Users Guide

Directory

The first step: Data Preparation..... 1
(1). Data preparation for TCGA dataset..... 1
I.Analysis introduction..... 1
II.Parameters setting guides..... 2
i.Basic Settings..... 2
ii. TCGA Datasets Integration..... 6
iii.Clinical & Survival..... 7
iv.mRNA & lncRNA..... 9
v.DNA methylation..... 11
vi.copy number alterations..... 13
vii.binary somatic mutation..... 15
viii.radiomics..... 17
III.Examples with results interpretation..... 20
Parameter Settings:..... 20
(2). Data preparation for validation dataset..... 23
I.Analysis introduction..... 23
II.Parameters setting guides..... 23
i.Validation Datasets Preparation..... 23
ii.Clinical & Survival (Validation)..... 26
iii.mRNA & lncRNA (Validation)..... 28
iv.DNA methylation (Validation)..... 30
v.copy number alterations (Validation)..... 32
vi.binary somatic mutation (Validation)..... 34
vii.radiomics (Validation)..... 36
viii.Validation Datasets Integration..... 38
III.Examples with results interpretation..... 39
The second step: GET Module..... 41
(1). Get Elites for TCGA dataset and validation dataset..... 41
I.Analysis introduction..... 41
II.Parameters setting guides..... 41
i.GET Module..... 41
ii.Get Elites..... 42
iii.Get Elites settings for mRNA dataset..... 43
iv.Get Elites settings for lncRNA dataset..... 48
v.Get Elites settings for DNA methylation dataset..... 52
vi.Get Elites settings for copy number alterations dataset..... 56
vii.Get Elites settings for binary somatic mutation dataset..... 61
viii.Get Elites settings for radiomics dataset..... 64
ix.Process Get Elites..... 68
III.Examples with results interpretation..... 69

(2). Get Clustering Number for TCGA dataset..... 74
I.Analysis introduction..... 74
II.Parameters setting guides..... 74
Get Clustering Number..... 74
III.Examples with results interpretation..... 76
(3). Consensus Clustering for TCGA dataset..... 77
I.Analysis introduction..... 77
II.Parameters setting guides..... 78
i.Consensus Clustering..... 78
ii.iClusterBayes..... 79
iii. SNF..... 81
iv.PINSPlus..... 82
v. NEMO..... 83
vi. COCA..... 85
vii.LRAcluster..... 86
viii.ConsensusClustering..... 87
ix.IntNMF..... 89
x.CIMLR..... 90
xi.MoCluster..... 91
xii. Consensus Clustering..... 93
xiii.Consensus Heatmap..... 94
III.Examples with results interpretation..... 96
(4). Silhouette for TCGA dataset..... 100
I.Analysis introduction..... 100
II.Parameters setting guides..... 100
III.Examples with results interpretation..... 102
(5). Multi-omics Heatmaps for TCGA dataset..... 104
I.Analysis introduction..... 104
II.Parameters setting guides..... 104
Multi-omics Heatmaps..... 104
III.Examples with results interpretation..... 108
The third step: COMP Module (TCGA dataset)..... 113
(1). Compare survival outcome for TCGA dataset..... 114
I.Analysis introduction..... 114
II.Parameters setting guides..... 114
i.COMP Module..... 114
ii.Compare survival outcome..... 115
III.Examples with results interpretation..... 119
(2). Compare clinical features for TCGA dataset..... 122
I.Analysis introduction..... 122
II.Parameters setting guides..... 122
Compare clinical features..... 122
III.Examples with results interpretation..... 126
(3). Compare mutational frequency for TCGA dataset..... 129

I.Analysis introduction..... 129
II.Parameters setting guides..... 130
Compare mutational frequency..... 130
III.Examples with results interpretation..... 135
(4). Compare total mutation burden for TCGA dataset..... 139
I.Analysis introduction..... 139
II.Parameters setting guides..... 139
Compare total mutation burden..... 139
III.Examples with results interpretation..... 143
(5). Compare fraction genome altered for TCGA dataset..... 146
I.Analysis introduction..... 146
II.Parameters setting guides..... 147
Compare fraction genome altered..... 147
III.Examples with results interpretation..... 150
(6). Compare drug sensitivity for TCGA dataset..... 152
I.Analysis introduction..... 152
II.Parameters setting guides..... 153
Compare drug sensitivity..... 153
III.Examples with results interpretation..... 156
(7). Compare agreement with other subtypes for TCGA dataset..... 160
I.Analysis introduction..... 160
II.Parameters setting guides..... 160
Compare agreement with other subtypes..... 160
III.Examples with results interpretation..... 163
The fourth step: RUN Module..... 166
(1). Run differential expression analysis for TCGA dataset..... 167
I.Analysis introduction..... 167
II.Parameters setting guides..... 167
i.RUN Module..... 167
ii.Run differential expression analysis..... 169
III.Examples with results interpretation..... 170
(2). Run biomarker identification procedure for TCGA dataset..... 172
I.Analysis introduction..... 172
II.Parameters setting guides..... 172
Run biomarker identification procedure..... 172
III.Examples with results interpretation..... 178
(3). Run gene set enrichment analysis for TCGA dataset..... 183
I.Analysis introduction..... 183
II.Parameters setting guides..... 183
i.Prepare a gene set background file..... 183
ii.Run gene set enrichment analysis..... 184
III.Examples with results interpretation..... 188
(4). Run gene set variation analysis for TCGA dataset..... 194
I.Analysis introduction..... 194

II.Parameters setting guides..... 195
i.Prepare a gene set list of interest..... 195
ii.Run gene set variation analysis..... 196
III.Examples with results interpretation..... 200
(5). Run nearest template prediction for TCGA dataset and validation datasets 205
I.Analysis introduction..... 205
II.Parameters setting guides..... 206
Run nearest template prediction..... 206
III.Examples with results interpretation..... 209
(6). Run partition around medoids classifier for TCGA dataset and validation datasets..... 214
I.Analysis introduction..... 214
II.Parameters setting guides..... 214
Run partition around medoids classifier..... 214
III.Examples with results interpretation..... 217
(7). Run consistency evaluation using Kappa statistics for TCGA dataset and validation datasets..... 221
I.Analysis introduction..... 221
II.Parameters setting guides..... 221
i.Run consistency evaluation using Kappa statistics..... 221
ii. Run Kappa on tcga cohort (CMOIC vs NTP)..... 223
iii.Run Kappa on tcga cohort (CMOIC vs PAM)..... 224
iv.Run Kappa on tcga cohort (NTP vs PAM)..... 226
v.Run Kappa on validation cohort (NTP vs PAM)..... 227
vi.Finish..... 229
III.Examples with results interpretation..... 230
The fifth step: COMP Module (Validation dataset)..... 236
(1). Compare survival outcome for validation dataset..... 236
I.Analysis introduction..... 236
II.Parameters setting guides..... 236
III.Examples with results interpretation..... 236
(2). Compare clinical features for validation dataset..... 240
I.Analysis introduction..... 240
II.Parameters setting guides..... 240
III.Examples with results interpretation..... 240
(3). Compare drug sensitivity for validation dataset..... 244
I.Analysis introduction..... 244
II.Parameters setting guides..... 244
III.Examples with results interpretation..... 244
(4). Compare agreement with other subtypes for validation dataset..... 250
I.Analysis introduction..... 250
II.Parameters setting guides..... 250
III.Examples with results interpretation..... 250

MOVICShiny is an interactive website designed for multi-omics integration and visualization in cancer subtyping without coding based on MOVICS R package. To utilize this website smoothly, we need to follow the corresponding procedures and some rules, and the use of high-performance server saves a lot of time and improves analysis efficiency. This document will introduce the use of the website through three aspects, including analysis introduction, parameters setting guides and examples with results interpretation.

