# reverse-zMAP

**reverse-zMAP module is primarily designed for handling MS runs with relatively large sample sizes. The major concern is that, with the increase of the number of samples, the odds that outlier measurements are involved for each specific protein increase, giving rise to excessively large variances. Consequently, in the sliding-window process of the zMAP module, the proportion of proteins that are suitable to use for the quantile regression in each window can be very small. Besides, fitting a single MVC for a large number of samples may not be flexible enough to allow for the variation of mean-variance trend across samples. In practice, large-scale proteomic studies have frequently applied the strategy of adding a biologically identical reference sample to each individual MS run. For example, in cancer studies, a mixed sample is typically generated by pooling tumor samples and/or normal adjacent tissues (NATs) from several related patients in equal protein amounts. The proteome of this mixed sample is then profiled in every MS run separately. reverse-zMAP module alleviates the influence of outliers by repeatedly making pairwise comparisons, which in the meanwhile allows the modeling of sample-specific mean-variance trend, but it requires a biologically identical reference sample in each MS run for a subsequent integration across MS runs.**

|  |  |  |
| --- | --- | --- |
| Name | Flag | Description |
| protein intensity file\* | --intensity\_data | A tab-delimited file containing raw gene-level protein intensity with samples in columns, and gene symbols in rows. Note: 1. The protein intensity matrice does not require normalization. 2. Sample names can only consist of letters, numbers, and underscores. |
| sample information file\* | --sample\_information | Sample information file is a four-column, tab-delimited file with first line identifying the columns. The column names are ‘Sample\_id’ ‘MS\_run’ ‘Sample\_condition’ and ‘internal\_ref’. Note: To avoid code conflicts, ‘Sample\_condition’ consist only of letters, numbers, and underscores. |
| window size\* | --window\_size | Protein number in each sliding window, the default is 400. |
| step size\* | --step\_size | Step size for moving the window, the default is 100 proteins. |
| percent\* | --percent | To avoid the influence of differential proteins, only a certain proportion (50% by default) of the proteins with the middle are used for quantile regression. |
| method\* | --method | Method used for non-linear fitting,’ exponential\_function’ or ‘natural\_cubic\_spline’ |
| email address\* | --outdir | After the operation is completed, the download link for the result files will be sent to the user’s email. |

## Input Files

### 1.protein intensity file (--intensity\_data)

A tab-delimited file containing raw gene-level protein intensity with samples in columns, and gene symbols in rows.

Note:

1. The protein intensity matrice does not require normalization.

2. Sample names can only consist of letters, numbers, and underscores.

### 2.sample information file (--sample\_information)

Sample information file is a four-column, tab-delimited file with the first line identifying the columns. The column names are ‘Sample\_id’ ‘MS\_run’ ‘Sample\_condition’ and ‘internal\_ref’.

Note: To avoid code conflicts, ‘Sample\_condition’ consists only of letters, numbers, and underscores.

The column headers are:

Sample\_id (Sample\_id can only consist of letters, numbers, and underscores.)

MS\_run (The MS run in which the proteins of the sample were quantified.)

Sample\_condition (Sample\_condition consists only of letters, numbers, and underscores)

internal\_ref ("Yes" indicates that the sample is an internal reference sample, while "No" means it is not.)

## Output Files

### 1. Table of z-statistic:

z\_statistic\_table.txt

It is crucial for downstream analysis and visualization.

### 2. Chi-square statistics and corresponding statistical p-values:

reverse\_zmap\_chi\_square\_pvalue.txt

This file contains 5 columns: protein, number of samples where this protein was not detected, chi-square statistic, p-value, BH-corrected p-value, and Bonferroni-corrected p-value. This file will be utilized to identify hypervariable proteins across samples.

