Delta Chemometrics Tool Installation and Documentation (Chemospec Package)

July 2026

Contents

1. Installation and Loading Packages 2

2. File Loading 4

3. Preprocessing 5

a. Bucket Integration 5

b. ChemPack Analysis 7

4. PCA Processing 8

a. Plots 9

b. Loading 9

c. Scree 11

d. Diagnosis 12

e. Graph Options 13

5. HCA Processing 14

a. Distance 15

b. Cluster 16

c. Auto Scale 17

d. Heatmap 17

e. Graph Options 18

1. Installation and Loading Packages

1. Install R: https://www.r-project.org/
2. Install Delta and activate licence. See instructions here.
3. Open R.
4. Open Delta and go to “Options” > “Preferences” > “External”, scroll down to “RScript Executable” and select the RScript.exe file.
    a. If you are having a hard time finding the Rscript.exe, use the file search function.

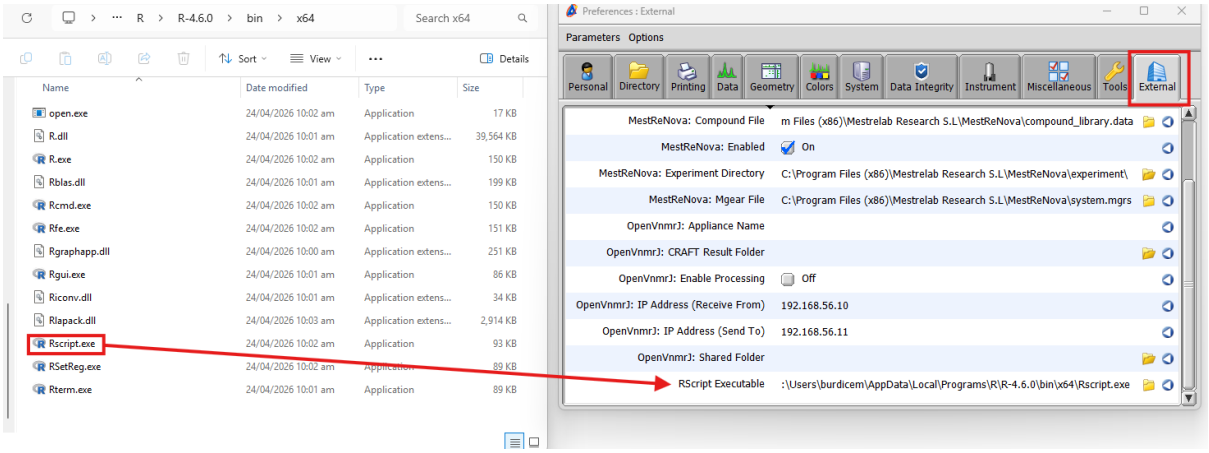


Figure 1. Selecting RScript.exe.

5. In Delta, open the chemometrics unit (Analyse > chemometrics).

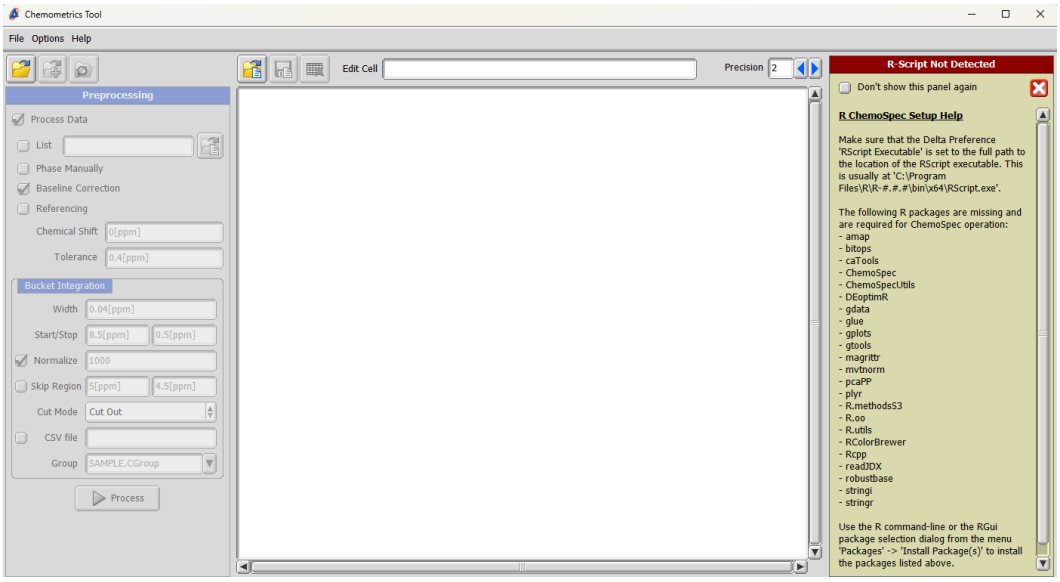


Figure 2. Chemometrics tool before extra R packages are loaded.

6. Delta will say that some R packages are needed to begin, go to the R Script window and copy and paste the below lines, this will take a moment to begin, and a CRAN proxy must be selected (this does not matter that much, just pick one near to your location).

```
install.packages("amap")
install.packages("bitops")
install.packages("caTools")
install.packages("ChemoSpec")
install.packages("ChemoSpecUtils")
install.packages("DEoptimR")
install.packages("gdata")
install.packages("glue")
install.packages("gplots")
install.packages("gtools")
install.packages("magrittr")
install.packages("mvtnorm")
install.packages("pcaPP")
install.packages("plyr")
install.packages("R.methodsS3")
install.packages("R.oo")
install.packages("R.utils")
install.packages("RColorBrewer")
install.packages("Rcpp")
install.packages("readJDX")
install.packages("robustbase")
install.packages("stringi")
install.packages("stringr")
```