The first step: Data Preparation

Before multi-omics anaysis, we need to finish data preparation first. “Data Preparation” is divided into TCGA dataset preparation and external validation dataset preparation, providing several ways for data preparation, which can be chosen by users according to actual situations.

(1). Data preparation for TCGA dataset

I.Analysis introduction

For the preparation of TCGA dataset, the software plans to provide several ways, including direct acquisition from built-in downloaded datasets, automatic downloading from websites, downloading from specified website links, and uploading locally. The software will preprocess and integrate the obtained data according to certain rules. In addition, users can also preprocess and integrate the data themselves, then upload the

integrated dataset to the software directly. Users can choose different data preparation ways based on actual situation. If there is no need to make too many personalized adjustments on data, it is most convenient to download the data automatically from websites, and then preprocess and integrate them by software. On the contrary, it is most appropriate for users to preprocess and integrate the data locally first, and then directly upload the integrated dataset to the software.

II.Parameters setting guides

i.Basic Settings

Welcome to MOVICShiny. Below, we will begin using this platform for multi-omics analysis. First, we need to configure some basic information, including the user's account details, the TCGA cancer type for research, data type, and the method of data preparation. The following are guidelines for configuring each parameter.

Start by entering the user's account information:

Username (E-mail Address): Please enter the email address you used when applying. To facilitate user management, applying for access is necessary before using this platform. If you haven't applied yet, please send an email to xlu.cpu@foxmail.com, providing your name, affiliation, and the purpose of using this platform.

User Account: Please enter your login account. If your application is successful, you will receive a confirmation email containing your login

account. The validity period of this login account is 30 days. Please complete the corresponding analysis within this timeframe and ensure to promptly save the obtained data and results locally.

Whether log in for the first time using this user account: If this is your first time logging in with this account, please select "Yes"; otherwise, choose "No". After configuring these three parameters, click the "Create" button to log in. The feedback on whether the login was successful will be provided in the Data Preparation (TCGA) results module on the right side of the webpage.

Next, configure the TCGA cancer type you want to study:

The Number of Cancer Types: Number of TCGA Cancer Types to Study: Typically set to 1 (default). However, if a specific cancer has multiple subtypes in TCGA, you can choose to study several subtypes together. For example, NSCLC in TCGA includes LUAD and LUSC.

Input Cancer Types From TCGA: Specify the types of cancer according to the specified number of cancer types. Enter the cancer types using the four-letter abbreviations as per TCGA, such as BLCA.

Next, specify the data type:

Multi-omics Types (Choose at least two types): By default, five data types, excluding imaging genomics, have been selected. Users can customize their selection based on their specific requirements, but they must choose at least two data types.

Finally, specify the data preparation method:

①. Use the built-in TCGA dataset within the software.

Data Sources: Choose "Internal resources".

②. Obtain the dataset from external sources (including downloading from official websites, specified links, or uploading files locally).

Data Sources: choose “External resources”.

Use integrated datasets (.RData) or not: choose “No”.

③. Upload the pre-processed .RData dataset from your local storage.

Data Sources: choose “External resources”.

Use integrated datasets (.RData) or not: choose “Yes”.

The approach to obtain integrated datasets: If obtaining the .RData dataset from a specified link, choose "Download from specified URLs"; if uploading the .RData dataset locally, choose "Upload manually."

Download url for the integrated datasets (.RData): If obtaining the .RData dataset from a specified link, please enter the download link here.

Upload the integrated datasets (.RData): If uploading the .RData dataset locally, click the "Browse" button here to select the .RData dataset from your local storage and upload it to the platform.

The three data preparation methods each have their advantages and disadvantages. Users can choose based on their specific needs. Method one, using the built-in TCGA dataset, is the fastest but may not provide the most up-to-date data (we update the dataset every 30 days). Method two

allows access to the latest data but involves a longer data preparation time. Method three allows users to personalize the dataset in advance, ensuring that subsequent analysis results meet expectations. Therefore, the platform recommends using method three as the default data preparation method.

As for the question of whether to use external datasets for validation:

External validation or not: If you need to use an external dataset for validation, choose "Yes" (default); otherwise, choose "No".

If you chose method one for data preparation and processing, click the "Process" button. Once completed, feedback will be provided in the Data Preparation (TCGA) results module on the right side of the webpage. If the "Multi-omics Types" parameter includes "radiomics," you'll need to separately prepare the TCGA dataset for radiomics. The platform doesn't include a built-in TCGA radiomics dataset, and you can find detailed parameter settings in the guidance document for the radiomics parameter module. After completing the preparation of all TCGA data types, you need to go to the TCGA Datasets Integration parameter module and click the "Integrate" button for data integration, as detailed in the guidance document for that parameter module.

If you chose method two for data preparation and processing, you need to further set the preparation method for each data type. Refer to the guidance documents for the Clinical & Survival, mRNA & lncRNA, DNA

methylation, copy number alterations, binary somatic mutation, and radiomics parameter modules.

If you chose method three for data preparation and processing, click the "Process" button. Once completed, feedback will be provided in the Data Preparation (TCGA) results module on the right side of the webpage. Afterward, you need to go to the TCGA Datasets Integration parameter module and click the "Integrate" button for data integration, as detailed in the guidance document for that parameter module.

ii. TCGA Datasets Integration

If you're using the software's built-in TCGA dataset or obtained a dataset from external sources (including downloading from official websites, specified links, or uploading files locally), after completing the data preparation for all types, you need to go to the TCGA Datasets Integration parameter module and click the "Integrate" button to complete the data integration. Once integration is complete, feedback will be provided in the Data Integration (TCGA) results module on the right side of the webpage.

If the "External validation or not" parameter in the Basic Settings module is set to "Yes", you'll need to proceed to prepare external validation data (refer to the guidance document for the Validation Datasets Preparation parameter module). Otherwise, click the "GET Module" button at the top of the webpage to begin multi-omics analysis (refer to the guidance document for the Steps parameter module under the GET Module).

iii.Clinical & Survival

If obtaining the dataset from external sources (including downloading from official websites, specified links, or uploading files locally), you need to specify the data preparation method for each type of data.