### 3. Summary of proteomic datasets at different levels.

qc\_boxplot.pdf

qc\_table.pdf

### 4. Goodness of fit for the application of reverse-zMAP to your data set.

4.1 Box plot of the R2 values of all quantile regressions.

r2\_linear\_model\_boxplot.pdf

4.2 The distribution of R2 values from the fitting of M-A curves. The associated three quartiles are marked.

r2\_distribution\_of\_nonlinear\_model\_fitting\_for\_estimated\_variance.pdf.pdf

4.3 The distribution of R2 values from the fitting of MVCs.

r2\_distribution\_of\_nonlinear\_fitting\_for\_intercept\_u.pdf

### 5. The comparison of cumulative distributions of m values and z-statistics, with each line representing a sample.

cumulative\_distribution.pdf

# Sample quality control

Performing hierarchical clustering and principal component analysis on the z-statistic matrix of samples provides a concise visualization of the overall impact of sample conditions and MS runs.

|  |  |  |
| --- | --- | --- |
| Name | Flag | Description |
| z-statistic file\* | --z\_statistic\_matrix | Output file z\_statistic\_table.txt from reverse-zMAP |
| sample information file\* | --sample\_info | The sample information file is a three-column, tab-delimited file with the first line identifying the columns. The column names are ‘Sample\_id’ ‘Sample\_condition’ and ‘MS\_run’. Note: To avoid code conflicts, ‘Sample\_condition’ consists only of letters, numbers, and underscores. |
| email address | --outdir | After the operation is completed, the download link for the result files will be sent to the user’s email. |

## Input Files

### 1. z statistic file (--z\_statistic\_matrix)

Output file z\_statistic\_table.txt from reverse-zMAP as input.

### 2. sample information file (--sample\_info)

As above.

## Output Files

### **1. Table of PCC between each pair of samples based on z-statistic:**

pearsonr\_correlation\_coefficient\_of\_z\_statistic.txt

### 2. Sample-based hierarchical cluster:

pearsonr\_correlation\_coefficient\_of\_z\_statistic.pdf.

### 3. Principal components matrix:

pca\_df.txt

### 4. A principal component plot displaying samples in a 2D plane defined by their first two principal components. This plot provides a concise visualization of the overall impact of sample conditions and MS runs.

PC1\_PC2\_scatterplot.pdf

# Sample subgrouping

For delineating molecular subtypes at the protein level, this function is crafted to curate a subset of highly variable proteins across samples. It then undertakes unsupervised clustering on the samples, offering quantitative evidence to ascertain both the number and composition of potential clusters within the dataset. For in-depth algorithmic insights, please consult the details provided in *Consensus Clustering.1,2*

|  |  |  |
| --- | --- | --- |
| Name | Flag | Description |
| z-statistic file\* | --z\_statistic\_matrix | Output file z\_statistic\_table.txt from reverse-zMAP |
| sample information file\* | --sample\_info | Sample information file is a three-column, tab-delimited file with the first line identifying the columns. The column names are ‘Sample\_id’ ‘Sample\_condition’ and ‘MS\_run’. Note: To avoid code conflicts, ‘Sample\_condition’ consists only of letters, numbers, and underscores. |
| Sample condition\* | --sample\_condition | perform clustering on samples under specific conditions. |
| Protein number\* | --top\_n | Proteins were ranked based on their variance across samples, and the z-statistic matrices of the top\_n proteins were used for clustering. By default, the top\_n is set to 3000. |
| email address | --outdir | After the operation is completed, the download link for the result files will be sent to the user’s email. |

## Input Files

### 1.z statistic file (--z\_statistic\_matrix)

Output file z\_statistic\_table.txt from reverse-zMAP as input.

### 2. sample information file (--sample\_info)

As above.

## Output File

### 1. Sample clustering results:

cluster\_XX.csv (XX is the number of clusters)

### 2. Graphic output:

consensus.pdf

We invoked the ConsensusClusterPlus R package at the underlying level. For interpretation of the output graphics, please refer to <https://bioconductor.org/packages/release/bioc/vignettes/ConsensusClusterPlus/inst/doc/ConsensusClusterPlus.pdf>

# Association with clinical and molecule features

In the prior step, samples underwent grouping using the z-statistic matrix derived from highly variable proteins, with the intent of correlating proteomic subgroups to clinical and molecular features. We then employed either the Chi-square test or Fisher's exact test to examine the association between proteomic subgroups and discrete sample features. This choice was made based on the number of categories within each feature. For continuous features, we utilized Student’s t-test or ANOVA.

