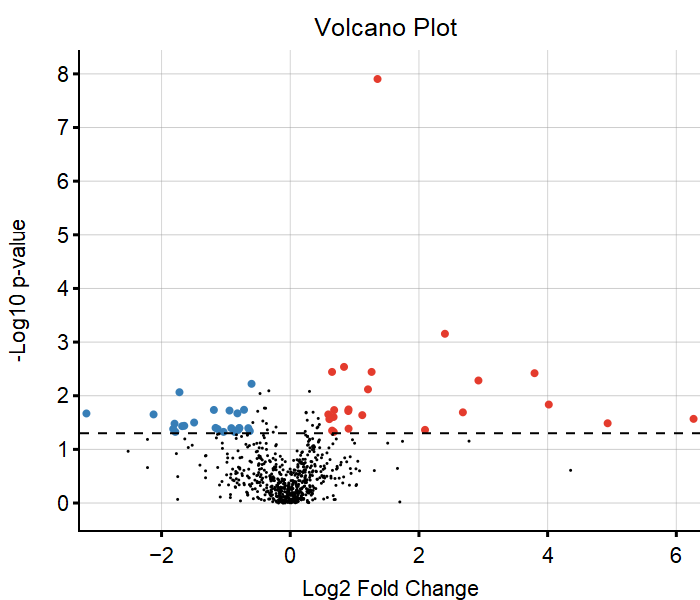


Figure 2: volcano plot.

# Principal Component Analysis (PCA) Summary:

The high-dimensional and wide data tables that are encountered in metabolomics can be difficult to analyse, but the metabolomics community is now routinely applying many techniques to interrogate these large data sets and increase our understanding of the changes in metabolism. The use of data reduction or dimension reduction methods to reduce the size of the data table (while minimizing information loss) before further statistical analysis takes place is extremely important in this respect. The previous step discussed a popular dimension reduction method, namely principal component analysis (PCA). [https://www.futurelearn.com/courses/metabolomics/0/steps/25039]

PCA is an example of a so-called unsupervised technique. This means that the method does not use class label information (i.e. to which group does each sample in the data table belong). This has important consequences for dimension reduction. Dimension reduction is achieved by a rotation of the data followed by mathematical projection to a lower dimension resulting in a small data table. PCA rotates (i.e. linearly transforms) the variables (i.e. compound values) such that the largest differences between the samples are highlighted. This is very useful for explorative analysis of the data, for example, to detect outliers.

### Input Summary:

Input Dataset: test7 -> Missing Value Imputation -> e.csv

Output Datasets and Files: test7 -> Missing Value Imputation -> PCA -> scree\_plot.svg -- loading\_plot.svg -- score\_plot.svg -- compound\_loadings.csv -- sample\_scores.csv

**- Scaling Method:** Auto Scaling. Scaling methods are data pretreatment approaches that divide each variable by a factor, the scaling factor, which is different for each variable. They aim to adjust for the differences in fold differences between the different metabolites by converting the data into differences in concentration relative to the scaling factor. It is highly recommended for multivariate statistical analyses. More information is available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1534033/.

### Result Summary:

The dataset was first scaled using Auto Scaling. Then the Principal Component Analysis (PCA) was performed on the scaled dataset. See Figure 4, Figure 5 and Figure 6 for more details.

Figure 4

Scores plot. It summarizes the linear relationship/similarity between the samples. Samples colors/shapes/sizes with 95% confidence intervals can be added afterwards to visualize the sample clusters. The confidence interval can also be used for outlier detection.

Figure 5

Loadings plot. It summarizes the linear relationship/similarity between the compounds. Compounds colors/shapes/sizes with 95% confidence intervals can be added afterwards to visualize the compound clusters. Together with the scores plot (Figure 4), loadings plot can help to understand the relationship between the compounds and samples. For example, the compounds with loadings in the first quadrant in the loadings plot is positively correlated with the samples with scores in the first quadrant in the scores plot. The further the loadings from the origin, the higher the correlation. On the other hand, the compounds with loadings in the third quadrant are negatively correlated with samples in the first quadrant in the score plot.

Figure 6

Screes plot. It visualizes the percentage of variance explained by each of the principal components. Variance can be deemed as 'information' in the dataset. The first two principal components summarize a total of 47.03% variation in the dataset.

