Biogenity

#Demo / June 2022

Statistical Data Analysis

Cell study

Your Global Partner in Omics

A short introduction

This report contains an overview of the analysis made on your data. We have based the analysis on our state of the art analytic pipeline designed to give you the best foundation for understanding your data.

Reading guide

A few tips to ensure the best understanding of the analysis:

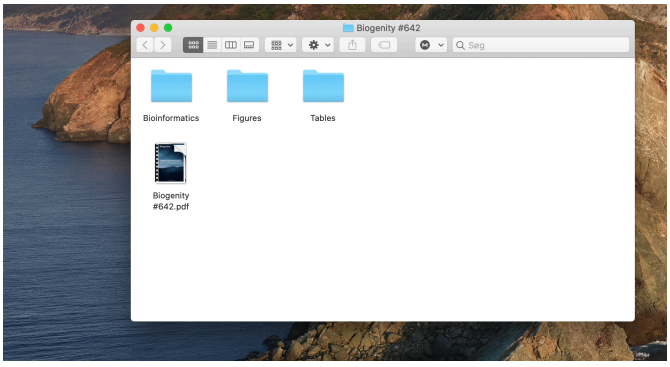
1. The report functions as a walk-through, where every page shows a result of the analysis or pre-analysis.

2. Results are displayed by either a table or a figure followed by a thorough description.

3. You can always examine the original file. For each figure or table shown in the analysis, we have integrated the 'Pathfinder.' It lets you know where to find the shown object in the compressed file received. Example of such is displayed below:

"Figures/Mean log2 (Expression) across samples histogram.pdf"

Explained: The "Mean log2 (Expression) across samples histogram.pdf" can be located in the figures-folder provided in the compressed file.



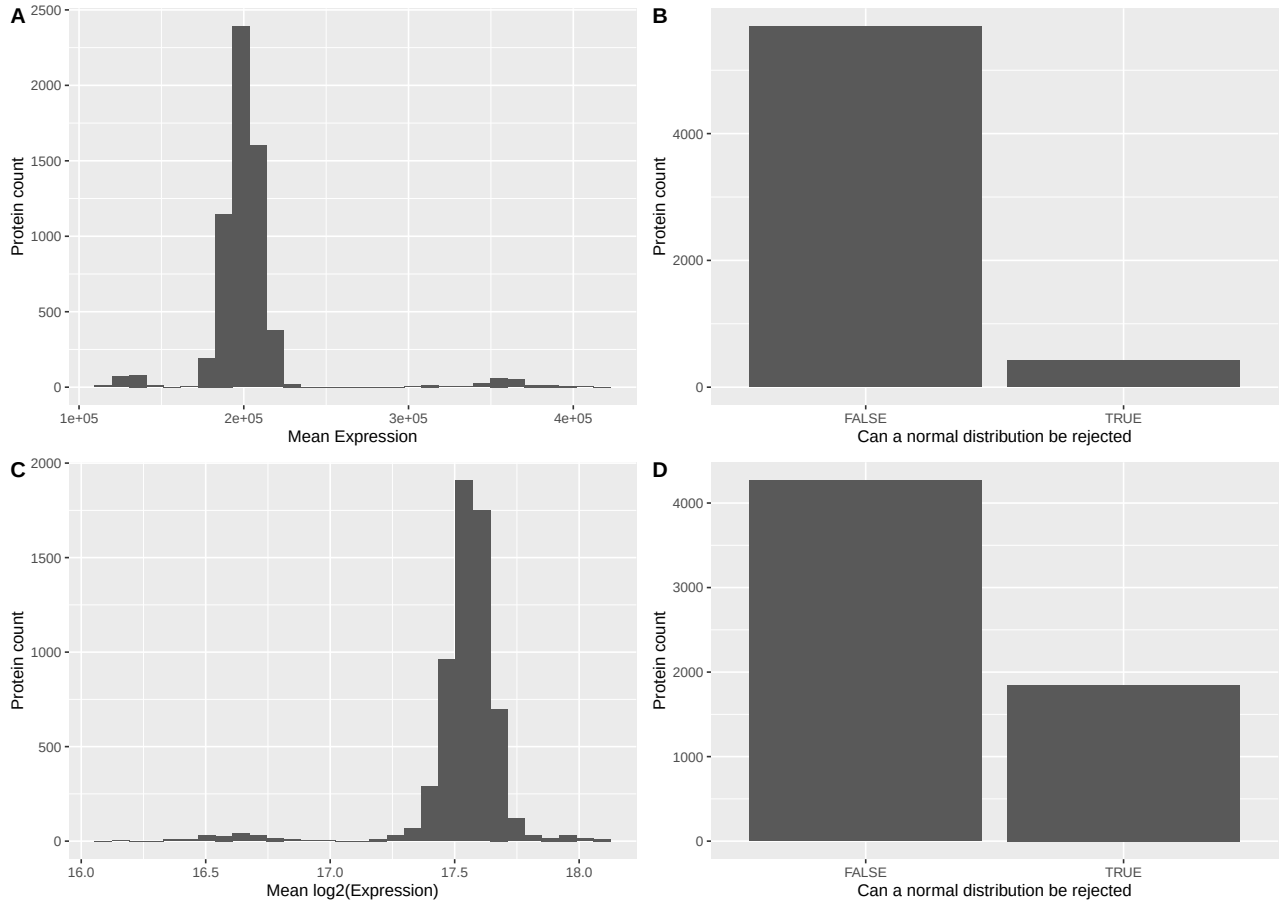
Example of "compressed file"

Thank you for choosing us
We appreciate the confidence you have placed in us, and we look forward to providing you with the best possible service in the future.

