# CBioProfiler user tutorial

Xiao-Ping Liu, Zisong Wang, Hongjie Shi, Sheng Li, Xing-Huan Wang

2023-10-30

## Contents



# 1 Introduction

**CBioProfiler** (**C**ancer **Bio**marker and subtype **Profiler**) was developed to facilitate researchers and clinicians to screen, characterize, annotate and translate cancer biomarkers and subtypes from molecular level to clinical settings more comfortably with graphical user interfaces (GUI), which will help implement targeted clinical diagnosis and treatment measures for different patients to achieve precision medicine.

CBioProfiler integrated a novel R package CuratedCancerPrognosisData that reviewed, curated and integrated the gene expression data and corresponding clinical data of 47,210 clinical samples from 268 gene expression studies of 43 common blood and solid tumors.

The whole pipeline of CBioProfiler includes two main pipelines: cancer biomarker pipeline and cancer subtype pipeline. The cancer biomarker pipeline includes 5 modules: (1) dimensionality reduction using three methods of weighted gene co-expression network analysis (WGCNA), univariate Cox proportional

hazards regression model (CoxPH), differentially expressed gene (DEG) analysis, (2) benchmark experiment with 6 machine learning learners (Lasso, Ridge, Elastic net, Glmboost, Coxboost, Randomforest) using cross validation (CV) and nested cross validation (nCV) based on R package mlr, (3) prediction model construction using Cox proportional hazards regression model and nomogram, (4) clinical annotation using a variety of clinical approaches (correlation with clinical features, Kaplan-Meier curve, CoxPH model, time-dependent ROC, most correlated genes, correlation with specific gene, gene expression in different groups, correlation with immune infiltration, correlation with stemness score, correlation with ESTIMATE score, correlation with immune checkpoint, correlation with IFN-gamma score, correlation with cytolytic activity, correlation with cancer pathway, correlation with metabolism pathway, correlation with hallmark signature, correlation with drug response, and (5) biological annotation using over-representation analysis (ORA) and gene set enrichment analysis (GSEA).

The subtype pipeline includes 3 modules: (1) data preprocessing (feature selection based on variance, median absolute deviation (MAD),CoxPH model, and principal component analysis (PCA), (2) subtype identification (integration of multiple unsupervised machine learning methods (K-means clustering (K-means) (Hartigan and Wong 1979), hierarchical clustering (Maimon and Rokach 2005), partitioning around medoids (PAM) clustering (Van der Laan, Pollard, and Bryan 2003), etc.) using two popular consensus clustering methods (ConsensusClusterPlus (Wilkerson and Hayes 2010) and M3C (John et al. 2020)), (3) subtype evaluation and validation.

The overview of the CBioProfiler is summarized in Figure 1:

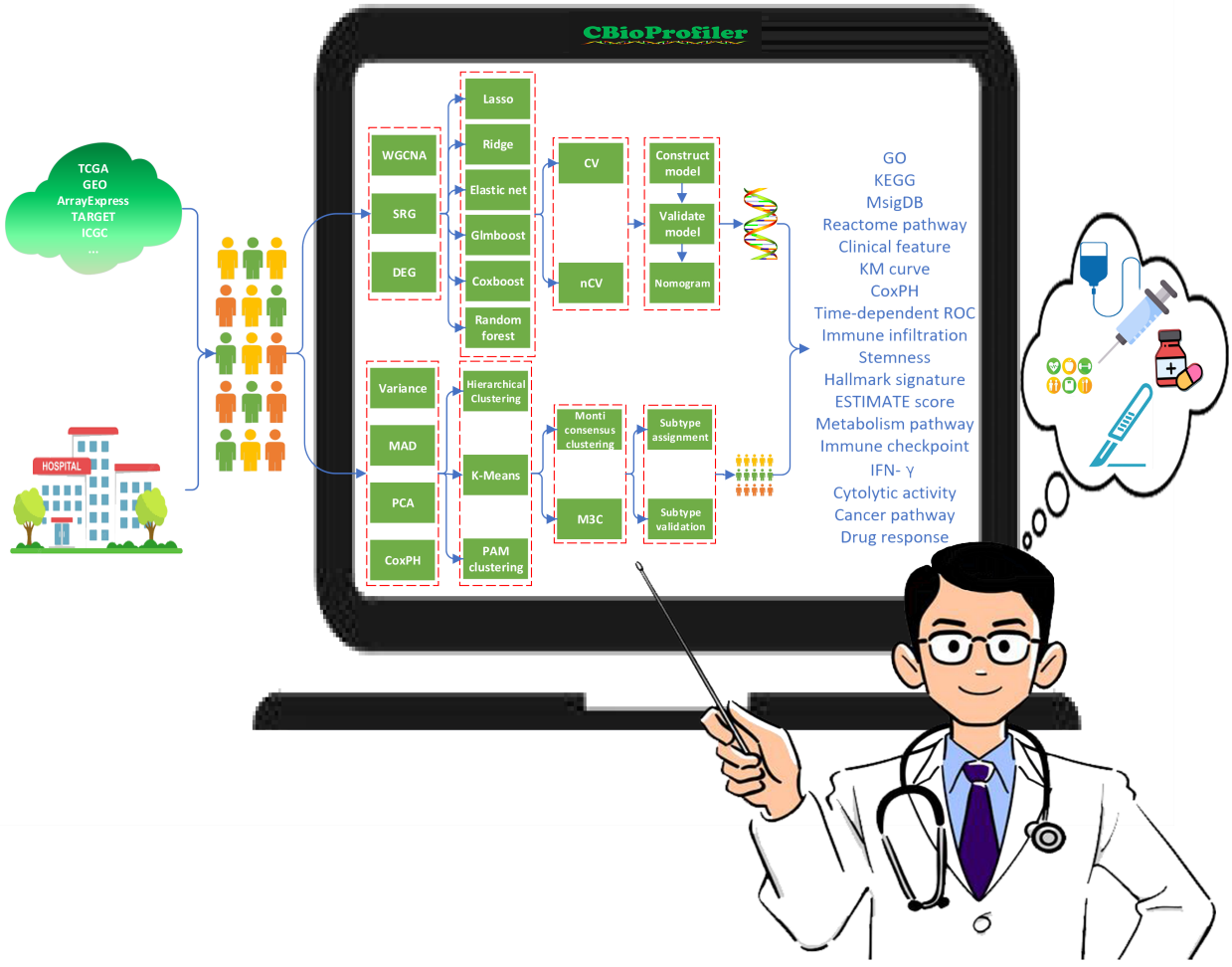


Figure 1: Overview of CBioProfiler.

The CBioProfiler GUI was created using Shiny, a web application framework for R, as well as several useful packages to provide more advanced features that help enhance the CBioProfiler GUI, including shinyjs to add JavaScript actions for the app, shinydashboard to add dashboards and shinyBS to add additional functionality and interactivity to the app.

Correspondence: Any suggestions and questions regarding the development and use of the CBioProfiler are welcome, which can be reached by e-mailing to the first author of the work: Xiao-Ping Liu; Department of Urology, Zhongnan Hospital of Wuhan University; Email: liuxiaoping@whu.edu.cn.

# 2 Data

CBioProfiler supports public open access data and user customized data. The public data was curated open access gene expression profile studies (both gene expression profile and corresponding clinical information of the participants) from GEO, TCGA, ICGC, ArrayExpress, TARGET, CGGA and some research center and journal website. More details can be found at the Data source module of the CBioProfiler.

## 2.1 Data input

In the menu, there are two window: Output window and parameter setting window. The users can input both public data and personally customized data through specifying '**Data type**'. For public data, the users need to select one type of '**Cancer**' and then specify the '**Accession**' of the associated gene expression profile study. Finally, the data can be loaded by pressing the button '**Submit dataset**'.

CBioProfiler has reviewed, curated and integrated the gene expression data and corresponding clinical data of 43 common blood and solid tumors from GEO, TCGA, ICGC, TARGET, ArrayExpress, CGGA and other public databases. These public data come from 47,210 clinical samples from 268 gene expression studies.

To promote the use of these public data, we have develop '**CuratedCancerPrognosisData**' that encompasses these public data. Thus, CuratedCancerPrognosis is required by CBioProfiler when running public data analysis. For more details on implementation of the two packages, please refers to https://github.com/liuxiaoping2020/CBioProfiler and https://zenodo.org/record/5728447#.YdqjDcj1dk4.

If you use these public data when conducting your own research, we strongly recommend that you cite these public data. For a detailed introductions regarding these public data, please refer to: https://liuxiaoping2020.github.io/CBioProfilerDatasource/. Figure 2 shows the public data input interface.

For personally customized data, the users need to prepare their own gene expression profile data and the associated clinical data. Gene expression profile data represents a data matrix in '*csv*' format where genes (with official gene symbols) in rows and sample/patient names in columns, while,the clinical data represents a data matrix in '*csv*' format where the sample/patient names in rows and clinical variables in columns. The sample/patient names in the gene expression profile data matrix and clinical data matrix should be identical. Figure 3 shows customized data input interface.

If the users do not know how to prepare either the gene expression data or clinical data matrix, then they can download sample data from certain public data via the download button in the lower left corner of the output window.

# 3 Biomaker analysis

The '**Analysis**' module includes five sequential sub-modules: (1) **Dimensionality reduction**; (2) **Benchmark experiment**; (3) **Prediction model** (4) **Clinical annotation**; and (5) **Biological annotation**.

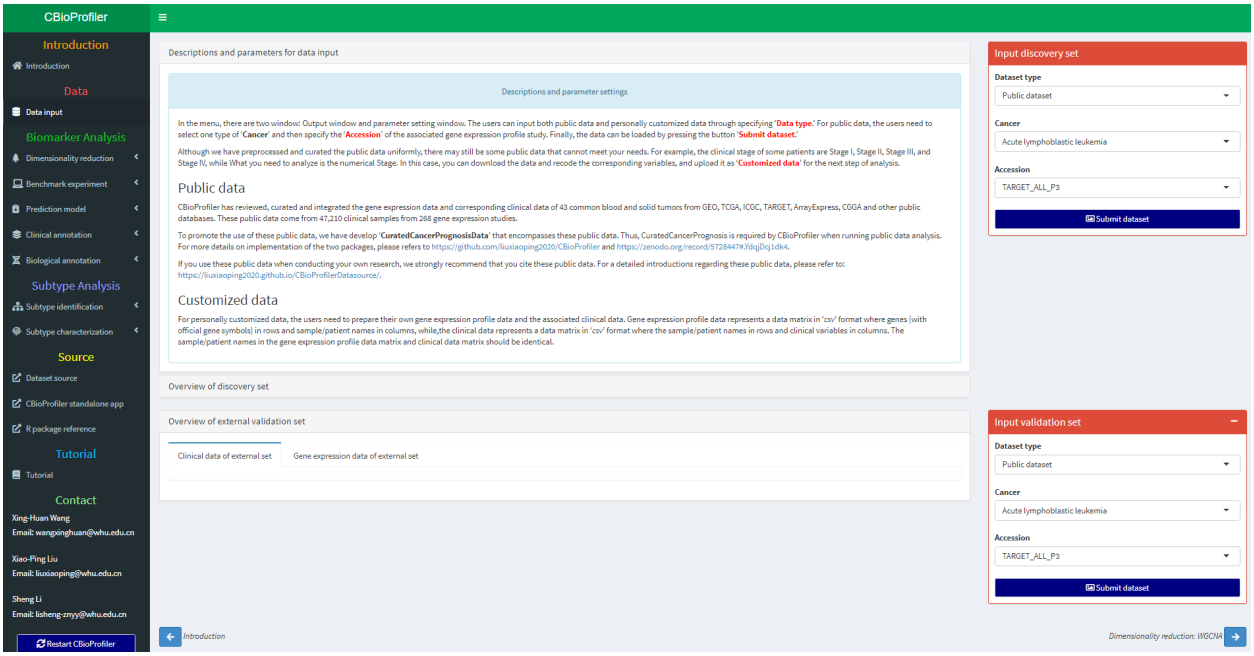


Figure 2: Data input menu: Public data input main window.

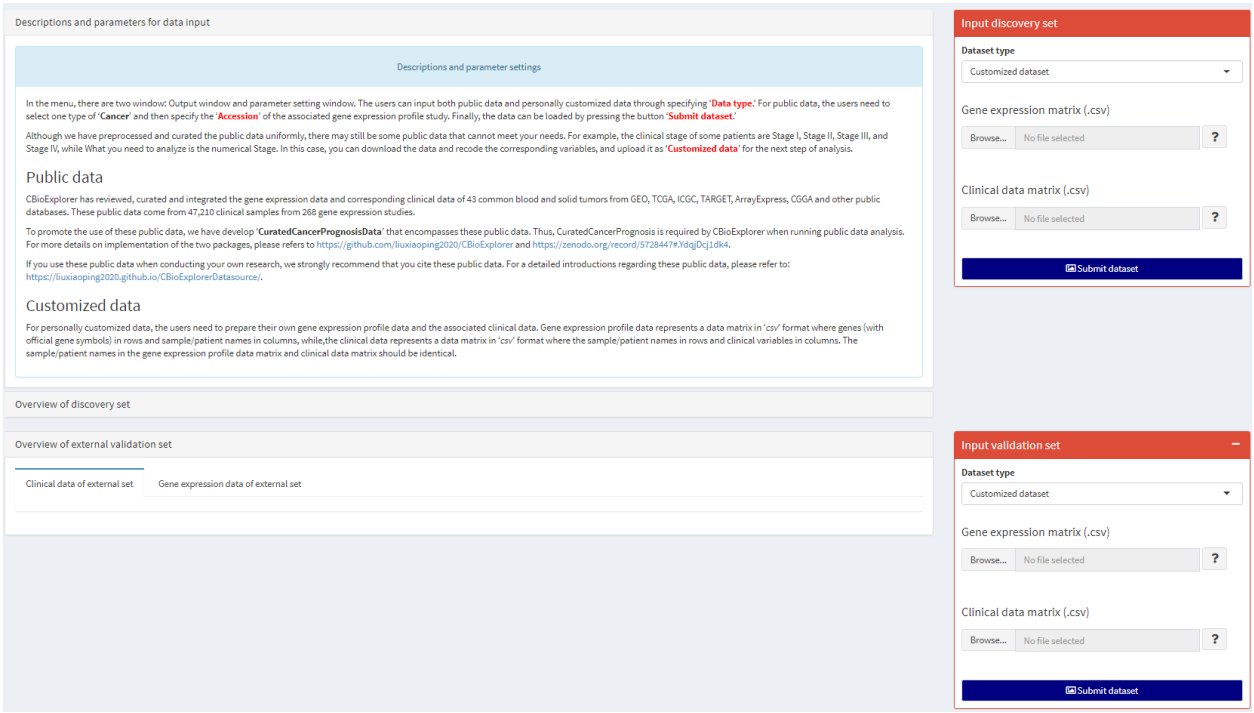


Figure 3: Data input menu: Customized data input main window.

## 3.1 Dimensionality reduction

CBioProfiler uses the 3 most commonly used bioinformatics analysis methods to reduce the dimensionality of data: (1) WGCNA Langfelder et al. (2020); (2) Survival related genes (conducted based on univariate Cox proportional hazards regression model using R package survival (Therneau 2020)); (3) Differentially expressed genes (conducted using R package limma (Smyth et al. 2020)).

### 3.1.1 WGCNA

The WGCNA module includes 3 sequential pipelines:(1) **Data Preprocessing**; (2) **Network construction and module detection**; (3) **Module-trait relationships**.

For **Data preprocessing**, CBioProfiler uses sample network methods for finding outlying samples,Specifically, the Euclidean distance based sample network is simply the canonical Euclidean distance based network. Figure 4 shows data preprocessing of WGCNA interface.

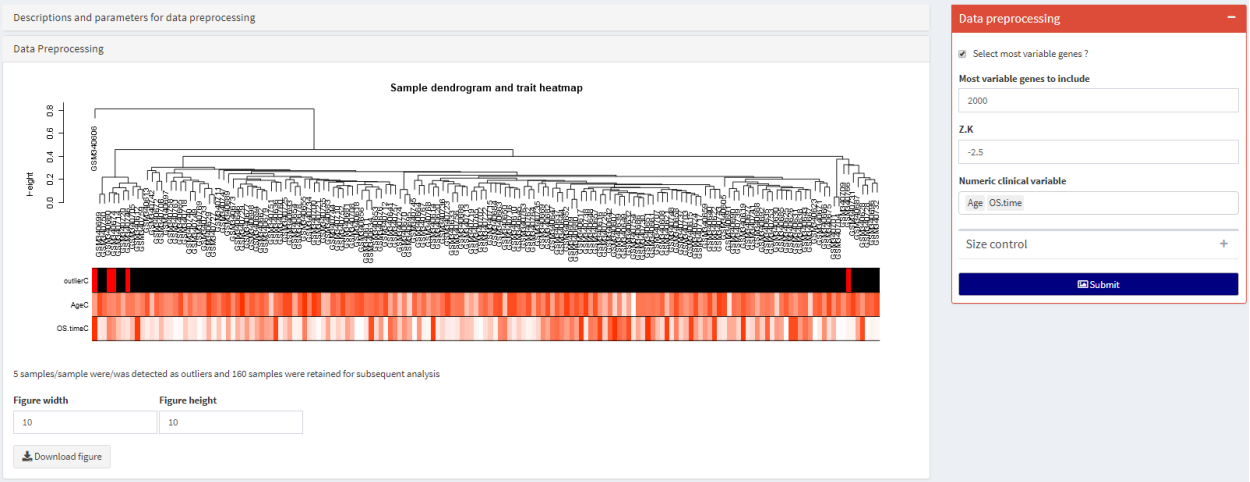


Figure 4: WGCNA menu: Data Preprocessing main window.

In the main window, users can determine which samples are outliers based on the sample dendrogram and trait heatmap.

Parameters:

- **Most variable genes to include:** Whether select top n variable genes to be included for network construction based on the variances of genes cross all samples.
- **Z.K:** Cutoff in sample network for outlier detection, please refer to (Horvath 2011) for more details.
- **Numeric clinical variable:** Select ***2 or more numeric*** clinical variables to be included for correlation analysis. **Please note that the clinical trait should be numeric value, if users want to identify the correlation between modules and categorical clinical traits, then the user can transform the categorical clinical trait to dummy variables**
- **Heatmap Width (%):** Set the width of the heatmap.
- **Heatmap Height (px):** Set the height of the heatmap.

For **Network construction and module detection**, CBioProfiler integrates Langfelder's (Langfelder and Horvath 2008) 1-step automatic construction of the gene network and identification of modules. Figure 5 shows network construction and module detection of WGCNA interface.

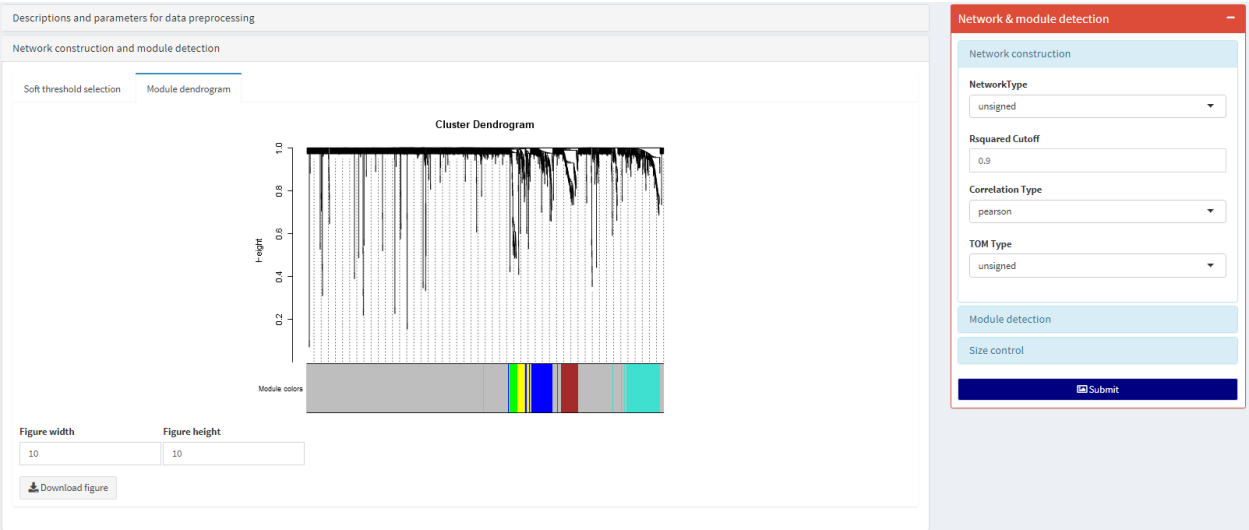


Figure 5: WGCNA menu: Network construction and module detection main window.

In the main window, users can (1) determine soft-threshold based on the **Rsquared Cutoff** they specified, and (2) see how many modules they identified.

Parameters:

- **NetworkType:** network type. Allowed values are (unique abbreviations of) 'signed','unsigned','signed hybrid'. More detail can be found at https://cran.r-project.org/web/packages/WGCNA/WGCNA.pdf.
- **Rsquared Cutoff:** Desired minimum scale free topology fitting index R^2.
- **Correlation Type:** Specifying the correlation type, 'pearson' means 'Pearson's Correlation','bicor' means 'Bidweight midcorrelation.
- **TOM Type:** Specifying the Topology Overlap Matrix (TOM) type.
- **Deep split:** Provides a simplified control over how sensitive module detection should be to module splitting, with 0 least and 4 most sensitive.
- **Maximum joining heights:** For method=='hybrid' it defaults to 99% of the range between the 5th percentile and the maximum of the joining heights on the dendrogram.
- **Minimum Module size:** Minimum module size for module detection.
- **P value threshold:** p-value ratio threshold for reassigning genes between modules.
- **Threshold to Merge Modules:** Dendrogram cut height for module merging.
- **Should modules be labeled numbers ?:** Should the returned modules be labeled by colors (FALSE), or by numbers (TRUE)?
- **PAM-like stage ?:** Only used for method 'hybrid'. If TRUE, the second (PAM-like) stage will be performed.
- **PAM stage will respect the dendrogram ?:** Only used for method 'hybrid'. If TRUE, the PAM stage will respect the dendrogram in the sense that objects and small clusters will only be assigned to clusters that belong to the same branch that the objects or small clusters being assigned belong to.
- **Size control** Change the size of the Soft threshold selection plot and Module dendrogram.

For **Module-trait relationships**, CBioProfiler allows users (1) to analyze **Module-trait relationships** and screen modules significantly correlated with the clinical variables they specified, (2) to identify genes with high gene significance (GS) and module membership (MM) (**GS vs MM**), and then (3) to **Output genes** in a specific module or all non-grey modules. Figure 6 shows module-trait relationships of WGCNA interface.

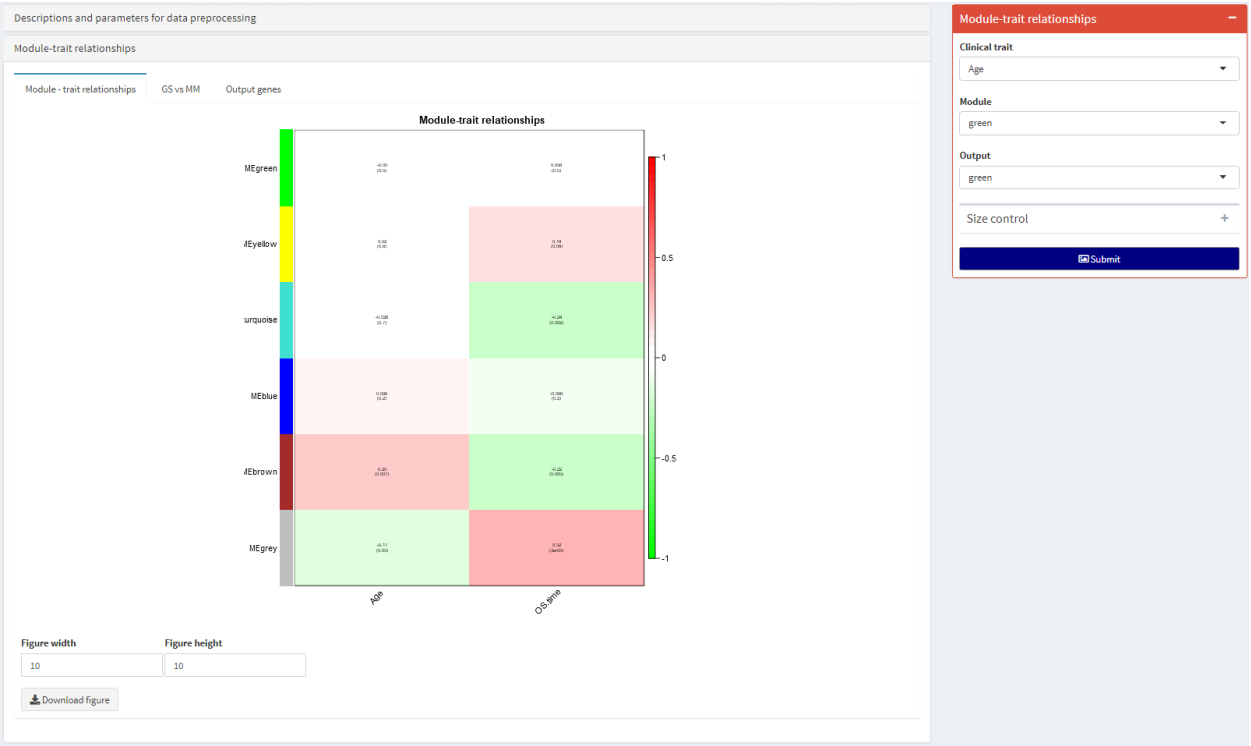


Figure 6: WGCNA menu: Module-trait relationships main window.

Parameters:

- **Clinical trait:** Select the clinical trait you are interested and to identify correlated modules and genes. **Please note that the clinical trait should be numeric value, if users want to identify the correlation between modules and categorical clinical traits, then the user can transform the categorical clinical trait to dummy variables**.
- **Module:** Select the module you are interested.
- **Output:** Output genes in modules for downstream analysis.
- **Size control:** Set the width, height, and margins of plot from **Module-trait relationships** and **GS vs MM**.

### 3.1.2 Survival related genes

The **Survival related genes (SRG)** module allows users to identify most survival related genes based on univariate Cox proportional hazards regression model (CoxPH) using R package 'survival' (Therneau 2020).

In the **Univariate CoxPH table** main window, the users can obtain a full table of the result of univariate CoxPH result. Figure 7 shows univariate CoxPH table interface.

Parameters:

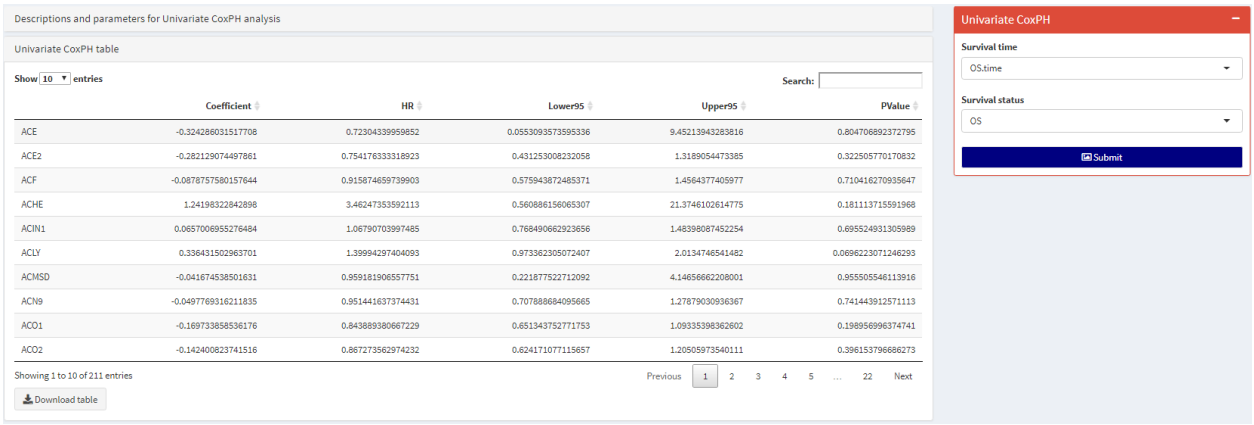


Figure 7: Survival related genes menu: Univariate CoxPH table main window.