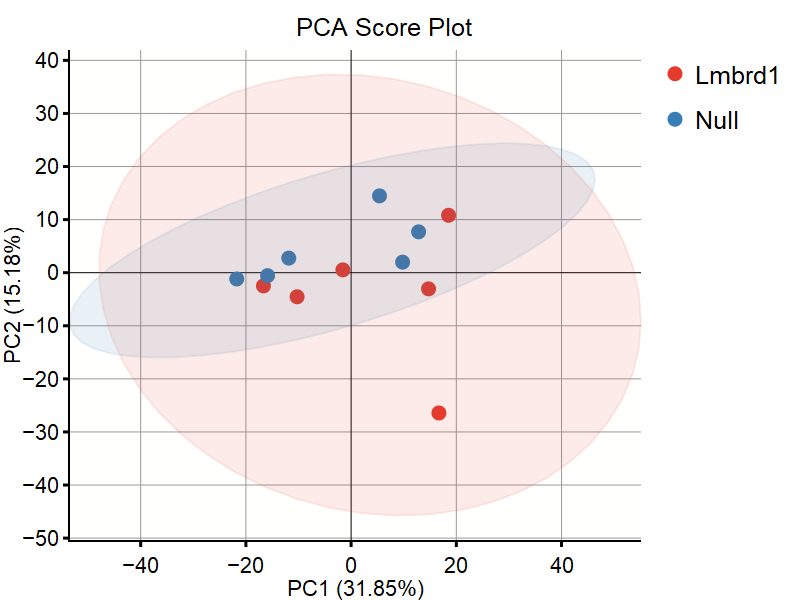


Figure 4: PCA Scores Plot.

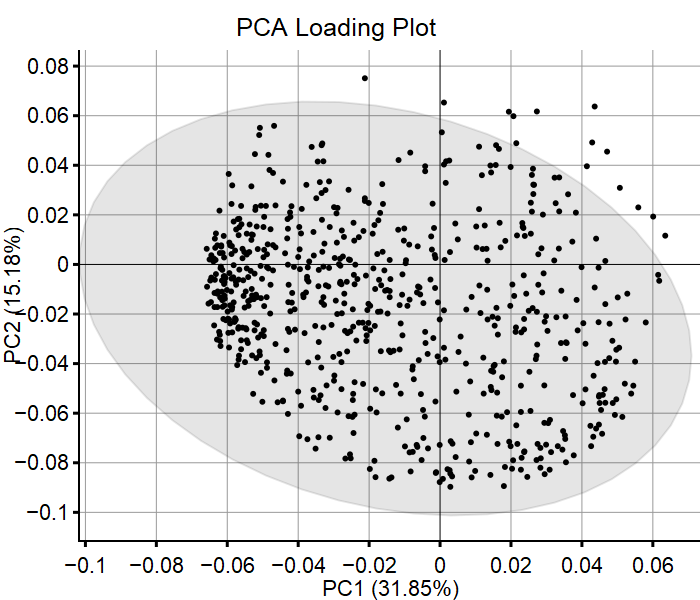


Figure 5: PCA Loadings Plot.

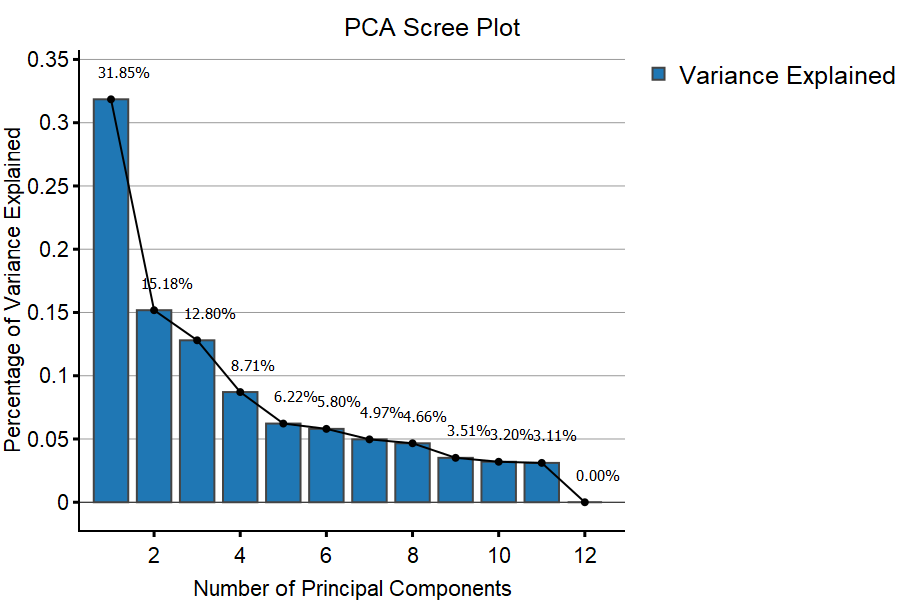


Figure 6: PCA Scree Plot.

