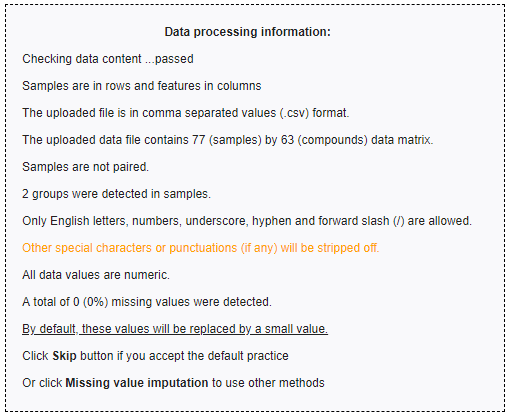
**Project 4 of BIO316 – Metabolomics Data Analysis**

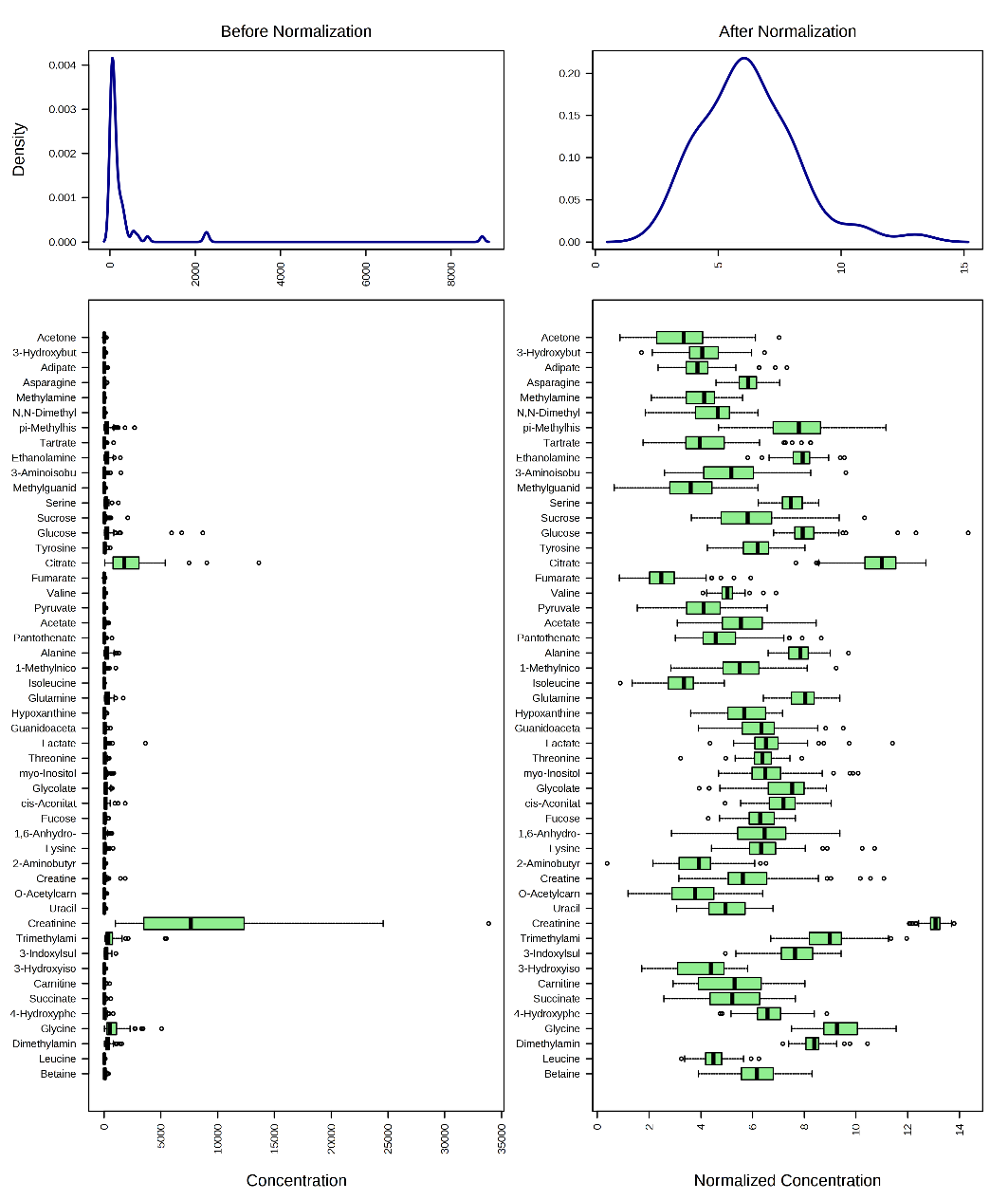
**Task 1:** Data processing and normalization (15 points)