- **Survival time:** Select the survival time column for univariate CoxPH analysis.
- **Survival status:** Select the survival Status column for univariate CoxPH analysis.

In the **Significant univariate CoxPH result** main window, the users can draw a forest plot of survival related genes based on the significance cutoff they set. Figure 8 shows forest plot for significant univariate CoxPH result.

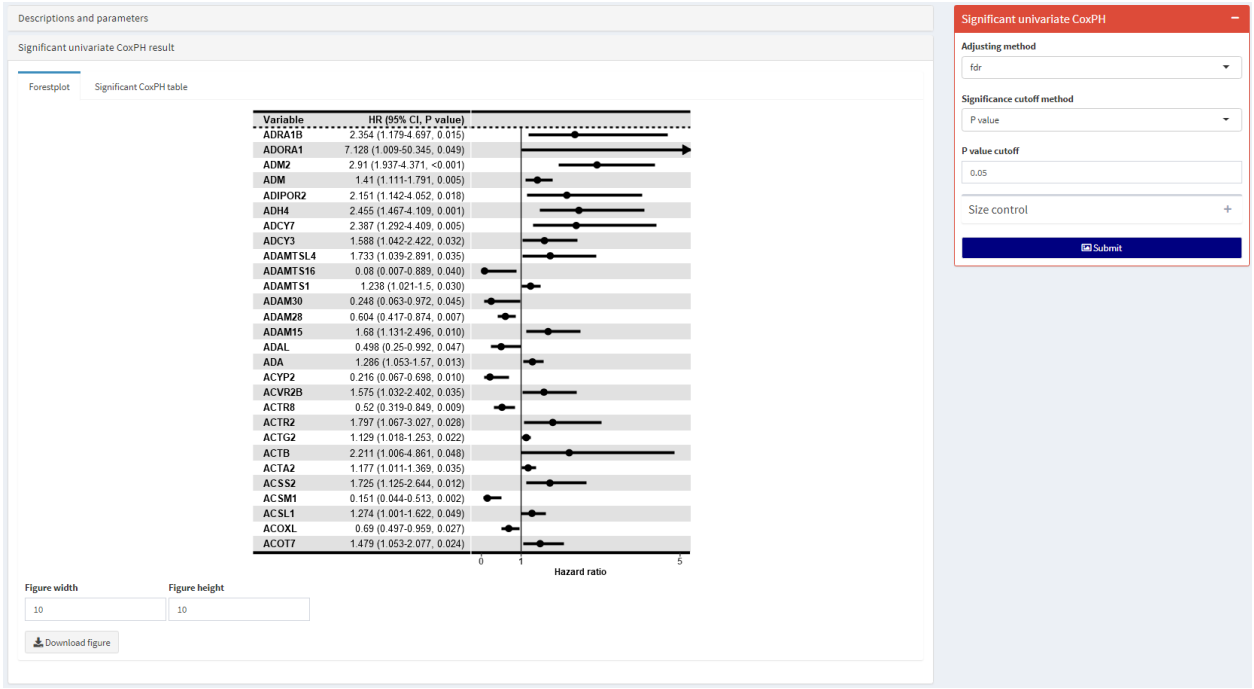


Figure 8: Survival related genes menu: Significant univariate CoxPH result main window.

Parameters:

- **Adjusting method:** Specify the correction method.
- **Significance cutoff method:** Specify the method to select to most significant survival related genes, either 'P value' or 'Adjusted P value'.

- **Adjusted P cutoff:** Specify cutoff value based on adjusted P value.
- **P value cutoff:** Specify cutoff value based on P value.

### 3.1.3 Differentially expressed genes

CBioProfiler uses linear models provided by the R package 'limma' (Smyth et al. 2020) to identify differentially expressed genes (DEGs) between two or more biological groups. In the **Differentially expressed gene table**, the users can access the whole table of limma-based DEG result. Figure 9 shows differentially expressed gene table interface.

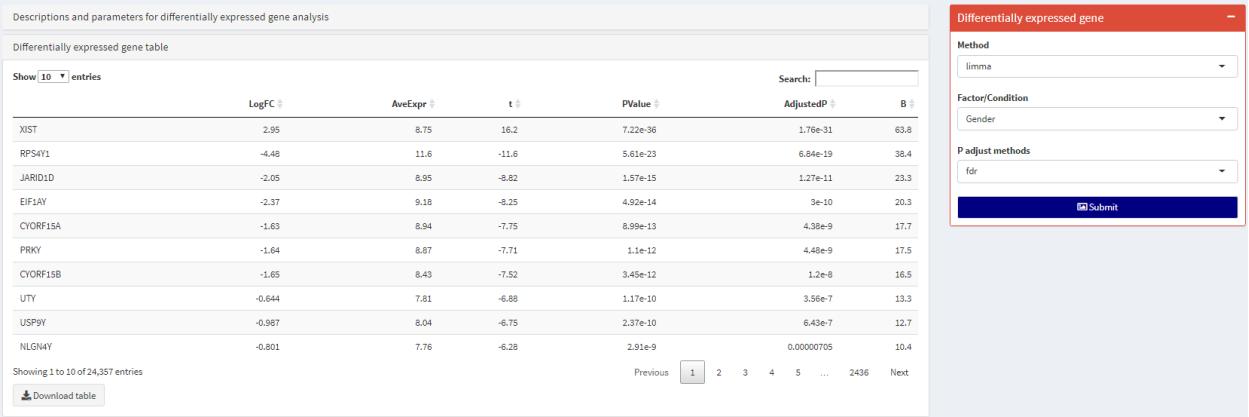


Figure 9: Differentially expressed genes menu: Differentially expressed gene table main window.

Parameters:

- **Method:** 'limma' means conducting moderated contrast t-test for each gene in limma".
- **Factor/Condition:** Specify the group variable for DEG analysis.
- **P adjust methods:** Specify the correction method for P values. For more details, please refer to https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/p.adjust.

After the users finished calculating limma-based DEG analysis, they can further screen significant DEGs based on the screening method and cutoff the specified. Figure 10 shows significant DEG output interface.

In the **DEG output**, the users can screen, output, and visualize significant DEGs they specified.

General parameters:

- **DEG cutoff method:** Specify the DEG cutoff method: "Adjusted P", "LogFC", "Adjusted P & LogFC".
- **Visualization method:** Specify visualization method of DEGs, which includes heatmap, Volcano plot, MA plot, Adjusted P plot.

Heatmap is generated by using R package 'ComplexHeatmap' (Gu 2021). Figure 11 shows DEG heatmap output interface.

Parameters for **Heatmap**:

- **Heatmap name:** Names for the heatmap, by default the heatmap name is used as the title of the heatmap legend.

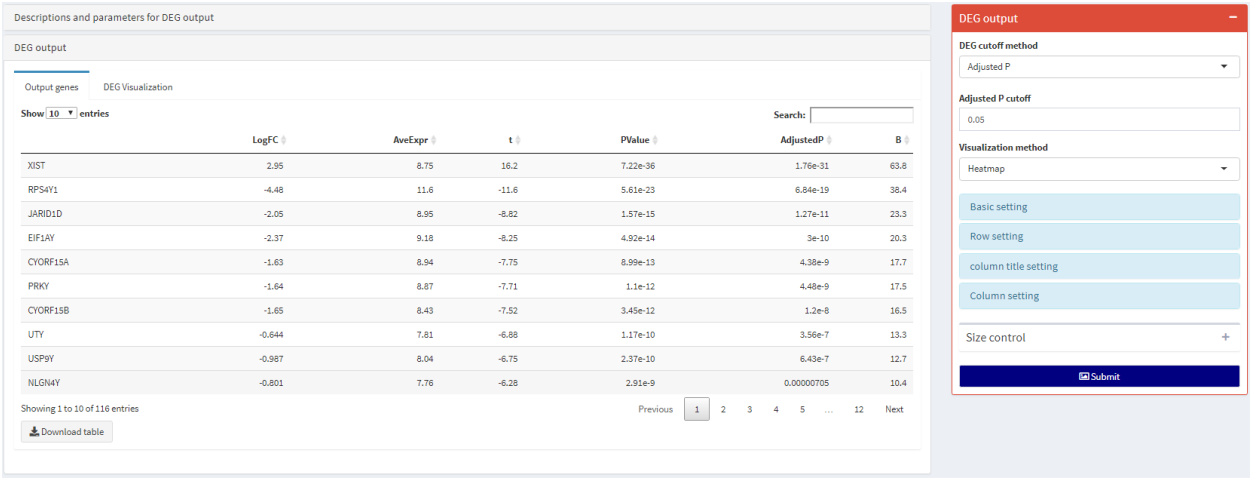


Figure 10: Differentially expressed genes menu: DEG output main window.

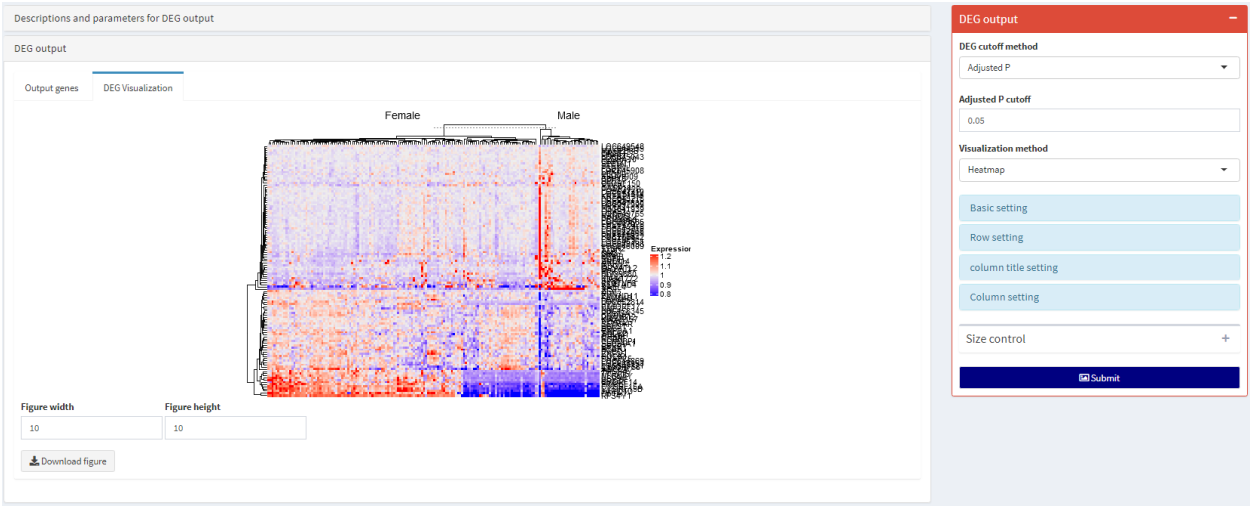


Figure 11: Differentially expressed genes menu: DEG heatmap output main window.

- **Min colour, Median colour, Max colour :** Color range for the heatmap.
- **Normalize the heatmap ?:** Whether to normalize the expression level of the heatmap.
- **Normalization method:** Specify a method to normalize the expression level of the heatmap, which includes “Scale”, “Center”, “Log”, “Z-score”, “0-1 normalization”.
- **Cluster on rows ?:** Whether to make a cluster on rows.
- **Cluster on row slice ?:** If rows are split into slices, whether perform clustering on the slice means?
- **Cluster distance on rows:** It can be a pre-defined character which is in (“euclidean”, “maximum”, “manhattan”, “canberra”, “binary”, “minkowski”, “pearson”, “spearman”, “kendall”).
- **Cluster method on rows:** Method to perform hierarchical clustering, pass to hclust.
- **Row dendrogram side:** Should the row dendrogram be put on the ‘left’ or ‘right’ of the heatmap?
- **Show row names ?:** Whether should row names of the heatmap.
- **Row name side:** Should the row names be put on the left or right of the heatmap?
- **Show adjusted P value ?:** Whether show the adjusted P value for DEGs
- **Show logFC ?:** Whether show the logFC for DEGs.
- **Column title font size:** Specify the font size for the column title
- **Cluster on columns ?:** Whether to make a cluster on columns.
- **Cluster on column slices?:** If columns are split into slices, whether perform clustering on the slice means?
- **Cluster distance on columns:** It can be a pre-defined character which is in (“euclidean”, “maximum”, “manhattan”, “canberra”, “binary”, “minkowski”, “pearson”, “spearman”, “kendall”).
- **Cluster method on columns:** Method to perform hierarchical clustering, pass to hclust.
- **Column dendrogram side:** Should the column dendrogram be put on the top or bottom of the heatmap?
- **Show column names ?:** Whether show the column names for the heatmap.
- **Column name side:** Should the column names be put on the top or bottom of the heatmap?

volcano plot is generated by using R package ‘EnhancedVolcano’ (Blighe, Rana, and Lewis 2020). Figure 12 shows DEG volcano plot output interface.

Parameters for **Volcano plot**:

- **Adjusted P cutoff:** Adjusted P value for DEGs to be visualized using volcano plot.
- **LogFC cutoff:** LogFC cutoff for DEGs to be visualized using volcano plot.
- **X-axis label:** Set the label for X-axis.
- **Y-axis label:** Set the label for y-axis.
- **Legend position:** Position of legend (‘top’, ‘bottom’, ‘left’, ‘right’). DEFAULT = ‘top’.

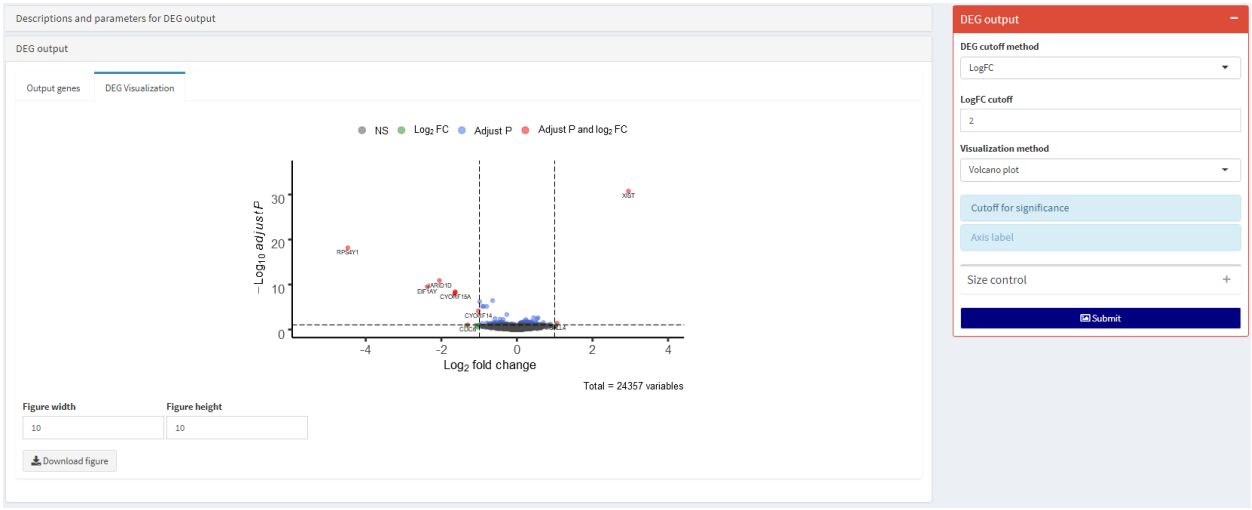


Figure 12: Differentially expressed genes menu: DEG volcano plot output main window.

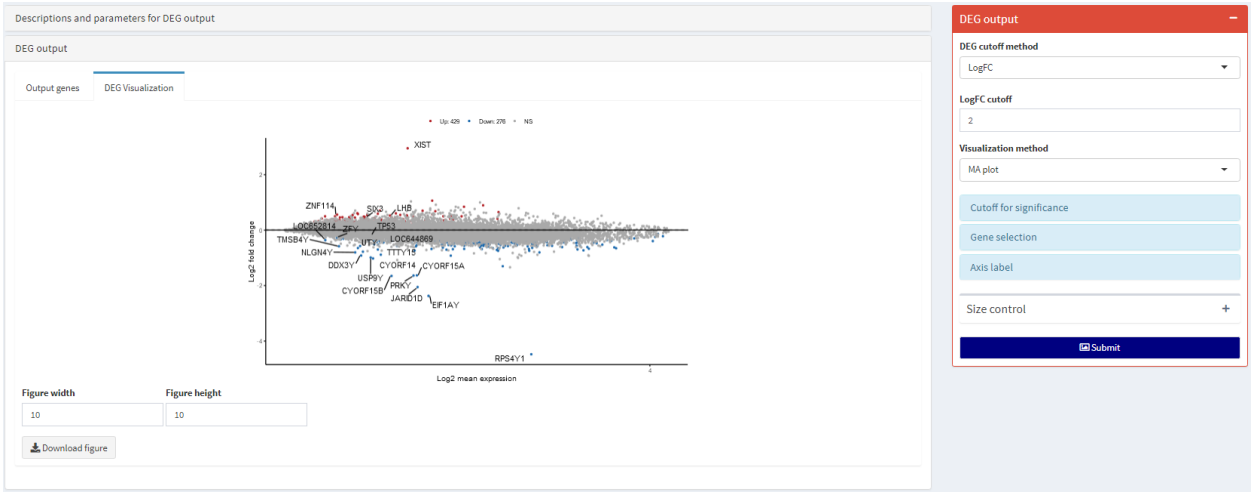


Figure 13: Differentially expressed genes menu: DEG MA plot output main window.

MA plot is generated using R package 'ggpubr'(Kassambara 2020). Figure 13 shows DEG MA plot output interface

Parameters for **MA plot**:

- **Adjusted P cutoff:** Adjusted P value for DEGs to be visualized using volcano plot.
- **LogFC cutoff:** LogFC cutoff for DEGs to be visualized using volcano plot.
- **Selection methods:** Specify the method of label genes: Top genes, Cutoff, Gene symbol.
- **Top gene:** Set the number of top genes that will show gene symbol based on the cutoff set
- **Select top method:** The method used to select top genes.
- **Gene symbol:** Character vector specifying some gene labels to show.
- **X-axis label:** The label for X-axis.
- **Y-axis label:** The label for Y-axis.

Adjusted P plot is generated using R basic function. Figure 14 shows DEG adjusted p plot output interface.

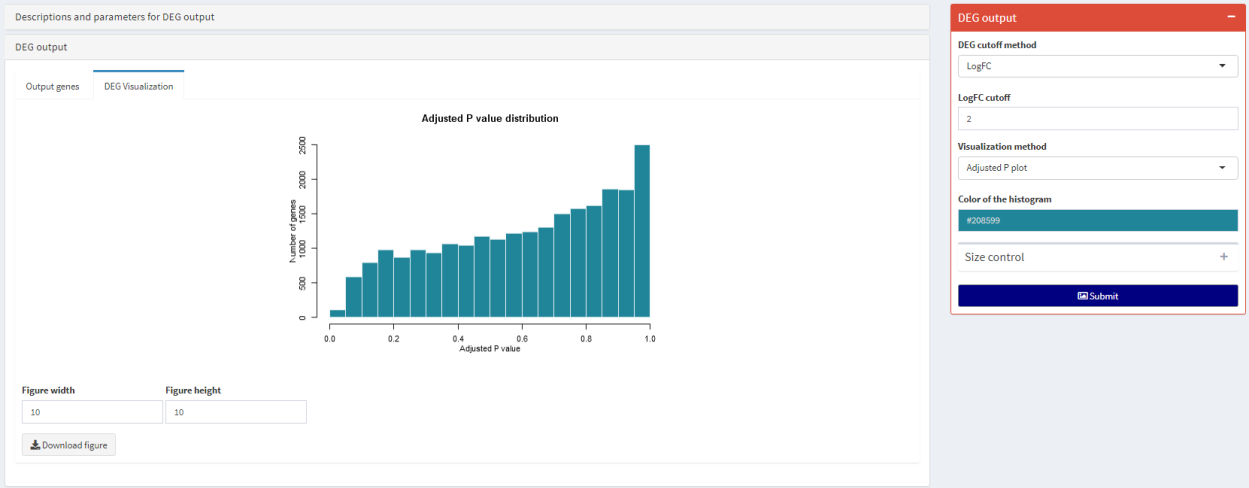


Figure 14: Differentially expressed genes menu: DEG adjusted p plot output main window.

Parameter for **Adjusted P plot**:

- **Color of the histogram:** Set the color for the Adjusted P plot.

## 3.2 Benchmark experiment

For the **Benchmark experiment**, CBioProfiler includes 6 machine learning algorithms (**LASSO**, **Ridge**, **ElasticNet**, **Glmboost**, **Coxboost**, **RandomForestSRC**) for survival analysis and applies cross validation (CV) and nested cross validation (nCV) to train and validate the above survival models. The Benchmark experiment is supported by the R package 'mlr' (Bischl et al. 2020).

During CV, the whole **Data set** is randomly split (If the users desired to split the whole dataset) into **Training set** and **Test set**, then k-fold CV is applied to the training set: (1) Divide the **Training set** into equal K folds; (2) Use the first fold as inner test set, and the rest as inner training set. (3) Train the model and calculate the C-index of the model on the inner test set; (4) Use a different fold as inner test set each

time, and repeat steps (2) and (3) K times. (5) Apply the best model to **Test set** and external independent validation cohort.

The workflow for CV is depicted in figure 15:

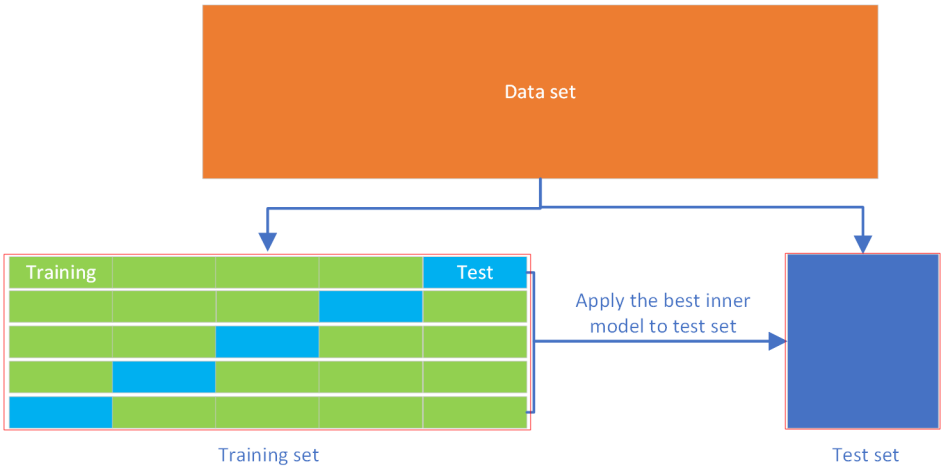


Figure 15: Flow chart of 5-fold cross validation.

During nCV, the whole data set is divided into n outer folds, and then each outer fold is divided into **Training set** and **Test set**. The workflow of nCV: (1) Divide the **Training set** into equal K folds; (2) Use the first fold as inner test set, and the rest as inner training set. (3) Train the model and calculate the C-index of the model on the inner test set; (4) Use a different fold as inner test set each time, and repeat steps (2) and (3) K times. (5) Apply the best model to outer fold **Test set**. (6) Select the best outer model features and parameters and train on the whole data set to get final model. (7) If the users have divided the whole data set into two parts, one is for training nCV, the other is for validation, then they can validate the final model on the validation part and external validation cohort, otherwise, they can validate the final model on external cohort.

The workflow for nCV is depicted figure 16:

The users can select one or more survival learners (**LASSO**, **Ridge**,**ElasticNet**, **Glmboost**, **Coxboost**, **RandomForestSRC**) to conduct benchmark experiment using CV or nCV. For CV the users can perform bootstrap resampling on the test set to compare the performance of the survival learners they selected. For nCV, they can use the performance of each outer fold to compare the survival learners they selected. Figure 17 shows C-index comparison of survival models based on benchmark experiment.

Parameters:

- **Learner:** Specify the survival learner for benchmark experiment.
- **Method:** Specify the method for benchmark experiment: Cross validation or nested cross validation.
- **Random Seed:** Set the random seed for benchmark experiment.
- **Survival time:** Select survival time column. Example: “OS.time”, “RFS.time”, “PFS.time”, etc.
- **Survival status:** Select survival time column. Example: “OS”, “RFS”, “PFS”, etc.
- **Optimization algorithm:** Choose an optimization algorithm to search an appropriate set of parameters from a given search space for given learners.
- **Resolution:** Resolution of the grid for each numeric/integer parameter in par.set. For vector parameters, it is the resolution per dimension. Either pass one resolution for all parameters, or a named vector. Default is 10.

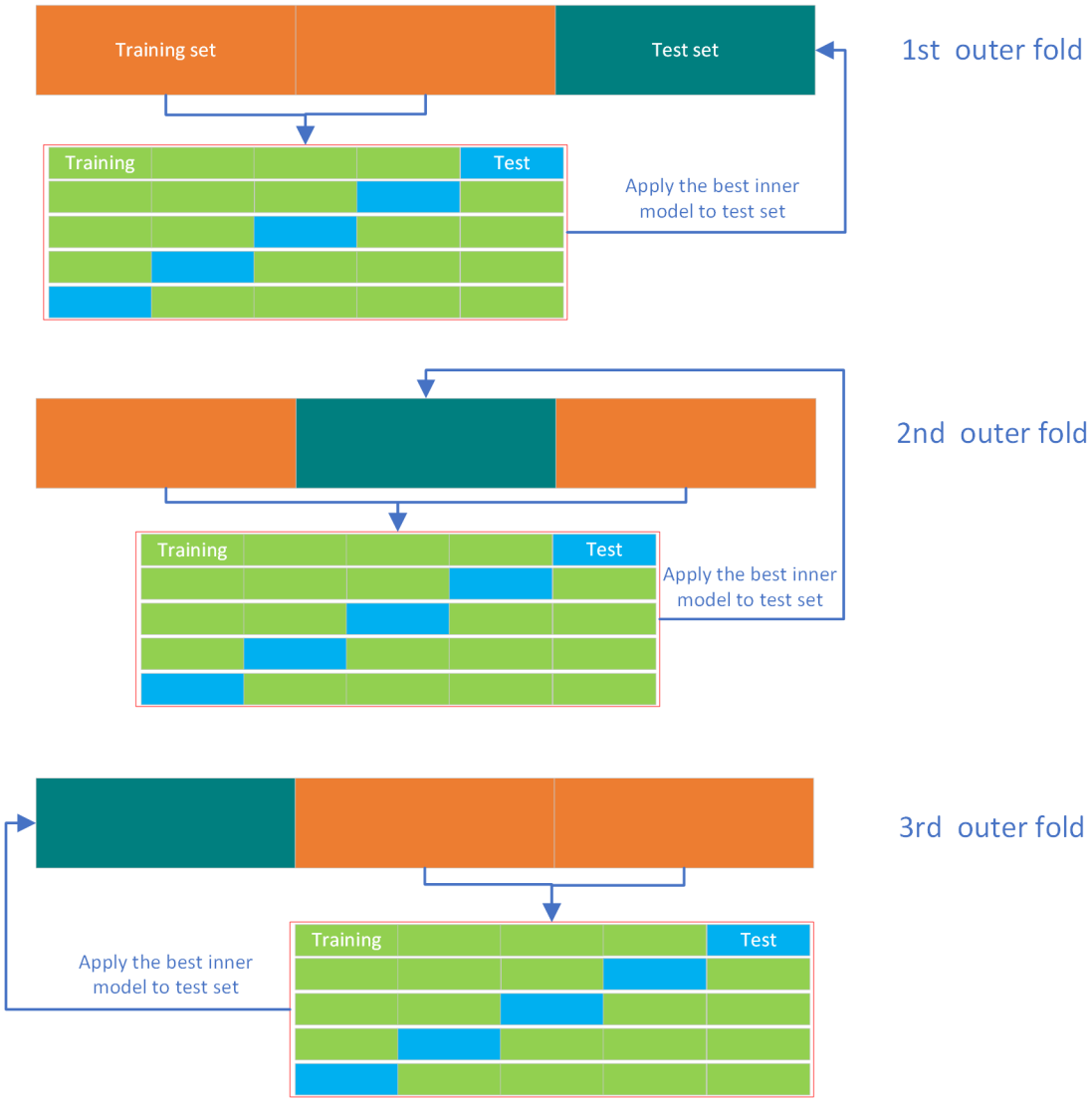


Figure 16: Flow chart of nested cross validation (3 outer folds and 5 inner folds).