# Partial Least Square - Discriminant Analysis (PLSDA) Summary:

PLS-DA is a chemometrics technique used to optimise separation between different groups of samples, which is accomplished by linking two data matrices X (i.e., raw data) and Y (i.e., groups, class membership etc.). The method is in fact an extension of PLS1 which handles single dependent continues variable whereas PLS2 (called PLS-DA) can handle multiple dependent categorical variables. This approach aims to maximize the covariance between the independent variables X (sample readings; that is to say the metabolomics data) and the corresponding dependent variable Y (classes, groups; that is to say the targets that one wants to predict) of highly multidimensional data by finding a linear subspace of the explanatory variables. This new subspace permits the prediction of the Y variable based on a reduced number of factors (PLS components, or what are also known as latent variables). These factors describe the behavior of dependent variables Y and they span the subspace onto which the independent variables X are projected.

The main advantage of this PLS-DA approach is the availability and handling of highly collinear and noisy data, which are very common outputs from metabolomics experiments. In addition, this provides several statistics such as loading weight, variable importance on projection (VIP) and regression coefficient that can be used to identify the most important variables. This technique provides a visual interpretation of complex datasets through a low-dimensional, easily interpretable scores plot that illustrates the separation between different groups. Comparison of loadings and scores plot supports investigations in terms of the relationship between important variables that can be specific to the group of interest.

### Input Summary:

Input Dataset: test7 -> Missing Value Imputation -> e.csv

Output Datasets and Files: test7 -> Missing Value Imputation -> PLSDA -> perm\_plot.svg -- vip\_plot.svg -- scree\_plot.svg -- loading\_plot.svg -- score\_plot.svg -- compound\_loadings.csv -- sample\_scores.csv

**- Treatment Group:** Genotype. The PLS-DA model will be performed to find a linear transformation on the compounds to discriminant the Genotype group.

**- Scaling Method:** Auto Scaling. Scaling methods are data pretreatment approaches that divide each variable by a factor, the scaling factor, which is different for each variable. They aim to adjust for the differences in fold differences between the different metabolites by converting the data into differences in concentration relative to the scaling factor. It is highly recommended for multivariate statistical analyses. More information is available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1534033/.

### Result Summary:

The dataset was first scaled using Auto Scaling. Then the Partial Least Square - Discriminant Analysis (PLS-DA) was performed on the scaled dataset. When the number of predictive components equals to 9, the model achieves the highest Q2 score. In this model, the R2X (variances explained), R2Y, and Q2 (cum) is 5.8038, 1 and 1, respectively. See Figure 7, Figure 8 and Figure 9 for more details.

Figure 7

Scores plot. It can be used to visually validate the fitness of the PLS-DA model. If the between-group variation is obvious, it indicates a goodness of fit. Samples colors/shapes/sizes with 95% confidence intervals can be added to visualize the sample clusters.

Figure 8

Loadings plot. It summarizes the linear relationship/similarity between the compounds. Compounds colors/shapes/sizes with 95% confidence intervals can be added afterwards to visualize the compound clusters. Together with the scores plot (Figure 7), loadings plot can help toidentify compounds contributing to between-group variability based on separations observed between groups in the scores plot.

Figure 9

Screes plot. Commonly, R2X and R2Y represent the fraction of variance of the X and Y matrix, respectively, and Q2 represents the predictive accuracy of the model, with cumulative (cum) values of R2X, R2Y and Q2 equating to ~1 indicating an effective model. Please note, when the number of components increase, the total sum of the R2X and R2Y increase, but not necessarily the Q2 as it is cross-validated R2Y. Generally, the higher the Q2 the better the model is. If R2X and R2Y is high but Q2 is too low, it means a crisis of overfitting.

This model achieves a Q2 of 1 at the 9 component.

Figure 10

Vip score plot. The VIP (Variable Importance in Projection) quantifies the contribution of a compound when building the model. Usually, a VIP score greater than one is considered important and (positively) affect classification between the groups.

On the right hand side of the vip score plot is a simple heatmap, indicating the changing direction of the compounds.

This model achieves a Q2 of 1 at the 9 component.

Figure 11

Permutation test plot. A permutation test can evaluate whether the PLS-DA classification in the designed groups is significantly better than any other random classification in arbitrary groups.

In our case, 0 permutations was performed on 9 components (which achieved the highest Q2 score). The p-value of the R2Y is 1, while the Q2 (cum) is 0.96. At least one of the p-values of R2Y and Q2 (cum) less than 0.05 indicates a valid model. Otherwise, it is hard to justify whether the designed dataset is different from a random arbitrarty dataset.