Approaches: Specify the data preparation method, including downloading automatically from the official website, downloading from specified URLs, and uploading manually from local files.

①. If choosing Download automatically

Simply click the "Process" button to initiate the download and processing of TCGA data. Upon completion, the Data Preparation (TCGA) results module on the right side of the webpage will provide feedback.

②. If choosing Download from specified urls

Url sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Download url for combined clinical dataset: If "Url sources" is selected as "Combined dataset", input the download link for the combined dataset in this parameter.

Input clinical dataset url for each cancer type: If "Url sources" is selected as "Separated datasets", enter the download links for each dataset one by one. Finally, click the "Process" button to start downloading and

processing TCGA data. Upon completion, the Data Preparation (TCGA) results module on the right side of the webpage will provide feedback.

③. If choosing Upload manually

Upload sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Upload the combined clinical dataset (a .txt file): If "Upload sources" is selected as "Combined dataset", click the "Browse" button to upload the combined dataset from your local storage.

Upload the clinical dataset sequentially for each cancer type at once (a .txt file): If "Upload sources" is selected as "Separated datasets", click the "Browse" button sequentially to upload each dataset from your local storage. Please note: you can only upload one dataset at a time. Wait for the first dataset to finish uploading before clicking the "Browse" button again to upload another dataset.

Finally, click the "Process" button to initiate the upload and processing of TCGA data. Upon completion, the Data Preparation (TCGA) results module on the right side of the webpage will provide feedback.

After the preparation and processing of each data type is completed, feedback will be provided in the Data Preparation (TCGA) results module on the right side of the webpage. Once the TCGA data preparation for all

data types is finished, proceed to the TCGA Datasets Integration parameter module and click the "Integrate" button for data integration. Refer to the guidance document for that parameter module for detailed instructions.

iv.mRNA & lncRNA

If obtaining the dataset from external sources (including downloading from official websites, specified links, or uploading files locally), you need to specify the data preparation method for each type of data.

Approaches: Specify the data preparation method, including downloading automatically from the official website, downloading from specified URLs, and uploading manually from local files.

①. If choosing Download automatically

Simply click the "Process" button to initiate the download and processing of TCGA data. Upon completion, the Data Preparation (TCGA) results module on the right side of the webpage will provide feedback.

②. If choosing Download from specified urls

Url sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Download url for combined clinical dataset: If "Url sources" is selected as "Combined dataset", input the download link for the combined dataset in

this parameter.

Input clinical dataset url for each cancer type: If "Url sources" is selected as "Separated datasets", enter the download links for each dataset one by one. Finally, click the "Process" button to start downloading and processing TCGA data. Upon completion, the Data Preparation (TCGA) results module on the right side of the webpage will provide feedback.

③. If choosing Upload manually

Upload sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Upload the combined clinical dataset (a .txt file): If "Upload sources" is selected as "Combined dataset", click the "Browse" button to upload the combined dataset from your local storage.

Upload the clinical dataset sequentially for each cancer type at once (a .txt file): If "Upload sources" is selected as "Separated datasets", click the "Browse" button sequentially to upload each dataset from your local storage. Please note: you can only upload one dataset at a time. Wait for the first dataset to finish uploading before clicking the "Browse" button again to upload another dataset.

Finally, click the "Process" button to initiate the upload and processing of TCGA data. Upon completion, the Data Preparation (TCGA) results module

on the right side of the webpage will provide feedback.

After the preparation and processing of each data type is completed, feedback will be provided in the Data Preparation (TCGA) results module on the right side of the webpage. Once the TCGA data preparation for all data types is finished, proceed to the TCGA Datasets Integration parameter module and click the "Integrate" button for data integration. Refer to the guidance document for that parameter module for detailed instructions.

v.DNA methylation

If obtaining the dataset from external sources (including downloading from official websites, specified links, or uploading files locally), you need to specify the data preparation method for each type of data.

Approaches: Specify the data preparation method, including downloading automatically from the official website, downloading from specified URLs, and uploading manually from local files.

①. If choosing Download automatically

Simply click the "Process" button to initiate the download and processing of TCGA data. Upon completion, the Data Preparation (TCGA) results module on the right side of the webpage will provide feedback.

②. If choosing Download from specified urls

Url sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose

"Separated datasets". For a dataset of a single cancer type, either option can be selected.

Download url for combined clinical dataset: If "Url sources" is selected as "Combined dataset", input the download link for the combined dataset in this parameter.

Input clinical dataset url for each cancer type: If "Url sources" is selected as "Separated datasets", enter the download links for each dataset one by one. Finally, click the "Process" button to start downloading and processing TCGA data. Upon completion, the Data Preparation (TCGA) results module on the right side of the webpage will provide feedback.

③. If choosing Upload manually

Upload sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Upload the combined clinical dataset (a .txt file): If "Upload sources" is selected as "Combined dataset", click the "Browse" button to upload the combined dataset from your local storage.

Upload the clinical dataset sequentially for each cancer type at once (a .txt file): If "Upload sources" is selected as "Separated datasets", click the "Browse" button sequentially to upload each dataset from your local storage. Please note: you can only upload one dataset at a time. Wait for

the first dataset to finish uploading before clicking the "Browse" button again to upload another dataset.

Finally, click the "Process" button to initiate the upload and processing of TCGA data. Upon completion, the Data Preparation (TCGA) results module on the right side of the webpage will provide feedback.

After the preparation and processing of each data type is completed, feedback will be provided in the Data Preparation (TCGA) results module on the right side of the webpage. Once the TCGA data preparation for all data types is finished, proceed to the TCGA Datasets Integration parameter module and click the "Integrate" button for data integration. Refer to the guidance document for that parameter module for detailed instructions.

vi.copy number alterations

If obtaining the dataset from external sources (including downloading from official websites, specified links, or uploading files locally), you need to specify the data preparation method for each type of data.

Approaches: Specify the data preparation method, including downloading automatically from the official website, downloading from specified URLs, and uploading manually from local files.

①. If choosing Download automatically

Simply click the "Process" button to initiate the download and processing of TCGA data. Upon completion, the Data Preparation (TCGA) results

module on the right side of the webpage will provide feedback.

②. If choosing Download from specified urls

Url sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Download url for combined clinical dataset: If "Url sources" is selected as "Combined dataset", input the download link for the combined dataset in this parameter.

Input clinical dataset url for each cancer type: If "Url sources" is selected as "Separated datasets", enter the download links for each dataset one by one. Finally, click the "Process" button to start downloading and processing TCGA data. Upon completion, the Data Preparation (TCGA) results module on the right side of the webpage will provide feedback.

③. If choosing Upload manually

Upload sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Upload the combined clinical dataset (a .txt file): If "Upload sources" is selected as "Combined dataset", click the "Browse" button to upload the combined dataset from your local storage.