Figure 17: Benchmark experiment menu: C-index comparison of survival models based on benchmark experiment.

- **Maxit:** Number of iterations for random search. Default is 100.
- **Inner fold:** Set the fold number for inner K-fold cross-validation.
- **Outer fold:** Set the fold number for outer K-fold cross-validation.
- **Bootstrap iterations:** Set the iteration for bootstrap validation.
- **Split the dataset ?:** Whether split the whole dataset into internal training and test set based on stratified sampling.
- **Split ratio:** The percentage of samples that goes to training set.

## 3.3 Prediction model

After benchmark experiment, CBioProfiler calculates risk score of each patients based on the coefficient (for Lasso, ridge, elastic net, glmboost, and Coxboost) or cumulative hazard function (for randomforestSRC) and the expression levels of the genes derived from the final model. Based on the risk score, CBioProfiler constructs and validates prediction model and draw a nomogram for the prediction of survival probability of patients.

### 3.3.1 Construct model

CBioProfiler constructs and evaluates prediction model using **Survival ROC curve**, **KM plot**, **Forestplot**, and **CoxPH table**.

Common parameters for prediction model:

- **Survival time:** Select survival time column. Example: “OS.time”, “RFS.time”, “PFS.time”, etc.

- **Survival status:** Select survival status column. Example: “OS”, “RFS”, “PFS”, etc.

**Survival ROC curve** is generated using R package “survivalROC” (Heagerty and Paramita Saha-Chaudhuri 2013).

Figure 18 shows Survival ROC curve for the prediction model in the training set (A) and test set (B).

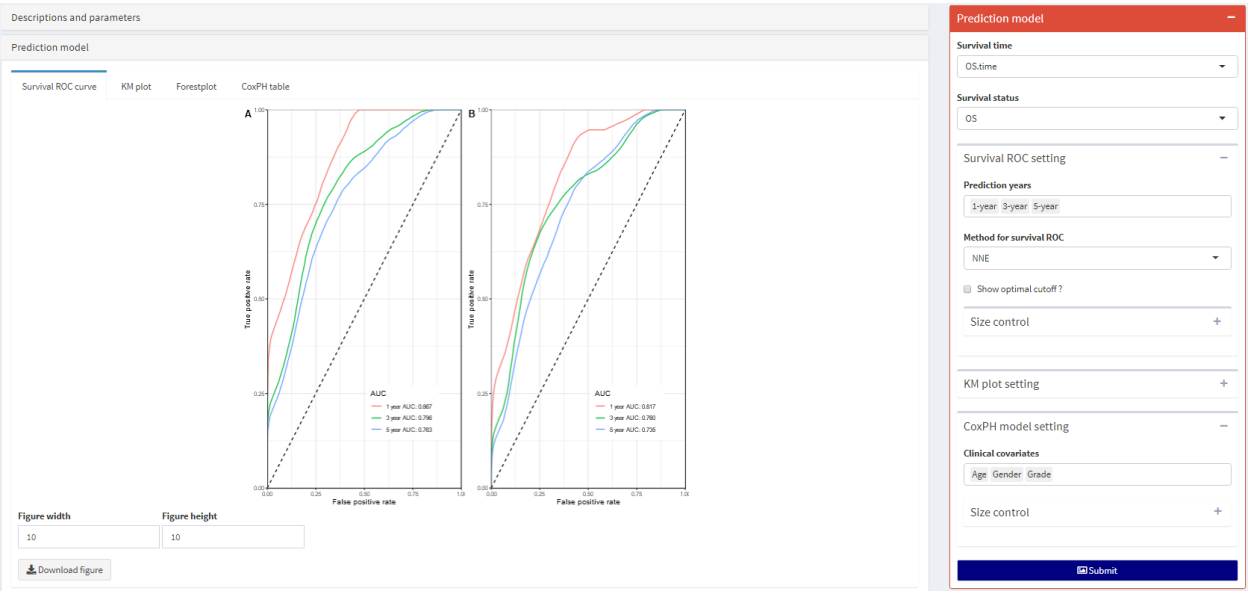


Figure 18: Construct model menu: Survival ROC curve.

Parameters for Survival ROC curve are:

- **Prediction years:** Define the time points in years the users want to predict based on time dependent ROC analysis. The longest time point should not exceed the max of survival (relapse) duration.
- **Method for survival ROC:** Method for fitting joint distribution of (marker,t), either of KM or NNE, the default method is NNE.
- **Show optimal cutoff ?:** Whether show the optimal cutoff of the risk score, which can be used to divided patients into different groups.

**KM plot** is generated using R package “survminer” (Kassambara, Kosinski, and Biecek 2020). Figure 19 shows KM plot for the prediction model in the training set (A) and test set (B).

Parameters specific for KM plot are:

- **Group by:** The method for dividing patients into different groups. “Percentage” means that patients are divided into different groups based on percentage. “Value” mean patients are divided into different groups based on a specific value of the risk score of patients, such as the optimal cutoff calculated according to survival ROC curve.
- **Cutoff percentage:** Input a cutoff (range:0-1) to categorize the samples into low and high risk group. For example if 0.25: 0%-25% = Low, 25%-100% high.
- **Cutoff value:** Input a specific cutoff value to categorize the samples into low and high risk groups. The cutoff value should exceed the range of the risk score.
- **Label of X axis:** Define the label of X axis of the KM plot.

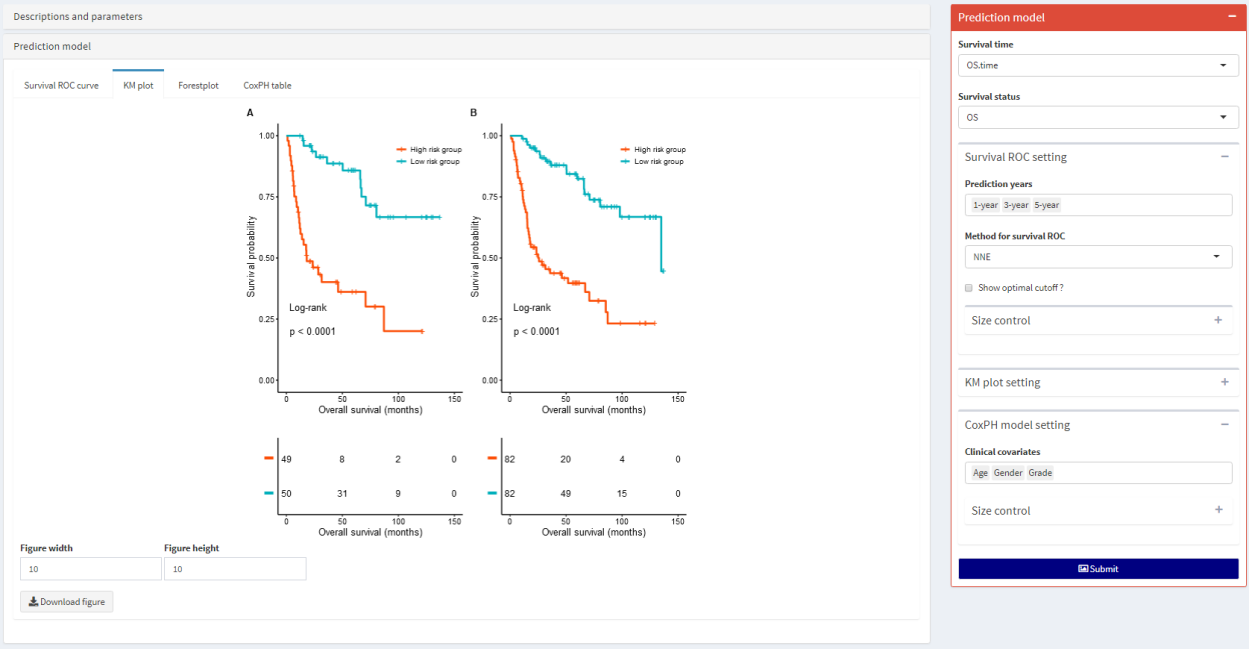


Figure 19: Construct model menu: KM plot.

- **Show p-value?:** Whether show the P value for the KM plot.
- **Show risk table?:** Whether show the risk table for the KM plot.
- **Show confidence interval?:** Whether show confidence interval for the KM plot.
- **Color of group 1:** Specify the color for the group 1 patient in the KM plot.
- **Color of group 2:** Specify the color for the group 2 patient in the KM plot.

**Forestplot** for the CoxPH model is drawn using the R package "ggplot2" (Wickham et al. 2020). Figure 20 shows KM plot for the prediction model in the training set (A) and test set (B).

Parameters specific for forest plot are:

- **Clinical covariates:** Select clinical variables to included to CoxPH model.
- **Variable names:** Define the variable names you interested,the number of variable names should be identical with the number of variables you included in the CoxPH model and use '|' to separate multiple variable names.
- **Maximum of xticks:** Define the maximum of xticks and clip to adjust the size of the forestplot.
- **Figure legend position:** Set the coordinates of the legend box. Their values should be between 0 and 1. c(0,0) corresponds to the 'bottom left' and c(1,1) corresponds to the 'top right' position.

### 3.3.2 Validate model

To validate the prediction model, the users need to select and load an external validation cohort. Moreover, the users should ensure that **all the selected biomarkers (features) identified by "Dimensionality reduction" should be included in the gene expression matrix of the external validation cohort**.

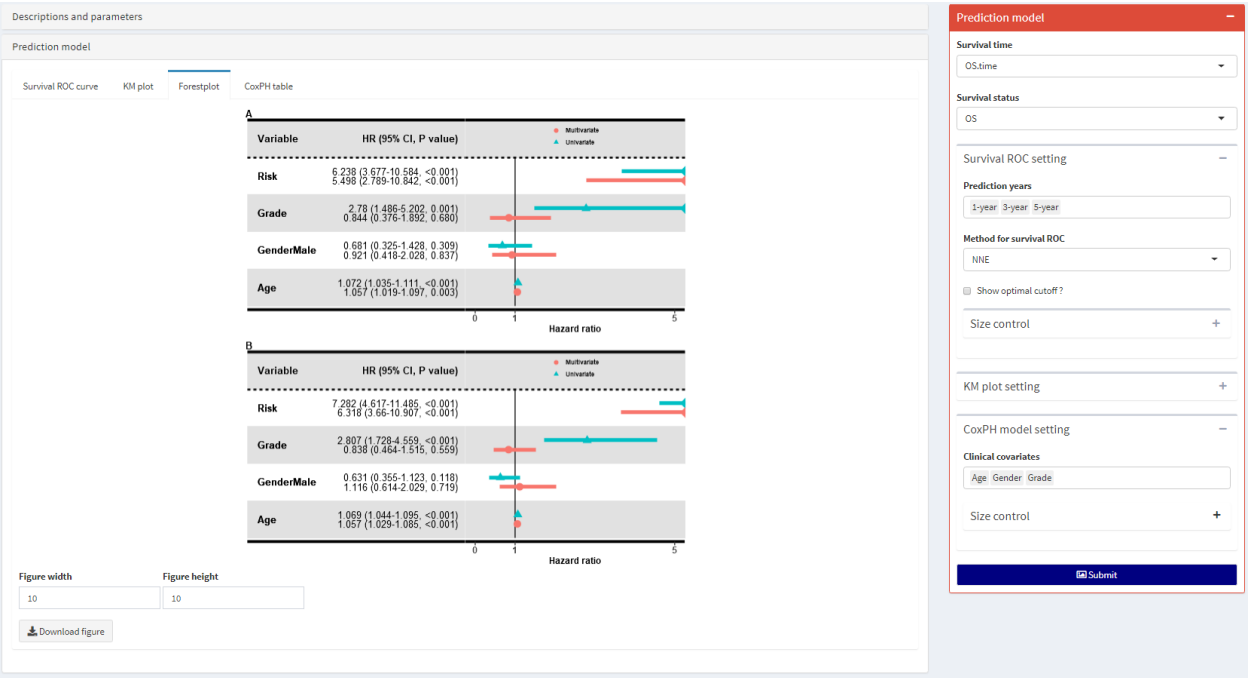


Figure 20: Construct model menu: Forest plot.

The validation process uses **Survival ROC curve**, **KM plot**, **Forestplot**, and **CoxPH table** to validate the prediction model, which is similar to **Construct model**, thus, the corresponding parameter settings are also similar. Figure 21 shows survival ROC curve for the prediction model in the validation set.

### 3.3.3 Nomogram

In order to maximize the clinical significance of biomarkers, CBioProfiler further draws nomoram on the basis of the biomarker calculated based on the benchmark experiment, so that researchers or clinicians can predict the long-term survival rate of tumor patients. The nomogram is generated using R package "rms" (Harrell, Jr. 2020). Figure 22 shows nomogram for the prediction model in the training set.

Parameters:

- **Survival time:** Select survival time column. Example: "OS.time", "RFS.time", "PFS.time", etc.
- **Survival status:** Select survival status column. Example: "OS", "RFS", "PFS", etc.
- **Clinical variable:** Select survival variable you want to included in the nomogram.
- **Variable names:** Define the variable names you interested,the number of variable names should be identical with the number of variables you included in the CoxPH model to contrust the nomogram and use '|' to separate multiple variable names.
- **Prediction years:** Define the time points in years you want to predict based on the nomogram. The longest time point should not exceed the max of survival (relapse) duration.

CBioProfiler allows the users to internally validate and calibrate the nomogram based on bootstrap and calibration analysis. Figure 23 shows internal validation for the prediction model in the training set and test set.

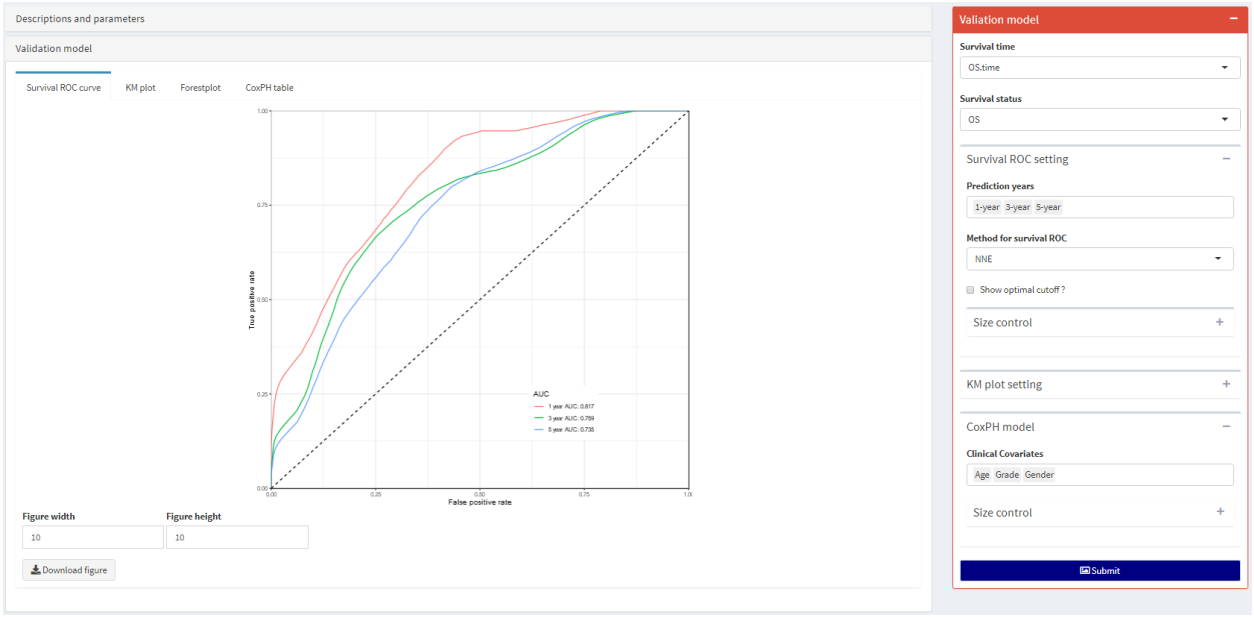


Figure 21: Validate model menu: Survival ROC curve.


Figure 22: Nomogram menu: Nomogram.


Figure 23: Nomogram menu: Internal validation.


Figure 24: Nomogram menu: Internal calibration.

Figure 24 shows internal calibration curve for the prediction model in the training set (A-C) and test set (D-F).

Parameters:

- **Prediction years:** Define the time points in years you want to predict based on the nomogram. The longest time point should not exceed the max of survival (relapse) duration.
- **Bootstrap iterations:** Number of bootstrap resamplings.
- **Ratio:** Ratio of resamplings.

Moreover, CBioProfiler allows the users to externally validate and calibrate the nomogram based on bootstrap and calibration analysis. For external validation, the users need to load an independent validation cohort.

Figure 25 shows external validation.

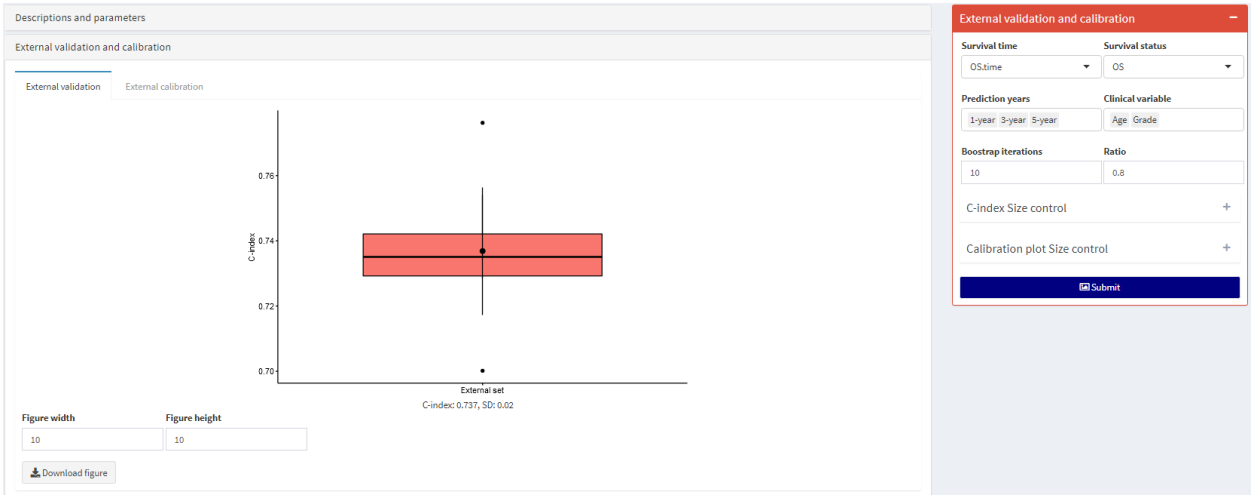


Figure 25: Nomogram menu: External validation.

Figure 26 shows external calibration.

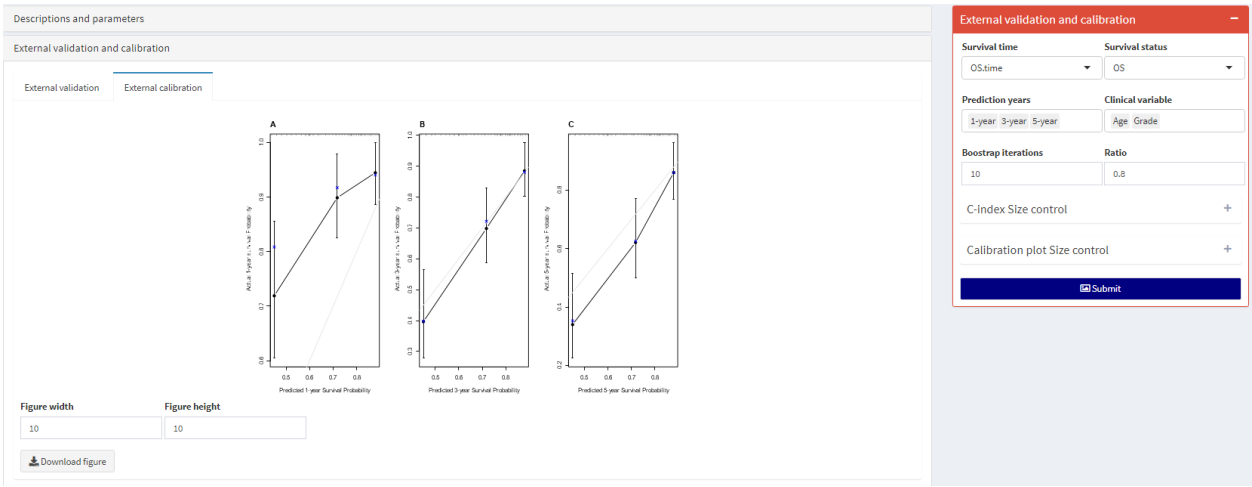


Figure 26: Nomogram menu: External calibration.

Parameters:

- **Survival time:** Select survival time column. Example: “OS.time”, “RFS.time”, “PFS.time”, etc.
- **Survival status:** Select survival status column. Example: “OS”, “RFS”, “PFS”, etc.
- **Prediction years:** Define the time points in years the users want to predict based on the nomogram. The longest time point should not exceed the max of survival (relapse/disease free) duration.
- **Clinical variable:** Select survival variable the users want to included.
- **Boostrap iterations:** Number of resamplings.
- **Ratio:** Ratio of resamplings.

## 3.4 Clinical annotation

CBioProfiler helps users investigate the clinical relevance of the biomarkers they identified through the above benchmark experiment.

### 3.4.1 Correlation with clinical features

The users can identify the correlation between the biomarker(gene) they are interested and the clinical features of patients through descriptive statistics table. The table is generated using R package “table1” (Rich 2020). When the table is generated, the users can copy it to Excel or OpenOffice. Figure 27 shows table for the correlations between gene expression and clinical features.

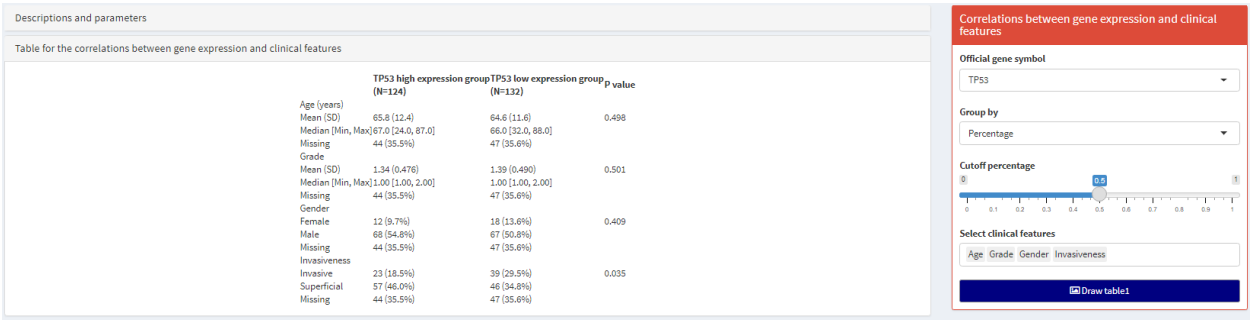


Figure 27: Correlation with clinical features menu: Table for the correlations between gene expression and clinical features main window.

Parameters:

- **Official gene symbol:** Input a gene with official gene symbol.
- **Group by:** Method for dividing the samples (patients) in to different groups. “Percentage” means dividing patients into different groups via a specific percentage. “Value” means groups patients into different groups via a specific value of the expression level of the gene.
- **Cutoff percentage:** Input a cutoff percentage (range:0-1) to categorize the samples into low and high expression group regarding to your interested gene. Example if 0.25: 0%-25% = Low, 25%-100% high.
- **Cutoff value:** Input a specific cutoff value to categorize the samples into low and high expression group regarding to your interested gene.
- **Select clinical features:** Select clinical features you want to include in the table.

### 3.4.2 Kaplan-Meier curve

Kaplan-Meier (KM) curve is generated using R package survminer (Kassambara, Kosinski, and Biecek 2020), through which, the users can analyze the survival difference between different groups based on the expression level of the biomarkers. Figure 28 shows Kaplan-Meier plot for TP53.

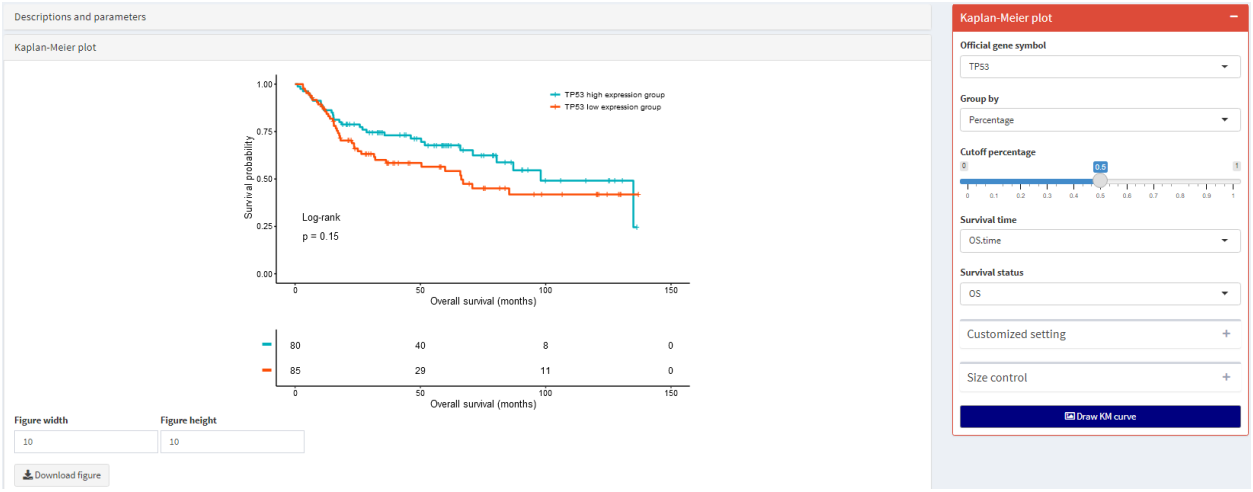


Figure 28: Kaplan-Meier curve menu: Kaplan-Meier plot main window.