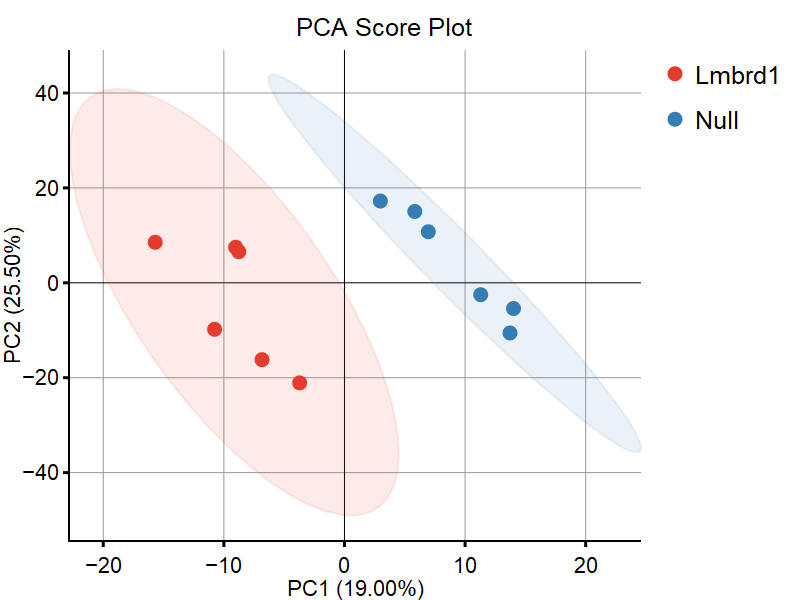


Figure 7: PLS-DA Scores Plot.

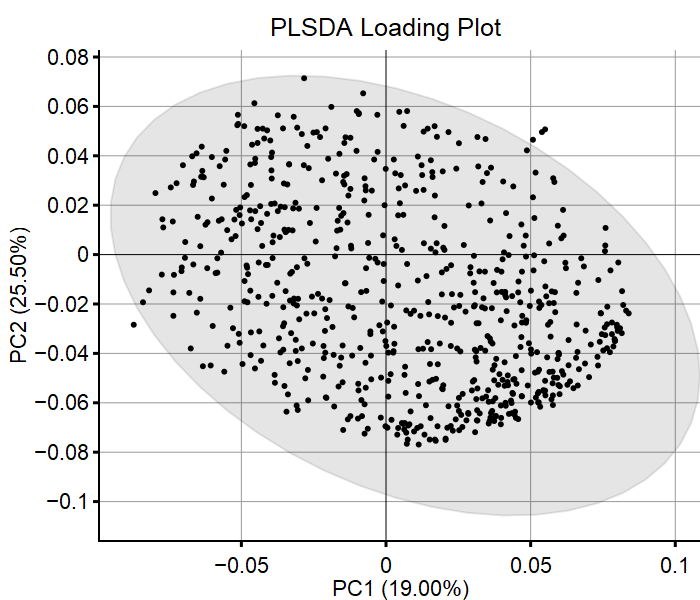


Figure 8: PLS-DA Loadings Plot.

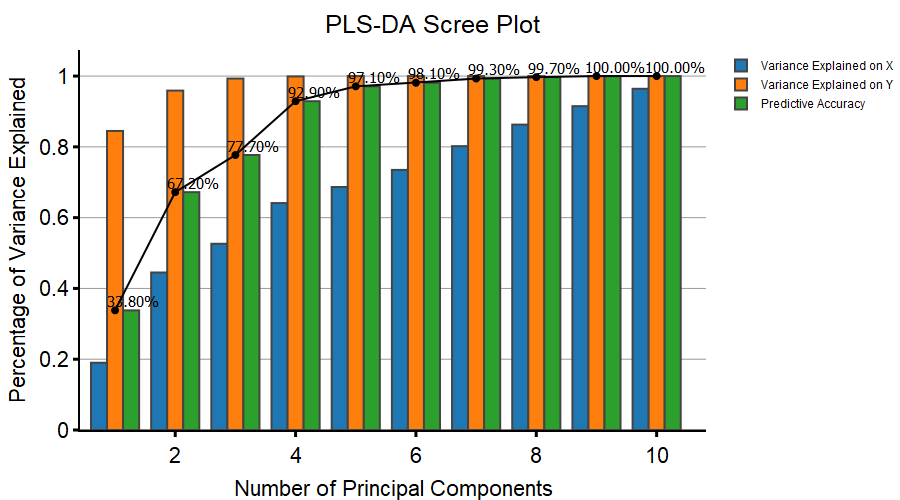


Figure 9: PLS-DA Scree Plot.