1. The data integrity check run automatically and the result is shown below. For lists of concentrations the data integrity check will assess the content (look for consistent formatting and the presence of two groups), determine whether the data is paired or determine if negative numbers exists. In this case, no missing values were identified in the data. Since zero values may cause some algorithms not to work properly, MetaboAnalyst will replace these values with a small positive value (the half of the minimum positive number detected in the data).



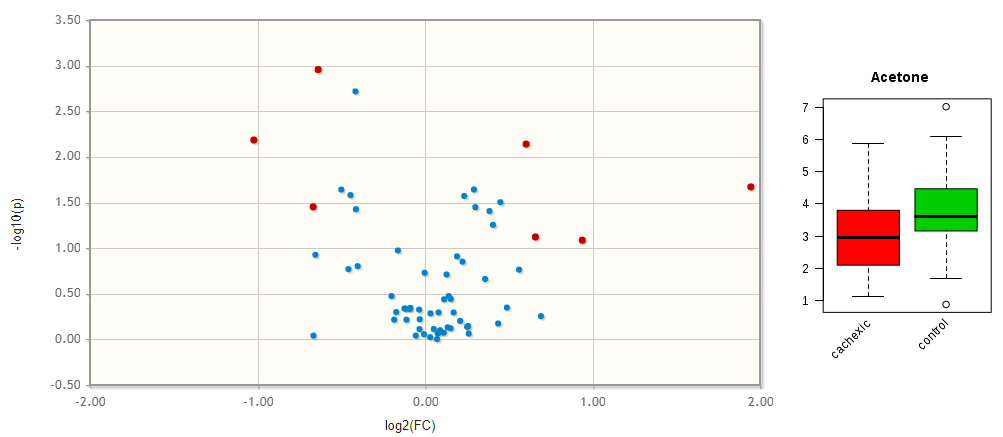
1. The distribution of normalized and original concentrations is shown below. A general rule of reference sample selection in normalization is to choose a sample in the control group with the fewest missing values. In this case, NETCR\_005\_V1 was used as the reference sample for row-wise normalization.

The internal data structure is transformed now to a table with each row representing a urine sample (from a patient) and each column representing a feature (a compound with a concentration). Row-wise normalization and column-wise normalization are often applied sequentially to reduce systematic variance and to improve the performance for downstream statistical analysis. Row-wise normalization aims to normalize each sample (row) so that that they are comparable to each other. In contrast to row-wise normalization, column-wise normalization aims to make each feature (column) more comparable in magnitude to each other.