Best regards
The Biogenity team

Table of contents

Pre-analysis



You can find the figure at:
Cell_study analysis/Figures/Transformation overview.pdf

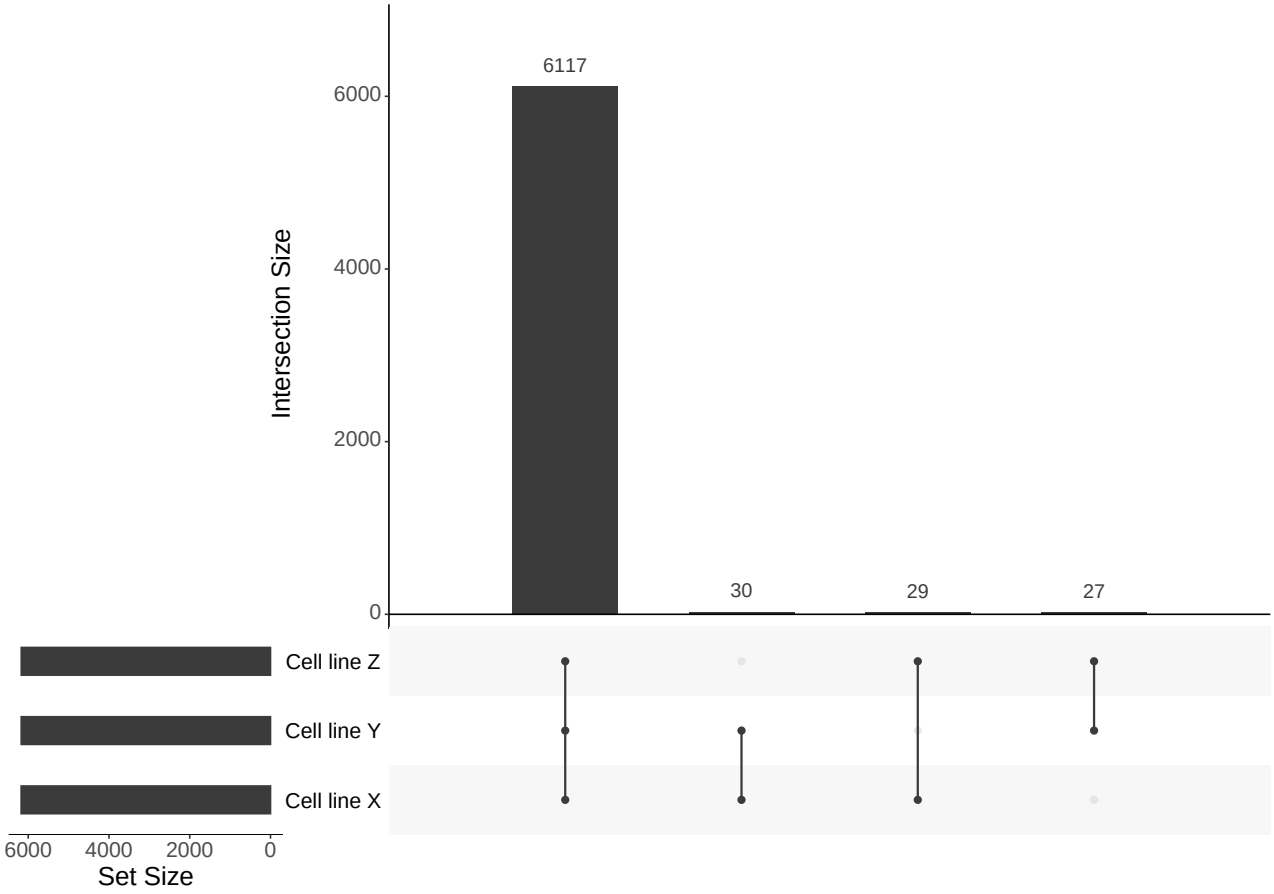
Data distribution and Transformation

The data was filtered, so only proteins with at least 2 unique peptide were included. Log2 transformation is a commonly used method in proteomics to increase the number of expressions following a normal distribution, which is preferred. Normal distribution is often a requirement for some tools in data science used to analyze proteomic data. To test if transformation would increase the number of expressions that followed a normal distribution, a Shapiro Wilk normality test was performed for each expression before and after transformation.

Figures A and C are histograms of the mean expression of the different proteins in the original data and log2 transformed data, respectively. Figures B and D are bar charts showing the number of expressions for which a normal distribution could be rejected using the Shapiro Wilk normality test. As depicted in the figure above, the normality test showed that most variables followed a normal distribution. The normal distribution could be rejected for 425 proteins before log2 transformation and 1848 after log2 transformation. After log2 transformation, a higher number of protein expressions were rejected from following a normal distribution. Thus, data was not log2 transformed before further processing.

The individual plots and this figure (combined plots) can be found in the *Figures* folder.