Upload the clinical dataset sequentially for each cancer type at once (a .txt file): If "Upload sources" is selected as "Separated datasets", click the "Browse" button sequentially to upload each dataset from your local storage. Please note: you can only upload one dataset at a time. Wait for the first dataset to finish uploading before clicking the "Browse" button again to upload another dataset.

Finally, click the "Process" button to initiate the upload and processing of TCGA data. Upon completion, the Data Preparation (TCGA) results module on the right side of the webpage will provide feedback.

After the preparation and processing of each data type is completed, feedback will be provided in the Data Preparation (TCGA) results module on the right side of the webpage. Once the TCGA data preparation for all data types is finished, proceed to the TCGA Datasets Integration parameter module and click the "Integrate" button for data integration. Refer to the guidance document for that parameter module for detailed instructions.

vii.binary somatic mutation

If obtaining the dataset from external sources (including downloading from official websites, specified links, or uploading files locally), you need to specify the data preparation method for each type of data.

Approaches: Specify the data preparation method, including downloading automatically from the official website, downloading from specified URLs,

and uploading manually from local files.

①. If choosing Download automatically

Simply click the "Process" button to initiate the download and processing of TCGA data. Upon completion, the Data Preparation (TCGA) results module on the right side of the webpage will provide feedback.

②. If choosing Download from specified urls

Url sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Download url for combined clinical dataset: If "Url sources" is selected as "Combined dataset", input the download link for the combined dataset in this parameter.

Input clinical dataset url for each cancer type: If "Url sources" is selected as "Separated datasets", enter the download links for each dataset one by one. Finally, click the "Process" button to start downloading and processing TCGA data. Upon completion, the Data Preparation (TCGA) results module on the right side of the webpage will provide feedback.

③. If choosing Upload manually

Upload sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option

can be selected.

Upload the combined clinical dataset (a .txt file): If "Upload sources" is selected as "Combined dataset", click the "Browse" button to upload the combined dataset from your local storage.

Upload the clinical dataset sequentially for each cancer type at once (a .txt file): If "Upload sources" is selected as "Separated datasets", click the "Browse" button sequentially to upload each dataset from your local storage. Please note: you can only upload one dataset at a time. Wait for the first dataset to finish uploading before clicking the "Browse" button again to upload another dataset.

Finally, click the "Process" button to initiate the upload and processing of TCGA data. Upon completion, the Data Preparation (TCGA) results module on the right side of the webpage will provide feedback.

After the preparation and processing of each data type is completed, feedback will be provided in the Data Preparation (TCGA) results module on the right side of the webpage. Once the TCGA data preparation for all data types is finished, proceed to the TCGA Datasets Integration parameter module and click the "Integrate" button for data integration. Refer to the guidance document for that parameter module for detailed instructions.

viii.radiomics

If obtaining the dataset from external sources (including downloading

from official websites, specified links, or uploading files locally), you need to specify the data preparation method for each type of data.

Approaches: The preparation methods for the radiomics dataset include only "Download from specified URLs" and "Upload manually". It's important to note that this file is not the original image but rather an imageomics feature matrix extracted from the original images. A common method for extraction is using PyRadiomics.

①. If choosing Download from specified urls

Url sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Download url for combined clinical dataset: If "Url sources" is selected as "Combined dataset", input the download link for the combined dataset in this parameter.

Input clinical dataset url for each cancer type: If "Url sources" is selected as "Separated datasets", enter the download links for each dataset one by one.

Finally, click the "Process" button to start downloading and processing TCGA data. Upon completion, the Data Preparation (TCGA) results module on the right side of the webpage will provide feedback.

②. If choosing Upload manually

Upload sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Upload the combined clinical dataset (a .txt file): If "Upload sources" is selected as "Combined dataset", click the "Browse" button to upload the combined dataset from your local storage.

Upload the clinical dataset sequentially for each cancer type at once (a .txt file): If "Upload sources" is selected as "Separated datasets", click the "Browse" button sequentially to upload each dataset from your local storage. Please note: you can only upload one dataset at a time. Wait for the first dataset to finish uploading before clicking the "Browse" button again to upload another dataset.

Finally, click the "Process" button to initiate the upload and processing of TCGA data. Upon completion, the Data Preparation (TCGA) results module on the right side of the webpage will provide feedback.

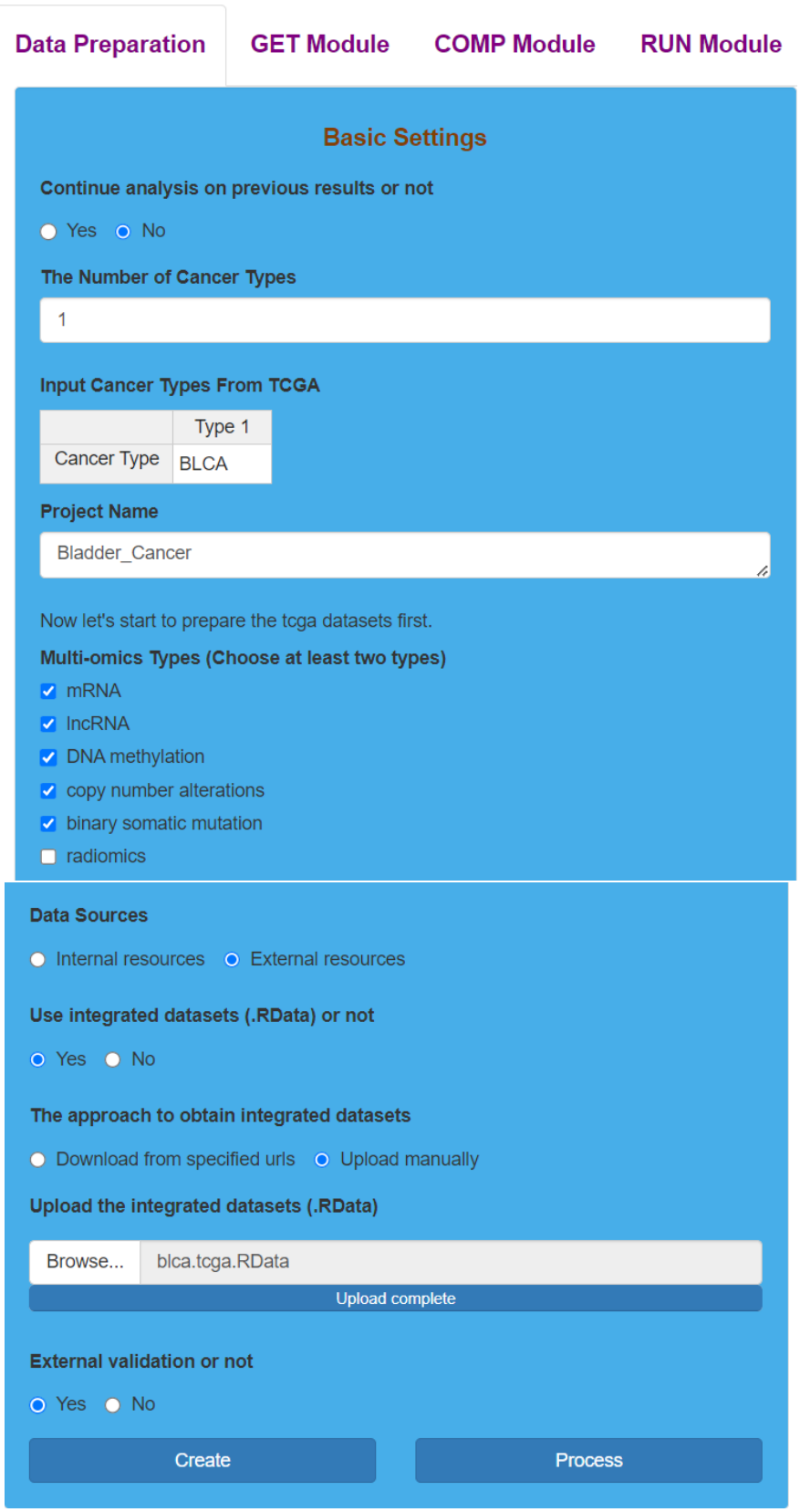
After the preparation and processing of each data type is completed, feedback will be provided in the Data Preparation (TCGA) results module on the right side of the webpage. Once the TCGA data preparation for all data types is finished, proceed to the TCGA Datasets Integration parameter module and click the "Integrate" button for data integration. Refer to the guidance document for that parameter module for detailed