Parameters:

- **Official gene symbol:** Input a gene with official gene symbol.
- **Group by:** Method for dividing the samples (patients) in to different groups. "Percentage" means dividing patients into different groups via a specific percentage. "Value" means groups patients into different groups via a specific value of the expression level of the gene.
- **Cutoff percentage:** Input a cutoff (range:0-1) to categorize the samples into low and high expression group regarding to your interested gene. For example if 0.25: 0%-25% = Low, 25%-100% high.
- **Cutoff value:** Input a specific cutoff value to categorize the samples into low and high expression group regarding to your interested gene.
- **Survival time:** Select survival time column. Example: "OS.time", "RFS.time", "PFS.time", etc.
- **Survival status:** Select survival time column. Example: "OS", "RFS", "PFS", etc.
- **Label of X axis:** Define the X axis label for the KM plot.
- **Show p-value?:** Whether show the p value for log-rank test for the KM plot.
- **Show risk table?:** Whether show the risk table for log-rank test for the KM plot.
- **Show confidence interval?:** Whether show the confidence interval for the KM plot.
- **Color of group 1/2:** Set the colors of group 1/2 for the KM plot.

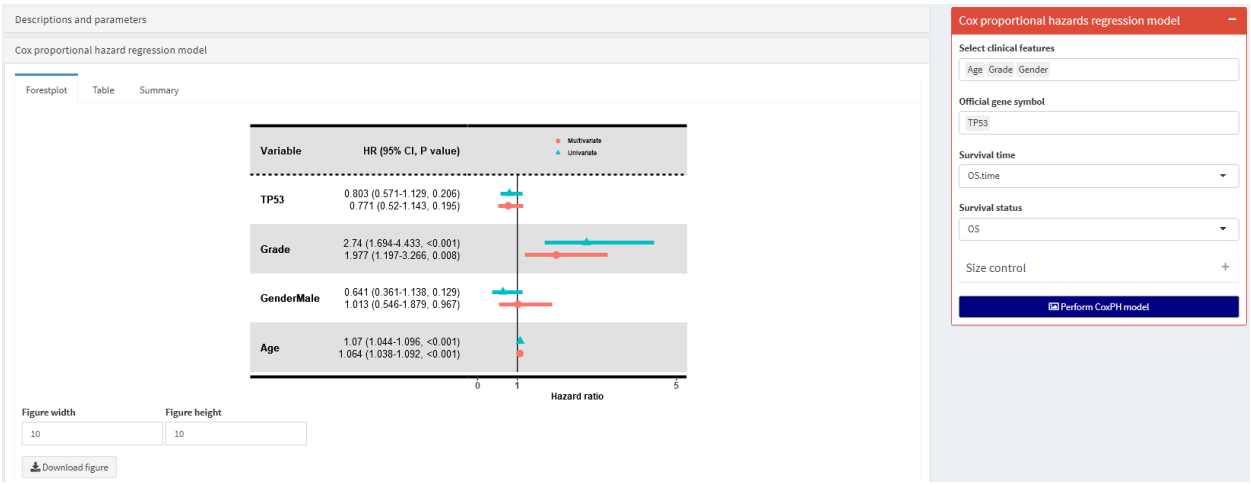


Figure 29: CoxPH model menu: Cox proportional hazard regression model main window.

### 3.4.3 CoxPH model

CBioProfiler uses Cox proportional hazards regression (CoxPH) model to help the users to characterize the prognostic value of the biomarker after adjusting for other clinical factors. CoxPH model was visualized using R package 'ggforestplot'(https://nightingalehealth.github.io/ggforestplot/articles/ggforestplot.html). Figure 29 shows forest plot for CoxPH model.

Parameters:

- **Select clinical features:** Select clinical variables to included to CoxPH model.
- **Official gene symbol:** Input one or more genes with official gene symbol that you want to included in the CoxPH model.
- **Survival time:** Select survival time column. Example: "OS.time", "RFS.time", "PFS.time", etc.
- **Survival status:** Select survival status column. Example: "OS", "RFS", "PFS", etc.
- **Figure legend position:** Set the coordinates of the legend box. Their values should be between 0 and 1. c(0,0) corresponds to the 'bottom left' and c(1,1) corresponds to the 'top right' position.
- **Variable names:** Define the variable names you interested,the number of variable names should be identical with the number of variables you included in the CoxPH model and use '|' to separate multiple variable names.
- **Maximum of xticks:** Define the maximum of xticks and clip to adjust the size of the forestplot.

### 3.4.4 Time-dependent ROC

Time-dependent ROC (survival ROC) analysis (Heagerty and Paramita Saha-Chaudhuri 2013) helps the users investigate the accuracy of the survival prediction of the biomarker they are interested in. Figure 30 shows time-dependent ROC analysis.

Parameters:

- **Official gene symbol:** Input a gene with official gene symbol.
- **Survival time:** Select survival time column. Example: "OS.time", "RFS.time", "PFS.time", etc.

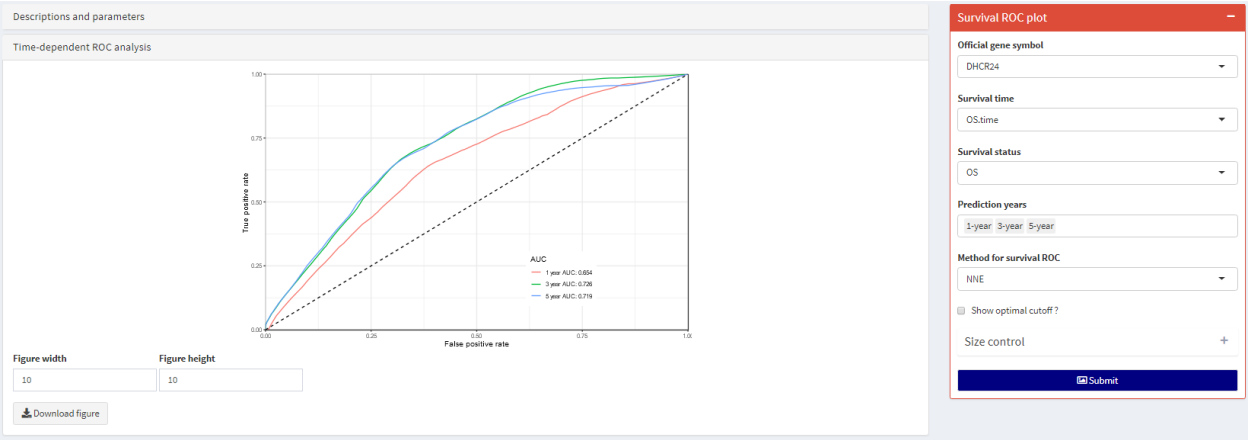


Figure 30: Time-dependent ROC menu: Time-dependent ROC analysis main window.

- **Survival status:** Select survival status column. Example: “OS”, “RFS”, “PFS”, etc.
- **Prediction years:** Define the time points in years you want to predict based on time dependent ROC analysis. the longest time point should not the max of survival (relapse/disease free) duration.
- **Method for survival ROC:** Define the method for survival ROC analysis, the default is “NNE”.

### 3.4.5 Most correlated genes

In the Most correlated gene module, the users can identify genes that are correlated with the gene they are interested in through Pearson’s correlation, Spearman’s correlation, and Kendall’s rank correlation. The correlation result can be visualized using **Bar plot**, **Bubble plot** and **Table**. Figure 31 shows most correlated genes with bar plot.

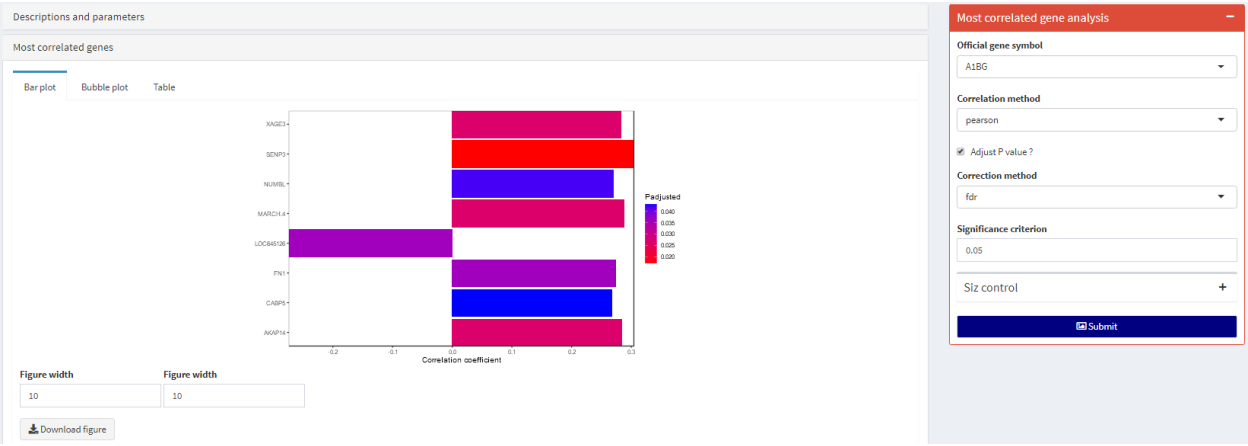


Figure 31: Most correlated genes menu: Most correlated genes main window.

Parameters:

- **Official gene symbol:** Select an official gene symbol you interested in.
- **Correlation method:** Specify the correction method, “pearson”,“spearman”,“kendall” means Pearson’s correlation, Spearman’s correlation, and Kendall’s rank correlation, respectively.

- **Adjust P value ?:** Whether to adjust the P value.
- **Correction method:** Specify the correction method for P values. For more details, please refer to https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/p.adjust.
- **Significance criterion:** Specify the criterion for significant correlation.

### 3.4.6 Correlation with specific gene

After identifying the most correlated genes, the users can visualize the correlation between a specific gene and the gene they are interested. Figure 32 shows the correlation between two genes.

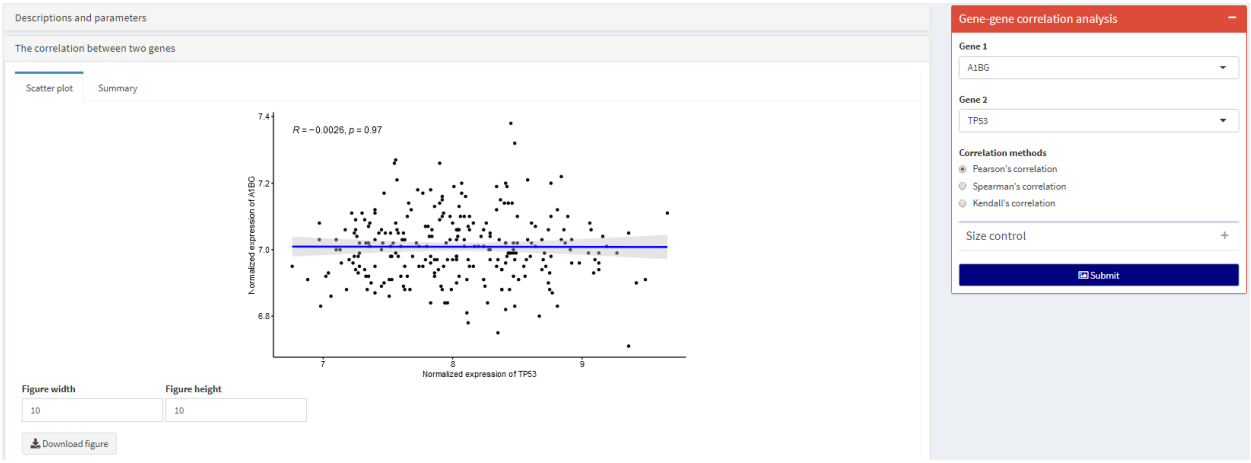


Figure 32: Correlation with specific gene: The correlation between two genes main window.

Parameters:

- **Gene 1:** Official gene symbol of gene 1.
- **Gene 2:** Official gene symbol of gene 2.
- **Correlation methods:** Specify the correlation method.

### 3.4.7 Gene expression in different groups

Users can also view the expression levels of related genes in different groups, which is achieved through the R package "ggpubr" (Kassambara 2020). The expression levels can be visualized using **Bar plot**, **Box plot**, and **Violin plot**. Figure 33 shows A1BG expression in different groups.

Parameters:

- **Official gene symbol:** Select an official gene symbol interested in.
- **Group:** Select a clinical variable to divide the gene expression into different groups.
- **Perform subgroup analysis ?:** Whether to perform subgroup analysis.
- **Subgroup:** Select a clinical variable as a subgroup indicator to compare the expression of the interested gene in subgroups.
- **Show P value ?:** Whether show the P value.

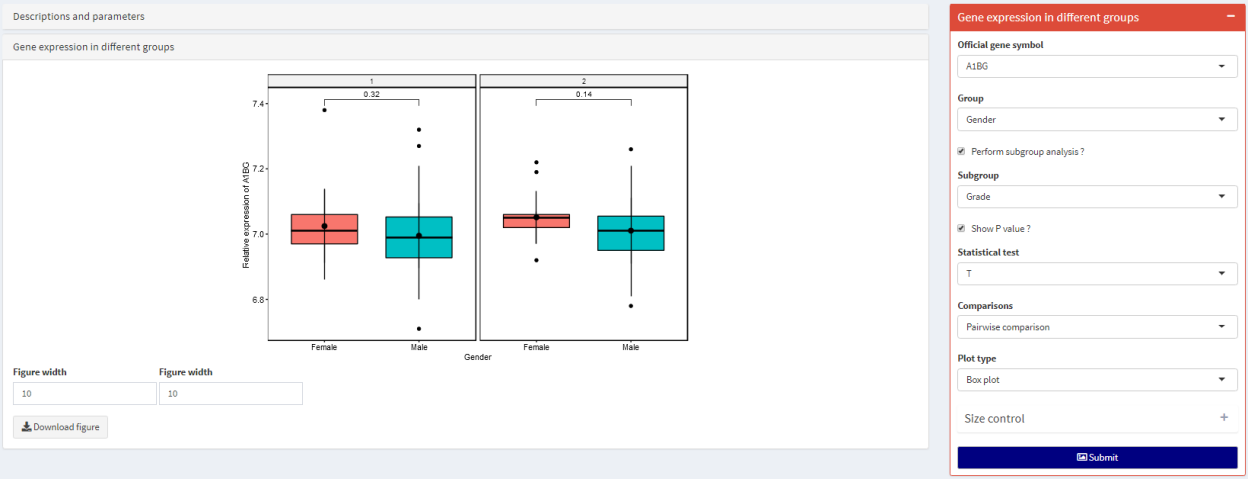


Figure 33: Gene expression in different groups menu: Gene expression in different groups main window.

- **Statistical test:** Select a statistical test for the comparisons, 'T' means 'T test', and 'wilcoxon' means 'wilcoxon test'.
- **Comparisons:** Specify the comparison pattern.
- **Comparison reference:** Specify the comparison reference.
- **Plot type:** Specify the type of the plot, one of **Bar plot**, **Box plot**, and **Violin plot**.

### 3.4.8 Correlation with immune infiltration

Immune cell infiltration, as an important component of the tumor microenvironment, exerts an important influence on the occurrence, development and outcome of tumors (Thorsson et al. 2018). CBioProfiler applies the R package ConsensusTME (Jiménez-Sánchez, Cast, and Miller 2019) which uses a consensus approach to estimate enrichment scores for multiple immune cells found within the tumor microenvironment using ssGSEA (Hänzelmann, Castelo, and Guinney 2013).CBioProfiler calculates enrichment score of immune cell types based on gene signatures from Bindea and colleagues (Bindea et al. 2013), Danaher and colleagues (Danaher et al. 2017), and Davoli and colleagues (Davoli et al. 2017), xCell(Aran, Hu, and Butte 2017), MCP-counter (Becht et al. 2016). Figure 34 shows the correlation between TP53 and immune cell infiltration.

Parameters:

- **Reference immune geneset:** Immune gene signatures from Bindea and colleagues, Danaher and colleagues, Davoli and colleagues, MCP-Counter, and xCell, respectively.
- **Official gene symbol:** Input one gene (biomarker candidate) with official gene symbol that you want to analyze the correlation between the gene and tummor microenvironment.
- **Correlation method:** Specify the correlation methods. "pearson" means Pearson's Correlation, "spearman" means Spearman's Correlation.
- **Adjust method for P:** Specify the adjusting methods for p value.
- **Select most correlated cells ?:** Select the most correlated immune cells based on the cutoff of adjusted p value?
- **Adjusted P cutoff:** Set the cutoff for adjusted P value.
- **Plot type:** Select the plot type for visualization.

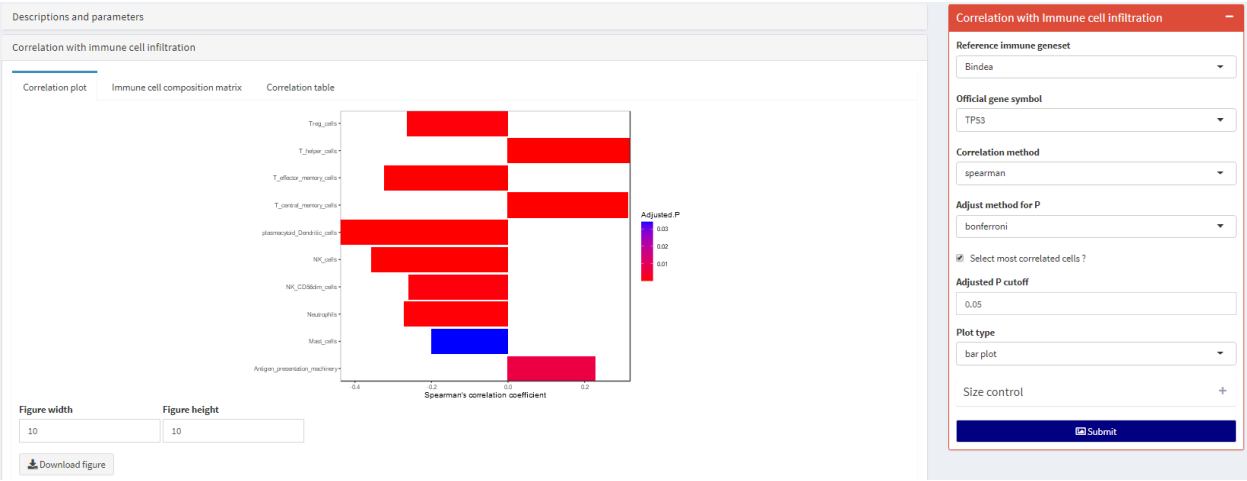


Figure 34: Correlation with immune infiltration menu: Correlation with Immune cell infiltration main window.

### 3.4.9 Correlation with stemness score

Malta et al (Malta et al. 2018). introduced a stemness score for assessing the degree of oncogenic dedifferentiation of cancer cells. Colaprico et al (Colaprico et al. 2016) developed R function to calculate the stemness score of samples in TCGA. We extend it to gene expression studies including but not limited to TCGA. CBioProfiler provides function to help users investigate the relationship between stemness score and the expression level of the genes they are interested in. Figure 35 shows the correlation between TP53 and stemness score.

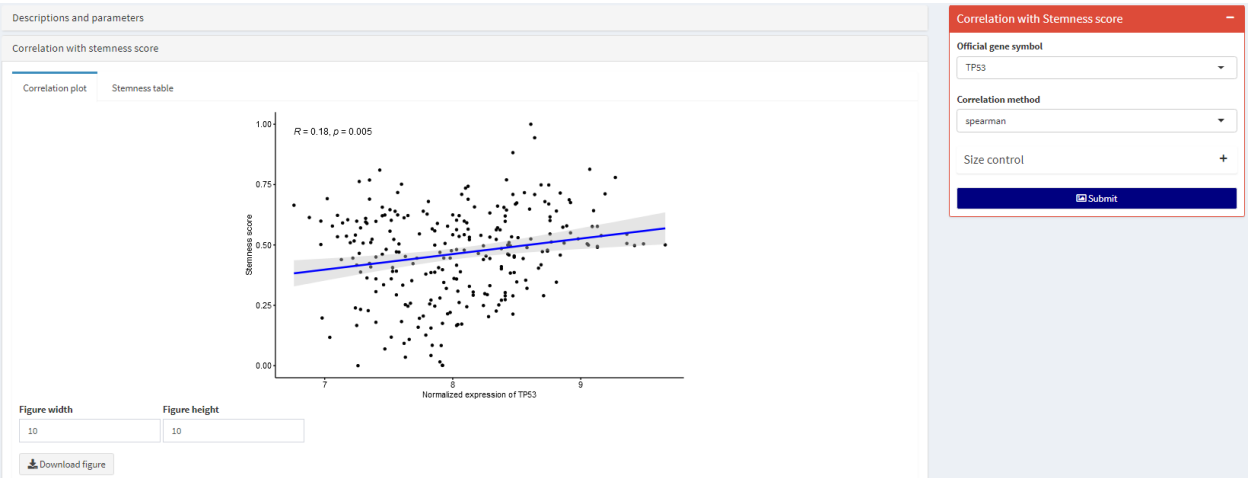


Figure 35: Correlation with stemness score menu: Correlation with stemness score main window.

Parameters:

- **Official gene symbol:** Input one gene (biomarker candidate) with official gene symbol that you want to analyze the correlation between the gene and stemness score.
- **Correlation method:** Specify the correlation methods. "pearson" means Pearson's Correlation, "spearman" means Spearman's Correlation.

### 3.4.10 Correlation with ESTIMATE score

Yoshihara et al.(Yoshihara et al. 2013) developed an R package ESTIMATE (Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression data) aiming at predicting tumor purity, and the presence of infiltrating stromal/immune cells in tumor tissues using gene expression data based on single sample Gene Set Enrichment Analysis and generates three scores:

- *1)* stromal score (that captures the presence of stroma in tumor tissue),
- *2)* immune score (that represents the infiltration of immune cells in tumor tissue), and
- *3)* estimate score (that infers tumor purity).

Thus, we applied the ESTIMATE algorithm to calculate the relationship between tumor biomarkers and stromal cells and immune cells in the tumor microenvironment. Figure 36 shows the correlation between TP53 and ESTIMATE score.

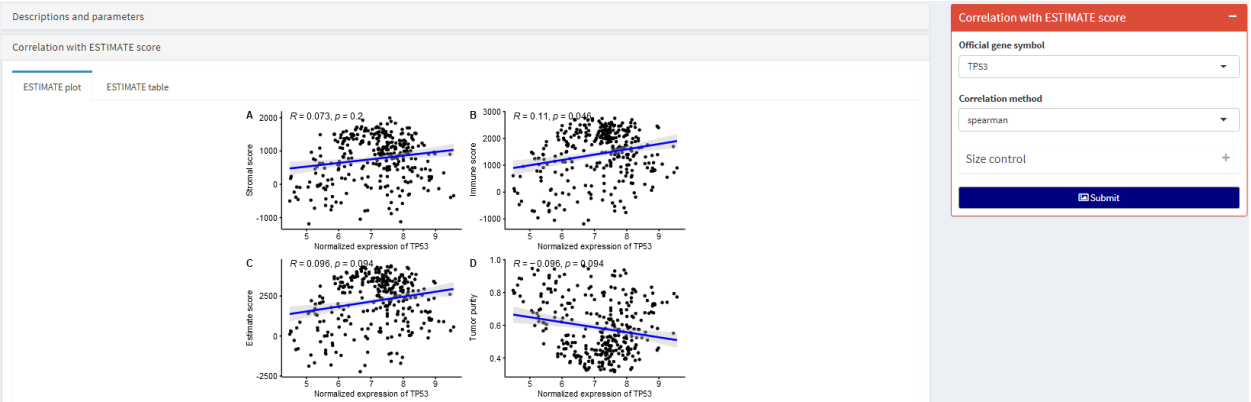


Figure 36: Correlation with ESTIMATE score menu: Correlation with ESTIMATE score main window.

Parameters:

- **Official gene symbol:** Input one gene (biomarker candidate) with official gene symbol that you want to analyze the correlation between the gene and ESTIMATE score.
- **Correlation method:** Specify the correlation methods. "pearson" means Pearson's Correlation, "spearman" means Spearman's Correlation.

### 3.4.11 Correlation with immune checkpoint molecules

Immune checkpoint molecules are regulators of the immune system. These pathways are crucial for self-tolerance, which prevents the immune system from attacking cells indiscriminately. However, some cancers can protect themselves from attack by stimulating immune checkpoint targets.(Pardoll 2012)

Inhibitory checkpoint molecules (Nirschl and Drake 2013; Sharma and Allison 2015), including cytotoxic T lymphocyte antigen-4 (CTLA-4), programmed death-1 (PD-1), lymphocyte activation gene-3 (LAG-3), T-cell immunoglobulin and mucin protein-3 (TIM-3),etc., are targets for cancer immunotherapy due to their potential for use in multiple types of cancers. The expression of these checkpoint molecules on T cells represents an important mechanism that the immune system uses to regulate responses to self-proteins. Recent clinical data show that these Checkpoint molecules play a critical role in objective tumor responses and improved overall survival. Therefore, CBioProfiler help users identify the correlation between the immune checkpoint molecules and the cancer biomarker they selected. Figure 37 shows the correlation between TP53 and immune checkpoint molecules.

Parameters:

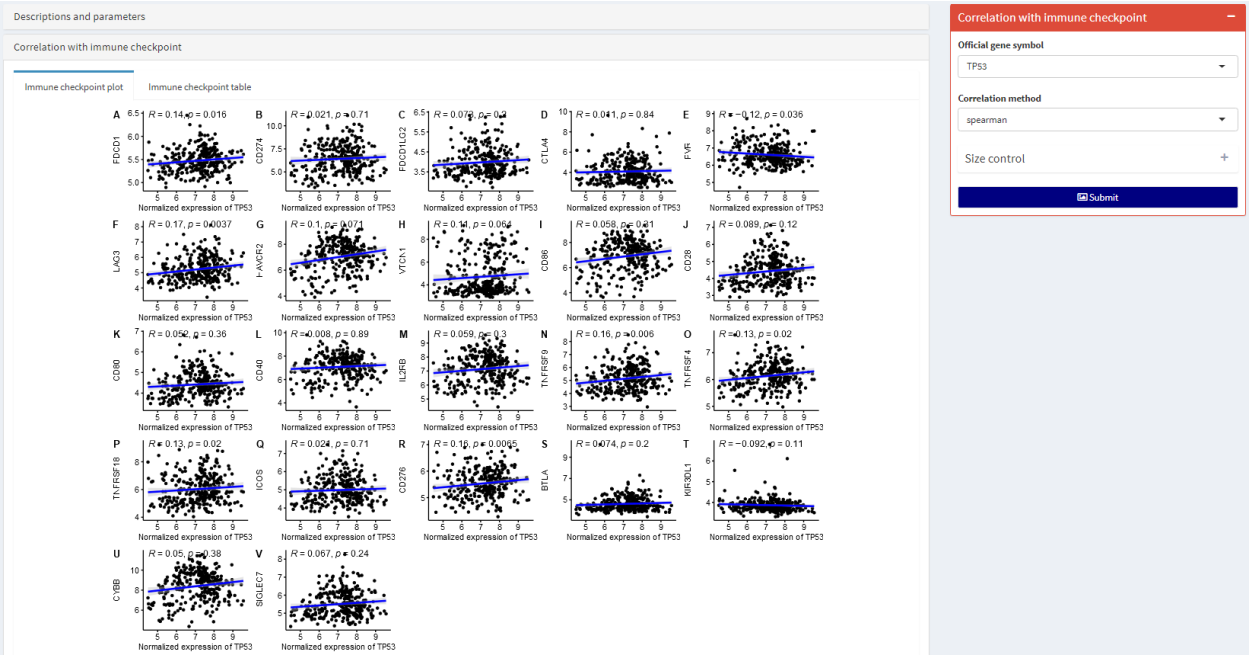


Figure 37: Correlation with immune checkpoint molecules menu: Correlation with immune checkpoint molecules main window.