**Task 2:** Identification of significantly different metabolites (50 points)

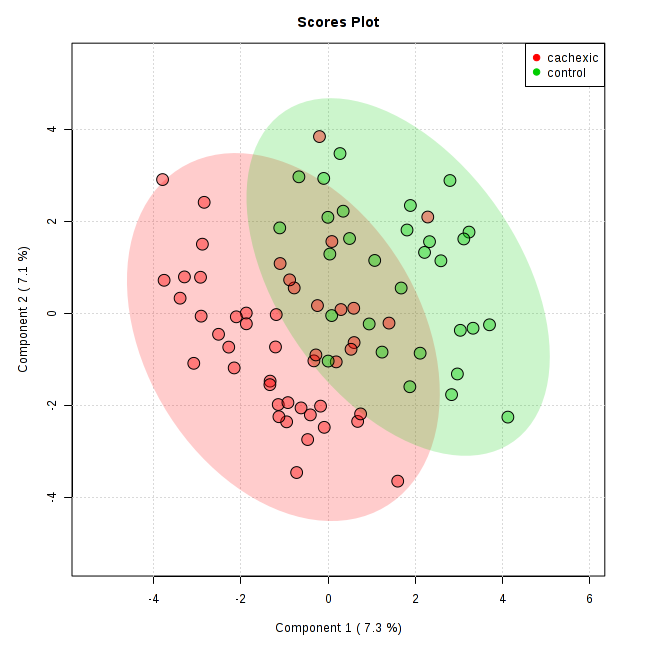
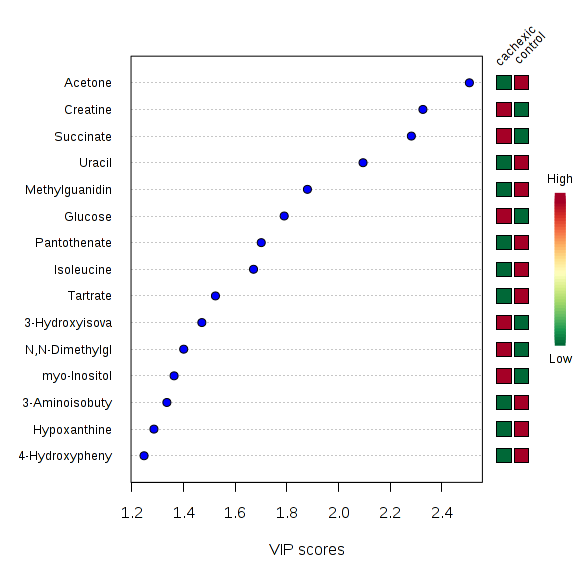
1. Volcano Plot is a commonly used univariate analysis method that is often first used to obtain an overview of the data or a rough ranking of potentially important features. Volcano plots are used to compare the size of the fold change to the statistical significance level. The X axis plots the fold change between the two groups (on a log scale), while the Y axis represents the p-value for a t-test of differences between samples (on a negative log scale). As a result, seven metabolites were identified as significantly different between the two groups by Volcano Plot analysis (see the table below).



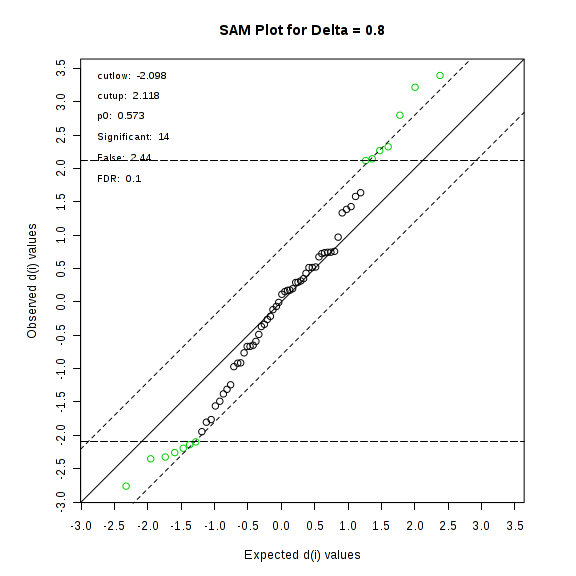
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | FC | log2(FC) | raw.pval | -LOG10(p) |
| Uracil | 0.64258 | -0.63804 | 0.001104 | 2.9569 |
| Acetone | 0.49240 | -1.0221 | 0.006531 | 2.1850 |
| Succinate | 1.51940 | 0.60348 | 0.007237 | 2.1404 |
| Glucose | 3.85100 | 1.94520 | 0.021325 | 1.6711 |
| Pantothenate | 0.62941 | -0.66793 | 0.035254 | 1.4528 |
| Adipate | 1.57910 | 0.65909 | 0.075384 | 1.1227 |
| myo-Inositol | 1.91740 | 0.93914 | 0.081861 | 1.0869 |