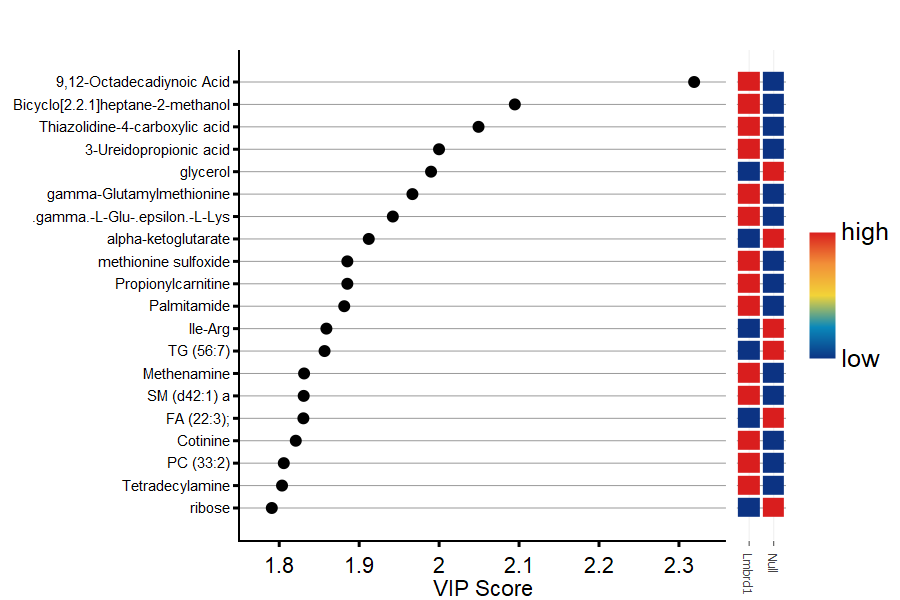


Figure 10: PLS-DA VIP Score Plot.

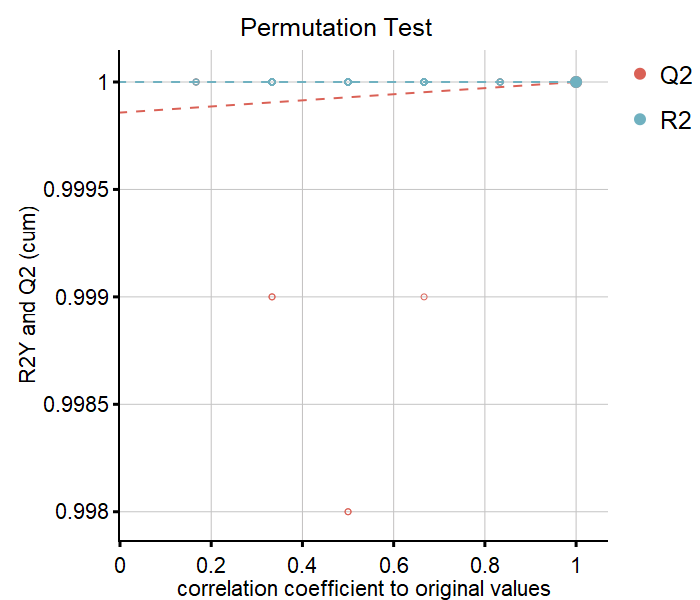


Figure 11: PLS-DA Permutation Plot.

# Data Subsetting:

This module can subset a dataset based on sample information and/or compound information. For example, you could subset dataset with Gender == 'Female' only. Or you could subset compounds with p-value less than 0.05 and get the dataset with significant compounds only.

### Input Summary:

Input Dataset: test7 -> Missing Value Imputation -> e.csv

Output Datasets and Files: test7 -> Missing Value Imputation -> Significant Compounds Only -> e.csv

**- Subset Data By Samples:** FALSE.

**- Subset Data By Compounds:** TRUE.

Compounds having all of the following criterions will be subset.

1. Compound Info Data: test7 -> Missing Value Imputation -> Student t-test -> student\_t\_test.csv. Column: p\_values. Type: numeric. Range: 0 to 0.05.

### Result Summary:

Data was subset according to the sample and compound criterion.

**- Subset Data By Samples:** FALSE.

**- Subset Data By Compounds:** TRUE.

Compounds having all of the following criterions are subset.

1. Compound Info Data: test7 -> Missing Value Imputation -> Student t-test -> student\_t\_test.csv. Column: p\_values. Type: numeric. Range: 0 to 0.05.

As a result, the subset data contains 12 samples and 70 compounds.

# Heatmap -- Dendrogram Summary:

Heatmaps are an effective tool for displaying feature variation among groups of samples. The basic concept of a heatmap is to represent relationships among variables as a color image. Rows and columns typically are reordered according to the dendrograms so that variables and/or samples with similar profiles are closer to one another, making these profiles more visible. Each value in the data matrix is displayed as a color, making it possible to view the patterns graphically.

Heatmaps uses an agglomerative hierarchical clustering algorithm to order and display the data as a dendrogram. Two important factors to consider when constructing a heatmap are the type of distance measure and the agglomeration method used. For details on the various methods available see [https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-S16-S10].