- **Official gene symbol:** Input one gene (biomarker candidate) with official gene symbol that you want to analyze the correlation between the gene and immune checkpoint molecules.
- **Correlation method:** Specify the correlation methods. "pearson" means Pearson's Correlation, "spearman" means Spearman's Correlation.

### 3.4.12 Correlation with interferon-gamma score

Interferon-gamma (IFN-gamma) plays a crucial role in the regulation of antitumor immunity: mainly secreted by activated lymphocytes such as CD8 cytotoxic T-cells or CD4 T-helper cells type I (Th1), IFN-gamma can enhance Th1-mediated antitumor immune response in terms of a positive feedback loop.(Castro et al. 2018) IFN-gamma is also known to play a protumorigenic role by transmitting antiapoptotic and proliferative signals, resulting in immune-escape of tumor cells. CBioProfiler allows users to estimate the relationship between the biomarker they selected and the IFN-gamma score calculated based on ssGSEA (Hänzelmann, Castelo, and Guinney 2013). Figure 38 shows the correlation between TP53 and interferon-gamma score.

Parameters:

- **Official gene symbol:** Input one gene (biomarker candidate) with official gene symbol that you want to analyze the correlation between the gene and interferon-gamma score.
- **Correlation method:** Specify the correlation methods. "pearson" means Pearson's Correlation, "spearman" means Spearman's Correlation.

### 3.4.13 Correlation with cytolytic activity

Based on the notion that effective natural anti-tumor immunity requires a cytolytic immune response, Rooney et al. (Rooney et al. 2015) quantified cytolytic activity using a simple expression metric of effector molecules that mediate cytolysis. They demonstrated that cytolytic activity was associated with MHC Class

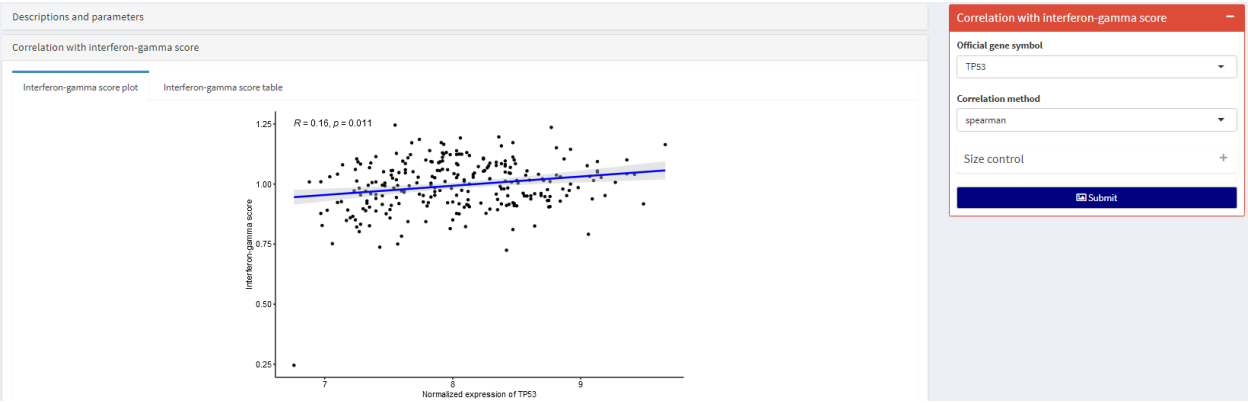


Figure 38: Correlation with interferon-gamma score menu: Correlation with interferon-gamma score main window.

I-associated neoantigens, gene mutations (including including beta-2-microglobulin (B2M), HLA-A, -B and -C and Caspase 8 (CASP8)) that highlighted loss of antigen presentation and blockade of extrinsic apoptosis, and genetic amplifications (PDL1/2 and ALOX12B/15B). Herein, CBioProfiler allows users calculate the correlation between the biomarker and cytolytic activity. Figure 39 shows the correlation between TP53 and cytolytic activity.

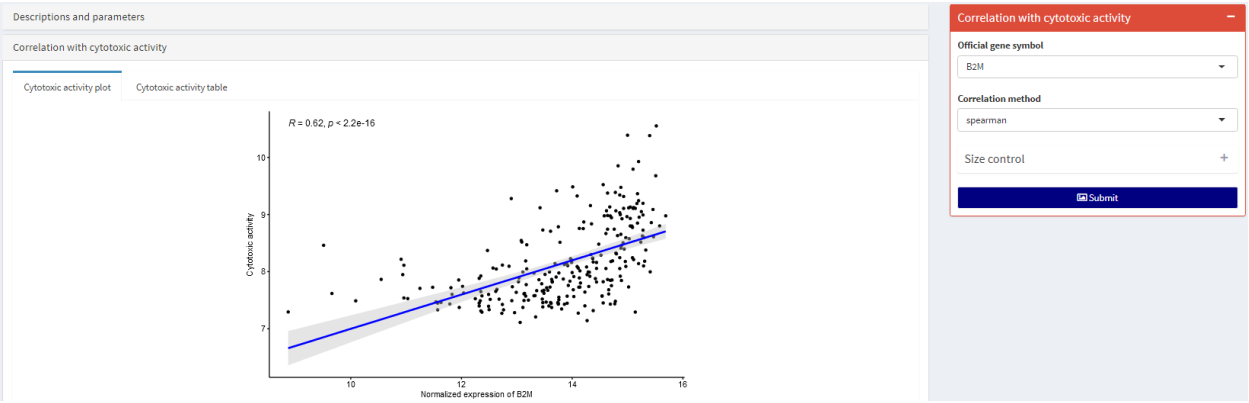


Figure 39: Correlation with cytolytic activity menu: Correlation with cytolytic activity main window.

Parameters:

- **Official gene symbol:** Input one gene (biomarker candidate) with official gene symbol that you want to analyze the correlation between the gene and cytolytic activity.
- **Correlation method:** Specify the correlation methods. "pearson" means Pearson's Correlation, "spearman" means Spearman's Correlation.

### 3.4.14 Correlation with cancer pathway

Cancer related pathways, including Cell cycle, Chromatin remodeling, Differentiation and development, DNA damage, Immune regulation, MAPK and PI3K pathway, Metabolism, PI3K pathway, RAS pathway, RNA metabolism, RTK pathway, TGFB pathway, Transcription regulation, WNT signaling, have been demonstrated to play an important role in the oncogenesis and progression of multiple human cancers. Davoli et al.(Davoli et al. 2017) curated key genes in these cancer related pathways, and we utilize these

gene signatures to help users to identify the relationship between the biomarker and the cancer related pathways based on ssGSEA algorithm (Hänzelmann, Castelo, and Guinney 2013). Figure 40 shows the correlation between TP53 and cancer related pathway.

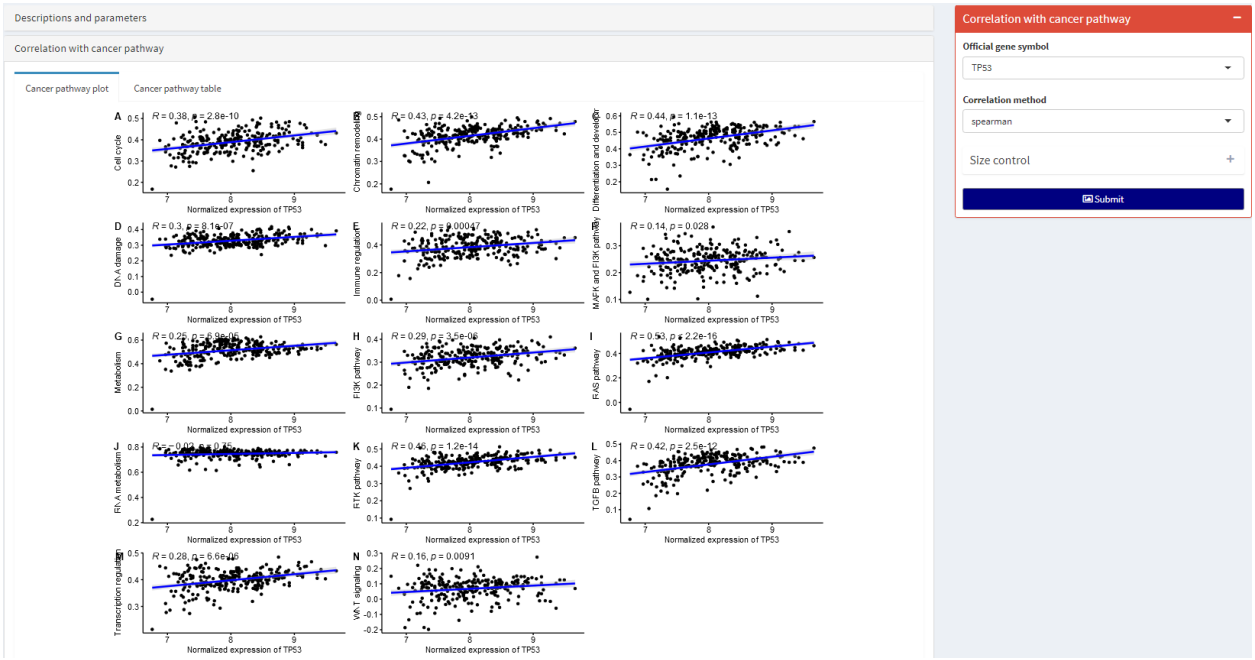


Figure 40: Correlation with cancer pathway menu: Correlation with cancer pathway main window.

Parameters:

- **Official gene symbol:** Input one gene (biomarker candidate) with official gene symbol that you want to analyze the correlation between the gene and cancer related pathways.
- **Correlation method:** Specify the correlation methods. "pearson" means Pearson's Correlation, "spearman" means Spearman's Correlation.

### 3.4.15 Correlation with metabolic pathway

Tumor initiation and progression depend on cellular metabolism. A common feature of tumor cell metabolism is the ability to obtain essential nutrients from a nutrient-poor environment and use these nutrients to maintain viability and generate new biomass. Alterations in intracellular and extracellular metabolites can have profound effects on gene expression, cell differentiation, and the tumor microenvironment along with tumor-associated metabolic reprogramming. There are six main characteristics of changes in tumor metabolism, but only a few tumors exhibit these six characteristics at the same time. According to the specific characteristics exhibited by tumors, it may help to better guide tumor classification and treatment. Ricketts et al.(Ricketts et al. 2018) curated common tumor metabolic pathways, including, AMPK Complex, Complex I, Complex III, Complex IV - cytochrome C, Complex V, Fatty Acid Synthesis, Glycogen Metabolism, Glycolysis, Krebs Cycle - Cyto, Krebs Cycle - Mito, NADH->NADPH, PDC Activation, PDC Suppression, Ribose Sugar Metabolism, and Serine Metabolism. In the present study, CBioProfiler helps the users to characterize the relationship between the biomarker they selected and the common metabolic pathways based on ssGSEA algorithm (Hänzelmann, Castelo, and Guinney 2013). Figure 41 shows the correlation between TP53 and metabolic pathway.

Parameters:

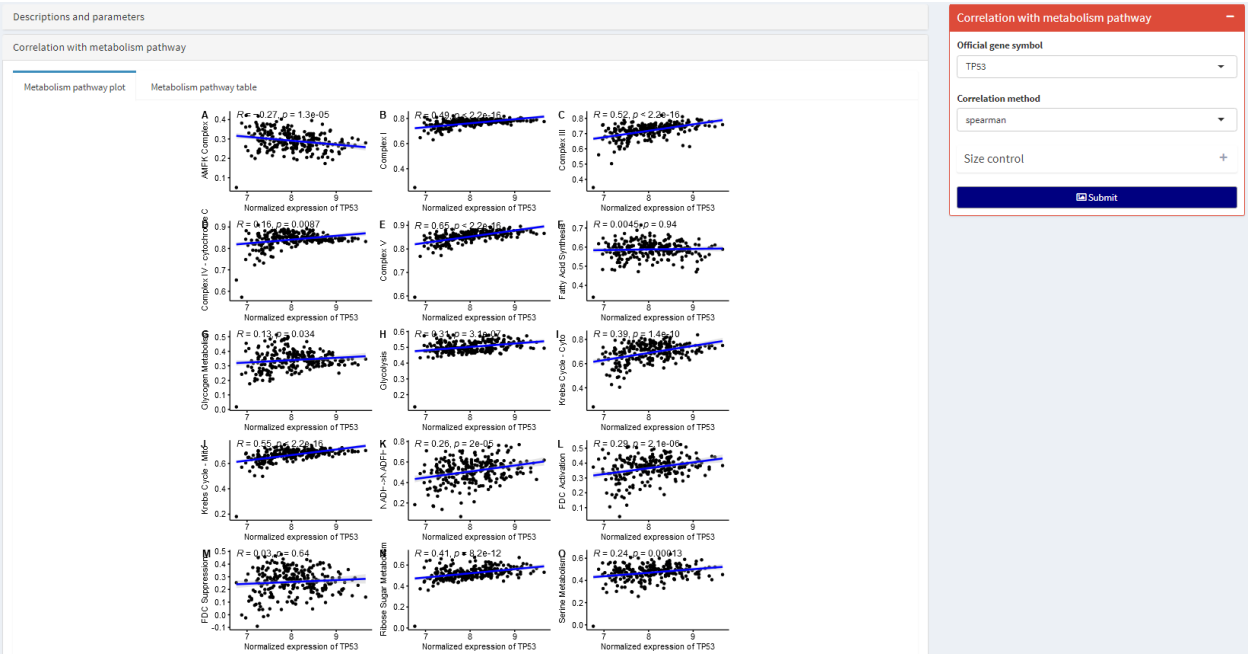


Figure 41: Correlation with metabolic pathway menu: Correlation with metabolic pathway main window.

- **Official gene symbol:** Input one gene (biomarker candidate) with official gene symbol that you want to analyze the correlation between the gene and metabolic pathways.
- **Correlation method:** Specify the correlation methods. "pearson" means Pearson's Correlation, "spearman" means Spearman's Correlation.

### 3.4.16 Correlation with halllmark gene signature

Hallmark gene sets are a group of 50 gene set that summarize and represent specific well-defined biological states or processes and display coherent expression. These gene sets were generated by a computational methodology based on identifying gene set overlaps and retaining genes that display coordinate expression (Liberzon et al. 2015; Dolgalev 2020). CBioProfiler helps the users to characterize the relationship between the biomarker they selected and hallmark gene sets based on ssGSEA algorithm (Hänzelmann, Castelo, and Guinney 2013). Figure 42 shows the correlation between TP53 and halllmark gene signature.

Parameters:

- **Official gene symbol:** Input one gene (biomarker candidate) with official gene symbol that you want to analyze the correlation between the gene and hallmark gene sets.
- **Correlation method:** Specify the correlation methods. "pearson" means Pearson's Correlation, "spearman" means Spearman's Correlation.

### 3.4.17 Correlation with drug response

Geeleher et al. recently introduced a methodology that worked by training and building statistical models from gene expression profile and drug sensitivity data in a very large panel of cancer cell lines, then applying these models to gene expression data from primary tumor biopsies (Geeleher, Cox, and Huang 2014a). They also created an R package called pRRophetic (Geeleher, Cox, and Huang 2014b), which extends the previously described pipeline and allows prediction of clinical drug response for many cancer drugs in an

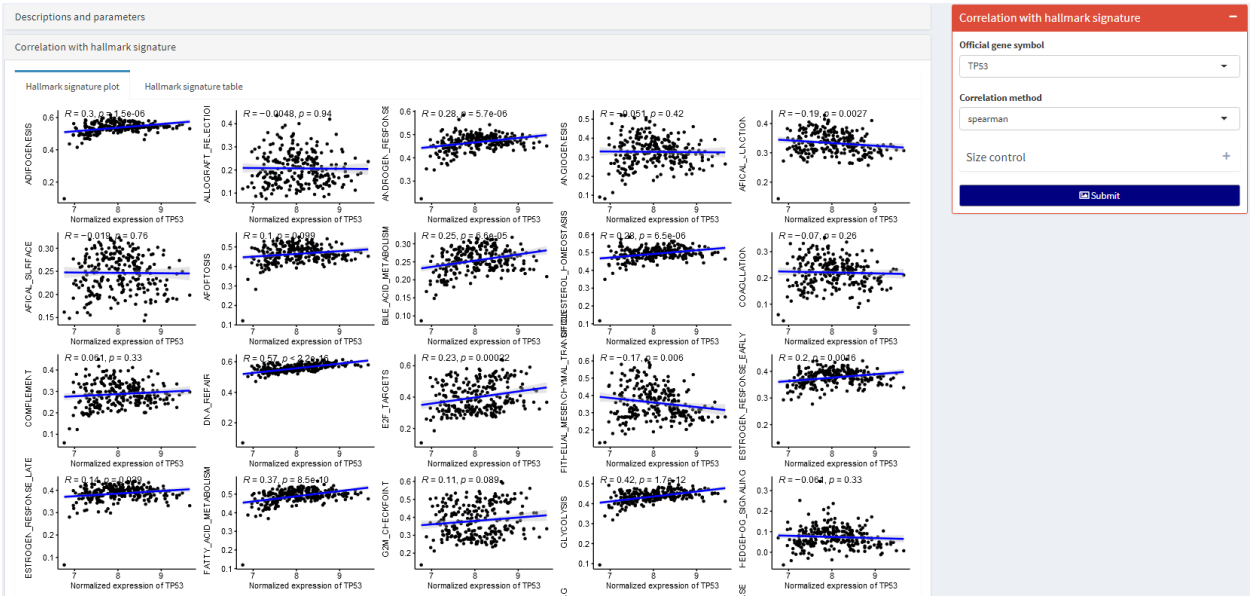


Figure 42: Correlation with halllmark gene signature menu: Correlation with halllmark gene signature main window.

R environment. CBioProfiler takes advantage of pRRophetic to help users predict the drug sensitivity and estimate the relationship between chemotherapy drugs and the biomarker they selected. Figure 43 shows the correlation between TP53 and drug response.

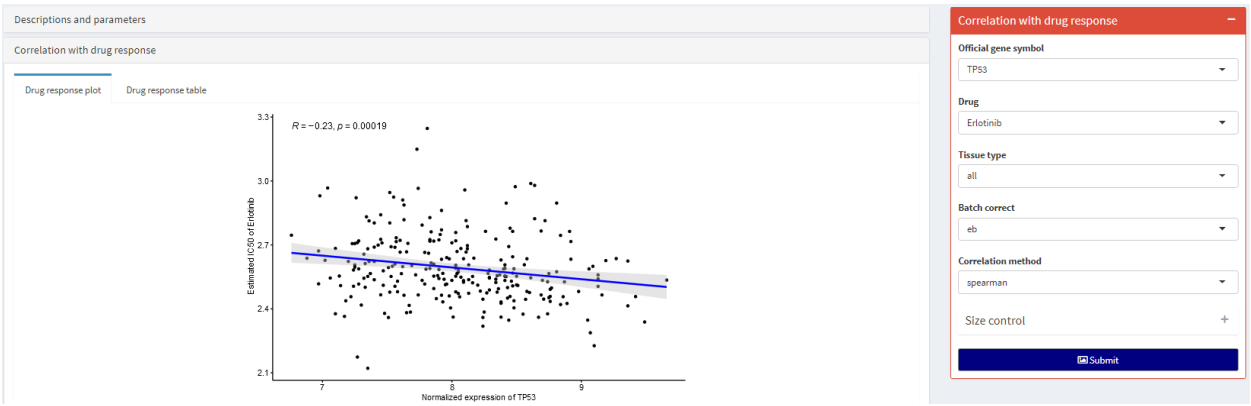


Figure 43: Correlation with drug response menu: Correlation with drug response main window.

Parameters:

- **Official gene symbol:** Input one gene (biomarker candidate) with official gene symbol that you want to analyze the correlation between the gene and drug response.
- **Drug:** The name of the drug for which you would like to predict sensitivity.
- **Tissue type:** Specify if you would like to traing the models on only a subset of the CGP cell lines (based on the tissue type from which the cell lines originated). This be one any of “all” (for everything, default option), “allSolidTumors” (everything except for blood), “blood”, “breast”, “CNS”, “GI tract” ,“lung”, “skin”, “upper aerodigestive”.

- **Batch correct:** How should training and test data matrices be homogenized. Choices are “eb” (default) for ComBat, “qn” for quantiles normalization or “none” for no homogenization.
- **Correlation method:** Specify the correlation methods. “pearson” means Pearson’s Correlation, “spearman” means Spearman’s Correlation.

## 3.5 Biological annotation

Yu and colleagues developed clusterProfiler (Yu 2021), a very outstanding R language package for gene functional annotation. CBioProfiler integrates some important functions of clusterProfiler to annotate tumor markers developed by users. CBioProfiler allows users to perform functional anntation of their biomarkers regarding gene ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Molecular Signatures Database (MSigDb), and Reactome pathway based on over representation analysis (ORA) and gene set enrichment analysis (GSEA) and to visualize their functional annotation results using **Bar plot**, **Dot plot**, **Gene-concept network**, **Heatmap**, **Enrichment Map**, **Ridgeline plot**, **Geneset enrichment plot1**, **Geneset enrichment plot2**. Figure 44 shows the output of functional Enrichment analysis.

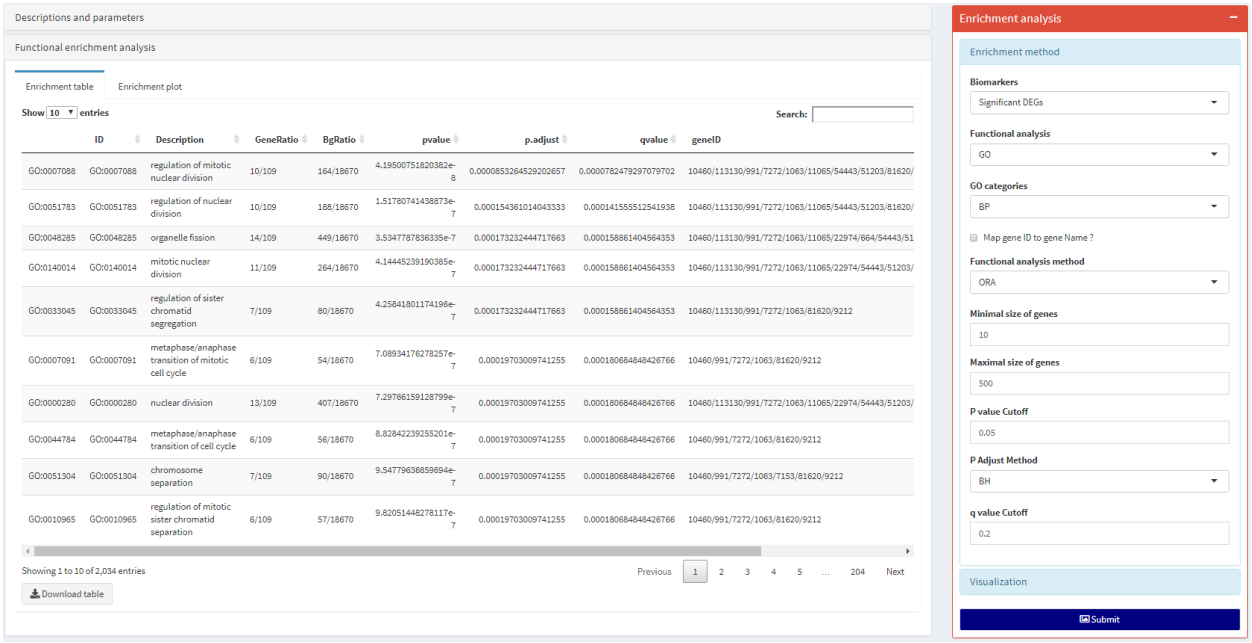


Figure 44: Biological annotation menu: Functional Enrichment analysis main window.

Parameters:

- **Biomarkers:** Specify the source of biomarkers. ‘Significant DEGs’ means significantly differential experssion genes at the cutoff you specified in the ‘Differentially expressed genes’ module; ‘Significant SRGs’ means significant survival-related genes at the cutoff you specified in the ‘Survival related genes’ module; ‘Network hub genes’ means genes in one or all non-grey modules you selected in the ‘WGCNA’ module; ‘Genes from benchmark experiment’ means genes derived from benchmark experiment based on cross-validation or nested cross-validation.
- **Functional analysis:** Define the functional enrichment analysis methods,GO, gene ontology analysis; KEGG, Kyoto Encyclopedia of Genes and Genomes analysis; MsigDb, Molecular Signatures Database analysis; Reactome Pathway, Reactome Pathway analysis.

- **GO categories:** Define the subcategory of gene ontology.BP, biological process; CC, cellular component;MF, molecular function; 'All' means performing functional GO analysis based on the all three subcategories.
- **Functional analysis method:** Define the method for functional enrichment analysis, "ORA" means over representation analysis, "GSEA" means gene set enrichment analysis.
- **GSEA algorithm:** Specify the GSEA method.
- **Maximal size of genes:** Maximal size of genes annotated by functional term for testing.
- **P value Cutoff:** Adjusted pvalue cutoff on enrichment tests to report.
- **P Adjust Method:** Adjusting method for P.
- **q value Cutoff:** Cutoff value for q.
- **Plot type:** visualization method for enrichment result. Please note! Only 'Bar plot', 'Dot plot', 'Gene-concept network', 'Enrichment Map' are suitable for visualizing the results of ORA analysis, while only 'Dot plot', 'Gene-concept network', 'Heatmap', 'Enrichment Map', 'Ridgeline plot', 'Geneset enrichment plot1', 'Geneset enrichment plot2 are suitable for visualizing the result of GSEA.

### Bar plot

Figure 45 shows the bar plot for output of functional Enrichment analysis.


Figure 45: Biological annotation menu: Bar plot.

### Dot plot

Figure 46 shows the dot plot for output of functional Enrichment analysis.

### Gene-concept network

Figure 47 shows the gene-concept network plot for output of functional Enrichment analysis.

### Heatmap

Figure 48 shows the heatmap for output of functional Enrichment analysis.

### Enrichment Map

Figure 49 shows the enrichment Map for output of functional Enrichment analysis.

### Ridgeline plot

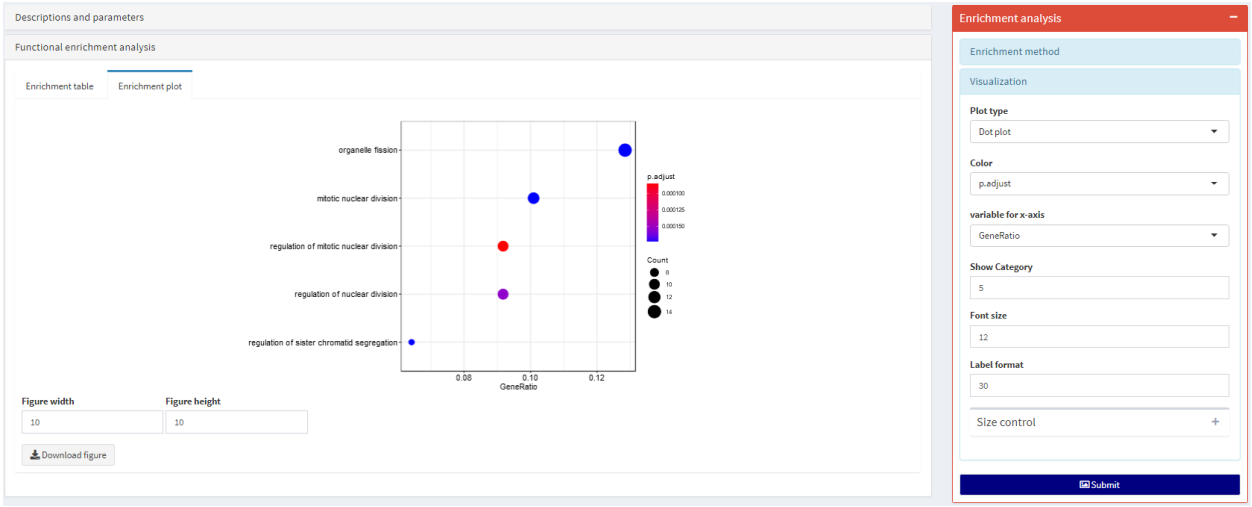


Figure 46: Biological annotation menu: Dot plot.