Pre-analysis



You can find the figure at:
Cell_study analysis/Figures/UpSet diagram.pdf

Unique Identifications

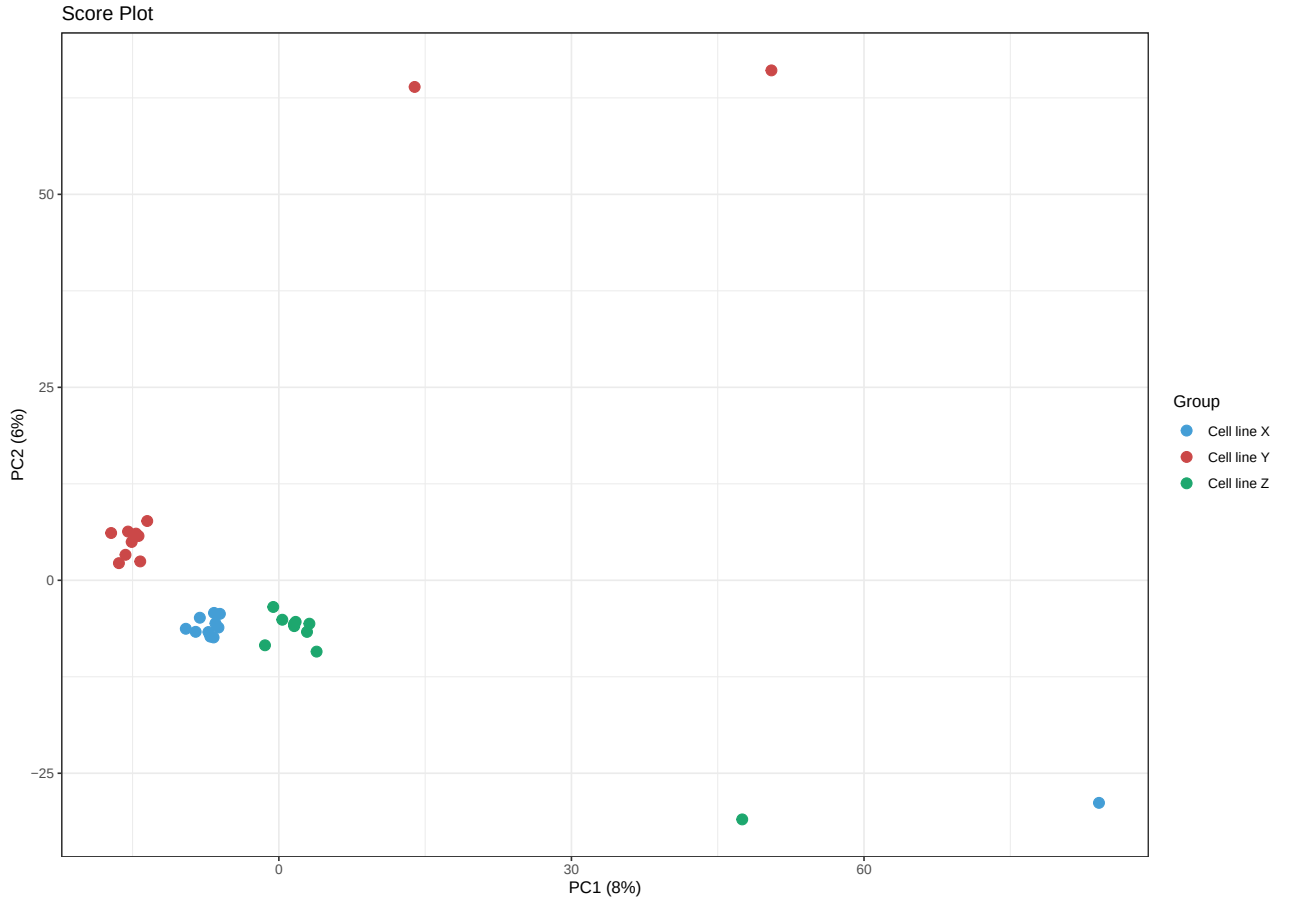
The identification of unique proteins within a single group can potentially indicate actual differences between the groups. Owing to the nature of mass spectrometers, proteins can be missing due to chance. Thus, the evaluation of unique identifications should be based on their abundance, their biological function, and the missingness in the dataset as a whole. The identifications should be treated as potential targets, that are identified in a less stringent manner than the targets that are statistically significant.

The unique identifications within groups were determined as uniquely identified when the proteins were identified in at least half of the samples within a group, and completely missing in another group.

The findings are illustrated in the plot above. This plot is known as an Upset plot which displays the various overlaps between different groups (The lower dots and lines), and how many proteins are in the different combinations of overlaps (The upper bar chart). It is an alternative to Venn diagrams as it provides an improved overview when comparing multiple groups. In this analysis 6203 identifications were found and 86 were identified as being unique.

Venn diagrams (color and greyscale) and Upset diagrams can be found in the *Figures* folder.