instructions.

III.Examples with results interpretation

Parameter Settings: Before data preparation, we should first write a mail to 'xlu.cpu@foxmail.com' (Name, Affiliation, purpose, etc.) to apply for a Username (E-mail Address) and a User Account. If we log in for the first time using the applied username and user account, “Whether log in for the first time using this user account” should choose “Yes”, otherwise it should choose “No” after inputting “Username (E-mail Address)” and “User Account”. Then, we fill in the name of the tumor in TCGA database (“Input Cancer Types From TCGA” enters “BLCA”). After that, we should specify a project name (“Project Name” entered “Bladder_Cancer”) and choose at least 2 kinds of omics data types (“Multi-omics Types” chooses all omics data types except “radiomics” by default). Then, “Data Sources” chooses “External resources” and we use integrated dataset for analysis directly (“Use integrated datasets or not” chooses “Yes”) due to many adjustments on clinical data. We click the “Browse” button to upload the integrated dataset, and “External validation or not” should choose “Yes” for validation. We should prepare the integrated dataset by ourselves and save it as a “.RData” file. The integrated dataset is a data list containing information which should be named as “mRNA.expr”, “lncRNA.expr”, “meth.beta”, “cna”, “mut.stauts”, “radiomics”, “count”, “tpm” or (”fpkm_mrna”+”fpkm_lncrna”), “maf”, “segment”, “clin.info” if exists.

Finally, we click the “Create” button to create an account, and then click the “Process” button to move the integrated dataset to the specified file path. If you want to continue the analysis using previous account, “Continue analysis on previous results or not” chooses “Yes” and “Please input the previously used user account” enters the previous account ID. Then, you should indicate the following parameters in turn: “The Number of Cancer Types”, “Input Cancer Types From TCGA”, “Project Name”, “Multi-omics Types” for TCGA dataset, “Multi-omics Types” for validation dataset, “Clustering algorithms” for “Consensus Clustering” in “GET Module” step. If you utilize user-defined clustering number, “The number of clusters” should choose “User defined” and you should indicate the clustering number for “User defined cluster number”.

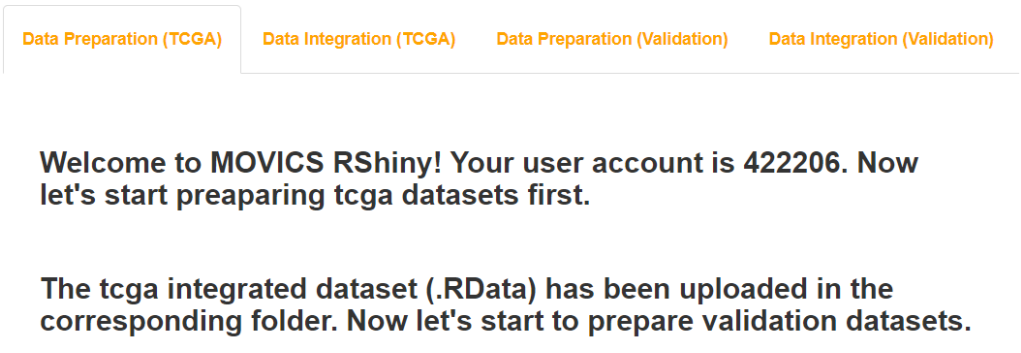


Parameter settings for TCGA dataset preparation

Result Display:

The software will give feedback to the users after finishing the process of

preparing TCGA dataset.



Result display for TCGA dataset preparation

(2). Data preparation for validation dataset

I.Analysis introduction

For the preparation of external validation dataset, two data preparation ways of obtaining from the built-in downloaded datasets as well as automatically downloading from websites are no longer provided due to the wide range of sources for external validation dataset. In terms of omics data types, the software plans to add radiomics data on the basis of five original data types from MOVICS R package, containing mRNA, lncRNA, DNA methylation, copy number alterations, and somatic mutation.

II.Parameters setting guides

i.Validation Datasets Preparation

If the "External validation or not" parameter in the Basic Settings module is set to "Yes", you need to proceed to prepare external validation data. Similarly, you will need to specify the data type and data preparation method.

Start by clicking the "Create" button to create a folder for storing the external validation dataset. Upon completion, feedback will be provided in the Data Preparation (Validation) results module on the right side of the webpage.

Then specify the data type:

Multi-omics Types: The default selection includes five data types, excluding imaging genomics. Users can customize their selection based on their specific needs and choose the data types they require.

Finally, you need to specify the data preparation method:

①. Obtain the dataset from external sources (including downloading from specified links or uploading files locally).

Use integrated datasets (.RData) or not: choose “No”.

②. If uploading an integrated .RData dataset locally, click the "Browse" button to select and upload the integrated .RData dataset from your local storage to the platform.

Use integrated datasets (.RData) or not: choose “Yes”.

The approach to obtain integrated datasets: If obtaining the .RData dataset from a specified link, choose "Download from specified URLs"; if uploading the .RData dataset locally, choose "Upload manually".

Download url for the integrated datasets (.RData): If obtaining the .RData dataset from a specified link, you need to enter the download link in this field.

Upload the integrated datasets (.RData): If uploading the .RData dataset locally, click the "Browse" button to select and upload the .RData dataset from your local storage to the platform.

Method one only requires users to provide various types of data files, and the platform will process the data according to certain rules. Method two requires users to handle and integrate various types of omics data themselves, allowing for personalized data processing to achieve analysis results that better align with expectations. Therefore, the platform recommends using method two and sets it as the default data preparation method for external validation data. Users can choose the most suitable method based on their specific circumstances.