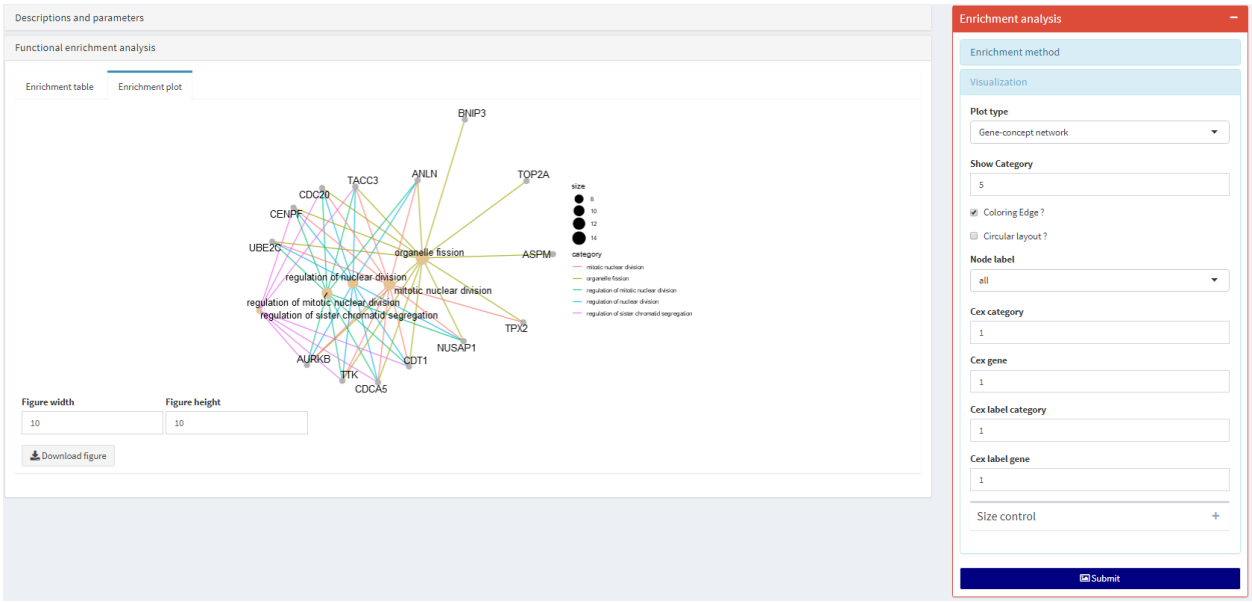


Figure 47: Biological annotation menu: Gene-concept network.

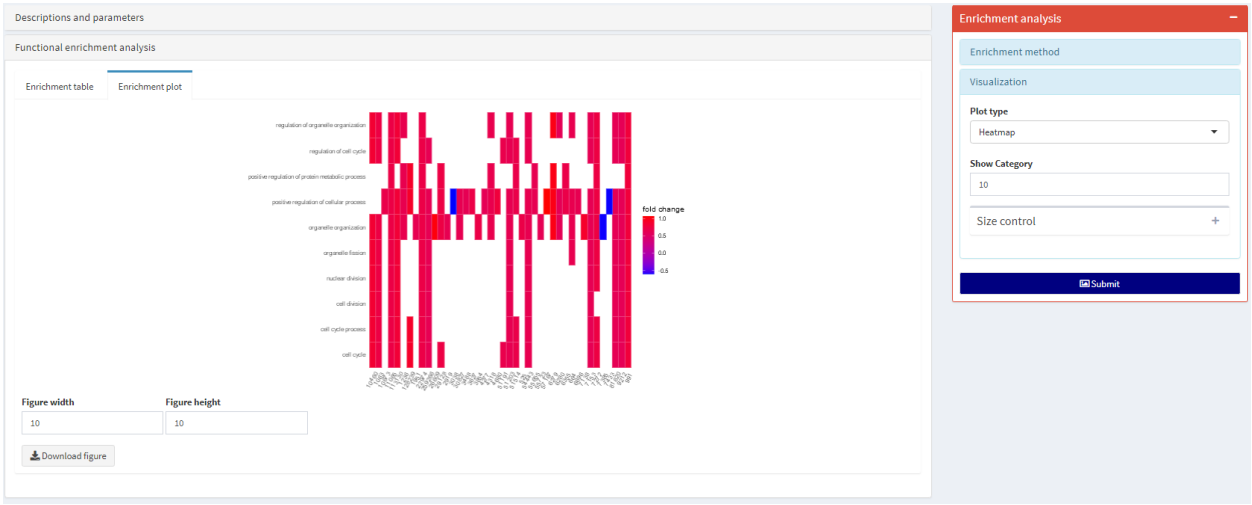


Figure 48: Biological annotation menu: Heatmap.


Figure 49: Biological annotation menu: Enrichment Map.

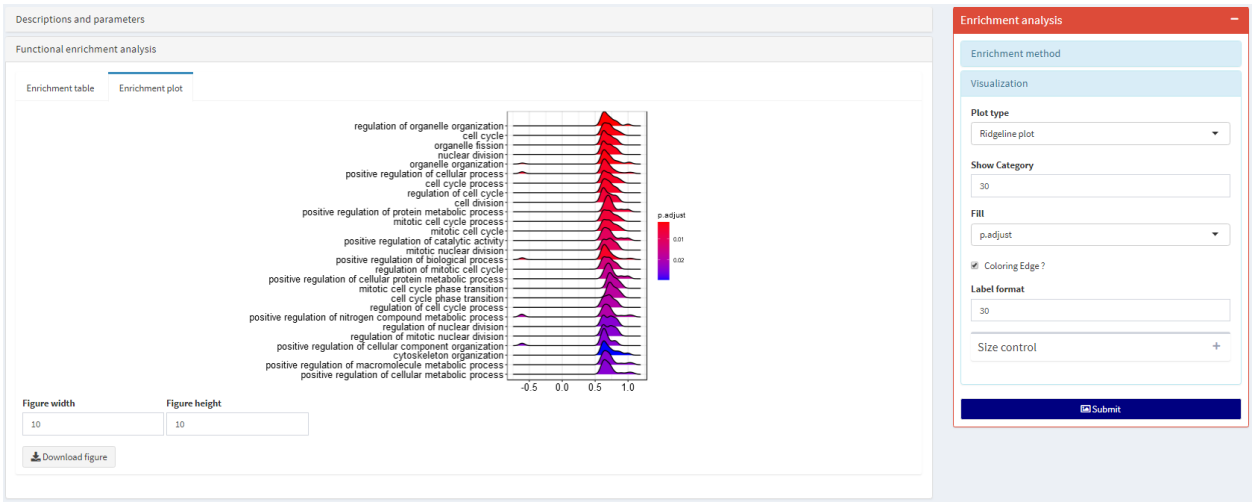


Figure 50: Biological annotation menu: Ridgeline plot.

Figure 50 shows the ridgeline plot for output of functional Enrichment analysis.

**Geneset enrichment plot1**

Figure 51 shows the geneset enrichment plot 1 for output of functional Enrichment analysis.


Figure 51: Biological annotation menu: Geneset enrichment plot1.

**Geneset enrichment plot2**

Figure 52 shows the geneset enrichment plot 2 for output of functional Enrichment analysis.

## 3.6 Meta-analysis

CBioProfiler also provides a Meta-analysis module to assist researchers in assessing the impact of biomarkers on the prognosis of a disease. The Meta-analysis module references the methods of (Schwarzer, Carpenter, and Rücker 2015), which first calculates the correlation between a specific gene and survival time of patients in a specific cohort through a univariate Cox proportional hazard model, and then conducts meta-analysis

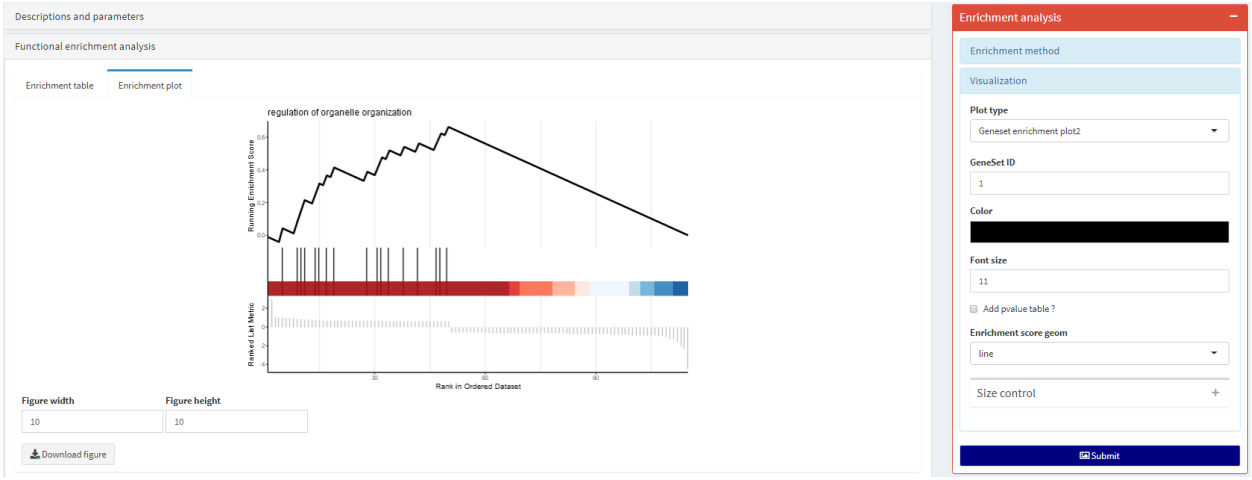


Figure 52: Biological annotation menu: Geneset enrichment plot2.

based on the hazard ratio and its 95% confidence interval of the patients. The results of the meta-analysis are displayed using a forest plot.

Parameters:

- **Official gene symbol:** Input one gene (biomarker candidate) with official gene symbol that you want to analyze the correlation between the gene and the prognosis information of the patients based on meta-analysis
- **Survival time:** Select survival time column. Example: "OS.time", "RFS.time", "PFS.time", etc.
- **Survival status:** Select survival status column. Example: "OS", "RFS", "PFS", etc.

Figure 53 shows the meta-analysis result of the survival impact of NEK6 on patients with glioma.

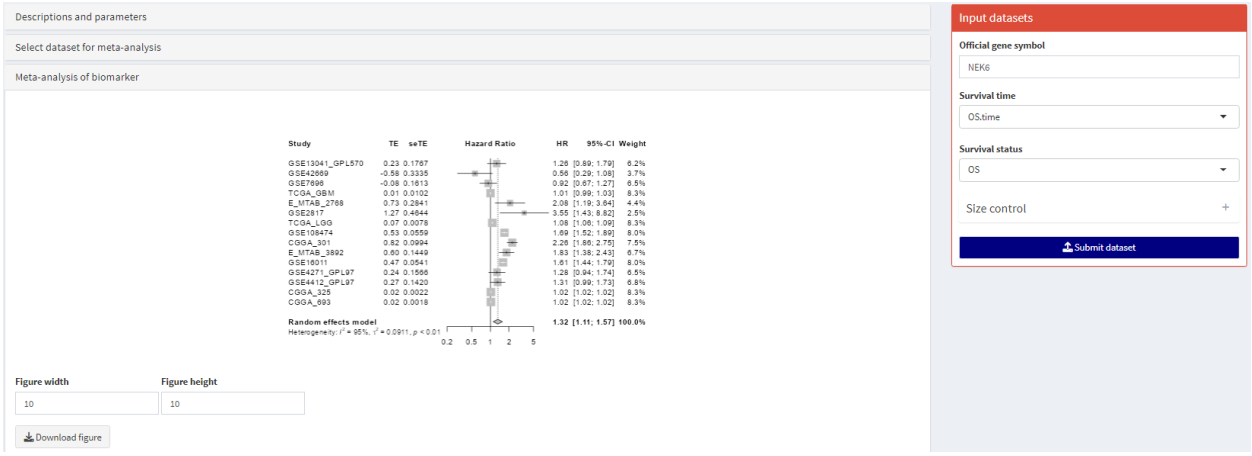


Figure 53: Meta-analysis of the survival impact of NEK6 on patients with glioma.

# 4 Subtype analysis

As we know, malignant tumors are a class of highly heterogeneous diseases (Dagogo-Jack and Shaw 2018). Almost all cells in the human body can become malignant tumors. Even in the same type of malignant

tumors, they may have obvious differences in biological behavior and clinical prognosis. A very important task of precision medicine is to classify patients into different subgroups or subtypes. Within these different subtypes, patients may have significant differences in pathogenesis, disease progression, treatment response, and disease outcome. After successfully classifying malignant tumor patients into different subtypes, doctors can conduct targeted prognosis evaluation and treatment for different subtypes of malignant tumor patients, so as to help patients obtain the most effective treatment effect at the least cost.

CBioProfiler implements the integration of multiple unsupervised machine learning methods (K-means clustering (Hartigan and Wong 1979), hierarchical clustering (Maimon and Rokach 2005), partitioning around medoids (PAM) clustering (Van der Laan, Pollard, and Bryan 2003), etc.) using two popular consensus clustering methods (ConsensusClusterPlus (Wilkerson and Hayes 2010) and M3C (John et al. 2020)). Based on this, CBioProfiler can assist users to classify tumor patients into different subtypes, and perform evaluation, validation, and biological and clinical annotation of these subtypes.

## 4.1 Subtype identification

The whole process of subtype identification mainly contains 5 steps: (1) data processing; (2) subtype identification; (3) subtype evaluation; and subtype validation.

### 4.1.1 Data preprocessing

Before performing subtype analysis, we usually need to preprocess the data. CBioProfiler integrates four methods provided by the CancerSubtype(Xu et al. 2017) package for data preprocessing (Variance, Median Absolute Deviation(MAD), Cox model (CoxPH), Principal Component Analysis (PCA)). Figure 54 shows top 1000 variant genes according to the variance-based method.

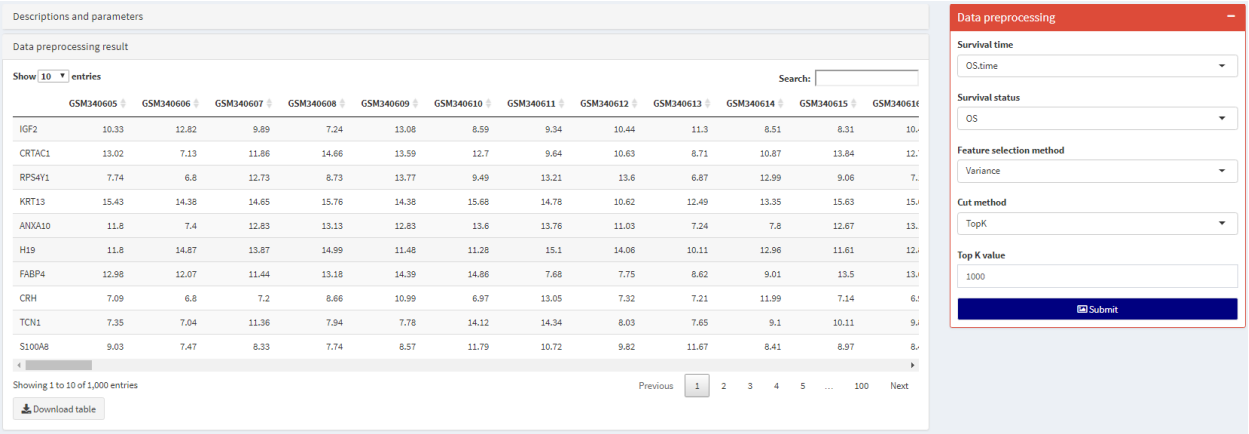


Figure 54: Data preprocessing menu: Data preprocessing result.

Parameters:

- **Survival time:** Select survival time column. Example: “OS.time”, “RFS.time”, “PFS.time”, etc.
- **Survival status:** Select survival time column. Example: “OS”, “RFS”, “PFS”, etc.
- **Feature selection method:** Specify the feature selection method, “Variance”, “MAD”, “PCA”, and “CoxPH” refer to “Variance”, “Median Absolute Deviation”(MAD)“,”Cox model”, and “Principal Component Analysis”, respectively.
- **Cut method:** A character value representing the selection type. “TopK” represents performing variable selection based on a specific number of genes. “Cutoff” performing variable selection based on specific thresholds.

### 4.1.2 Subtype identification

As mentioned above, CBioProfiler integrates multiple unsupervised machine learning methods (K-means clustering (Hartigan and Wong 1979), hierarchical clustering (Maimon and Rokach 2005), partitioning around medoids (PAM) clustering (Van der Laan, Pollard, and Bryan 2003), etc.) using two popular consensus clustering methods (ConsensusClusterPlus (Wilkerson and Hayes 2010) and M3C (John et al. 2020)). Based on this, CBioProfiler can assist users to classify tumor patients into different subtypes, and perform evaluation, validation, and biological and clinical annotation of these subtypes. Figure 55 shows result of consensus clustering (Wilkerson and Hayes 2010) based on hierarchical clustering.

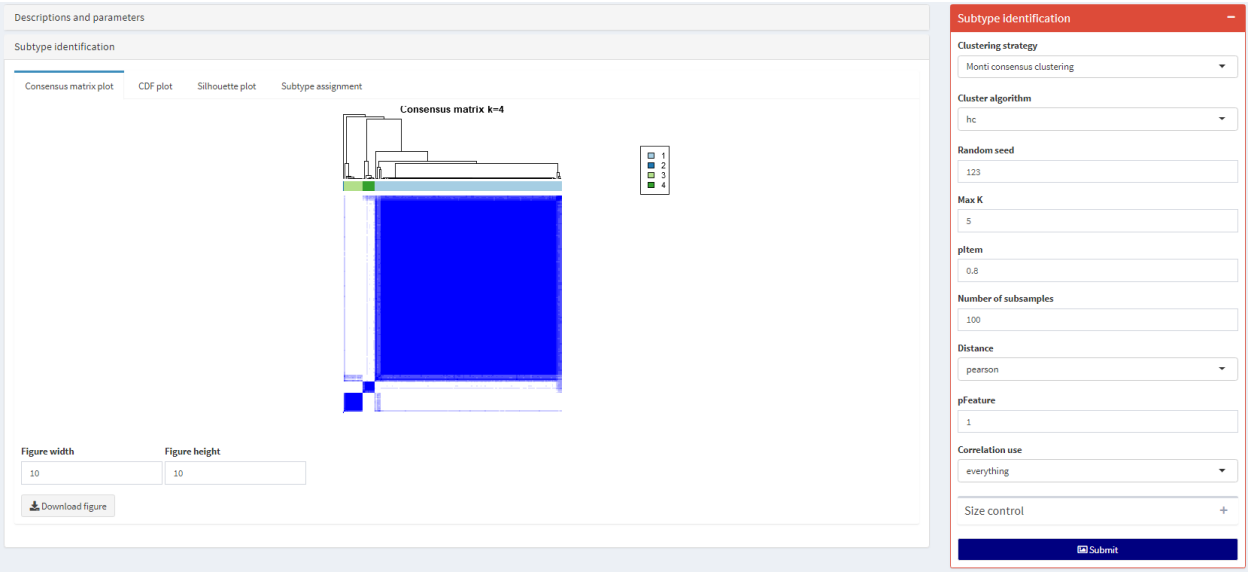


Figure 55: Subtype identification menu: Subtype identification result.

Parameters:

- **Clustering strategy:** Specify the clustering strategy, "Monti consensus clustering" and "M3C" refer to ConsensusClusterPlus (Wilkerson and Hayes 2010) and M3C (John et al. 2020), respectively.
- **Cluster algorithm:** Specify the cluster method. "hc" represents hierarchical clustering, "pam" represents partitioning around medoids and "km" represents K-means.
- **Random seed:** Specify the random seed.
- **Max K:** Integer value. maximum cluster number to evaluate.
- **pItem:** Numerical value. proportion of items to sample.
- **Number of subsamples:** Integer value. number of subsamples.
- **Distance:** Character value. 'pearson': (1 - Pearson correlation), 'spearman' (1 - Spearman correlation), 'euclidean', 'binary', 'maximum', 'canberra', and 'minkowski".
- **pFeature:** Numerical value. proportion of features to sample.
- **Correlation use:** Specifies how to handle missing data in correlation distances 'everything','pairwise.complete.obs', 'complete.obs' see cor() for description.
- **iteration:** Numerical value: how many Monte Carlo iterations to perform (default: 25, recommended: 5-100).

- **Reference method:** Character string: refers to which reference method to use.
- **Reference resamples:** Numerical value: how many resampling reps to use for reference (default: 100, recommended: 100-250).
- **Pacx1:** Numerical value: The 1st x co-ordinate for calculating the pac score from the CDF (default: 0.1).
- **Pacx2:** Numerical value: The 2nd x co-ordinate for calculating the pac score from the CDF (default: 0.1).
- **Objective function:** Character string: whether to use 'PAC' or 'entropy' objective function (default = entropy).
- **Simulation method:** 1 refers to the Monte Carlo simulation method, 2 to regularized consensus clustering.
- **Tune lambda ?:** Logical flag: whether to tune lambda or not.
- **Default lambda value:** Numerical value: if not tuning fixes the default (default: 0.1).

### 4.1.3 Subtype evaluation

The identified cancer subtypes by the computational methods should be in accordance with biological meanings and reveal the distinct molecular patterns. Silhouette (Si) analysis is a cluster validation approach that measures how well an observation is clustered and it estimates the average distance between clusters. Figure 56 shows the silhouette plot.

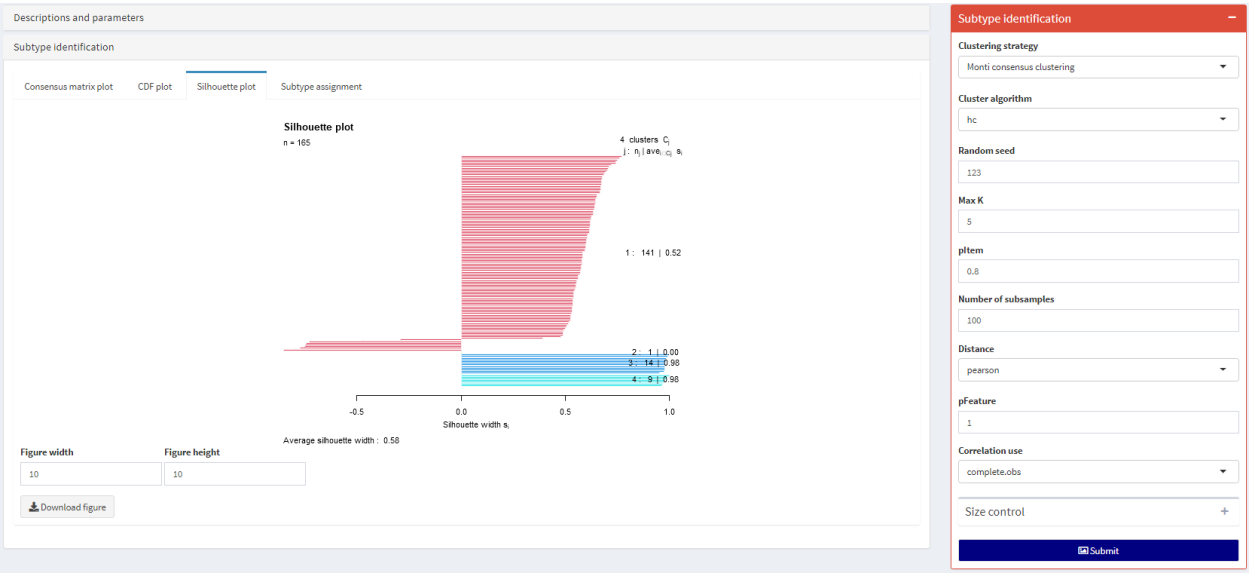


Figure 56: Subtype identification menu: Subtype evaluation silhouette plot.

### 4.1.4 Subtype validation

For subtype validation, CBioProfiler allows users assign the subtype on new patients. Specifically, calculate the Pearson's or Spearman's distances between the new patient's representation and the mean representation of each pre-identified patient subtype, and assign the new patient to its closest subtype yielding smallest Pearson's or Spearman's distance. Figure 57 shows the subtype validation result.

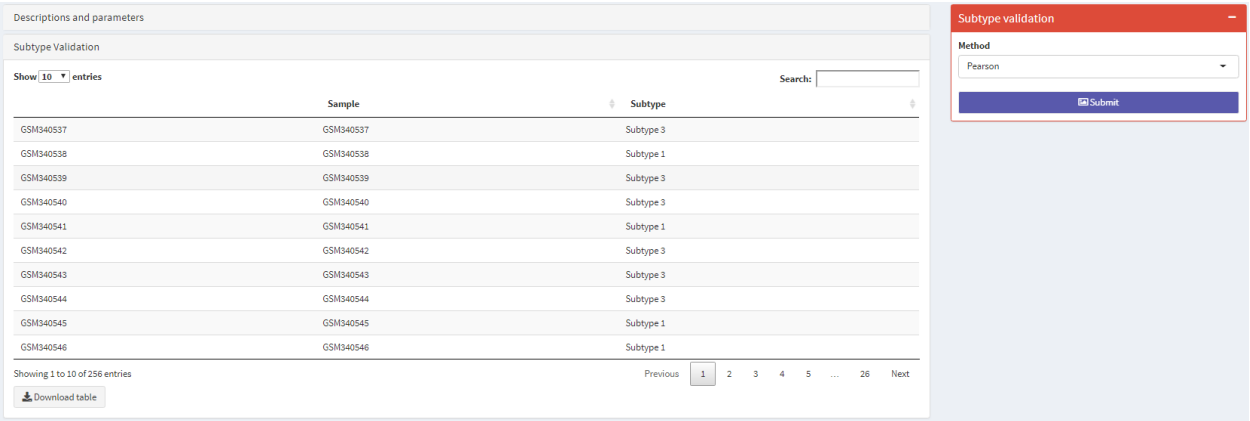


Figure 57: Subtype identification menu: Subtype validation on new data.

## 4.2 Subtype characterization

CBioProfiler can help users understand the clinical and biological significance of cancer subtypes from the following aspects: "Correlation with clinical features","Kaplan-Meier curve","CoxPH model","Time-dependent ROC","Differentially expressed genes","Immune infiltration among subtypes","Stemness score among subtypes","ESTIMATE score among subtypes","Immune checkpoints among subtypes","Interferon-gamma among subtypes","Cytolytic activity among subtypes","Cancer pathway score among subtypes","Metabolisim score among subtypes","Hallmark signature among subtypes",and "Drug response among subtypes".

### 4.2.1 Correlation with clinical features

The users can identify the correlation between the cancer subtype and the clinical features of patients through descriptive statistics table. The table is generated using R package "table1" (Rich 2020). When the table is generated, the users can copy it to Excel or OpenOffice. Figure 58 shows table for the correlations between cancer subtype and clinical features.