Pre-analysis



You can find the figure at:
Cell_study analysis/Figures/PCA of data with Group annotation.pdf

Data visualization with Principal Component Analysis (PCA)

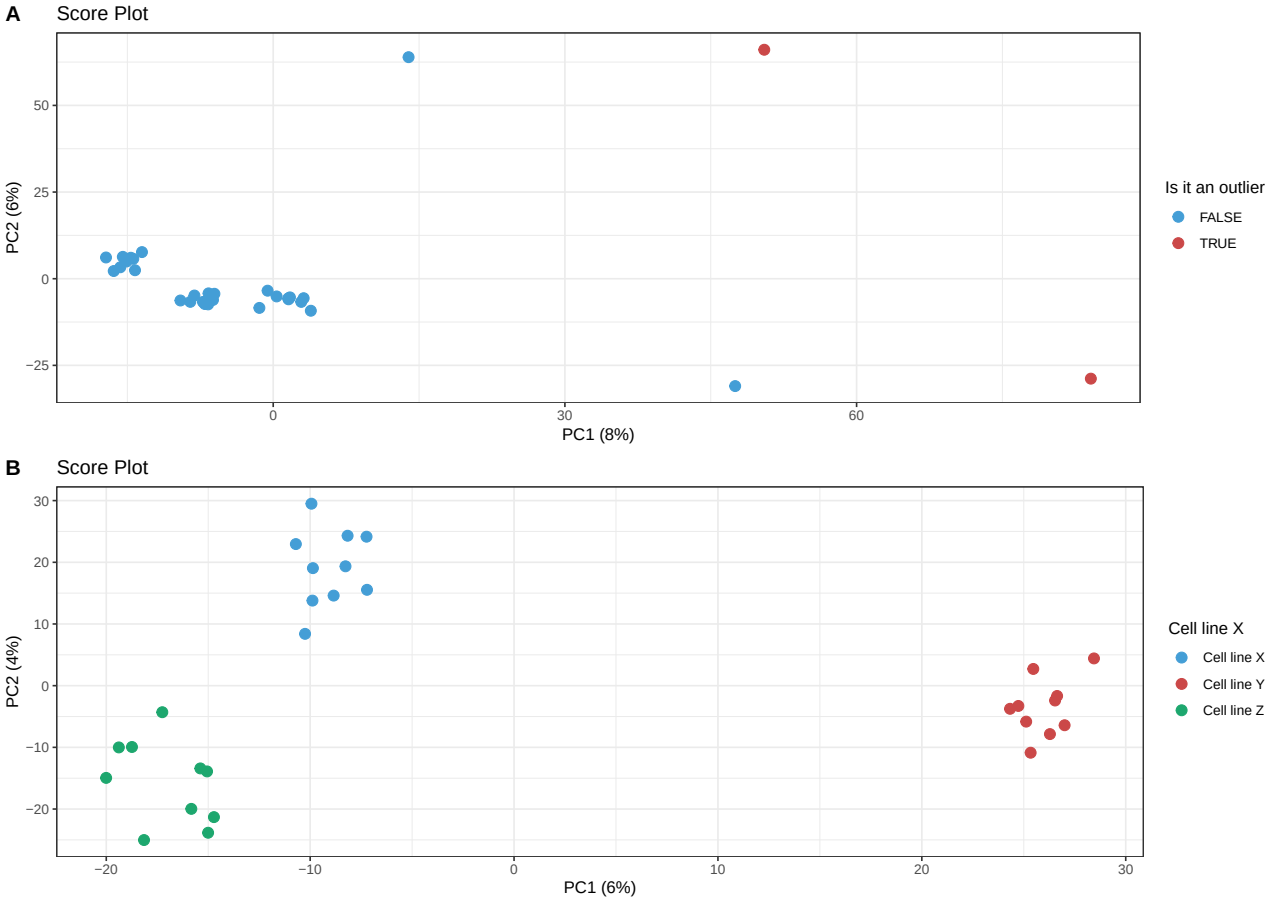
Principal Component Analysis is an excellent approach to improve the interpretability of large datasets. This method reduces the dimensionality of such large datasets with minimal loss of information by converting possibly correlated variables into a set of linearly uncorrelated variables that consecutively maximize variance.

The number of Principal Components (PC) is the same as the number of variables in our dataset. The PCs are calculated in such a way that PC1 will represent the largest possible variance in the dataset. PC2 is calculated as accounting for the next largest variance while uncorrelated with PC1 and so on for all the PCs. This analysis is useful in the visualization of strong patterns among subjects and potential outliers. PC1 and PC2 are plotted in the figure above. The respective percentage of variability for PC1 and PC2 are displayed in brackets.

This PCA plot was also constructed in an interactive form. Both plots can be found in the *Figures* folder.

Outlier identification was performed under the assumption that outlier presence would affect the total sum of expression intensities in each sample. Using a convenience function, potential outliers were detected based on the median absolute deviation. Thus, samples that presented values greater than 3 median absolute deviations away from the median were considered outliers. 2 outliers were found.