R Packages installation commands

7. Ensure all packages are installed and no errors have occurred.

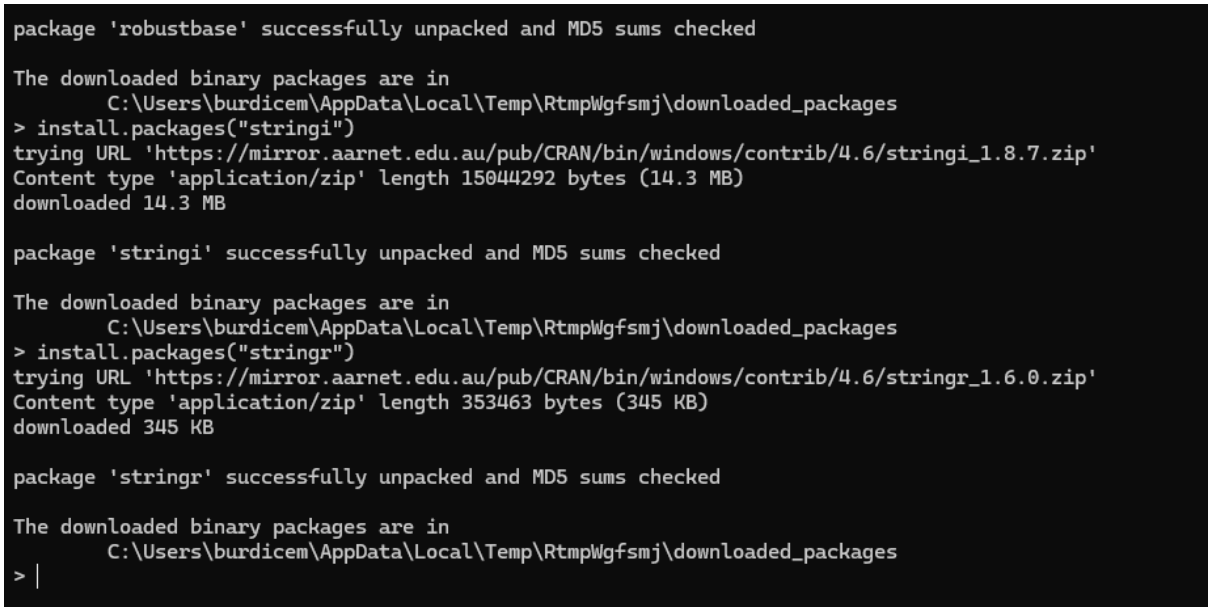


Figure 3. Successful Package Installation.

8. Close the chemometrics tab and reopen, the error should have disappeared, you are ready for data entry.

2. File Loading

To add spectral data to the chemometrics tool, select the open file button and select the spectra. To select multiple, hold the control button while selecting, or select the "file" dropdown menu and select the "open all (folder)" option to open all spectra within a file. Alternatively, if a csv was generated from previous processing, data can be loaded directly.

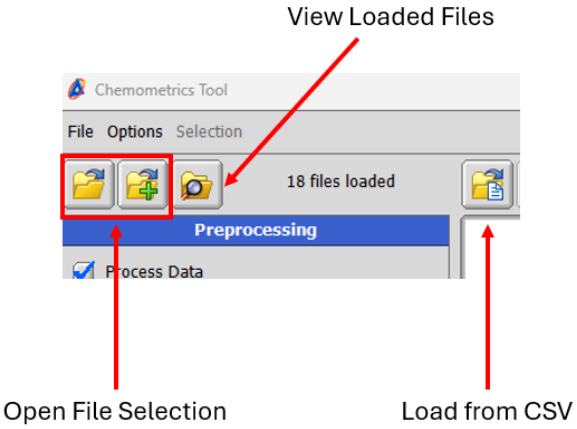


Figure 4. File Selection Options

The file selection window looks like this. You can also view loaded files once selected, allowing for selection/deletion of specific loaded files.

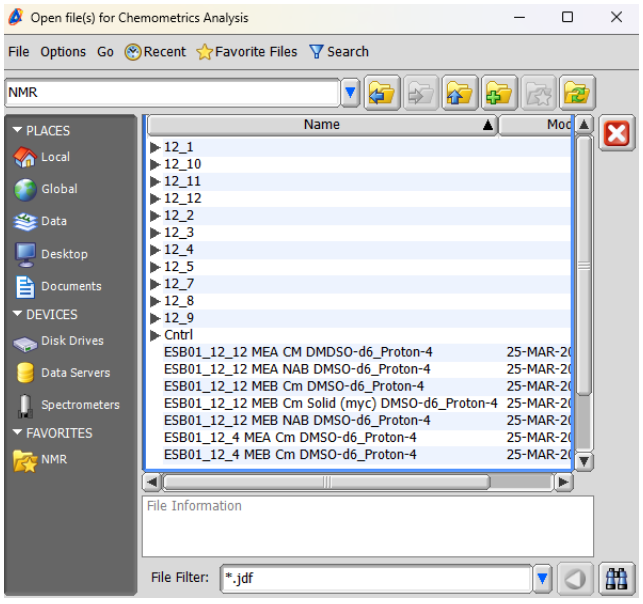


Figure 5. File Selection Window

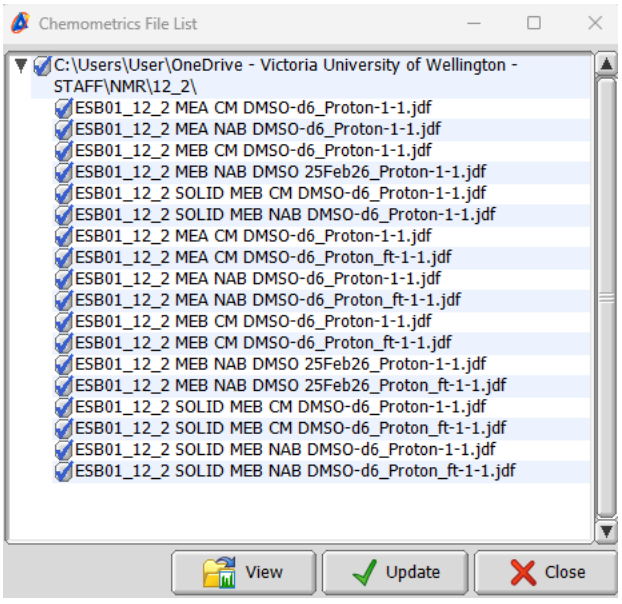


Figure 6. Loaded Files Window

3. Preprocessing

Once files are selected, it must be processed for analysis in delta, you can either process the data yourself before analysis or use the chemometrics unit preprocessing. To preprocess the data in the chemometrics unit, select the “process data” box. When no other boxes are checked the chemometrics tool perform a Fourier transform and automatic phase correction.

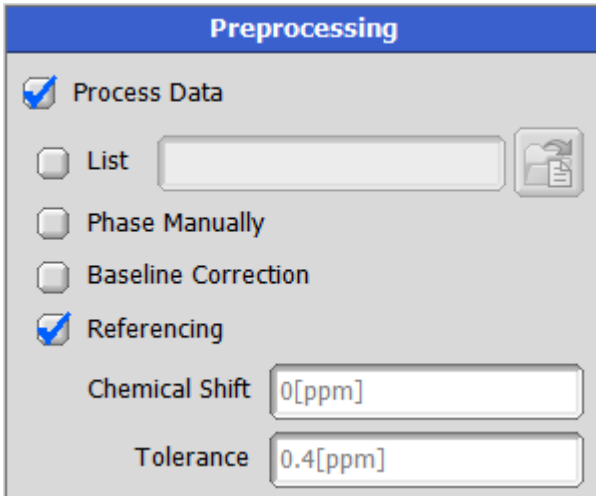


Figure 5. Preprocessing Selections.

List: Insert a processing list file from delta.

Phase manually: Opens a popup to phase each spectrum individually.

Baseline correction: Corrects wandering baselines.

Referencing: Allows for a chemical shift reference, the tool will select the highest peak at the entered shift within tolerance and set it to the entered shift.

a. Bucket Integration

Groups of frequencies are collapsed into one frequency value, and the corresponding intensities are summed up to create buckets/bins. Calculating peak areas within specified segments minimizes the effects of variations caused by feild strength or pH differences.