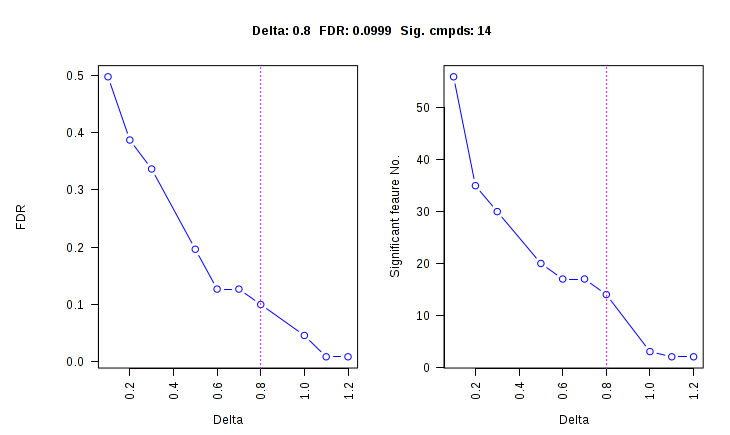
1. In PLS-DA, Variable Importance in Projection or VIP score is a weighted sum of squares of the PLS loadings. The weights are based on the amount of explained Y-variance in each dimension. VIP indicates the importance of the variable to the whole model. In many studies VIP values >2.0 are selected and used for further data analysis, but this cut-off depends on the number of variables used. A more relaxed VIP cutoff of around 1.0 can be used if only a small number of variables (e.g. <100) are involved in the study. As a result, four metabolites, **Acetone (VIP=2.5062), Creatine (VIP=2.3263), Succinate (VIP=2.2816) and Uracil (VIP=2.0946)**, were identified as significantly different between the two groups by PLS-DA.

Univariate analyses are often first used to obtain an overview of the data or a rough ranking of potentially important features before applying more sophisticated data analysis tools. Univariate analysis examines each variable separately without taking into account the effect of multiple comparisons. As a supervised method, PLS-DA can perform both classification and feature selection. The algorithm uses cross-validation to select an optimal number of components for classification.

1. The Delta plots are a visualization of the table generated by SAM that contains the estimated FDR and the number of identified metabolites for a set of Delta values. The default Delta value (0.8) has an FDR of 0.1 and identifies 14 significant compounds above this threshold. One can increase the Delta to reduce the FDR if one is seeking stricter results.