Pre-analysis

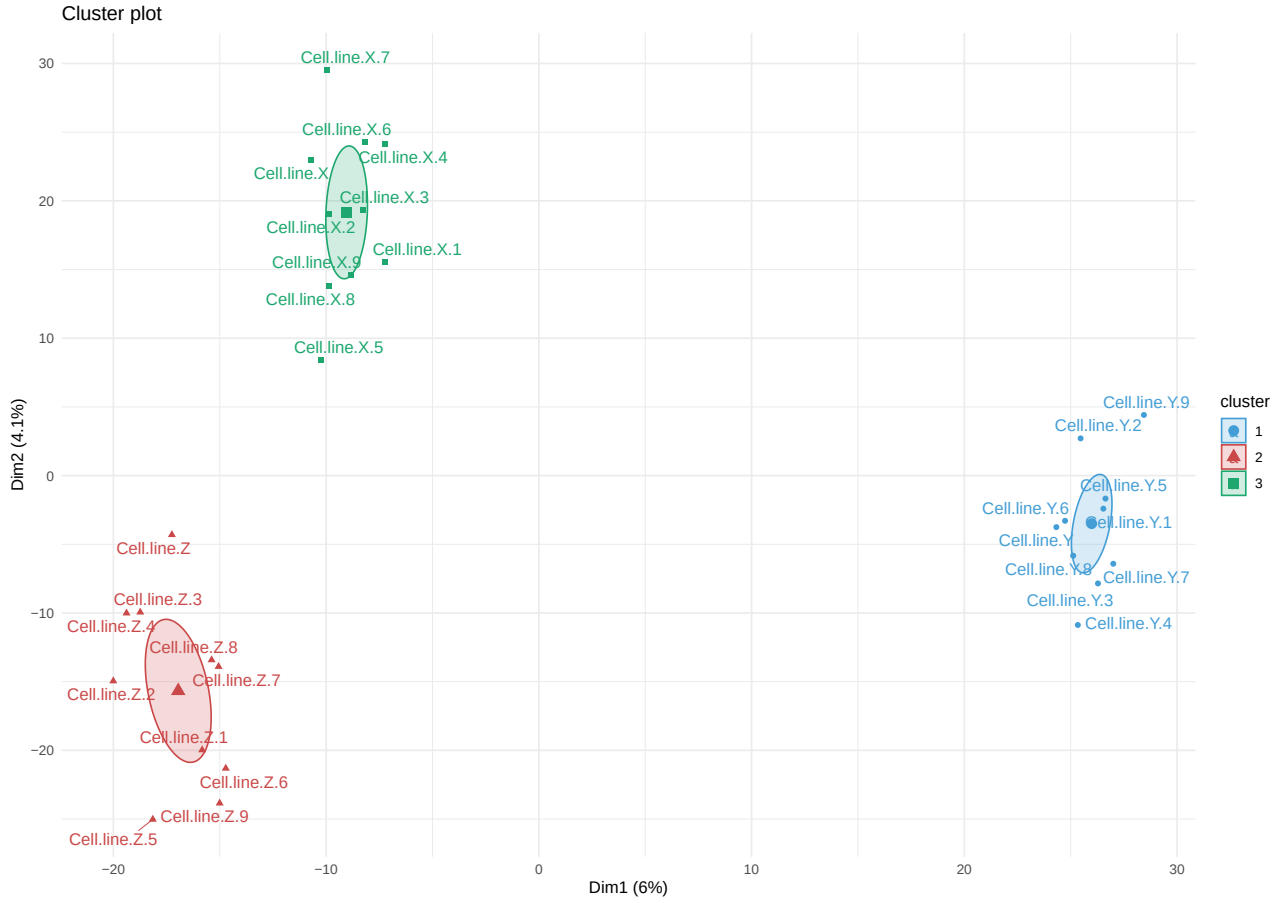


You can find the figure at:
Cell_study analysis/Figures/Outlier overview.pdf

Outlier analysis and filtration

All outliers were identified and plotted (Figure A). The identified outliers were filtered before a new PCA analysis was performed (Figure B). The plot with the outlier overview can be found in the *Figures* folder.

Cluster analysis



You can find the figure at:
Cell_study analysis/Figures/Kmeans clusters of data with Group annotation.pdf

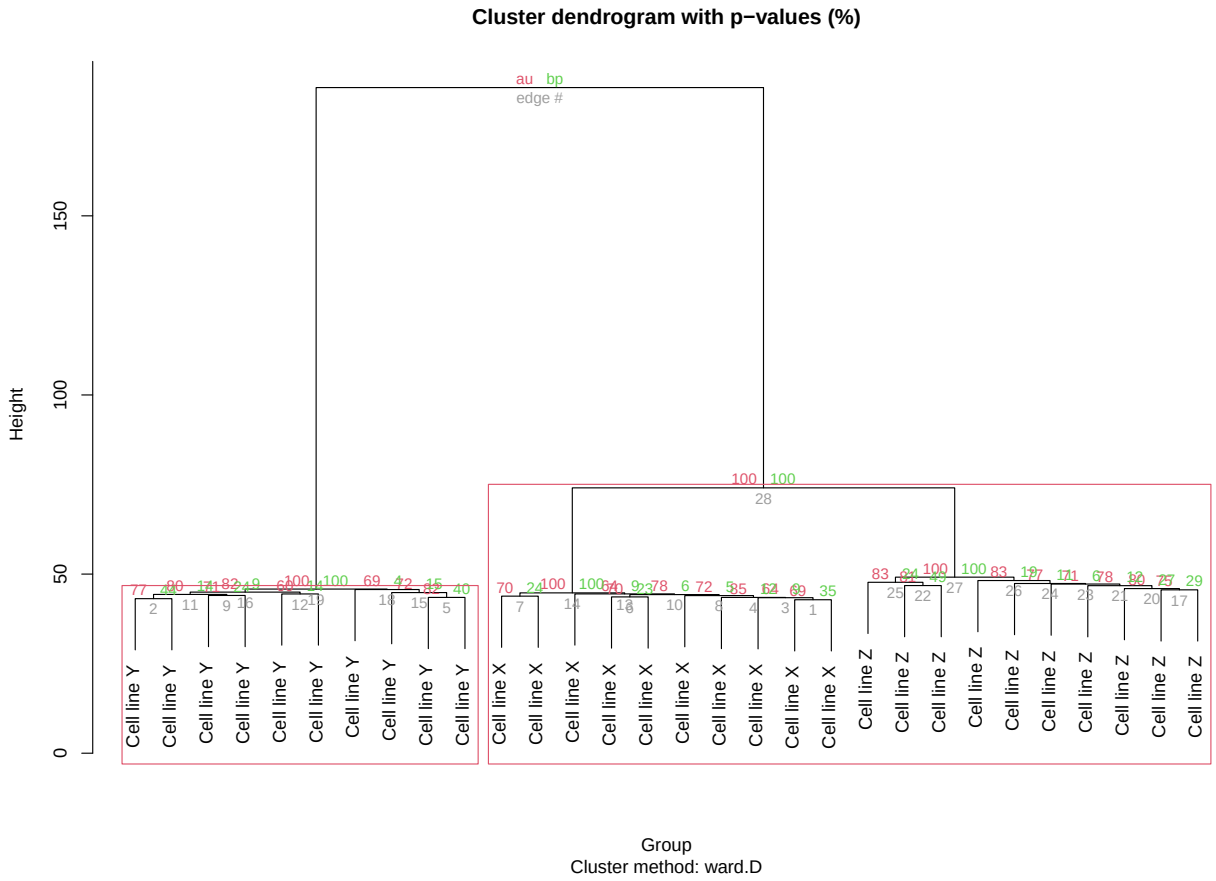
Partitioning Around Medoid - K-means clustering