If you choose method two for external validation data preparation, click the "Process" button to initiate data preparation and processing. Upon completion, feedback will be provided in the Data Preparation (Validation) results module on the right side of the webpage. Then, you can click the "GET Module" button at the top of the webpage to start multi-omics analysis (refer to the guidance document for the Steps parameter module under the GET Module module). If you choose method one, you need to further set the preparation method for each type of data. Refer to the guidance documents for the Clinical & Survival (Validation), mRNA & lncRNA (Validation), DNA methylation (Validation), copy number

alterations (Validation), binary somatic mutation (Validation), and radiomics (Validation) parameter modules.

ii.Clinical & Survival (Validation)

If choosing to obtain datasets from external sources (including downloading from specified links or uploading files locally), you need to specify the data preparation method for each type of data individually.

Approaches: Specify the data preparation method, including Download from specified urls and Upload manually.

①. If choosing “Download from specified urls”

Url sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Download url for combined clinical dataset: If "Url sources" is selected as "Combined dataset," input the download link for the combined dataset in this parameter.

Input clinical dataset url for each cancer type: If "Url sources" is selected as "Separated datasets", enter the download links for each dataset one by one.

Finally, click the "Process" button to start downloading and processing external validation data. Upon completion, the Data Preparation (Validation) results module on the right side of the webpage will provide

feedback.

② . If choosing “Upload manually”

Upload sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Upload the combined clinical dataset (a .txt file): If "Upload sources" is selected as "Combined dataset", click the "Browse" button to upload the combined dataset from your local storage.

Upload the clinical dataset sequentially for each cancer type at once (a .txt file): If "Upload sources" is selected as "Separated datasets," click the "Browse" button one by one to upload each dataset from your local storage. Please note: you can only upload one dataset at a time. Wait for the first dataset to finish uploading before clicking the "Browse" button again to upload another dataset.

Finally, click the "Process" button to start uploading and processing external validation data. Upon completion, the Data Preparation (Validation) results module on the right side of the webpage will provide feedback.

After the preparation and processing of each data type is completed, feedback will be provided in the Data Preparation (Validation) results module on the right side of the webpage. Once the external validation

data preparation for all data types is finished, proceed to the Validation Datasets Integration parameter module and click the "Integrate" button for data integration. Refer to the guidance document for that parameter module for detailed instructions.

iii.mRNA & lncRNA (Validation)

If choosing to obtain datasets from external sources (including downloading from specified links or uploading files locally), you need to specify the data preparation method for each type of data individually.

Approaches: Specify the data preparation method, including Download from specified urls and Upload manually.

③ . If choosing “Download from specified urls”

Url sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Download url for combined clinical dataset: If "Url sources" is selected as "Combined dataset," input the download link for the combined dataset in this parameter.

Input clinical dataset url for each cancer type: If "Url sources" is selected as "Separated datasets", enter the download links for each dataset one by one.

Finally, click the "Process" button to start downloading and processing

external validation data. Upon completion, the Data Preparation (Validation) results module on the right side of the webpage will provide feedback.

④ . If choosing “Upload manually”

Upload sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Upload the combined clinical dataset (a .txt file): If "Upload sources" is selected as "Combined dataset", click the "Browse" button to upload the combined dataset from your local storage.

Upload the clinical dataset sequentially for each cancer type at once (a .txt file): If "Upload sources" is selected as "Separated datasets," click the "Browse" button one by one to upload each dataset from your local storage. Please note: you can only upload one dataset at a time. Wait for the first dataset to finish uploading before clicking the "Browse" button again to upload another dataset.

Finally, click the "Process" button to start uploading and processing external validation data. Upon completion, the Data Preparation (Validation) results module on the right side of the webpage will provide feedback.

After the preparation and processing of each data type is completed,

feedback will be provided in the Data Preparation (Validation) results module on the right side of the webpage. Once the external validation data preparation for all data types is finished, proceed to the Validation Datasets Integration parameter module and click the "Integrate" button for data integration. Refer to the guidance document for that parameter module for detailed instructions.

iv.DNA methylation (Validation)

If choosing to obtain datasets from external sources (including downloading from specified links or uploading files locally), you need to specify the data preparation method for each type of data individually.

Approaches: Specify the data preparation method, including Download from specified urls and Upload manually.

⑤ . If choosing “Download from specified urls”

Url sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Download url for combined clinical dataset: If "Url sources" is selected as "Combined dataset," input the download link for the combined dataset in this parameter.

Input clinical dataset url for each cancer type: If "Url sources" is selected as "Separated datasets", enter the download links for each dataset one by

one.

Finally, click the "Process" button to start downloading and processing external validation data. Upon completion, the Data Preparation (Validation) results module on the right side of the webpage will provide feedback.

⑥ . If choosing “Upload manually”

Upload sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Upload the combined clinical dataset (a .txt file): If "Upload sources" is selected as "Combined dataset", click the "Browse" button to upload the combined dataset from your local storage.

Upload the clinical dataset sequentially for each cancer type at once (a .txt file): If "Upload sources" is selected as "Separated datasets," click the "Browse" button one by one to upload each dataset from your local storage. Please note: you can only upload one dataset at a time. Wait for the first dataset to finish uploading before clicking the "Browse" button again to upload another dataset.

Finally, click the "Process" button to start uploading and processing external validation data. Upon completion, the Data Preparation (Validation) results module on the right side of the webpage will provide

feedback.

After the preparation and processing of each data type is completed, feedback will be provided in the Data Preparation (Validation) results module on the right side of the webpage. Once the external validation data preparation for all data types is finished, proceed to the Validation Datasets Integration parameter module and click the "Integrate" button for data integration. Refer to the guidance document for that parameter module for detailed instructions.

v.copy number alterations (Validation)

If choosing to obtain datasets from external sources (including downloading from specified links or uploading files locally), you need to specify the data preparation method for each type of data individually.

Approaches: Specify the data preparation method, including Download from specified urls and Upload manually.

⑦ . If choosing “Download from specified urls”

Url sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Download url for combined clinical dataset: If "Url sources" is selected as "Combined dataset," input the download link for the combined dataset in this parameter.

Input clinical dataset url for each cancer type: If "Url sources" is selected as "Separated datasets", enter the download links for each dataset one by one.

Finally, click the "Process" button to start downloading and processing external validation data. Upon completion, the Data Preparation (Validation) results module on the right side of the webpage will provide feedback.

⑧ . If choosing “Upload manually”

Upload sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Upload the combined clinical dataset (a .txt file): If "Upload sources" is selected as "Combined dataset", click the "Browse" button to upload the combined dataset from your local storage.