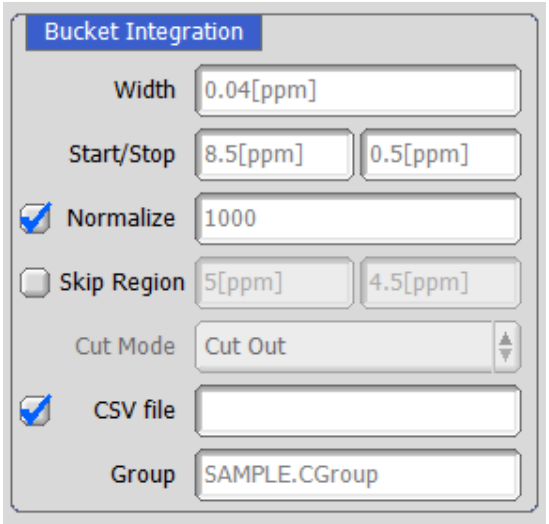


Figure 6. Bucket Integration Selections.

Width: Width of each bucket/bin, smaller bins will retain better peak information but will also extend the processing time and complication. Common range is between 0.01 and 0.04 ppm.

Start/stop: Range of ppm within the spectra to be binned.

Normalize: Each row (spectrum) is normalized by the sum of all bins. (See ChemPack Analysis for more information).

Skip region: Range of ppm that will not be binned, used when looking in specific regions.

Cut mode:

Cut out: Removes the bins in the indicated region, even if a bin overlaps into the next region. Will slightly cut off data on both the high and low end.

Cut Mid: Removes the bins in the indicated region, even if a bin overlaps into the next region. Will slightly cut off data at the high end.

Cut in: Removes exactly the region indicated.

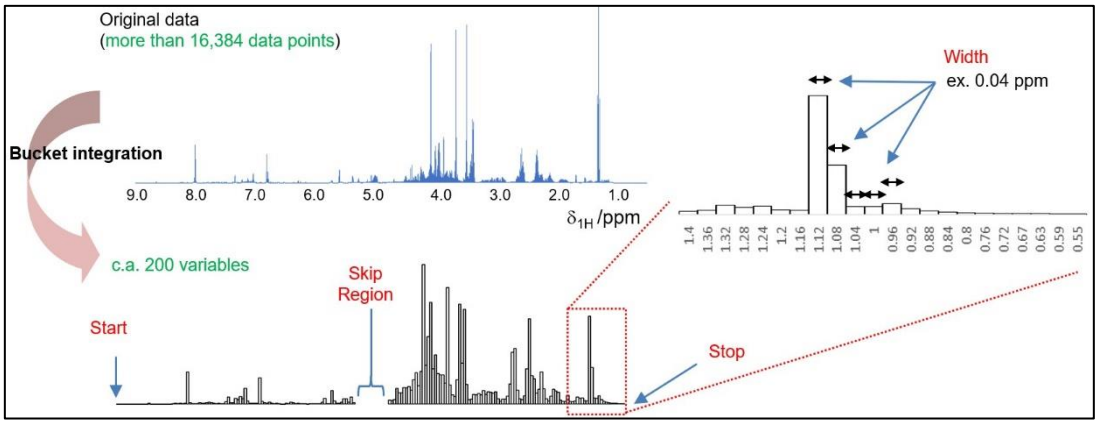


Figure 7. Bucket Integration Process Illustrated.

CSV file: Reports the created bins into a csv file, will be exported into the JEOL Delta reports file.

Group: Optional naming of samples into a category.

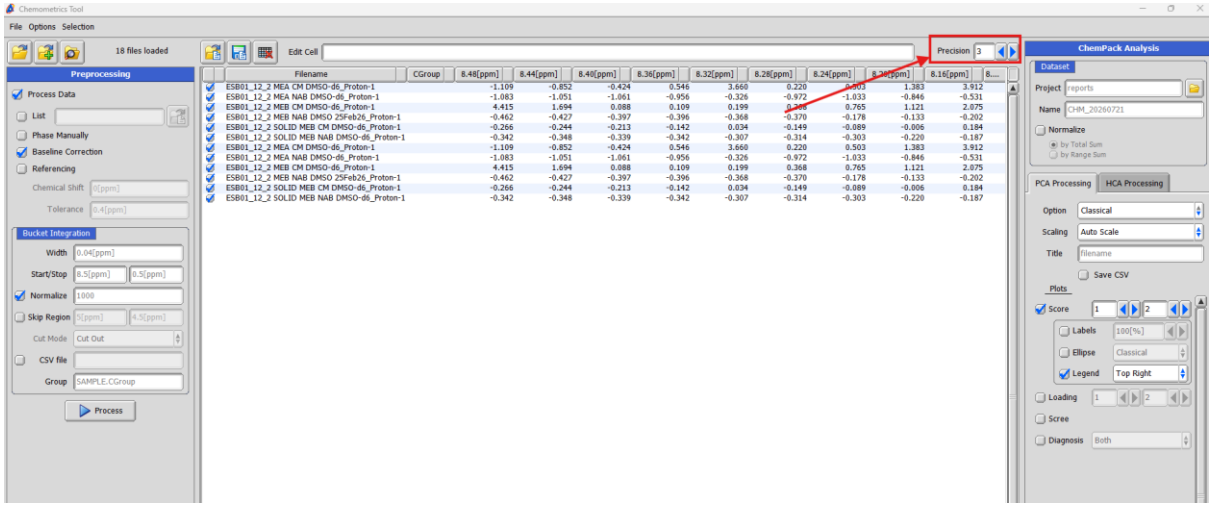


Figure 8. Loaded and Processed Data with Precision Selection Highlighted.

Once preprocessing and bucket integration settings are selected, the process button can be clicked. If phase manually is selected the spectra will pop-up for manual phase correction. Once completed the selected spectra and their bins will be reported in the center. If the csv options were selected this is when the file is created. In addition, the precision of the bin score measurements can be indicated in the top right. Groups can also be added to specific samples here.

b. ChemPack Analysis

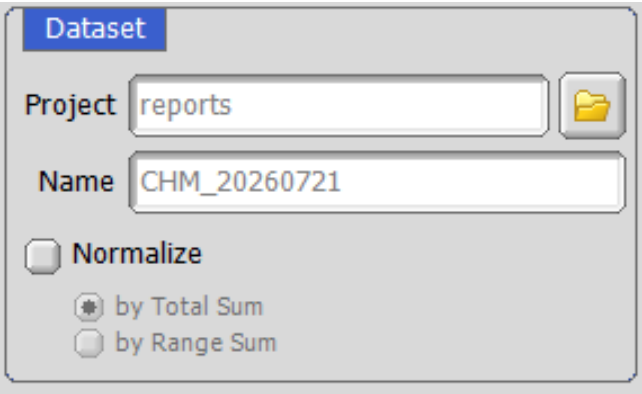


Figure 9. ChemPack Analysis Selections.