K-means clustering is a method frequently used in clustering analysis. The results of this clustering analysis are presented in the figure above, including the Group annotations. This clustering method takes a divisive (top-down) approach to generate clusters. Briefly, all points start off belonging to a single cluster, and subsequently, clusters are merged by looking at the distances between each observation and fixed points.

Cluster analysis was performed using Partitioning Around Medoids (PAM), which is a more robust K-means clustering version as it is less sensitive to outliers. PAM is an unsupervised machine learning algorithm that searches for k representative points (medoids) along the data set and constructs the clusters by designating each observation in the dataset to the nearest medoid. The K-means method does not determine the number of clusters present on the dataset. The optimal number of clusters for each dataset can be determined by performing a silhouette analysis. To measure the distances between samples, the Euclidean method was used. Silhouette analysis (search for optimal number of clusters) was performed for 1-10 clusters in order to determine the best number of clusters to describe the data. The ellipses illustrate the samples inside a 95% confidence interval for each cluster, if any.

K-means clusters (both with Group and Sample annotations) can be found in the *Figures* folder.

Cluster analysis



You can find the figure at:
Cell_study analysis/Figures/Ward clusters of data with Group annotation.pdf

Ward clustering

A complementary clustering method is hierarchical clustering, such as the Ward method illustrated as a dendrogram above. This method is known as a bottom-up method which initially considers all datapoints to be a cluster of their own. From there, it merges the points into clusters by the rule that it minimizes the within cluster variance as measured by the Euclidian distance. The length of the vertical branches indicates how similar the clusters are. Thus, clusters that have a short vertical line before merging with another cluster exhibit a higher degree of similarity compared to clusters that have a longer line prior to merging.

The plot contains two p-values: the red Approximately Unbiased (AU) and the green Bootstrap Probability (BP) values. Essentially, these two values are based on sampling the data in different ways 1000 times. Thus, a value of 0.31 means that a specific cluster appeared in 310 of the 1000 samplings. The AU utilizes a different sampling and scaling version of BP which is more unbiased than the BP. Clusters with an AU equal to or higher than 0.95 are marked by a red square and are considered highly robust.

Ward clusters were built both with Group and Sample annotations and can be found in the *Figures* folder. Interactive Heat Maps were also constructed to graphically represent the data. These can be found in the *Figures* folder as well.

Statistical analysis

Comparisons	p-value $\leq$ 0.05	Adj. p-value $\leq$ 0.05	Adj. p-value $\leq$ 0.05 + 30% regulation	|Log2 fold change| $\geq$1
Cell line Y vs Cell line X	500	283	283	227
Cell line Z vs Cell line X	517	233	231	103
Cell line Z vs Cell line Y	390	209	207	178