Upload the clinical dataset sequentially for each cancer type at once (a .txt file): If "Upload sources" is selected as "Separated datasets," click the "Browse" button one by one to upload each dataset from your local storage. Please note: you can only upload one dataset at a time. Wait for the first dataset to finish uploading before clicking the "Browse" button again to upload another dataset.

Finally, click the "Process" button to start uploading and processing

external validation data. Upon completion, the Data Preparation (Validation) results module on the right side of the webpage will provide feedback.

After the preparation and processing of each data type is completed, feedback will be provided in the Data Preparation (Validation) results module on the right side of the webpage. Once the external validation data preparation for all data types is finished, proceed to the Validation Datasets Integration parameter module and click the "Integrate" button for data integration. Refer to the guidance document for that parameter module for detailed instructions.

vi.binary somatic mutation (Validation)

If choosing to obtain datasets from external sources (including downloading from specified links or uploading files locally), you need to specify the data preparation method for each type of data individually.

Approaches: Specify the data preparation method, including Download from specified urls and Upload manually.

⑨ . If choosing “Download from specified urls”

Url sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Download url for combined clinical dataset: If "Url sources" is selected as

"Combined dataset," input the download link for the combined dataset in this parameter.

Input clinical dataset url for each cancer type: If "Url sources" is selected as "Separated datasets", enter the download links for each dataset one by one.

Finally, click the "Process" button to start downloading and processing external validation data. Upon completion, the Data Preparation (Validation) results module on the right side of the webpage will provide feedback.

⑩ . If choosing “Upload manually”

Upload sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Upload the combined clinical dataset (a .txt file): If "Upload sources" is selected as "Combined dataset", click the "Browse" button to upload the combined dataset from your local storage.

Upload the clinical dataset sequentially for each cancer type at once (a .txt file): If "Upload sources" is selected as "Separated datasets," click the "Browse" button one by one to upload each dataset from your local storage. Please note: you can only upload one dataset at a time. Wait for the first dataset to finish uploading before clicking the "Browse" button

again to upload another dataset.

Finally, click the "Process" button to start uploading and processing external validation data. Upon completion, the Data Preparation (Validation) results module on the right side of the webpage will provide feedback.

After the preparation and processing of each data type is completed, feedback will be provided in the Data Preparation (Validation) results module on the right side of the webpage. Once the external validation data preparation for all data types is finished, proceed to the Validation Datasets Integration parameter module and click the "Integrate" button for data integration. Refer to the guidance document for that parameter module for detailed instructions.

vii.radiomics (Validation)

If choosing obtaining the dataset from external sources (including downloading from official websites, specified links, or uploading files locally), you need to specify the data preparation method for each type of data.

Approaches: The preparation methods for the radiomics dataset include only "Download from specified URLs" and "Upload manually". It's important to note that this file is not the original image but rather an imageomics feature matrix extracted from the original images. A common method for extraction is using PyRadiomics.

③ . If choosing Download from specified urls

Url sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Download url for combined clinical dataset: If "Url sources" is selected as "Combined dataset", input the download link for the combined dataset in this parameter.

Input clinical dataset url for each cancer type: If "Url sources" is selected as "Separated datasets", enter the download links for each dataset one by one.

Finally, click the "Process" button to start downloading and processing external validation data. Upon completion, the Data Preparation (Validation) results module on the right side of the webpage will provide feedback.

④ . If choosing Upload manually

Upload sources: If it is a combined dataset for multiple cancers, choose "Combined dataset"; if it is separate datasets for multiple cancers, choose "Separated datasets". For a dataset of a single cancer type, either option can be selected.

Upload the combined clinical dataset (a .txt file): If "Upload sources" is selected as "Combined dataset", click the "Browse" button to upload the

combined dataset from your local storage.

Upload the clinical dataset sequentially for each cancer type at once (a .txt file): If "Upload sources" is selected as "Separated datasets", click the "Browse" button sequentially to upload each dataset from your local storage. Please note: you can only upload one dataset at a time. Wait for the first dataset to finish uploading before clicking the "Browse" button again to upload another dataset.

Finally, click the "Process" button to start downloading and processing external validation data. Upon completion, the Data Preparation (Validation) results module on the right side of the webpage will provide feedback.

After the preparation and processing of each data type is completed, feedback will be provided in the Data Preparation (Validation) results module on the right side of the webpage. Once the external validation data preparation for all data types is finished, proceed to the external validation Datasets Integration parameter module and click the "Integrate" button for data integration. Refer to the guidance document for that parameter module for detailed instructions.

viii.Validation Datasets Integration

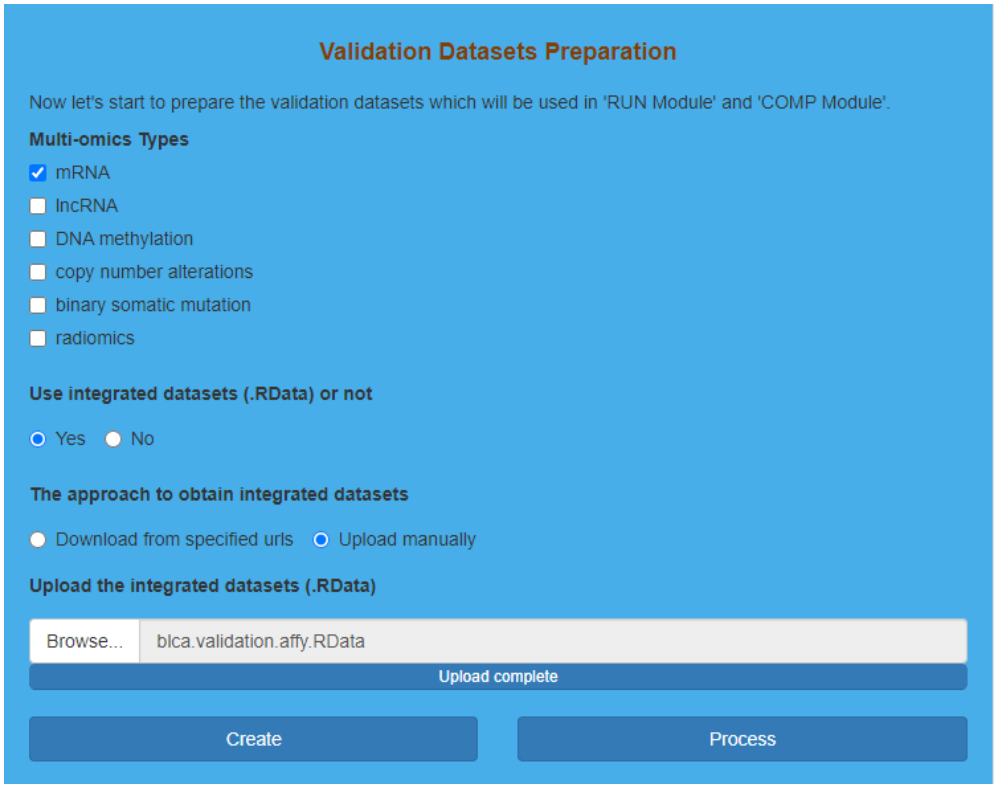
If choosing to obtain datasets from external sources (including downloading from specified links or uploading files locally), after completing the preparation for all data types, click the "Integrate" button

in the Validation Datasets Integration parameter module to complete data integration. Upon completion, the Data Integration (Validation) results module on the right side of the webpage will provide feedback.