Project: Where the project files are stored on your computer.

Name: To save the project and its file name.

Normalize: Makes data from all samples comparable to each other.

Total sum: Normalize every observation by dividing every element in a sample by the total row sum so each spectra sums to 1. This method assumes that the total concentration of all substances causing peaks does not vary across samples.

Range sum: Similar to total sum, but instead of using the total sum, uses the sum of a specific entered range. This is used when an internal standard is present and you want to normalize relative to it.

4. PCA Processing

"Principal Component Analysis" is a dimensionality reducing analysis technique that determines the minimum number of principal components to describe a data set – removing noise and redundant information. Able to reduce data to essential components to show patterns but does not correlate to concrete evidence of similarity.

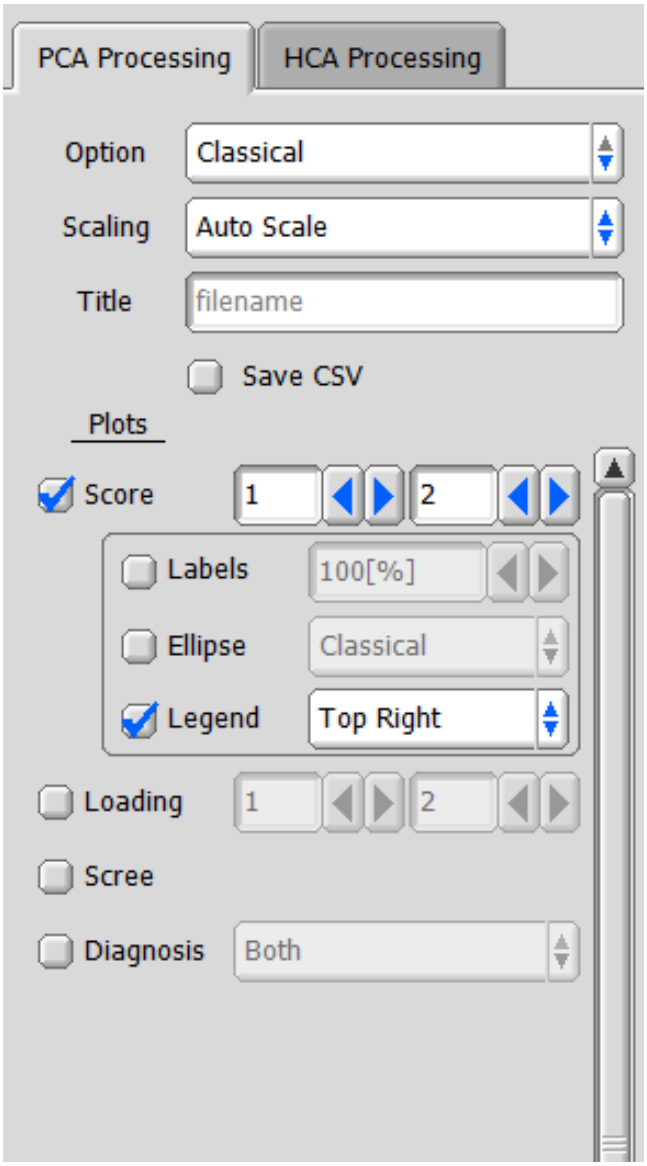


Figure 10. PCA Analysis Selections.

Option:

Classical: Uses all data provided to compute scores and loadings.

Scaling options: *No scaling, autoscaling, Pareto scaling.*

Robust: Focuses on the core of data, samples may be down weighted.

Scaling options: *No scaling, median absolute deviation.*

Scaling: Gives each peak the same weight in the PCA analysis based on method.

No scaling: Without scaling, larger peaks contribute more strongly.

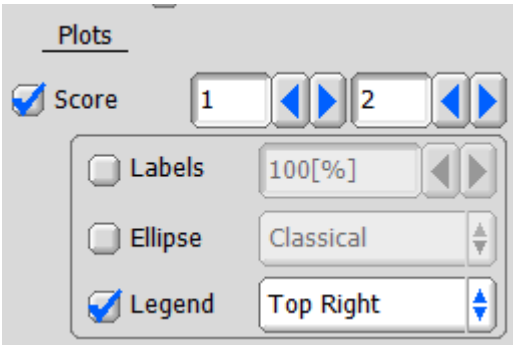
Autoscale: Each peak contributes equally to analysis, including "noise". Most common method but can "blow-up" smaller peaks. Scales by standard deviation.

Pareto scaling: A compromise between the two above, scales by the square root of standard deviation.

Median absolute deviation: Downweighs more extreme peaks, recommended if outliers are present (see Ellipses in Plots, Scree, or Diagnosis for more information on outliers).

Title and save CSV: the choice to save the PCA analysis as a csv file and the name of the file.

a. Plots

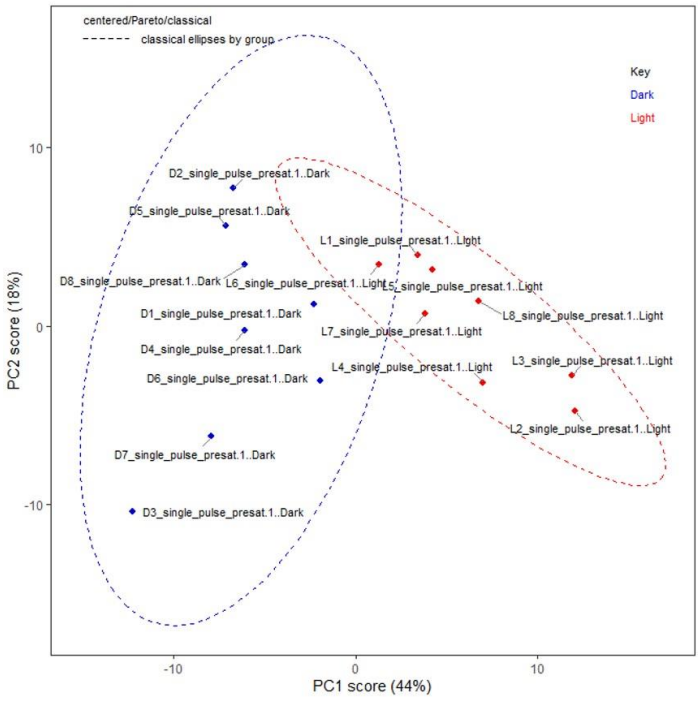


Scores: Generates a score plot and determines which principal components are shown (within graphs score relates to the coordinates of the principal component).

Labels: Labels the data points with the file name and group, with a choice of opacity level.

Figure 11. Plot Options.

Ellipse: Visual tool used to summarize the distribution and grouping of data points, data points outside the ellipse are considered outliers. An ellipse will be drawn for each specified group.



Note: The use of classical and robust is not related to the PCA algorithm, but it does follow the same calculations applied to the 2D array.

Not checked: No ellipses.

Classical: Classically computed confidence ellipses.

Robust: Robustly computed ellipses.

Both: Both classic and robust modes, used to compare the two methods.