You can find the figure at:
Cell_study analysis/Tables/Statistics.txt

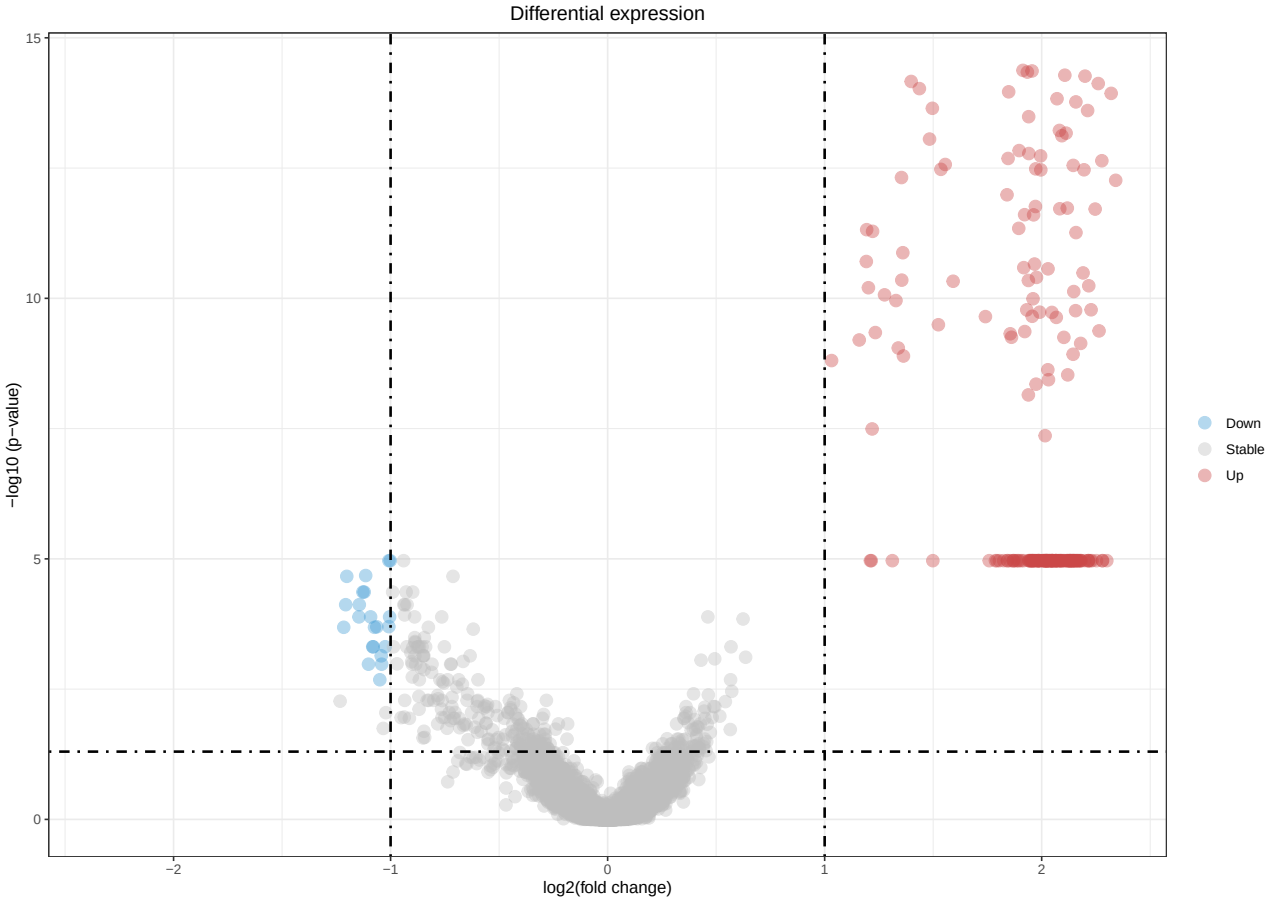
Overview of regulations

To ensure high stringency, the proteins were filtered prior to statistical testing. Only proteins that were identified in at least half of the samples of a single sample group were included in the analysis. Each protein expression was tested for normal distribution by the Shapiro-Wilk test.

If the Shapiro-Wilk test returned a p-value below 0.05, which means that the residuals are normally distributed, then a parametric test such as Analysis of Variance (ANOVA) test, followed by a Tukey's post hoc test can be applied. Conversely, if the residuals were not normally distributed (p-value ≤ 0.05), then a non-parametric test such as the Wilcoxon-Mann-Whitney test (referred to as Wilcox test in the report) was applied. In the statistical result spreadsheets, provided in the *Tables* folder, the recommended test to use according to the p-values obtained in the Shapiro-Wilk test is noted in the *"Recommended statistical test"* column.

In total, 18351 comparisons were made, 1407 of which had a p-value below or equal to 0.05. Due to the large number of comparisons, correction for multiple comparisons was performed using the Benjamini-Hochberg procedure. 725 proteins retained an adjusted p-value equal to or below 0.05, 721 of which were regulated by more than 30%, and 508 had an absolute log2 fold change higher or equal to 1. The regulations are distributed across 3 group comparisons (Cell line Y vs Cell line X, Cell line Z vs Cell line X, Cell line Z vs Cell line Y).

Statistical analysis



You can find the figure at:
Cell_study analysis/Figures/Volcano plot/Volcano plot of Cell line Y vs Cell line X statistical comparison.pdf