ANOVA was further applied to pinpoint differentially expressed proteins (DEPs) among sample groups. Hierarchical clustering of proteins ensued, driven by the z-statistic matrix, revealing distinctive expression signatures. Following this, pathway enrichment analysis was conducted for each set of proteins.

|  |  |  |
| --- | --- | --- |
| Name | Flag | Description |
| z-statistic file\* | --z\_statistic\_matrix | Output file z\_statistic\_table.txt from reverse-zMAP. |
| cluster file\* | --cluster\_f | A comma-separated file with the first column containing sample names and the second column indicating the respective sample groups. |
| Sample clinical and molecule file\* | --clinical\_info | A tab-separated file with the first column representing sample names, and the subsequent columns containing clinical and molecular features. Please note that column names can only include letters, numbers, and underscores. Allows NaN values. When plotting, NaN values are represented by white color by default. |
| discrete features\* | --discrete | This is a string where discrete features are comma-separated. For example, 'gender, stage'. |
| Continuous features\* | --continuous | This is a string where continuous features are comma-separated. For example, 'age,tumor\_size'. |
| color for discrete features\* | --color\_f | Tab-delimited file with no column names. The first column represents discrete sample features, and the second column corresponds to the hexadecimal color code for each feature. |
| color for continuous feature\* | --colorbar\_f | Tab-delimited file with no column names. The first column represents discrete sample features, and the second column corresponds to the colormaps in Matplotlib for each feature. |
| FDR cutoff\* | --fdr | FDR cutoff used for identifying DEPs. The default FDR is 0.05. |
| Number of protein clusters | --cluster\_n | Cluster number for DEPs. |
| email address | --outdir | After the operation is completed, the download link for the result files will be sent to the user’s email. |

## Input Files

### 1.z statistic file (--z\_statistic\_matrix)

Output file z\_statistic\_table.txt from reverse-zMAP as input.

### 2. cluster file (--cluster\_f)

A comma-separated file with the first column containing sample names and the second column indicating the respective sample groups.

### 3. Sample clinical and molecule file (--clinical\_info)

A tab-separated file with the first column representing sample names, and the subsequent columns containing clinical and molecular features. Please note that column names can only include letters, numbers, and underscores. Allows NaN values. When plotting, NaN values are represented by white color by default.

### 4. color for discrete features (--color\_f)

Tab-delimited file with no column names. The first column represents discrete sample features, and the second column corresponds to the hexadecimal color code (such as #808080) for each feature.

### 5. color for continuous feature (--colorbar\_f)

Tab-delimited file with no column names. The first column represents discrete sample features, and the second column corresponds to the colormaps in *Matplotlib* for each feature.

Colormaps: <https://matplotlib.org/stable/users/explain/colors/colormaps.html>

## Output File

### 1. Graphic output:

sample\_clustering\_association\_with\_clinical\_and\_molecule\_feature.pdf

### 2. pathway enrichment results:The directory enrichment\_results contains pathway enrichment results for each expression signature.

# Survival analysis for proteomic subgroups

Plots the Kaplan-Meier survival curve and assesses the significance of survival differences among subgroups.

|  |  |  |
| --- | --- | --- |
| Name | Flag | Description |
| survival and group file\* | --input\_file | Sample information file is a four-column, tab-delimited file with the first line identifying the columns. The column names are ‘Sample\_id’ ‘survival\_time’ ‘death\_or\_not’ and ‘group’. |
| email address | --outdir | After the operation is completed, the download link for the result files will be sent to the user’s email. |

## Input Files

### 1.survival and group file (--input\_file)

Sample information file is a four-column, tab-delimited file with the first line identifying the columns. The column names are ‘Sample\_id’ ‘survival\_time’ ‘death\_or\_not’ and ‘group’.