### Input Summary:

Input Dataset: test7 -> Missing Value Imputation -> Significant Compounds Only -> e.csv

Output Datasets and Files: test7 -> Missing Value Imputation -> Significant Compounds Only -> Heatmap -> heatmap\_plot.svg -- compound\_order.csv -- sample\_order.csv

**- Scaling Method:** Auto Scaling. Scaling methods are data pretreatment approaches that divide each variable by a factor, the scaling factor, which is different for each variable. They aim to adjust for the differences in fold differences between the different metabolites by converting the data into differences in concentration relative to the scaling factor. It is highly recommended for multivariate statistical analyses. More information is available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1534033/.

**- Agglomeration Method:** average. Agglomeration is the process by which clusters are merged into larger clusters.

**- Distance Function:** euclidean. A distance metric is a non-negative number which measures the difference between two objects (e.g. samples/compounds.)

### Result Summary:

The Heatmap and Dendrograms on the dataset with Auto Scaling. See Figure12 for more details.

Figure 12

shows the heatmap and dendrogram analysis on the dataset with Auto Scaling. The row displays compounds and the column represents the samples. The color represent the value scale for the corresponding compound and sample (see colorbar).

The dendrogram on the right is conducted using distance metric euclidean with agglomeration method of average on the compounds while the top is on the samples.

The order of samples is determined by the dendrogram, and the compounds by the dendrogram. The annotation of the samples (Genotype) were added on top of the plot. No annotation of the compounds were added to the plot.

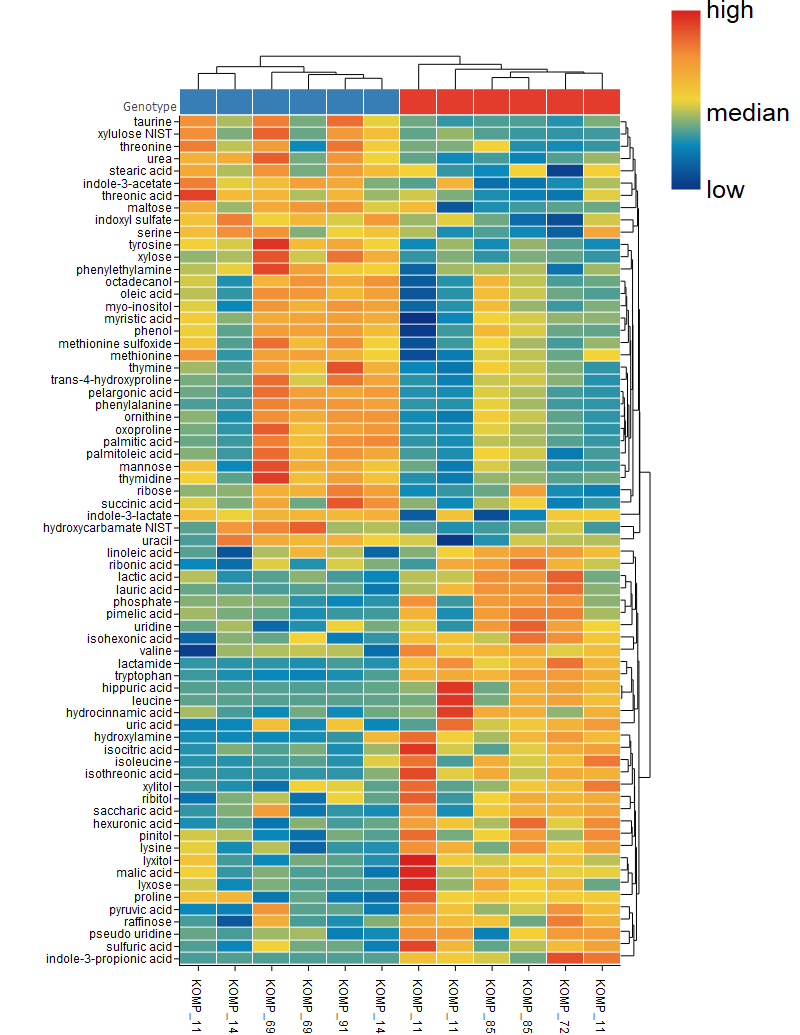


Figure 12: the heatmap of the dataset with Auto Scaling.

# Boxplot Visualization Summary:

Boxplots are good ways to visualize and compare compounds among different samples. A boxplot graphically depicts a vector through its five-number summary. The edges of the box represent the lower and upper quartiles while the whiskers represent the minimum and maximum values. The median is displayed as a line within the box. Outliers are represented as symbols outside of the whiskers. More information is available on https://www.nature.com/articles/nmeth.2813.