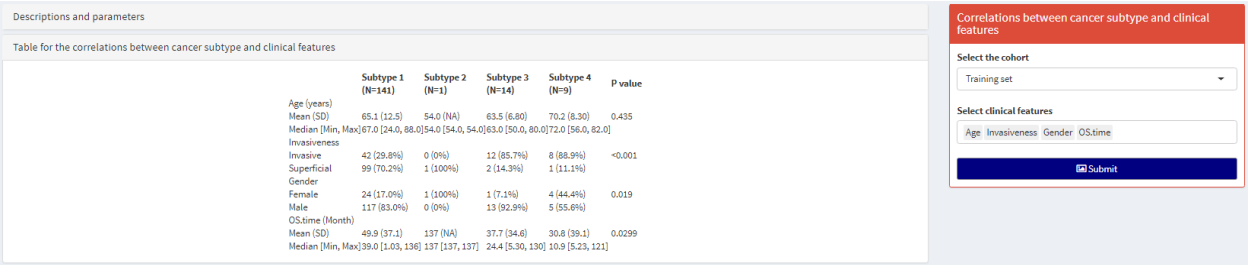


Figure 58: Correlation with clinical features menu: Table for the correlations between cancer subtype and clinical features main window.

Parameters:

- **Select the cohort:** Select the cohort that will be used for analysis.
- **Select clinical features:** Select clinical features you want to include in the table.

### 4.2.2 Kaplan-Meier curve

Kaplan-Meier (KM) curve is generated using R package survminer (Kassambara, Kosinski, and Biecek 2020), through which, the users can analyze the survival difference between different cancer subtypes. Figure 59 shows Kaplan-Meier plot for different cancer subtypes.

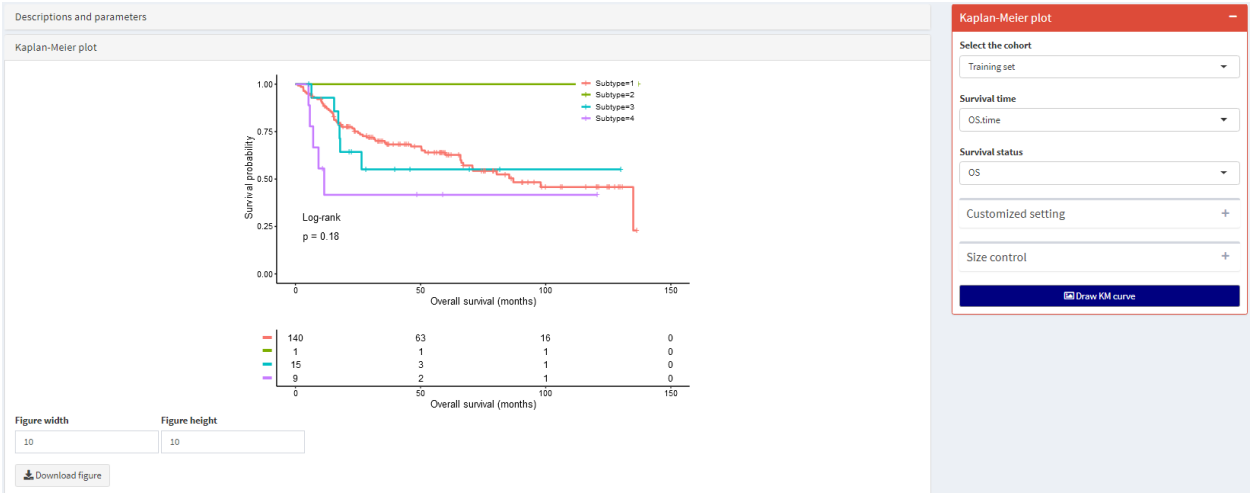


Figure 59: Correlation with clinical features menu: Table for the correlations between cancer subtype and clinical features main window.

Parameters:

- **Select the cohort:** Select the cohort that will be used for analysis.
- **Survival time:** Select survival time column. Example: "OS.time", "RFS.time", "PFS.time", etc.
- **Survival status:** Select survival time column. Example: "OS", "RFS", "PFS", etc.
- **Label of X axis:** Define the X axis label for the KM plot.
- **Show p-value?:** Whether show the p value for log-rank test for the KM plot.
- **Show risk table?:** Whether show the risk table for log-rank test for the KM plot.
- **Show confidence interval?:** Whether show the confidence interval for the KM plot.
- **Color of group 1/2:** Set the colors of group 1/2 for the KM plot.

### 4.2.3 CoxPH model

CBioProfiler uses Cox proportional hazards regression (CoxPH) model to help the users to characterize the prognostic value of the cancer subtype after adjusting for other clinical factors. CoxPH model was visualized using R package 'ggforestplot'(https://nightingalehealth.github.io/ggforestplot/articles/ggforestplot.html). Figure 60 shows forest plot for CoxPH model.

Parameters:

- **Select the cohort:** Select the cohort that will be used for analysis.
- **Survival time:** Select survival time column. Example: "OS.time", "RFS.time", "PFS.time", etc.
- **Survival status:** Select survival status column. Example: "OS", "RFS", "PFS", etc.

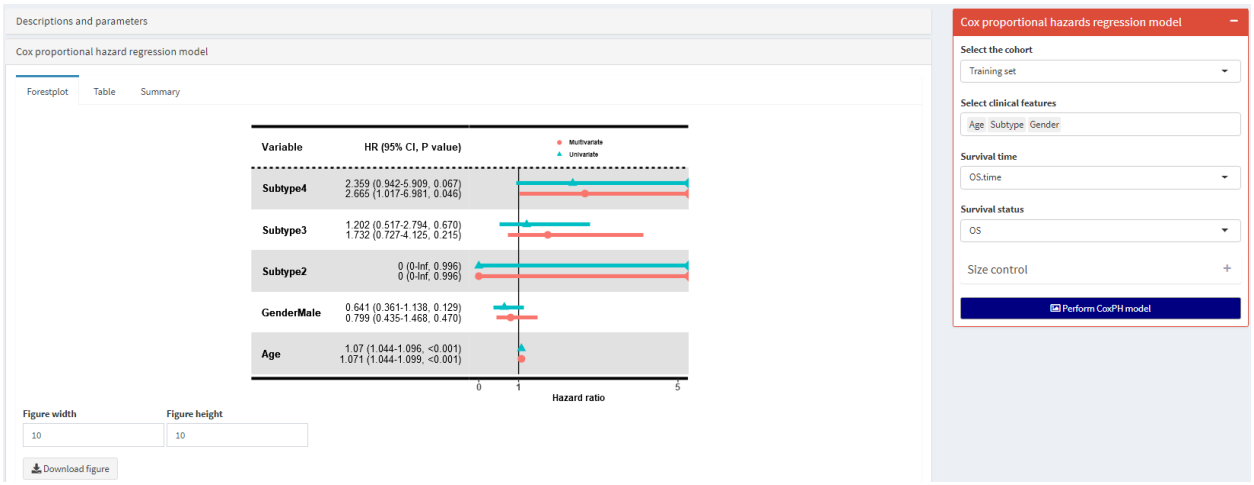


Figure 60: CoxPH model menu: Cox proportional hazard regression model main window.

- **Figure legend position:** Set the coordinates of the legend box. Their values should be between 0 and 1. c(0,0) corresponds to the 'bottom left' and c(1,1) corresponds to the 'top right' position.
- **Variable names:** Define the variable names you interested,the number of variable names should be identical with the number of variables you included in the CoxPH model and use '|' to separate multiple variable names.
- **Maximum of xticks:** Define the maximum of xticks and clip to adjust the size of the forestplot.

### 4.2.4 Time-dependent ROC

Time-dependent ROC (survival ROC) analysis (Heagerty and Paramita Saha-Chaudhuri 2013) helps the users investigate the accuracy of the survival prediction of the cancer subtype. Figure 61 shows time-dependent ROC analysis.

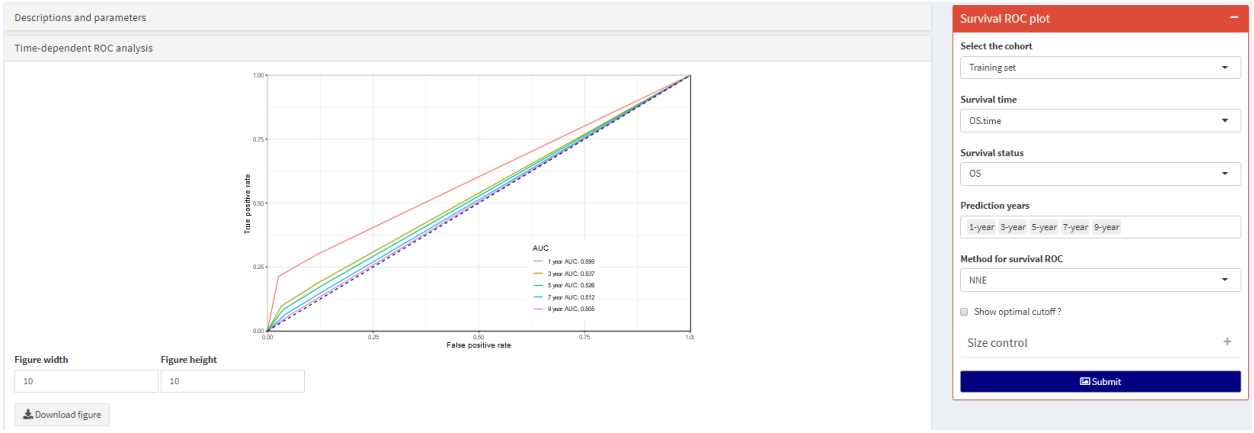


Figure 61: Time-dependent ROC menu: Time-dependent ROC analysis main window.

Parameters:

- **Select the cohort:** Select the cohort that will be used for analysis.
- **Survival time:** Select survival time column. Example: "OS.time", "RFS.time", "PFS.time", etc.

- **Survival status:** Select survival status column. Example: “OS”, “RFS”, “PFS”, etc.
- **Prediction years:** Define the time points in years you want to predict based on time dependent ROC analysis. the longest time point should not the max of survival (relapse/disease free) duration.
- **Method for survival ROC:** Define the method for survival ROC analysis, the default is “NNE”.

### 4.2.5 Differentially expressed genes

CBioProfiler uses linear models provided by the R package ‘limma’ (Smyth et al. 2020) to identify differentially expressed genes (DEGs) between two or more cancer subtypes. In the **Differentially expressed gene table**, the users can access the whole table of limma-based DEG result. Figure 62 shows differentially expressed gene table interface.

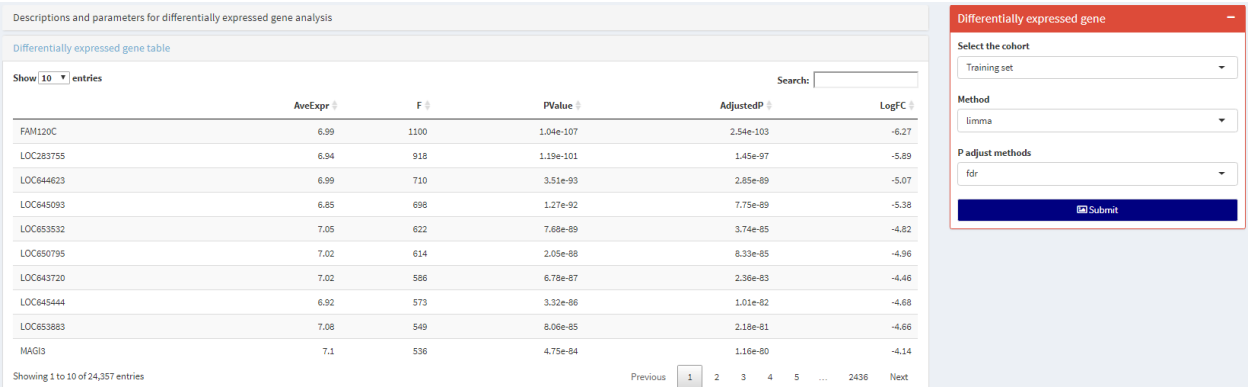


Figure 62: Differentially expressed genes menu: Differentially expressed gene table main window.

Parameters:

- **Select the cohort:** Select the cohort that will be used for analysis.
- **Method:** ‘limma’ means conducting moderated contrast t-test for each gene in limma”.
- **P adjust methods:** Specify the correction method for P values. For more details, please refer to https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/p.adjust.

After the users finished calculating limma-based DEG analysis, they can further screen significant DEGs based on the screening method and cutoff the specified. Figure 63 shows significant DEG output interface.

In the **DEG output**, the users can screen, output, and visualize significant DEGs they specified.

General parameters:

- **DEG cutoff method:** Specify the DEG cutoff method: “Adjusted P”, “LogFC”, “Adjusted P & LogFC”.
- **Visualization method:** Specify visualization method of DEGs, which includes heatmap, Volcano plot, MA plot, Adjusted P plot.

Parameters for **Heatmap**:

- **Heatmap name:** Names for the heatmap, by default the heatmap name is used as the title of the heatmap legend.

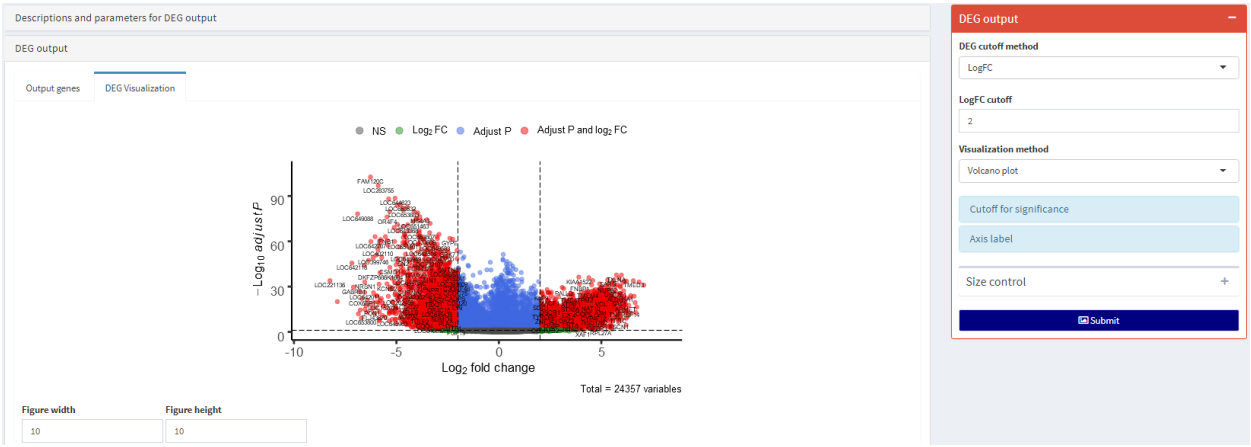


Figure 63: Differentially expressed genes menu: DEG output main window.

- **Min colour, Median colour, Max colour :** Color range for the heatmap.
- **Normalize the heatmap ?:** Whether to normalize the expression level of the heatmap.
- **Normalization method:** Specify a method to normalize the expression level of the heatmap, which includes “Scale”, “Center”, “Log”, “Z-score”, “0-1 normalization”.
- **Cluster on rows ?:** Whether to make a cluster on rows.
- **Cluster on row slice ?:** If rows are split into slices, whether perform clustering on the slice means?
- **Cluster distance on rows** It can be a pre-defined character which is in (“euclidean”, “maximum”, “manhattan”, “canberra”, “binary”, “minkowski”, “pearson”, “spearman”, “kendall”).
- **Cluster method on rows:** Method to perform hierarchical clustering, pass to hclust.
- **Row dendrogram side:** Should the row dendrogram be put on the ‘left’ or ‘right’ of the heatmap?
- **Show row names ?:** Whether should row names of the heatmap.
- **Row name side:** Should the row names be put on the left or right of the heatmap?
- **Show adjusted P value ?:** Whether show the adjusted P value for DEGs
- **Show logFC ?:** Whether show the logFC for DEGs.
- **Column title font size:** Specify the font size for the column title
- **Cluster on columns ?:** Whether to make a cluster on columns.
- **Cluster on column slices?:** If columns are split into slices, whether perform clustering on the slice means?
- **Cluster distance on rows:** It can be a pre-defined character which is in (“euclidean”, “maximum”, “manhattan”, “canberra”, “binary”, “minkowski”, “pearson”, “spearman”, “kendall”).
- **Cluster method on rows:** Method to perform hierarchical clustering, pass to hclust.
- **Row dendrogram side** Should the column dendrogram be put on the top or bottom of the heatmap?
- **Show column names ?:** Whether show the column names for the heatmap.
- **Column name side:** Should the column names be put on the top or bottom of the heatmap?

volcano plot is generated by using R package 'EnhancedVolcano' (Blighe, Rana, and Lewis 2020). Figure 12 shows DEG volcano plot output interface.

Parameters for **Volcano plot**:

- **Adjusted P cutoff:** Adjusted P value for DEGs to be visualized using volcano plot.
- **LogFC cutoff:** LogFC cutoff for DEGs to be visualized using volcano plot.
- **X-axis label:** Set the label for X-axis.
- **Y-axis label:** Set the label for y-axis.
- **Legend position:** Position of legend ('top', 'bottom', 'left', 'right'). DEFAULT = 'top'.

MA plot is generated using R package 'ggpubr'(Kassambara 2020). Figure 13 shows DEG MA plot output interface Parameters for **MA plot**:

- **Adjusted P cutoff:** Adjusted P value for DEGs to be visualized using volcano plot.
- **LogFC cutoff:** LogFC cutoff for DEGs to be visualized using volcano plot.
- **Selection methods:** Specify the method of label genes: Top genes, Cutoff, Gene symbol.
- **Top gene:** Set the number of top genes that will show gene symbol based on the cutoff set
- **Select top method:** The method used to select top genes.
- **Gene symbol:** Character vector specifying some gene labels to show.
- **X-axis label:** The label for X-axis.
- **Y-axis label:** The label for Y-axis.

Adjusted P plot is generated using R basic function. Figure 14 shows DEG adjusted p plot output interface.

Parameter for **Adjusted P plot**:

- **Color of the histogram:** Set the color for the Adjusted P plot.

#### 4.2.6 Immune infiltration among subtypes

Immune cell infiltration, as an important component of the tumor microenvironment, exerts an important influence on the occurrence, development and outcome of tumors (Thorsson et al. 2018). CBioProfiler applies the R package ConsensusTME (Jiménez-Sánchez, Cast, and Miller 2019) which uses a consensus approach to estimate enrichment scores for multiple immune cells found within the tumor microenvironment using ssGSEA (Hänzelmann, Castelo, and Guinney 2013).CBioProfiler calculates enrichment score of immune cell types based on gene signatures from Bindea and colleagues (Bindea et al. 2013), Danaher and colleagues (Danaher et al. 2017), and Davoli and colleagues (Davoli et al. 2017), xCell(Aran, Hu, and Butte 2017), MCP-counter (Becht et al. 2016). Figure 64 shows the correlation between cancer subtype and immune cell infiltration.

Parameters:

- **Select the cohort:** Select the cohort that will be used for analysis.
- **Reference immune geneset:** Immune gene signatures from Bindea and colleagues, Danaher and colleagues, Davoli and colleagues, MCP-Counter, and xCell, respectively.

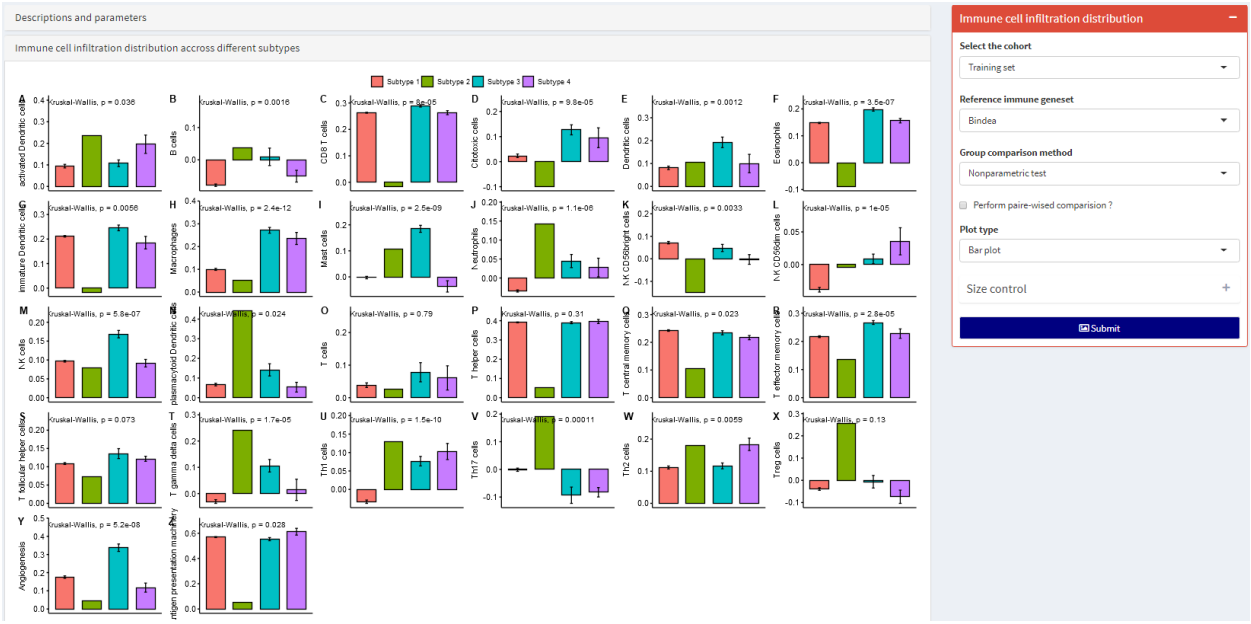


Figure 64: Correlation with immune infiltration menu: Correlation with Immune cell infiltration main window.

- **Group comparison method:** Specify the group comparison methods. Whether to use a parametric or nonparametric test for comparisons between groups.
- **Perform paired-wised comparison ?:** Whether to perform paired comparison?
- **Plot type:** Select the plot type for visualization.

### 4.2.7 Stemness score among subtypes

Malta et al (Malta et al. 2018). introduced a stemness score for assessing the degree of oncogenic dedifferentiation of cancer cells. Colaprico et al (Colaprico et al. 2016) developed R function to calculate the stemness score of samples in TCGA. We extend it to gene expression studies including but not limited to TCGA. CBioProfiler provides function to help users investigate the relationship between stemness score and the cancer subtpye. Figure 65 shows the correlation between cancer subtype and stemness score.


Figure 65: Correlation with stemness score menu: Correlation with stemness score main window.

Parameters:

- **Select the cohort:** Select the cohort that will be used for analysis.
- **Group comparison method:** Specify the group comparison methods. Whether to use a parametric or nonparametric test for comparisons between groups.
- **Perform paired-wised comparison ?:** Whether to perform paired comparison?
- **Plot type:** Select the plot type for visualization.

### 4.2.8 ESTIMATE score among subtypes

Yoshihara et al.(Yoshihara et al. 2013) developed an R package ESTIMATE (Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression data) aiming at predicting tumor purity, and the presence of infiltrating stromal/immune cells in tumor tissues using gene expression data based on single sample Gene Set Enrichment Analysis and generates three scores:

- *1)* stromal score (that captures the presence of stroma in tumor tissue),
- *2)* immune score (that represents the infiltration of immune cells in tumor tissue), and
- *3)* estimate score (that infers tumor purity).

Thus, we applied the ESTIMATE algorithm to calculate the relationship between cancer subtype and stromal cells and immune cells in the tumor microenvironment. Figure 66 shows the correlation between cancer subtype and ESTIMATE score.


Figure 66: ESTIMATE score among subtypes menu: ESTIMATE score among subtypes main window.

Parameters:

- **Select the cohort:** Select the cohort that will be used for analysis.
- **Group comparison method:** Specify the group comparison methods. Whether to use a parametric or nonparametric test for comparisons between groups.
- **Perform paired-wised comparison ?:** Whether to perform paired comparison?
- **Plot type:** Select the plot type for visualization.

### 4.2.9 Immune checkpoints among subtypes

Immune checkpoint molecules are regulators of the immune system. These pathways are crucial for self-tolerance, which prevents the immune system from attacking cells indiscriminately. However, some cancers can protect themselves from attack by stimulating immune checkpoint targets.(Pardoll 2012)

Inhibitory checkpoint molecules (Nirschl and Drake 2013; Sharma and Allison 2015), including cytotoxic T lymphocyte antigen-4 (CTLA-4), programmed death-1 (PD-1), lymphocyte activation gene-3 (LAG-3), T-cell immunoglobulin and mucin protein-3 (TIM-3),etc., are targets for cancer immunotherapy due to their potential for use in multiple types of cancers. The expression of these checkpoint molecules on T cells represents an important mechanism that the immune system uses to regulate responses to self-proteins. Recent clinical data show that these Checkpoint molecules play a critical role in objective tumor responses and improved overall survival. Therefore, CBioProfiler help users identify the correlation between the immune checkpoint molecules and cancer subtype. Figure 67 shows the immune checkpoints among subtypes.

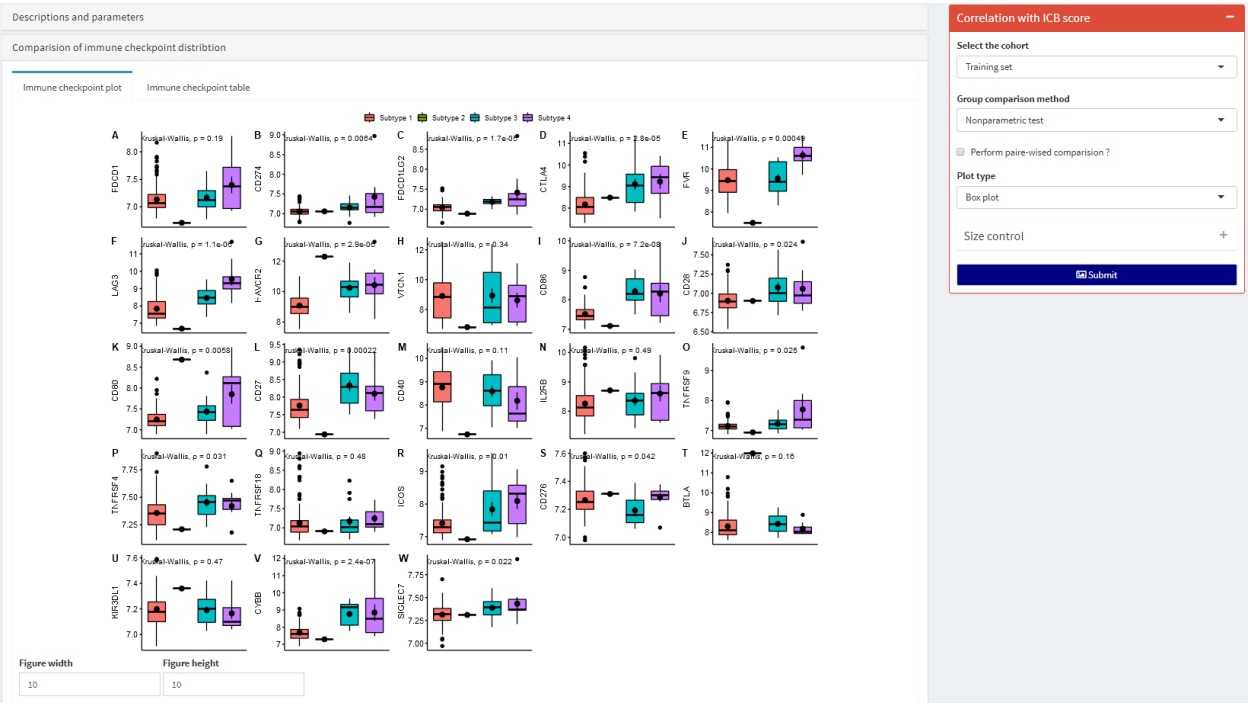


Figure 67: Immune checkpoints among subtypes menu: Immune checkpoints among subtypes main window.