## Output File

### 1. Graphic output:

survival\_analysis.pdf

# Association z-statistic with survival data

Cox proportional hazard regression model is used to identify prognostic markers based on z-statistic.3

|  |  |  |
| --- | --- | --- |
| Name | Flag | Description |
| z-statistic file\* | --z\_statistic\_matrix | Output file z\_statistic\_table.txt from reverse-zMAP |
| survival file\* | --input\_file | Sample information file is a three-column, tab-delimited file with the first line identifying the columns. The column names are ‘Sample\_id’ ‘survival\_time’, and ‘death\_or\_not’. |
| email address | --outdir | After the operation is completed, the download link for the result files will be sent to the user’s email. |

## Input Files

### 1. z-statistic file (--z\_statistic\_matrix)

Output file z\_statistic\_table.txt from reverse-zMAP.

### 2. survival file (--input\_file)

Sample information file is a three-column, tab-delimited file with the first line identifying the columns. The column names are ‘Sample\_id’ ‘survival\_time’ and ‘death\_or\_not’.

## Output File

### 1. Cox-regression results file. Proteins are categorized into three groups: favorable, unfavorable, and not prognostic.

results\_cox\_regression.txt

# Association z-statistic with mutation data

To pick out the proteins that were associated with non-silent somatic mutations, QTL analysis is used. Refer to the Methods section of the zMAP article for specific details of the methodology.

|  |  |  |
| --- | --- | --- |
| Name | Flag | Description |
| z-statistic file\* | --z\_statistic\_matrix | Output file z\_statistic\_table.txt from reverse-zMAP |
| mutation file\* | --mutation\_f | Tab-delimited file where rows represent genes, columns represent samples, and values are coded as 1 for non-silent mutations and 0 for no non-silent mutations. |
| covariates file\* | --covariates\_f | The file includes additional covariates. Columns represent samples, while rows encompass various additional covariates, including age, sex, and others. |
| gene position file \* | --gene\_tss\_location | The file containing transcription start site information for all genes in the genome. It contains three columns with the column names being 'gene', 'chromosome', and 'tss' respectively. |
| chromosome length file\* | --chr\_length | Tab-delimited file contains two columns, ‘chromosome’ and ‘length’. |
| FDR cutoff\* | --fdr | The default FDR is 0.05. |
| email address | --outdir | After the operation is completed, the download link for the result files will be sent to the user’s email. |

## Input Files

### 1. z-statistic file (--z\_statistic\_matrix)

Output file z\_statistic\_table.txt from reverse-zMAP.

### 2. mutation file (--input\_file)

Tab-delimited file where rows represent genes, columns represent samples, and values are coded as 1 for non-silent mutations and 0 for no non-silent mutations.

### 3. covariates file (--covariates\_f)

The file includes additional covariates. Columns represent samples, while rows encompass various additional covariates, including age, sex, and others.

### 4. gene position file (--gene\_tss\_location)

The file containing transcription start site information for all genes in the genome. It contains three columns with the column names being 'gene', 'chromosome', and 'tss' respectively.

### 5. chromosome length file (--chr\_length)

Tab-delimited file contains two columns, ‘chromosome’ and ‘length’.

## Output File

### 1. Graphic output:

pQTL\_results\_FDR\_XX.pdf (XX is FDR cutoff)

Two-dimensional plot displaying the significant gene-protein associations, with Y and X axes representing the locations of proteins and mutated genes in the genome, respectively. The total number of proteins associated with each mutated gene is also displayed. For a specific gene-protein association, Beta>0 suggests the protein has increased expression with the mutation of the gene, and vice versa.

### 2. proteins associated with mutation genes.

results\_association\_z\_statistic\_with\_mutation.txt

### 3. Number of proteins associated with mutated genes.

mutation\_associated\_protein\_count.txt

1 Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn 52, 91-118, doi:Doi 10.1023/A:1023949509487 (2003).

2 Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 26, 1572-1573, doi:10.1093/bioinformatics/btq170 (2010).

3 Cox, D. R. Regression models and lifetables. Journal of the Royal Statistical Society: Series B (Methodological) 34, 187-202 (1972).