Figure 12. Ellipses Example.

b. Loading

Generates a loading plot, a unit vector corresponding to projection coefficients which tell how each variable (frequencies in spectral data) affects the scores. You can choose which principal components to view, and which one will serve as the reference.

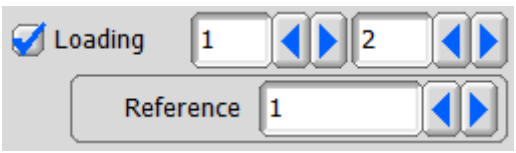


Figure 13. Loading Selections.

In this example PC1 (principal component 1) is more affected by the carbonyl peaks while PC2 is driven by many peaks with some opposing peaks in the hydrocarbon region.

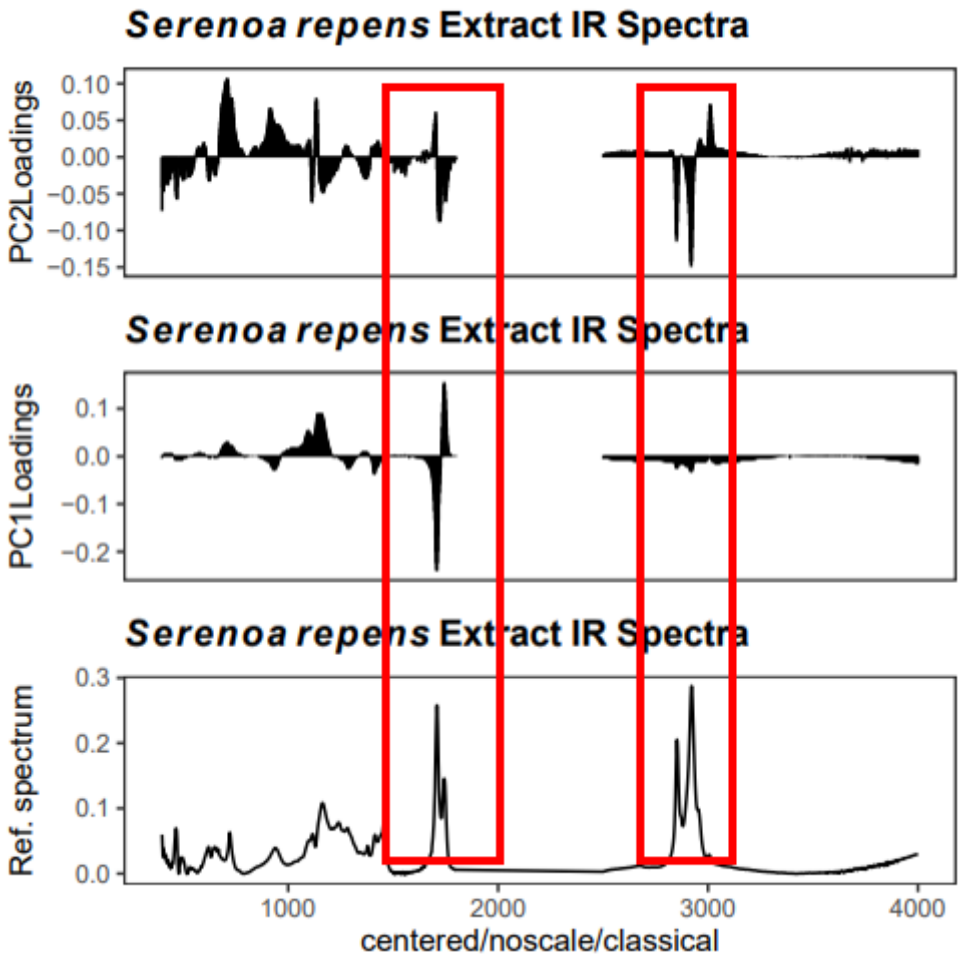


Figure 14. Loading Plot Example.

c. Scree

A line plot of eigenvalues of the principal components, used to determine the number of factors to retain to describe the data.

Traditional: Traditional line style of scree plot, interpreted by finding the "elbow" or where the slope transitions to a flat tail. In this example 2 PCs would be sufficient.

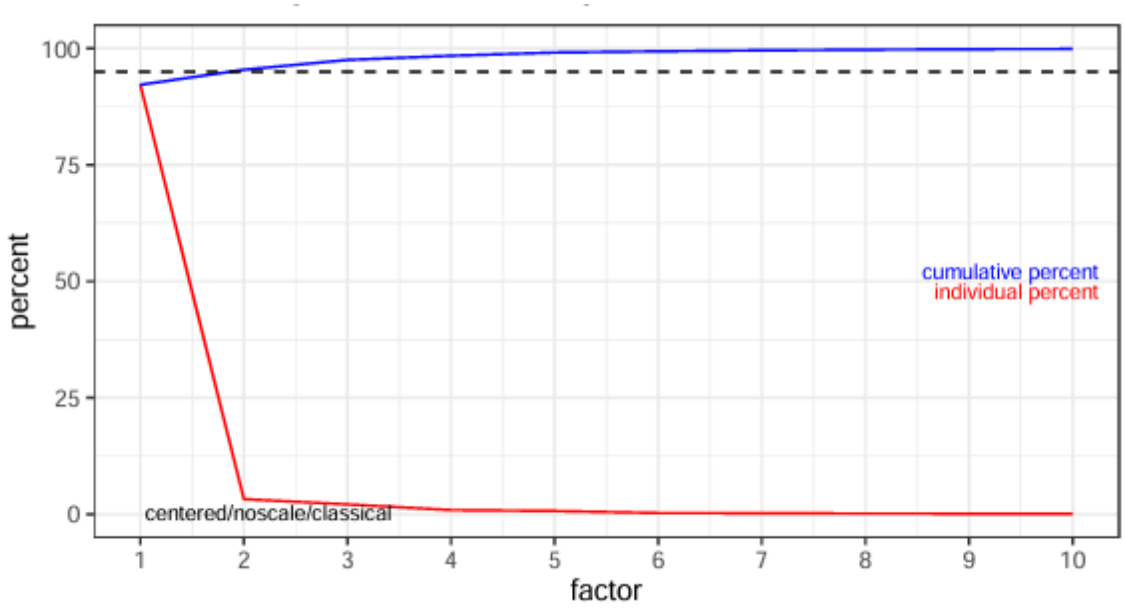


Figure 15. Traditional Scree Plot Example.

Alternative: Shows data points related to each component and their score, as well as the component variance percentage. In this example 2 PCs would be sufficient, however one might choose to go up to 5 to retain more variance.