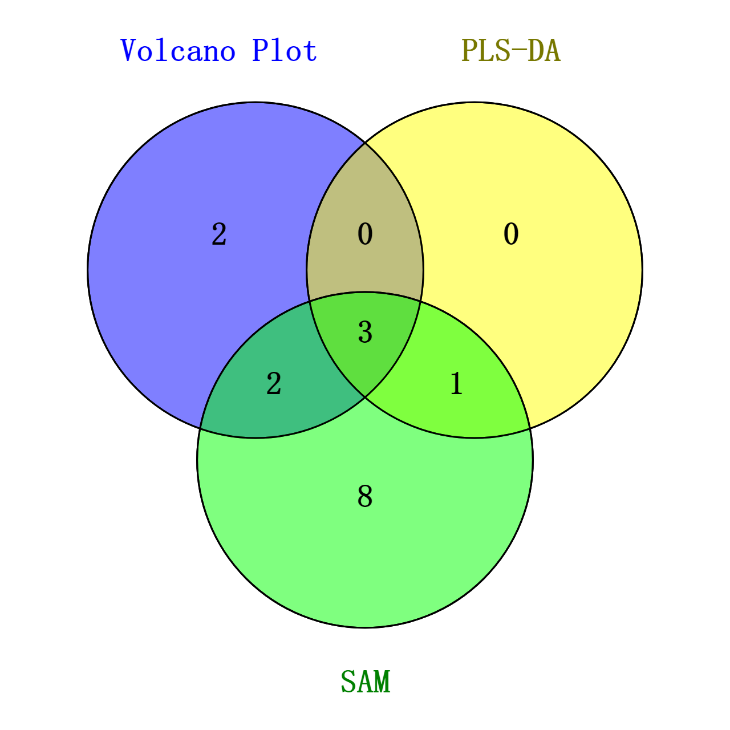


|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | d.value | stdev | rawp | q.value |
| Uracil | 3.3937 | 0.20475 | 0.000952 | 0.022938 |
| Isoleucine | 3.2167 | 0.17232 | 0.00127 | 0.022938 |
| Acetone | 2.798 | 0.29714 | 0.005714 | 0.055912 |
| Succinate | -2.7612 | 0.27413 | 0.006191 | 0.055912 |
| Glucose | -2.3515 | 0.25246 | 0.02 | 0.10004 |
| Glutamine | -2.3249 | 0.1466 | 0.02127 | 0.10004 |
| Methylguanidine | 2.3246 | 0.26821 | 0.02127 | 0.10004 |
| 4-Hydroxyphenylacetate | 2.2679 | 0.18244 | 0.024603 | 0.10004 |
| Creatine | -2.2587 | 0.34168 | 0.024921 | 0.10004 |
| cis-Aconitate | -2.194 | 0.18212 | 0.031429 | 0.10158 |
| Pantothenate | 2.1442 | 0.26312 | 0.036349 | 0.10158 |
| Alanine | -2.1406 | 0.1316 | 0.036349 | 0.10158 |
| Hypoxanthine | 2.1182 | 0.20133 | 0.037302 | 0.10158 |
| N,N-Dimethylglycine | -2.0984 | 0.22144 | 0.039365 | 0.10158 |

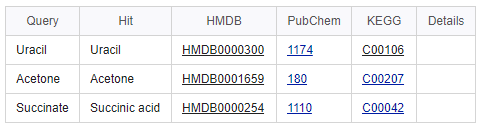
SAM is designed to address False Discovery Rate (FDR) problems when running multiple tests on high-dimensional data. It first assigns a significance score to each variable based on its change relative to the standard deviation of repeated measurements. Then it chooses variables with scores greater than an adjustable threshold and compares their relative difference to the distribution estimated by random permutations of the class labels. For each threshold, a certain proportion of the variables in the permutation set will be found to be significant by chance. This number is used to calculate the FDR.

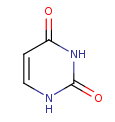
**Task 3:** Consensus results and functional analysis (15 points)

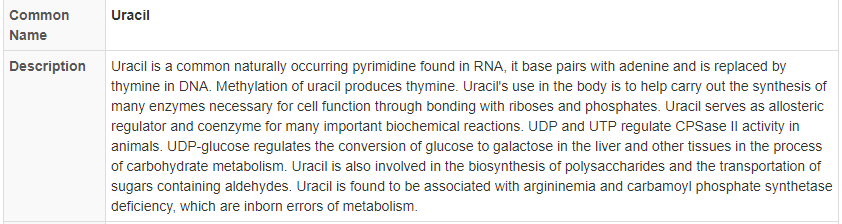
1. As shown in the Venn diagram, 3 metabolites were consistently identified by all the three approaches including **Uracil, Acetone** and **Succinate.**



1. By searching Pathway Analysis for Uracil, Acetone and Succinate, cross-links to HMDB, KEGG and PubChem are shown below.



Take **Uracil** as example, by checking the information from HMDB (Uracil HMDB ID: 0000300), we can find that Uracil serves as allosteric regulator and coenzyme for many important biochemical reactions. It is also involved in the biosynthesis of polysaccharides and the transportation of sugars containing aldehydes. Uracil is found to be associated with argininemia and carbamoyl phosphate synthetase deficiency, which are inborn errors of metabolism.



Besides, Uracil plays roles in Pyrimidine metabolism, beta-Alanine metabolism and Pantothenate and CoA biosynthesis processes.