### Input Summary:

Input Dataset: test7 -> Missing Value Imputation -> Significant Compounds Only -> e.csv

Output Datasets and Files: test7 -> Missing Value Imputation -> Significant Compounds Only -> Boxplot -> boxplot\_plot.zip

**- values:** The boxplot were generated to visualize the value of each compound. Samples were colored by Lmbrd1, Null.

### Result Summary:

Boxplot for each compound were generated and saved in a zip file.

**- values:** The boxplot were generated to visualize the value of each compound. Samples were colored by Lmbrd1, Null.

An example boxplot is shown in Figure 14.

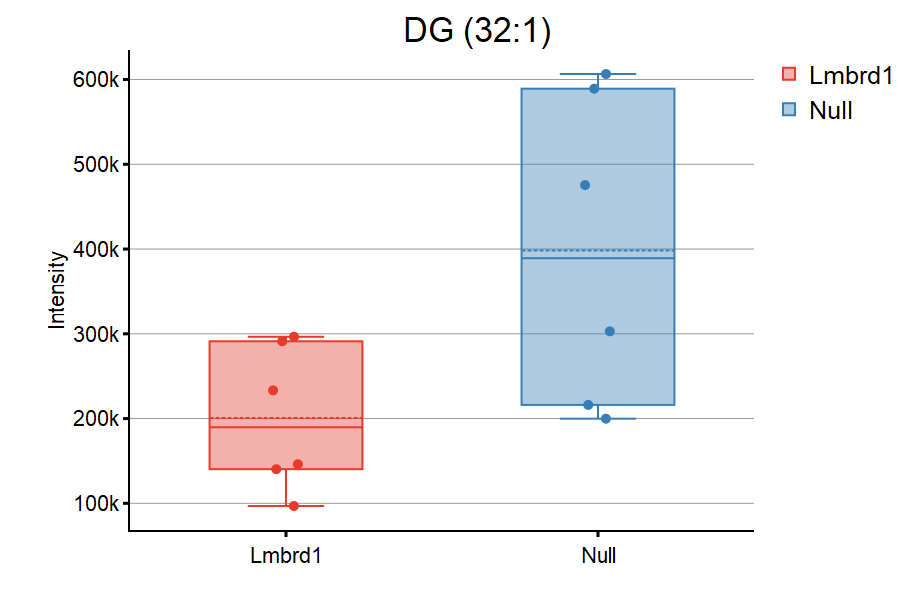


Figure 14: an example of boxplot.

# Sample Size Estimation -- Power Analysis Summary:

Warnes and Liu (2006) provide a simple method for computing sample size for microarray experiments, and reports on a series of simulations demonstrating its performance. Surprisingly, despite its simplicity, the method performs exceptionally well even for data with very high correlation between measurements.

The key component of this method is the generation of a cumulative plot of the proportion of compounds achieving a desired power as a function of sample size, based on simple gene-by-gene calculations. While this mechanism can be used to select a sample size numerically based on pre-specified conditions, its real utility is as a visual tool for understanding the trade off between sample size and power. In our consulting work, this latter use as a visual tool has been exceptionally valuable in helping scientific clients to make the difficult trade offs between experiment cost and statistical power.

Multiple comparison problem can also be taken into account with a correction of Benjamini-Hochberg Procedure. The proportion of true null hypothesis in your data can be estiamted using qvalue package.

### Input Summary:

Input Dataset: test7 -> Missing Value Imputation -> e.csv

Output Datasets and Files: test7 -> Missing Value Imputation -> post-hoc Sample Size Analysis -> power\_plot.svg -- ssize\_plot.svg -- ssize.csv

**- Test Type:** t-test. Compute the statistical power of the two-tailed student t-test given a sample size, and determines the sample size to obtain a target statistical power.

**- Treatment Group:** Genotype. The effect size, an input parameter for power analysis, will be estimated by the Genotype. Then the statistical power will be calculated based on the sample size in your dataset and the estimated sample size will be calculated based on the target statistical power.

**- Interested Number of Observations (per group):**  the given sample size. The statistical power will be estimated based on the given sample size.

**- Target Power (%):**  the target statistical power. The sample size to achieve the target power will be calculated.

**- Significance Level:**  Type I error rate (the significant criterion), the rate of falsely reject a true null hypothesis.

**- FDR Correction:**  perform the sample size estimation and power analysis considering the FDR (False Discovery Rate) correction

**- FDR Criterion:**  estimate the sample size and perform the power anlaysis while controlling the FDR at 0.05 level.

### Result Summary:

The statistical powers were estimated for each compound given a sample size of 12, and the sample size was estimated based on a target power of 0.8. The effect size of each compound was calculated from the dataset using the treatment group Genotype. A significant level of 0.05 was used. Multiple comparision (or false discovery rate, FDR) problem was also taken into account with Benjamini-Hochberg procedure. The FDR was controlled at the level of 0.05. The significant level for each compounds was adjusted accordingly. See Table 5, Figure 15 and 16 for more detail.