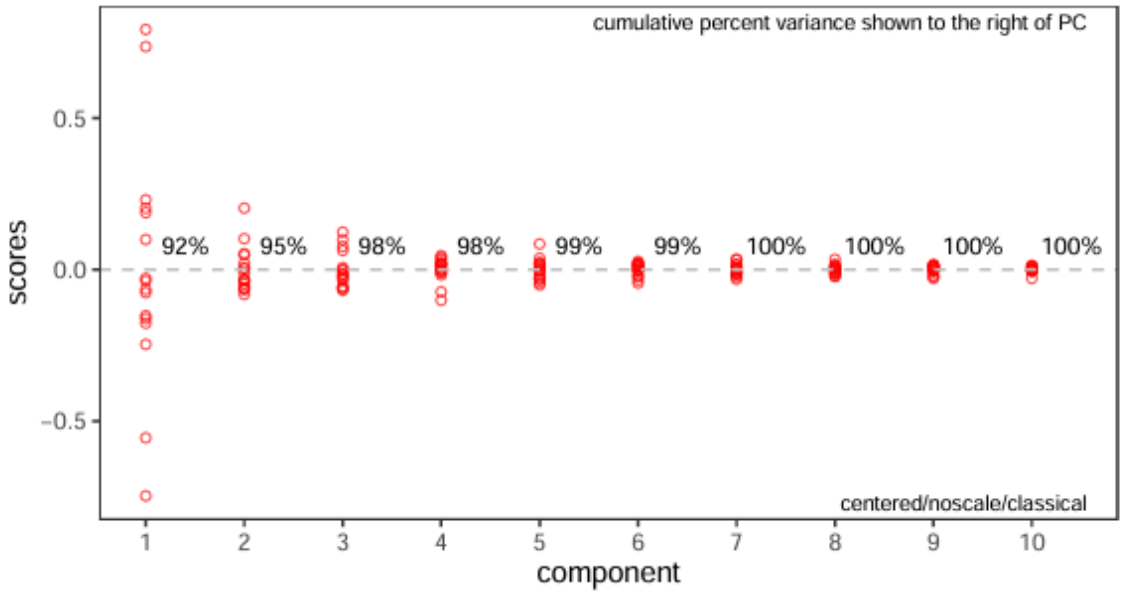


Figure 16. Alternative Scree Plot Example.

d. Diagnosis

Produces a diagnosis plot to determine outliers and allows for creation of this information into a csv file.

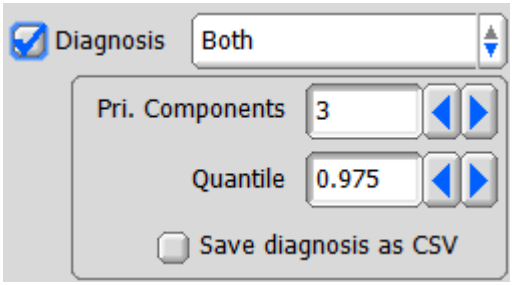


Figure 17. Diagnosis Options.

Orthogonal: Determines the distance of objects orthogonal from the principal component space (residual error norm) of each data set. Used to find orthogonal outliers.

Score: Determines the distance of each dataset in the principal component space and measures how outlying objects are. Used to find score based outliers (bad leverage outlier).

Both: Shows both orthogonal and score based diagnosis.

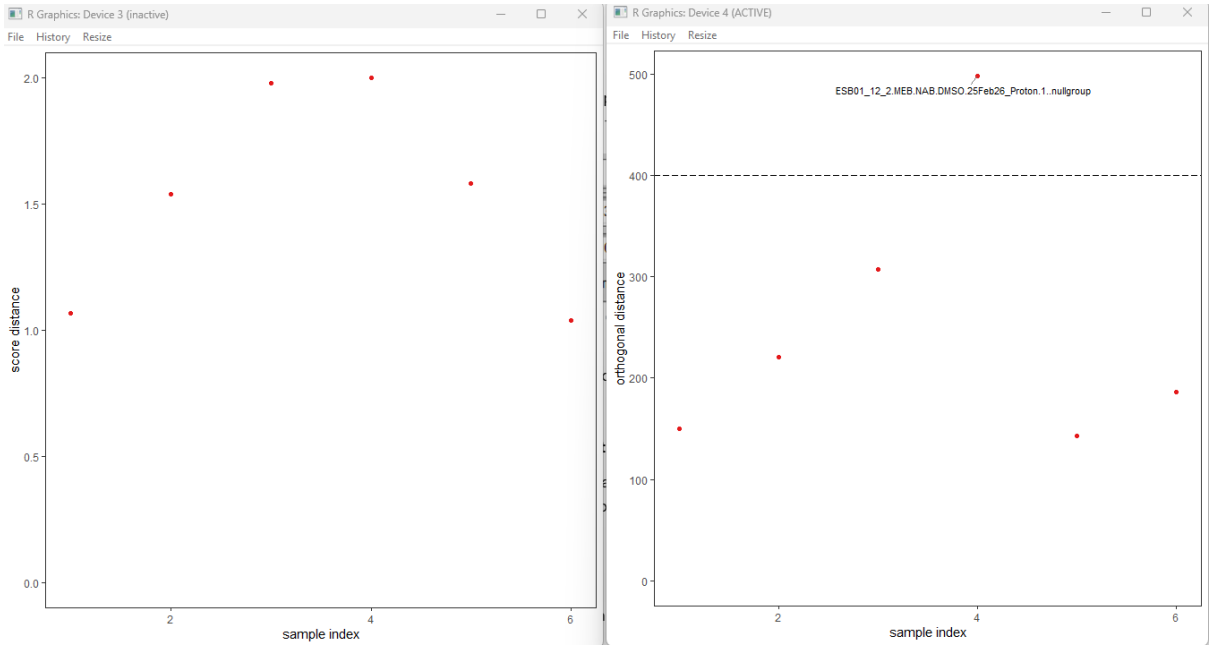


Figure 18. Both Diagnosis Plot Examples. In this example there are no score outliers, but there is one orthogonal outlier. The quantile cutoff is shown as a horizontal line.

Principal components: The number of principal components to include.

Quantile: The significance criteria to use as a cutoff. Distinguishes between outliers and regular observations, recommended to be 0.975.

e. Graph Options

Selection of how the analysis is presented. The choices are screen (as a pop up), PDF, SVG, and PNG.

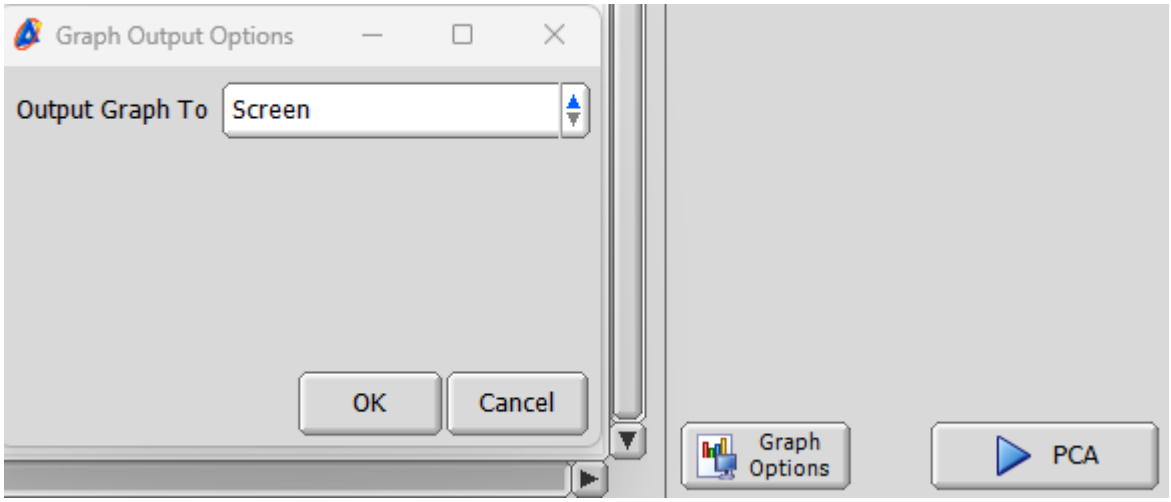


Figure 19. Graph Options and Start PCA Button

Once all the desired options are selected, the start PCA button can be clicked, this will output all the desired information in the form selected.