Volcano plot

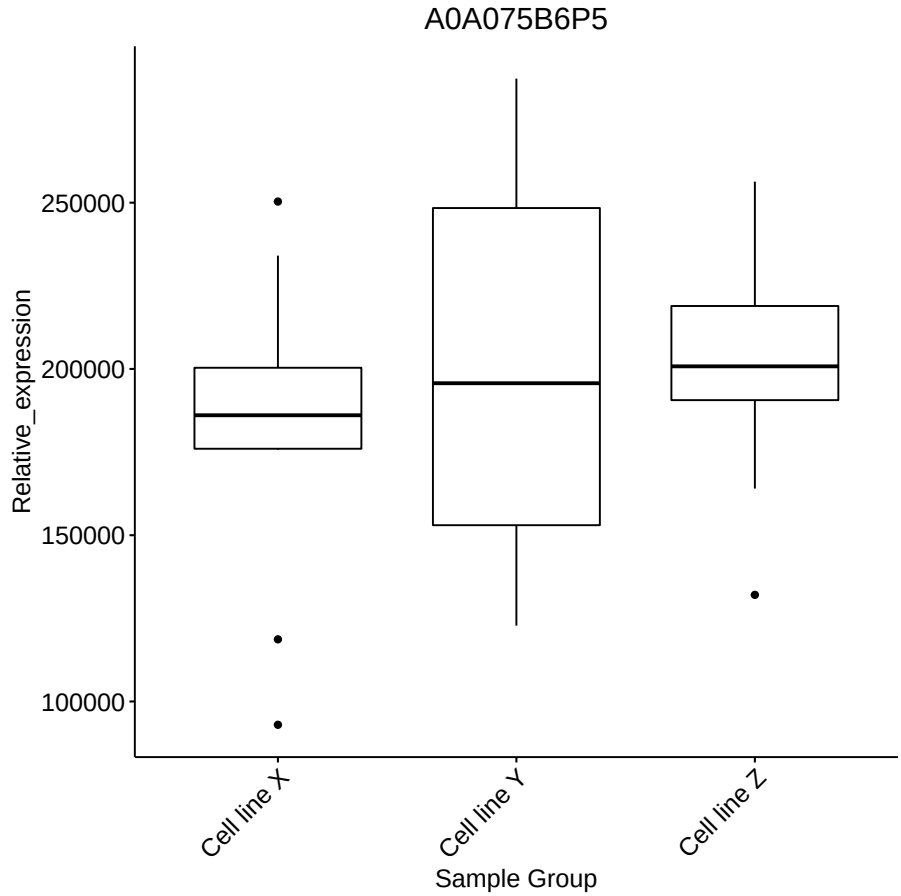
For each comparison, a Volcano plot has been made like the one illustrated above for Cell line Y vs Cell line X. P-values were adjusted for multiple comparisons using the Benjamini-Hochberg procedure.

In the Volcano plot, the dashed vertical lines represent the threshold for an absolute log2 fold change of one. The horizontal dashed line depicts where the unadjusted p-value is equal to 0.05. All the expressions that do not register an absolute log2 fold change equal or greater than one and an adjusted p-value < 0.05 are represented by the grey dots. The blue dots represent the expressions with an adjusted p-value < 0.05 and a log2 fold change < -1, if any. Likewise, red dots represent the expressions with an adjusted p-value < 0.05 and a log2 fold change > 1.

In the plot above 205 proteins were significantly upregulated and 22 proteins were significantly downregulated.

An interactive version of the volcano plot was also constructed. This figure and the interactive version can be found in the *Figures/Volcano plot* folder.

Statistical analysis



You can find the figure at:
Cell_study analysis/Figures/Box-Plots/Box plot of A0A075B6P5.pdf

Box plot

The relative protein expression for each identified protein across sample groups was plotted in Box-plots. The constructed box plots for all relative protein expressions can be found in the folder *Figures/Box-Plots*.

The figure above, a box-plot of A0A075B6P5 is one example of a box-plot that can be found in the folder. The sample groups are plotted in the x-axis, and the relative protein expression is plotted on the y-axis. The area inside the box represents the percentile Interquartile range (IQR), and the bold line inside the box corresponds to the median. The whiskers above and below the box represent 1.5 times the IQR. Values outside of this range are considered outliers and are represented as circles.

Statistical significance in the differences of relative protein expression between groups is also represented in the plots. "*" stands for p-values <= 0.05; "**" stands for p-values <=0.01, "***" stands for p-values <= 0.001 and "****" stands for p-values <= 0.0001.

Box-plots containing the information about all the performed statistical tests were plotted and can be found in the folder *Figures/Box-Plots*.

Thank you for choosing Biogenity

We appreciate the confidence you have placed in us by choosing Biogenity for your omics study. We are donating 15 trees in your name to the Amazon rainforest to acknowledge and celebrate our collaboration.

Partnering with One Tree Planted

The amazon rainforest is the host for the largest variety of species in the world, but it is threatened by deforestation. Deforestation, in general, contributes to increased CO 2 emissions, and 80,000 hectares of forest disappear from the Earth every day. More than 28,000 animal species will be extinct in 25 years if we continue this development, and the Amazon rainforest will be gone. We need to act now to try to rescue Earth's evolutional catalyzer.

Biogenity believes that we must counteract deforestation by planting new and preserving existing forests, which also counteract some of the arising factors of the climate crisis. We know that we will not solve all climate problems or secure the Amazon rainforest by ourselves. However, we believe that many kind-hearted acts will form a greener future and a hope for the Earth's ecosystem.

This is why we donate trees for reforestation of the Amazon rainforest for every complete work order using our partner One Tree Planted. Every time you receive a report on your data, we will attach a certificate showing how many trees we planted in the Amazon rainforest in honor of your project.

Let's help each other preserve the Earth to create a better future for generations to come.

Kindly,
The Biogenity Team

15 Trees Planted in the Amazon Rainforest

THE PRESENT CERTIFICATE IS AWARDED TO PRIVILEGED AND DISTINGUISHED PARTNERS OF ONE TREE PLANTED WHOSE CONTRIBUTIONS HAVE BEEN AND CONTINUE TO BE, ESSENTIAL TO THE REFORESTATION, CONSERVATION AND PROTECTION OF FORESTS AROUND THE WORLD.

Kenneth Kastaniegaard
PRESENTED BY
CEO of Biogenity

ONE TREE PLANTED

Certificate id: #Phospho DEMO

June 17, 2022
DATE
YOU CHANGED THE WORLD

Biogenity

Your Global Partner in Omics

Biogenity © 2022 | All Rights Reserved