Table Explanation.

- index: the index of compounds, mainly for sorting the table.

- label: compound labels.

- power (n=12): the estimated statistical power given sample size of 12.

- n (power=0.8): the estimated sample size for the target power of 80%.

Figure 15

answers the question of What is the necessary per-group sample size for 80% power with the observed effect size and at significant level of 0.05?.

The plot illustrates that smaple size of 27 ,45 ,68 is required to ensure that at least 10%, 20%, and 30% of compounds have a statistical power greater than 80%. It is also shown that a sample size of 6 is sufficient if 0.2946% of the compounds need to achieve a 80% power.

Figure 16

answers the question of What is the power for 12 samples per group with the observed effect size and significant level of 0.05?.

From the plot, 1.031% of compounds achieve at 80% statistical power at the sample size of 12 and significant level of 0.05.

| index | label | power (n=12) | n (power=0.8) |
| --- | --- | --- | --- |
| 1 | xylulose NIST | 0.0128434907 | 85.79958 |
| 2 | xylose | 0.0308160141 | 56.55924 |
| 3 | xylitol | 0.0094129432 | 100.70405 |
| 4 | valine | 0.0076155628 | 112.85576 |
| 5 | uridine | 0.0230052826 | 64.69849 |
| 6 | uric acid | 0.0002421252 | 3499.45507 |
| 7 | urea | 0.0005222776 | 951.01696 |
| 8 | uracil | 0.0006383546 | 745.80893 |
| 9 | tyrosine | 0.0090960826 | 102.54702 |
| 10 | tryptophan | 0.0006371618 | 747.41355 |

Table 5: First 10 compounds and their estimated statistial powers of having 12 samples and required sample size for 80% power.

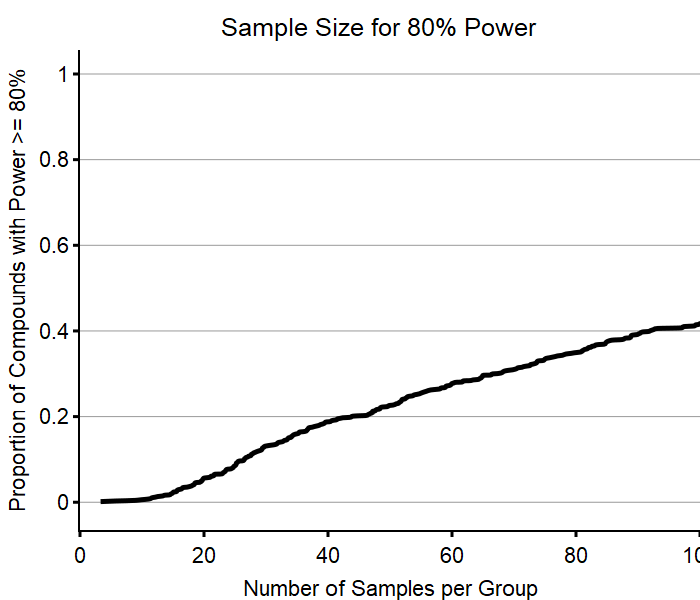


Figure 15: the proportion of compounds needing x samples to achieve a 80% statistical power.

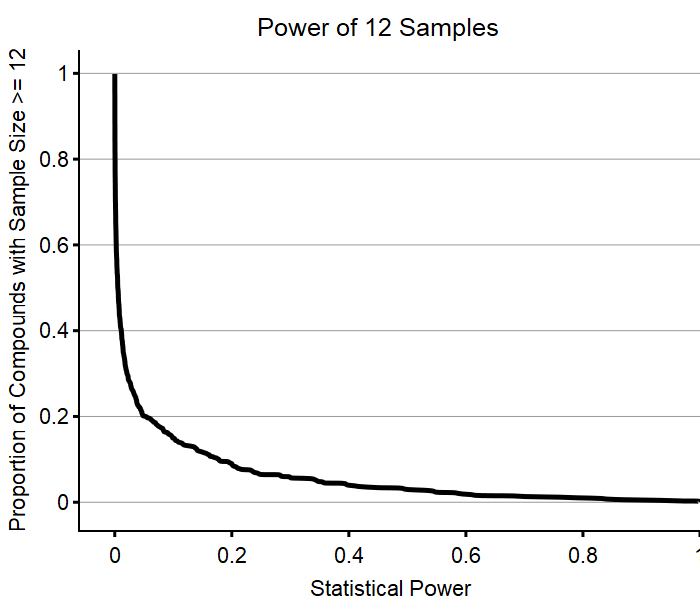


Figure 16: the proportion of compounds having x% statistical power when having 12 samples.