5. HCA Processing

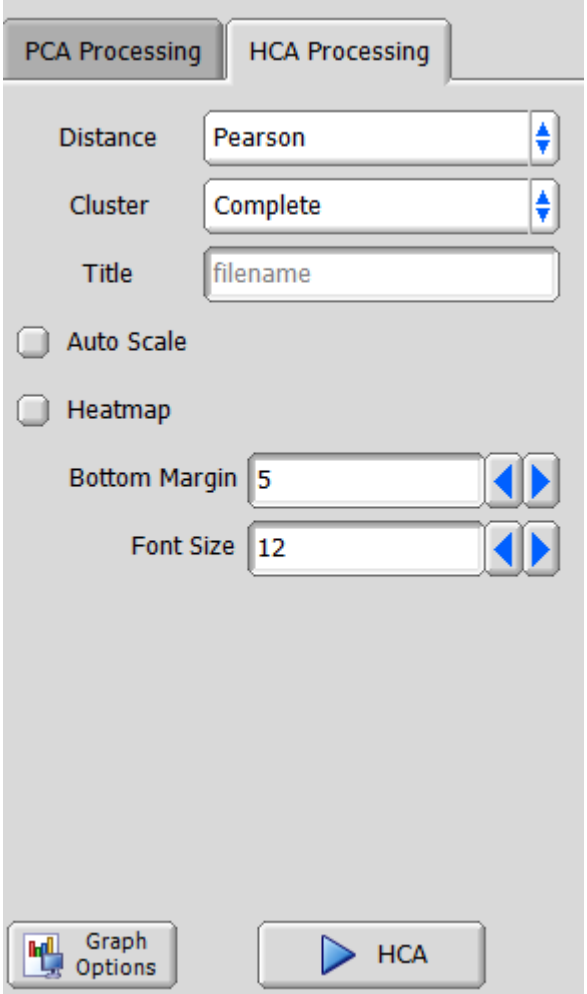


Figure 20. HCA Selections.

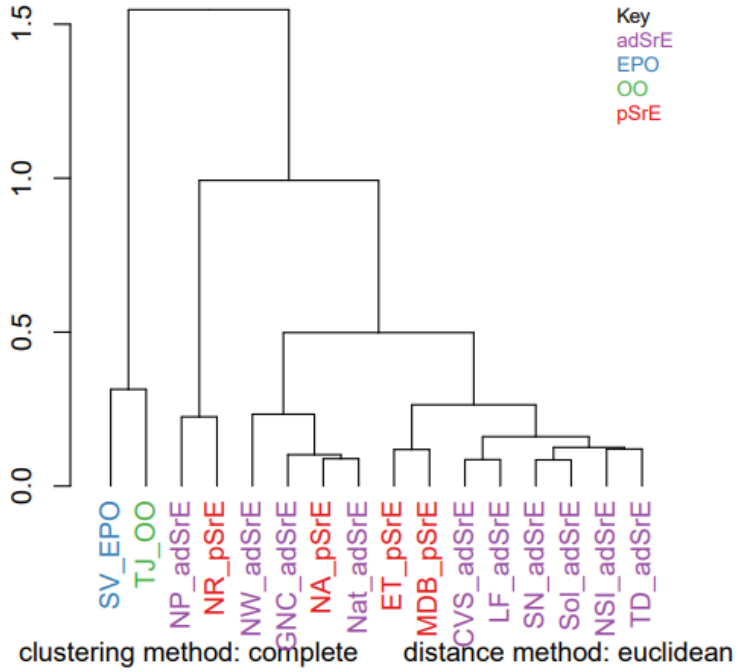


Figure 21. Example Dendrogram.

"Hierarchical Cluster Analysis" is a clustering method where "distances" between samples are calculated and presented in a dendrogram. The vertical scale represents the distance between samples.

a. Distance

Distance between objects in space is inverse to the similarity between objects (small distances equate to more similarity). Distance calculation method can impact if a sample is considered similar to another sample.

Binary: Calculated with binary data, expressed as a percentage of the number of dimensions. Useful for when looking for presence/absence of substructures.

Canberra: Sum of values obtained by normalizing the absolute difference between each dimension by the absolute value of their sum (weighted L1 norm). Useful for detecting trace metabolites, very sensitive to small differences.

Correlation: Indicator converted so the Pearson product-moment correlation coefficient treated the same as distance. Useful for uncovering complex interactions.

Cosine: Indicator converted so cosine similarity is treated the same as distance. Independent of variable scaling. Useful for NMR library development and metabolite identification, able to group based on compound ratios.

Euclidean: The square root of the sum of the squares of the differences between dimensions. (L2 norm). Default method, useful for comparing concentration changes of metabolites across samples.

Kendall: Indicator converted so Kendall's rank correlation coefficient treated the same as distance. More robust than Pearson. Similar to Spearman but does better with smaller sample sizes.

Manhattan: Sum of absolute values of the differences between dimensions (L1 norm). Useful when outliers or artifacts present.

Minkowski: Combination of Euclidean and Manhattan. Useful for algorithm development.

Maximum: Maximum value of the difference between dimensions (infinite norm). Useful for grouping based on single, dominant changes.

Pearson: Indicator converted so uncentered correlation coefficient treated the same as distance. Focuses on shape and pattern of spectra, useful for grouping metabolic trends.

Spearman: Indicator converted so Spearman's rank correlation coefficient treated the same as distance. More robust than Pearson. Useful for non-linear metabolic pathways or uncalibrated data.

b. Cluster

Starts with each object assigned its own cluster, each stage then joins the most similar clusters until there is just one cluster left. At each stage is distance between clusters (the similarity) is calculated based on the selected method.

Average: Group average method aka Unweighted Pair-Group Method using Arithmetic Mean (UPGMA), average distance between all samples belonging to each cluster is taken as the distance between clusters. Default method, useful for baseline sample classification.

Centroid: Centroid method aka Unweighted Pair-Group Method with Centroid (UPGMC), distance between clusters derived from the distance between the centers between clusters. Useful for comparing average metabolic "fingerprints" instead of individual samples.

