# 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: -> e.csv

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

**- Scaling Method:**

Figure 1

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 2

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 1), 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 3

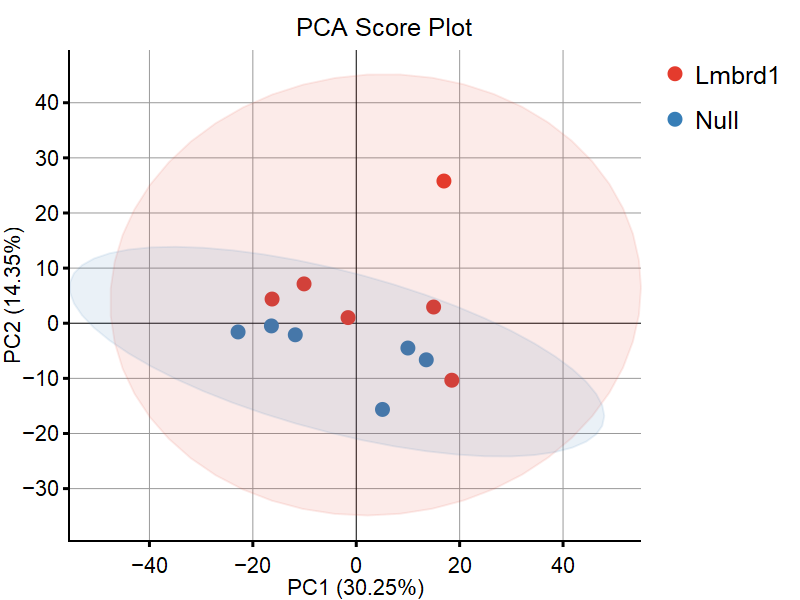


Figure 1: PCA Scores Plot.

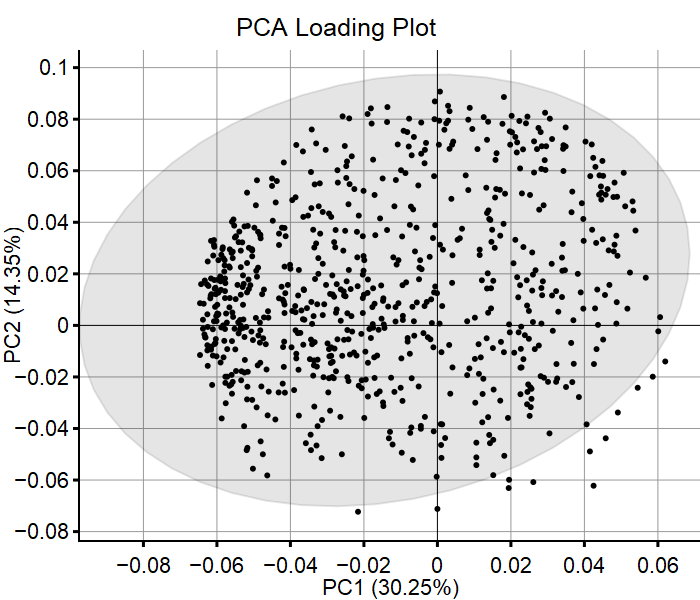


Figure 2: PCA Loadings Plot.

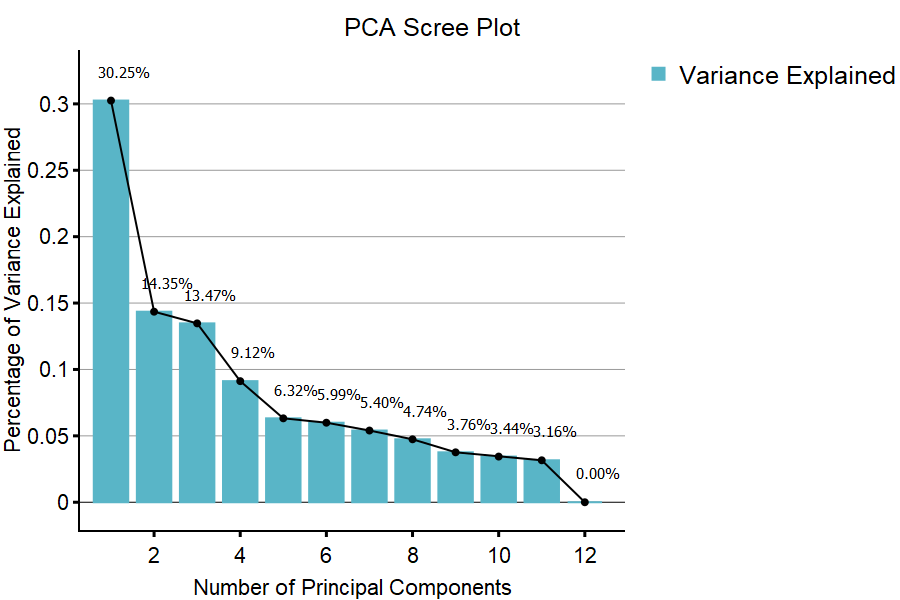


Figure 3: PCA Scree Plot.