Parameters:

- **Select the cohort:** Select the cohort that will be used for analysis.
- **Group comparison method:** Specify the group comparison methods. Whether to use a parametric or nonparametric test for comparisons between groups.
- **Perform paired-wised comparison ?:** Whether to perform paired comparison?
- **Plot type:** Select the plot type for visualization.

### 4.2.10 Interferon-gamma among subtypes

Interferon-gamma (IFN-gamma) plays a crucial role in the regulation of antitumor immunity: mainly secreted by activated lymphocytes such as CD8 cytotoxic T-cells or CD4 T-helper cells type I (Th1), IFN-gamma can

enhance Th1-mediated antitumor immune response in terms of a positive feedback loop.(Castro et al. 2018) IFN-gamma is also known to play a protumorigenic role by transmitting antiapoptotic and proliferative signals, resulting in immune-escape of tumor cells. CBioProfiler allows users to estimate the relationship between cancer subtype they identified and the IFN-gamma score calculated based on ssGSEA (Hänzelmann, Castelo, and Guinney 2013). Figure 68 shows the interferon-gamma score among different cancer subtypes.

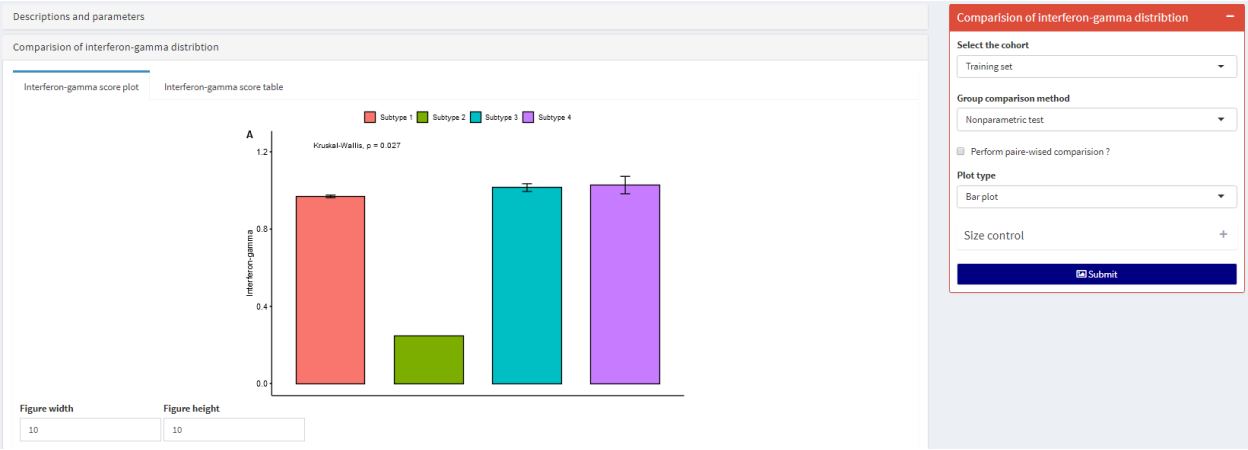


Figure 68: Interferon-gamma score among different cancer subtypes menu: Interferon-gamma score among different cancer subtypes main window.

Parameters:

- **Select the cohort:** Select the cohort that will be used for analysis.
- **Group comparison method:** Specify the group comparison methods. Whether to use a parametric or nonparametric test for comparisons between groups.
- **Perform paired-wised comparison ?:** Whether to perform paired comparison?
- **Plot type:** Select the plot type for visualization.

### 4.2.11 Cytolytic activity among subtypes

Based on the notion that effective natural anti-tumor immunity requires a cytolytic immune response, Rooney et al. (Rooney et al. 2015) quantified cytolytic activity using a simple expression metric of effector molecules that mediate cytolysis. They demonstrated that cytolytic activity was associated with MHC Class I-associated neoantigens, gene mutations (including including beta-2-microglobulin (B2M), HLA-A, -B and -C and Caspase 8 (CASP8)) that highlighted loss of antigen presentation and blockade of extrinsic apoptosis, and genetic amplifications (PDL1/2 and ALOX12B/15B). Herein, CBioProfiler allows users calculate the correlation between the subtype and cytolytic activity. Figure 69 shows the correlation between TP53 and cytolytic activity.

Parameters:

- **Select the cohort:** Select the cohort that will be used for analysis.
- **Group comparison method:** Specify the group comparison methods. Whether to use a parametric or nonparametric test for comparisons between groups.
- **Perform paired-wised comparison ?:** Whether to perform paired comparison?
- **Plot type:** Select the plot type for visualization.


Figure 69: Interferon-gamma score among different cancer subtypes menu: Interferon-gamma score among different cancer subtypes main window.

### 4.2.12 Cancer pathway score among subtypes

Cancer related pathways, including Cell cycle, Chromatin remodeling, Differentiation and development, DNA damage, Immune regulation, MAPK and PI3K pathway, Metabolism, PI3K pathway, RAS pathway, RNA metabolism, RTK pathway, TGFB pathway, Transcription regulation, WNT signaling, have been demonstrated to play an important role in the oncogenesis and progression of multiple human cancers. Davoli et al.(Davoli et al. 2017) curated key genes in these cancer related pathways, and we utilize these gene signatures to help users to identify the relationship between the biomarker and the cancer related pathways based on ssGSEA algorithm (Hänzelmann, Castelo, and Guinney 2013). Figure 70 shows the correlation between TP53 and cancer related pathway.

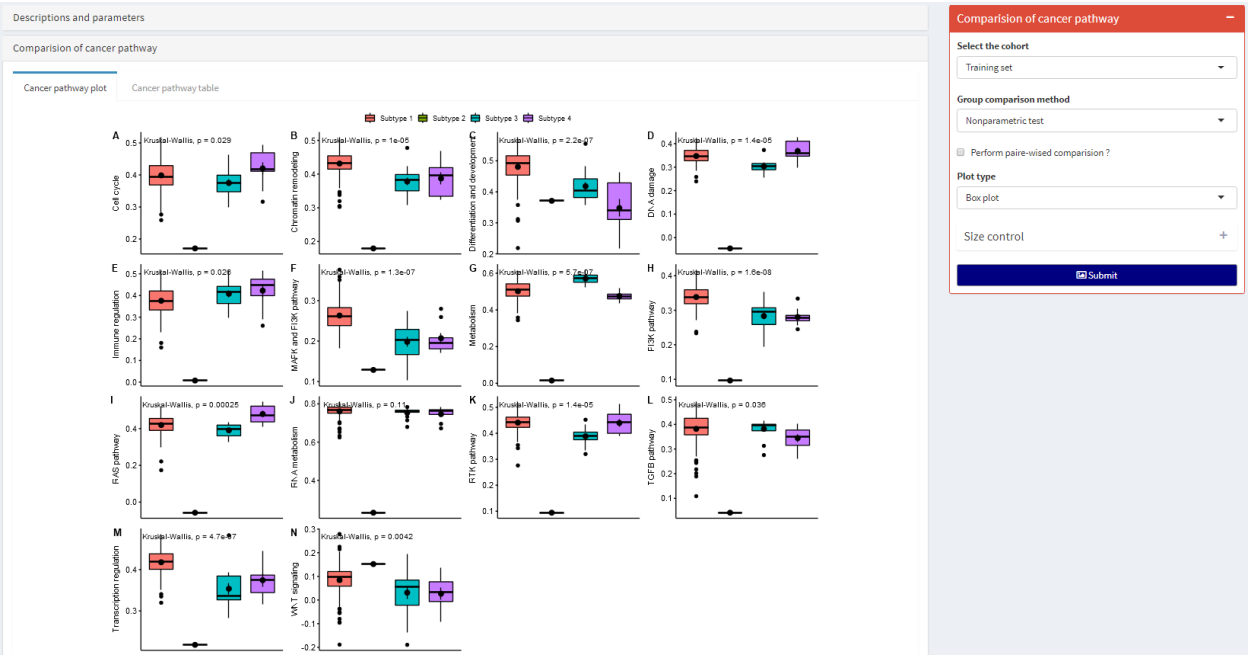


Figure 70: Interferon-gamma score among different cancer subtypes menu: Interferon-gamma score among different cancer subtypes main window.

Parameters:

- **Select the cohort:** Select the cohort that will be used for analysis.
- **Group comparison method:** Specify the group comparison methods. Whether to use a parametric or nonparametric test for comparisons between groups.
- **Perform paired-wised comparison ?:** Whether to perform paired comparison?
- **Plot type:** Select the plot type for visualization.

### 4.2.13 Metabolisim score among subtypes

Tumor initiation and progression depend on cellular metabolism. A common feature of tumor cell metabolism is the ability to obtain essential nutrients from a nutrient-poor environment and use these nutrients to maintain viability and generate new biomass. Alterations in intracellular and extracellular metabolites can have profound effects on gene expression, cell differentiation, and the tumor microenvironment along with tumor-associated metabolic reprogramming. There are six main characteristics of changes in tumor metabolism, but only a few tumors exhibit these six characteristics at the same time. According to the specific characteristics exhibited by tumors, it may help to better guide tumor classification and treatment. Ricketts et al.(Ricketts et al. 2018) curated common tumor metabolic pathways, including, AMPK Complex, Complex I, Complex III, Complex IV - cytochrome C, Complex V, Fatty Acid Synthesis, Glycogen Metabolism, Glycolysis, Krebs Cycle - Cyto, Krebs Cycle - Mito, NADH->NADPH, PDC Activation, PDC Suppression, Ribose Sugar Metabolism, and Serine Metabolism. In the present study, CBioProfiler helps the users to characterize the relationship between the cancer subtype and the common metabolic pathways based on ssGSEA algorithm (Hänzelmann, Castelo, and Guinney 2013). Figure 71 shows the metabolisim score among subtypes.

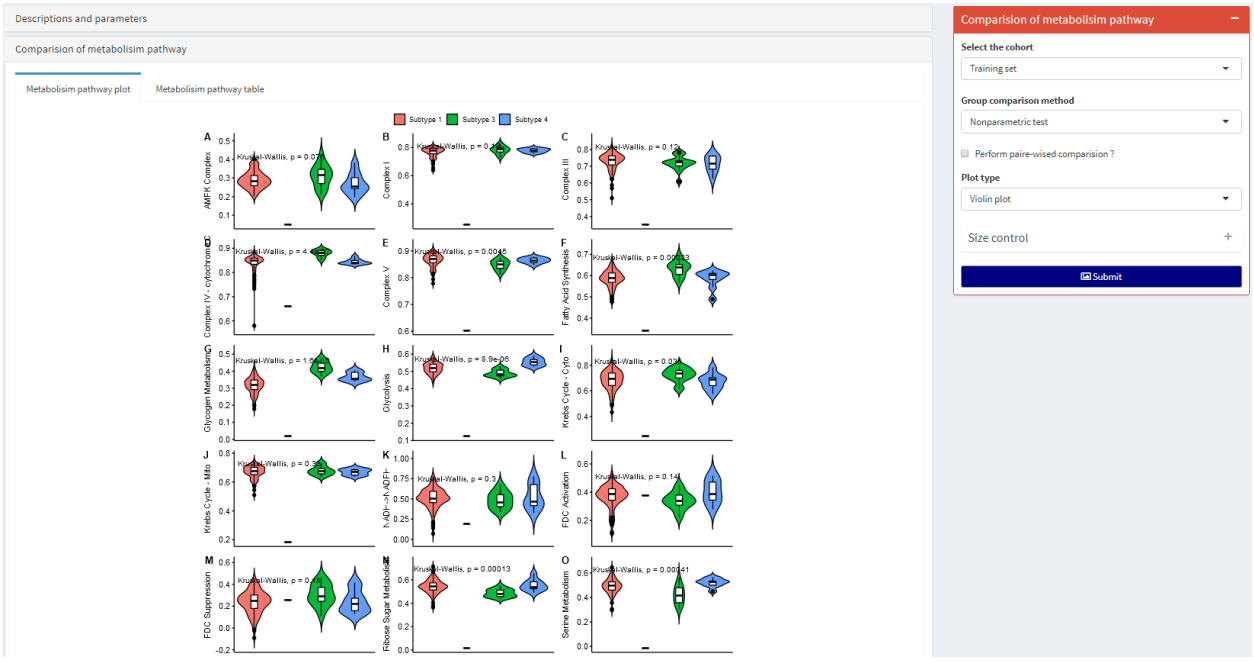


Figure 71: Metabolisim score among subtypes menu: Metabolisim score among subtypes main window.

Parameters:

- **Select the cohort:** Select the cohort that will be used for analysis.

- **Group comparison method:** Specify the group comparison methods. Whether to use a parametric or nonparametric test for comparisons between groups.
- **Perform paired-wised comparison ?:** Whether to perform paired comparison?
- **Plot type:** Select the plot type for visualization.

### 4.2.14 Hallmark signature among subtypes

Hallmark gene sets are a group of 50 gene set that summarize and represent specific well-defined biological states or processes and display coherent expression. These gene sets were generated by a computational methodology based on identifying gene set overlaps and retaining genes that display coordinate expression (Liberzon et al. 2015; Dolgalev 2020). CBioProfiler helps the users to characterize the relationship between the cancer subtype and hallmark gene sets based on ssGSEA algorithm (Hänzelmann, Castelo, and Guinney 2013). Figure 72 shows the hallmark signature among subtypes.


Figure 72: Hallmark signature among subtypes menu: Hallmark signature among subtypes main window.

Parameters:

- **Select the cohort:** Select the cohort that will be used for analysis.
- **Group comparison method:** Specify the group comparison methods. Whether to use a parametric or nonparametric test for comparisons between groups.
- **Perform paired-wised comparison ?:** Whether to perform paired comparison?
- **Plot type:** Select the plot type for visualization.

### 4.2.15 Drug response among subtypes

Geeleher et al. recently introduced a methodology that worked by training and building statistical models from gene expression profile and drug sensitivity data in a very large panel of cancer cell lines, then applying these models to gene expression data from primary tumor biopsies (Geeleher, Cox, and Huang 2014a). They also created an R package called pRRophetic (Geeleher, Cox, and Huang 2014b), which extends the previously described pipeline and allows prediction of clinical drug response for many cancer drugs in an R environment. CBioProfiler takes advantage of pRRophetic to help users predict the drug sensitivity and estimate the relationship between chemotherapy drugs and the biomarker they selected. Figure 73 shows the drug response among subtypes.

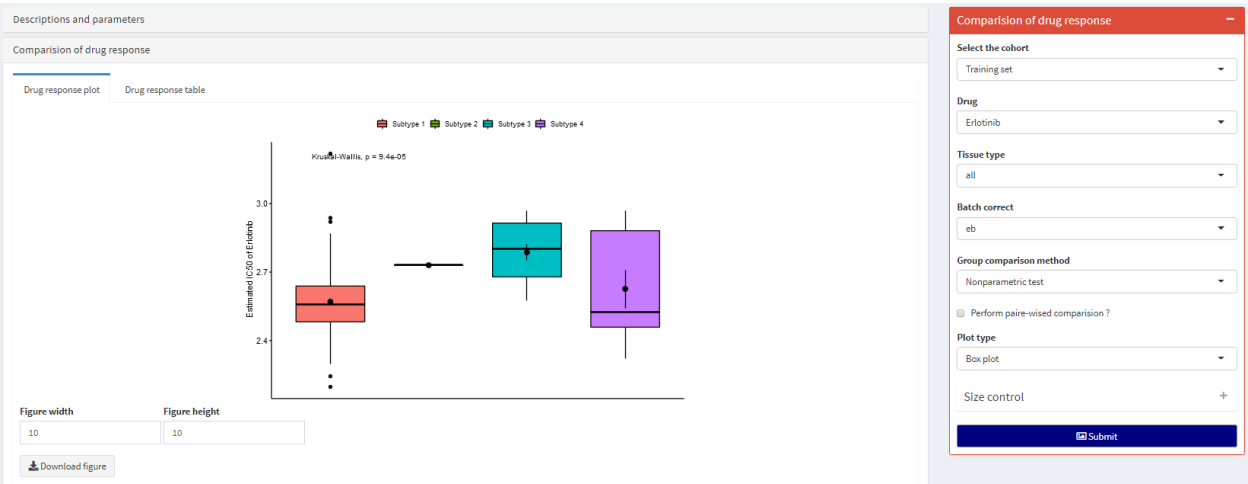


Figure 73: Hallmark signature among subtypes menu: Hallmark signature among subtypes main window.

Parameters:

- **Select the cohort:** Select the cohort that will be used for analysis.
- **Drug:** The name of the drug for which you would like to predict sensitivity.
- **Tissue type:** Specify if you would like to traing the models on only a subset of the CGP cell lines (based on the tissue type from which the cell lines originated). This be one any of "all" (for everything, default option), "allSolidTumors" (everything except for blood), "blood", "breast", "CNS", "GI tract" ,"lung", "skin", "upper aerodigestive".
- **Batch correct:** How should training and test data matrices be homogenized. Choices are "eb" (default) for ComBat, "qn" for quantiles normalization or "none" for no homogenization.
- **Group comparison method:** Specify the group comparison methods. Whether to use a parametric or nonparametric test for comparisons between groups.
- **Perform paired-wised comparison ?:** Whether to perform paired comparison?
- **Plot type:** Select the plot type for visualization.

# 5 References

Aran, Dvir, Zicheng Hu, and Atul J. Butte. 2017. "xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape." *Genome Biology* 18 (1): 220. https://doi.org/10.1186/s13059-017-1349-1.

Becht, Etienne, Nicolas A. Giraldo, Laetitia Lacroix, Bénédicte Buttard, Nabila Elarouci, Florent Petitprez, Janick Selves, et al. 2016. "Estimating the Population Abundance of Tissue-Infiltrating Immune and Stromal Cell Populations Using Gene Expression." *Genome Biology* 17 (1): 218. https://doi.org/10.1186/s13059-016-1070-5.

Bindea, Gabriela, Bernhard Mlecnik, Marie Tosolini, Amos Kirilovsky, Maximilian Waldner, Anna C. Obenauf, Helen Angell, et al. 2013. "Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer." *Immunity* 39 (4): 782–95. https://doi.org/10.1016/j.immuni.2013.10.003.

Bischl, Bernd, Michel Lang, Lars Kotthoff, Patrick Schratz, Julia Schiffner, Jakob Richter, Zachary Jones, Giuseppe Casalicchio, and Mason Gallo. 2020. *Mlr: Machine Learning in r*. https://CRAN.R-project.org/package=mlr.

Blighe, Kevin, Sharmila Rana, and Myles Lewis. 2020. *EnhancedVolcano: Publication-Ready Volcano Plots with Enhanced Colouring and Labeling*. https://github.com/kevinblighe/EnhancedVolcano.

Castro, Flávia, Ana Patrícia Cardoso, Raquel Madeira Gonçalves, Karine Serre, and Maria José Oliveira. 2018. "Interferon-Gamma at the Crossroads of Tumor Immune Surveillance or Evasion." *Frontiers in Immunology* 9 (May): 847. https://doi.org/10.3389/fimmu.2018.00847.

Colaprico, Antonio, Tiago C. Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S. Sabedot, et al. 2016. "TCGAbiolinks: An R/Bioconductor Package for Integrative Analysis of TCGA Data." *Nucleic Acids Research* 44 (8): e71–71. https://doi.org/10.1093/nar/gkv1507.

Dagogo-Jack, Ibiayi, and Alice T. Shaw. 2018. "Tumour Heterogeneity and Resistance to Cancer Therapies." *Nature Reviews Clinical Oncology* 15 (2): 81–94. https://doi.org/10.1038/nrclinonc.2017.166.

Danaher, Patrick, Sarah Warren, Lucas Dennis, Leonard D'Amico, Andrew White, Mary L. Disis, Melissa A. Geller, Kunle Odunsi, Joseph Beechem, and Steven P. Fling. 2017. "Gene Expression Markers of Tumor Infiltrating Leukocytes." *Journal for ImmunoTherapy of Cancer* 5 (1): 18. https://doi.org/10.1186/s40425-017-0215-8.

Davoli, Teresa, Hajime Uno, Eric C. Wooten, and Stephen J. Elledge. 2017. "Tumor Aneuploidy Correlates with Markers of Immune Evasion and with Reduced Response to Immunotherapy." *Science* 355 (6322): eaaf8399. https://doi.org/10.1126/science.aaf8399.

Dolgalev, Igor. 2020. *Msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format*. https://github.com/igordot/msigdbr.

Geeleher, Paul, Nancy J Cox, and R Stephanie Huang. 2014a. "Clinical Drug Response Can Be Predicted Using Baseline Gene Expression Levels and in Vitro Drug Sensitivity in Cell Lines." *Genome Biology* 15 (3): R47. https://doi.org/10.1186/gb-2014-15-3-r47.

Geeleher, Paul, Nancy Cox, and R. Stephanie Huang. 2014b. "pRRophetic: An R Package for Prediction of Clinical Chemotherapeutic Response from Tumor Gene Expression Levels." Edited by Jason D. Barbour. *PLoS ONE* 9 (9): e107468. https://doi.org/10.1371/journal.pone.0107468.

Gu, Zuguang. 2021. *ComplexHeatmap: Make Complex Heatmaps*.

Hänzelmann, Sonja, Robert Castelo, and Justin Guinney. 2013. "GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data." *BMC Bioinformatics* 14 (1): 7. https://doi.org/10.1186/1471-2105-14-7.

Harrell, Jr., Frank E. 2020. *Rms: Regression Modeling Strategies*. https://CRAN.R-project.org/package=rms.

Hartigan, J. A., and M. A. Wong. 1979. "Algorithm AS 136: A K-Means Clustering Algorithm." *Applied Statistics* 28 (1): 100. https://doi.org/10.2307/2346830.

Heagerty, Patrick J., and packaging by Paramita Saha-Chaudhuri. 2013. *survivalROC: Time-Dependent ROC Curve Estimation from Censored Survival Data*. https://CRAN.R-project.org/package=survivalROC.

Horvath, Steve. 2011. *Weighted Network Analysis*. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4419-8819-5.

Jiménez-Sánchez, Alejandro, Oliver Cast, and Martin L. Miller. 2019. "Comprehensive Benchmarking and Integration of Tumor Microenvironment Cell Estimation Methods." *Cancer Research* 79 (24): 6238–46. https://doi.org/10.1158/0008-5472.CAN-18-3560.

John, Christopher R., David Watson, Dominic Russ, Katriona Goldmann, Michael Ehrenstein, Costantino Pitzalis, Myles Lewis, and Michael Barnes. 2020. "M3c: Monte Carlo Reference-Based Consensus

Clustering." *Scientific Reports* 10 (1): 1816. https://doi.org/10.1038/s41598-020-58766-1.
Kassambara, Alboukadel. 2020. *Ggpubr: Ggplot2 Based Publication Ready Plots*. https://rpkgs.datanovia.com/ggpubr/.
Kassambara, Alboukadel, Marcin Kosinski, and Przemyslaw Biecek. 2020. *Survminer: Drawing Survival Curves Using Ggplot2*. http://www.sthda.com/english/rpkgs/survminer/.
Langfelder, Peter, Steve Horvath with contributions by Chaochao Cai, Jun Dong, Jeremy Miller, Lin Song, Andy Yip, and Bin Zhang. 2020. *WGCNA: Weighted Correlation Network Analysis*. http://horvath.genetics.ucla.edu/html/ CoexpressionNetwork/Rpackages/WGCNA/.
Langfelder, Peter, and Steve Horvath. 2008. "WGCNA: An R Package for Weighted Correlation Network Analysis." *BMC Bioinformatics* 9 (1): 559. https://doi.org/10.1186/1471-2105-9-559.
Liberzon, Arthur, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. 2015. "The Molecular Signatures Database Hallmark Gene Set Collection." *Cell Systems* 1 (6): 417–25. https://doi.org/10.1016/j.cels.2015.12.004.
Maimon, Oded, and Lior Rokach, eds. 2005. *Data Mining and Knowledge Discovery Handbook*. New York: Springer.
Malta, Tathiane M., Artem Sokolov, Andrew J. Gentles, Tomasz Burzykowski, Laila Poisson, John N. Weinstein, Bożena Kamińska, et al. 2018. "Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation." *Cell* 173 (2): 338–354.e15. https://doi.org/10.1016/j.cell.2018.03.034.
Nirschl, Christopher J., and Charles G. Drake. 2013. "Molecular Pathways: Coexpression of Immune Checkpoint Molecules: Signaling Pathways and Implications for Cancer Immunotherapy." *Clinical Cancer Research* 19 (18): 4917–24. https://doi.org/10.1158/1078-0432.CCR-12-1972.
Pardoll, Drew M. 2012. "The Blockade of Immune Checkpoints in Cancer Immunotherapy." *Nature Reviews Cancer* 12 (4): 252–64. https://doi.org/10.1038/nrc3239.
Rich, Benjamin. 2020. *Table1: Tables of Descriptive Statistics in HTML*. https://github.com/benjaminrich/table1.
Ricketts, Christopher J., Aguirre A. De Cubas, Huihui Fan, Christof C. Smith, Martin Lang, Ed Reznik, Reanne Bowlby, et al. 2018. "The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma." *Cell Reports* 23 (1): 313–326.e5. https://doi.org/10.1016/j.celrep.2018.03.075.
Rooney, Michael S., Sachet A. Shukla, Catherine J. Wu, Gad Getz, and Nir Hacohen. 2015. "Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity." *Cell* 160 (1-2): 48–61. https://doi.org/10.1016/j.cell.2014.12.033.
Schwarzer, Guido, James R. Carpenter, and Gerta Rücker. 2015. *Meta-Analysis with R*. Use R! Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-21416-0.
Sharma, Padmanee, and James P. Allison. 2015. "The Future of Immune Checkpoint Therapy." *Science* 348 (6230): 56–61. https://doi.org/10.1126/science.aaa8172.
Smyth, Gordon, Yifang Hu, Matthew Ritchie, Jeremy Silver, James Wettenhall, Davis McCarthy, Di Wu, et al. 2020. *Limma: Linear Models for Microarray Data*. http://bioinf.wehi.edu.au/limma.
Therneau, Terry M. 2020. *Survival: Survival Analysis*. https://github.com/therneau/survival.
Thorsson, Vésteinn, David L. Gibbs, Scott D. Brown, Denise Wolf, Dante S. Bortone, Tai-Hsien Ou Yang, Eduard Porta-Pardo, et al. 2018. "The Immune Landscape of Cancer." *Immunity* 48 (4): 812–830.e14. https://doi.org/10.1016/j.immuni.2018.03.023.
Van der Laan, Mark, Katherine Pollard, and Jennifer Bryan. 2003. "A New Partitioning Around Medoids Algorithm." *Journal of Statistical Computation and Simulation* 73 (8): 575–84. https://doi.org/10.1080/0094965031000136012.
Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2020. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.
Wilkerson, Matthew D., and D. Neil Hayes. 2010. "ConsensusClusterPlus: A Class Discovery Tool with Confidence Assessments and Item Tracking." *Bioinformatics* 26 (12): 1572–73. https://doi.org/10.1093/bioinformatics/btq170.
Xu, Taosheng, Thuc Duy Le, Lin Liu, Ning Su, Rujing Wang, Bingyu Sun, Antonio Colaprico, Gianluca Bontempi, and Jiuyong Li. 2017. "CancerSubtypes: An R/Bioconductor Package for Molecular Cancer Subtype Identification, Validation and Visualization." Edited by Inanc Birol. *Bioinformatics* 33 (19): 3131–33. https://doi.org/10.1093/bioinformatics/btx378.

Yoshihara, Kosuke, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, et al. 2013. "Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data." *Nature Communications* 4 (1): 2612. https://doi.org/10.1038/ncomms3612.

Yu, Guangchuang. 2021. *clusterProfiler: Statistical Analysis and Visualization of Functional Profiles for Genes and Gene Clusters*. https://yulab-smu.top/biomedical-knowledge-mining- book/.