Complete: Complete linkage method aka furthest neighbour method, the longest of distances between samples belonging to each cluster is taken as the distance between clusters. Useful for finding tight, well separated clusters and group separation i.e. control and treated samples.

Mcquitty: Mcquitty method aka Weighted Pair-Group Method using Arithmetic Method (WPGMA) or weighted average method – the weighted average of the distance between all samples belonging to each cluster is taken as the distance between clusters. Useful for finding INDIVDUAL similarities when sampling is imbalanced i.e. 10 of group A and 100 of group B.

Median: Median method aka Weighted Pair-Group Method with Centroid (WPGMC), distance between clusters is derived from the distance from the midpoint of the center of the cluster. Useful for finding GROUP similarities when sampling is imbalanced i.e. 10 of group A and 100 of group B.

Single: Single linkage method aka nearest neighbour method, the shortest of distances between samples belonging to each cluster is taken as the distance between clusters. Adopts a "friends of friends" clustering strategy, can result in "chaining" as inhomogeneous clusters can be linked and is sensitive to noise. Useful for identify outliers.

Ward.D: Simpler version of Ward.D2 that does not square dissimilarities. Aims to find compact, spherical clusters. Useful for profiling.

Ward.D2: Ward's method aka minimum variance method, merges clusters to minimize the increase in sum of squares in cluster – can only be used when Euclidian distance is selected. Most common for NMR - creates spherical, distinct clusters making it useful for finding specific metabolic phenotypes.

c. Auto Scale

Determines if the dendrogram is scaled and centred.

d. Heatmap

Generates a heat map, visualizing the distribution within clusters. When not selected the HCA is performed sample-wise (clustering based on similarity of samples).

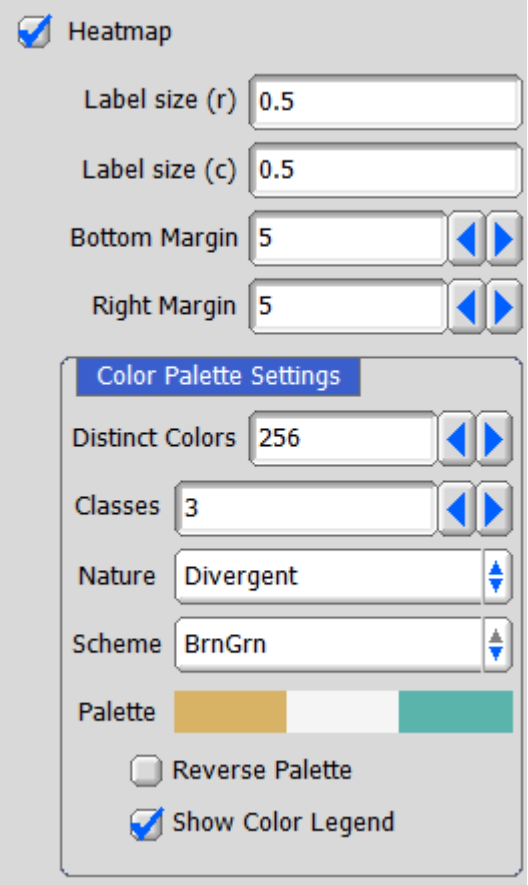


Figure 22. Heatmap Selections.

Label Size: r – right side label, c – bottom center label size.

Margins: Space from the bottom and right respectively.

Colour Palette Settings -

Distinct Colours: The number of distinct colours in the heat map.

Classes: The number of classes allowed, similar to distinct colours but based on sample groups.

Nature:

Sequential: Colours will progress from light, low-intensity to dark, high-intensity.

Divergent: Opposite ends have two contrasting colours and meet in the middle as a light, neutral colour. Good for representing midpoints.

Qualitative: Series of unrelated, distinct colours with no inherent ordering.

Scheme: A choice of several palettes, see next page.

Reverse Palette: Swaps the order of colours.

Show Colour Legend: Show the legend on the dendrogram.

Font Size: Size of the text in the dendrogram

e. Graph Options

Selection of how the analysis is presented. The choices are screen (as a pop up), PDF, SVG, and PNG. Once all the desired options are selected, the start HCA button can be clicked, this will output all the desired information in the form selected.

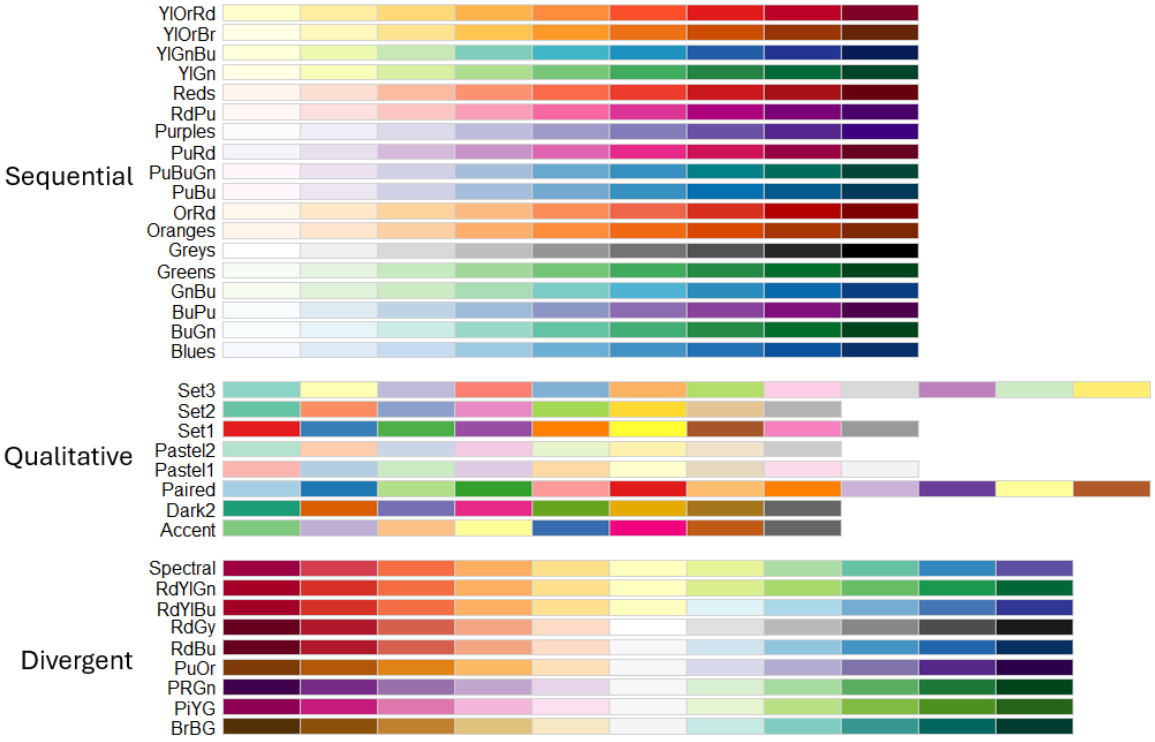


Figure 23. Heatmap Colour Scheme Options.