At this point, all data is prepared. Click the "GET Module" button at the top of the webpage to start multi-omics analysis (refer to the guidance document for the Steps parameter module under the GET Module).

III.Examples with results interpretation

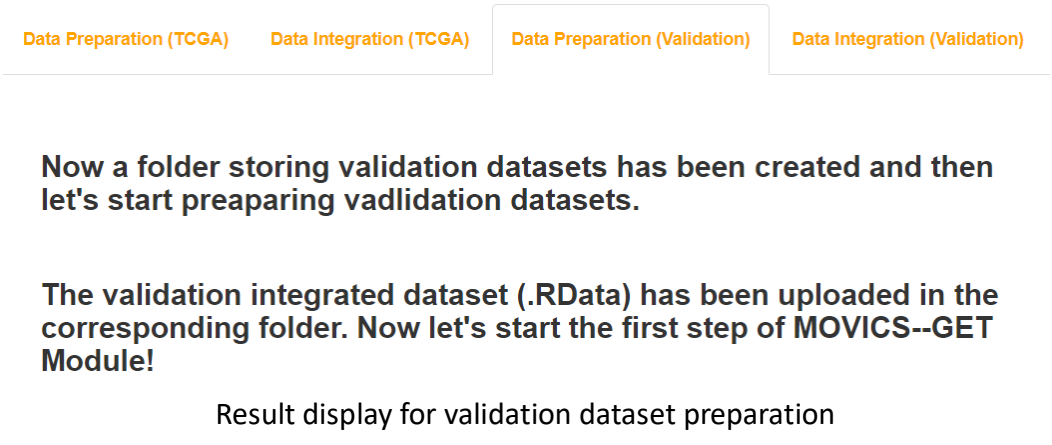
Parameter Settings: Since two validation cohorts only contain mRNA as well as clinical and survival information, “Multi-omics Types” should only choose “mRNA”. Then, we should upload the integrated dataset directly (“Use integrated datasets or not” chooses “Yes”) because some personalized processing needs to be done on the validation dataset. We select “Upload manually” for “The approach to obtain integrated datasets” and click the “Browse” button to upload the integrated validation dataset. Finally, we click the “Create” button to create a folder for validation dataset preparation, and then click the “Process” button to move the integrated validation dataset to the specified file path. It should be noted that only one validation dataset can be prepared at a time. We must finish the analysis of the current validation dataset, obtain as well as download and save the results before the preparation and analysis of another validation dataset. Otherwise, it will easily cause errors and confusion.



Parameter settings for validation dataset preparation

Result Display:

The software will also give feedback to the users after finishing the process of preparing validation dataset.



Result display for validation dataset preparation

The second step: GET Module

"GET Module" utilizes the obtained data to perform consensus clustering through specified algorithms to determine subtype for each sample. "GET Module" is divided into five sub-modules, including "Get Elites", "Get Optimal Clustering Number", "Consensus Clustering", "Silhouette" and "Multi-omics Heatmaps".

(1). Get Elites for TCGA dataset and validation dataset

I.Analysis introduction

"Get Elites" processes the obtained omics data sequentially and performs dimensionality reduction according to certain rules, which will be used for cluster analysis later.

II.Parameters setting guides

i.GET Module

After all the data preparation is complete, we will begin the multi-omics analysis. Firstly, the prepared data will undergo preprocessing to obtain elites. Next, we will determine the optimal clustering number using the Get Clustering Number method. Subsequently, we will perform consensus clustering using the appropriate technique. The next step involves evaluating the quality of the clustering results by calculating the sample-to-sample similarity of the obtained subtypes using Silhouette. Finally, we will generate multi-omics heatmaps based on the clustering results and preprocessed multi-omics data.

First, the prepared data needs to undergo preprocessing. For specific parameter settings, please refer to the guidance document in the Get Elites parameter module.

ii.Get Elites

Due to the typically large size of the prepared dataset, it is advisable to perform preprocessing to select valuable features and improve the efficiency of the multi-omics analysis before proceeding. We need to start by specifying the dataset object for preprocessing (TCGA dataset or external validation dataset). Then, for each data type present in the dataset, we will sequentially determine whether data preprocessing is necessary：

Get elites on tcga datasets or validation datasets: The default dataset object for preprocessing is the TCGA dataset.

Get Elites for mRNA dataset or not: The mRNA dataset undergoes data preprocessing by default. You can select "Yes" to confirm this preprocessing step.

Get Elites for lncRNA dataset or not: The lncRNA dataset also undergoes data preprocessing by default. You can select "Yes" to confirm this preprocessing step.

Get Elites for DNA methylation dataset or not: The DNA methylation dataset undergoes data preprocessing by default. You can select "Yes" to confirm this preprocessing step.

Get Elites for copy number alterations dataset or not: copy number alterations dataset undergoes data preprocessing by default. You can select "Yes" to confirm this preprocessing step.

Get Elites for binary somatic mutation dataset or not: The binary somatic mutation dataset undergoes data preprocessing by default. You can select "Yes" to confirm this preprocessing step.

Get Elites for radiomics dataset or not: The radiomics dataset undergoes data preprocessing by default. You can select "Yes" to confirm this preprocessing step.

Once you have specified the data preprocessing object and identified the data types that require preprocessing, you will need to set the data preprocessing parameters for each data type. For detailed instructions, please refer to the guidance document in the "Get Elites settings for mRNA dataset、Get Elites settings for lncRNA dataset、Get Elites settings for DNA methylation dataset、Get Elites settings for copy number alterations dataset、Get Elites settings for binary somatic mutation dataset and Get Elites settings for radiomics dataset" parameter module.

iii.Get Elites settings for mRNA dataset

The data can be categorized into two types, continuous and discrete. The mRNA dataset belongs to the continuous data type, with rows representing features and columns representing samples. To perform data preprocessing on the mRNA dataset, the relevant parameters need to be

set.

Firstly, filtering and log_2 normalization of samples and features are applied:

NA value action: The handling of NA (missing) values in the data can be done in three ways: Remove directly, KNN imputation, and No action. Remove directly: This option will delete samples that contain NA values. KNN imputation: It involves filling in NA values in samples using a K-nearest neighbors imputation method. No action: This option means no action will be taken to address NA values in the data.By default, the parameter uses "Remove directly," but users can choose the appropriate option based on the characteristics of their data.

Perform log2 transformation for data before calculating statistics or not: This parameter specifies whether log_2 normalization of the data is required to constrain larger expression values within a smaller range. The default choice is "No," meaning that log_2 normalization is not necessary. Users should make this determination based on the characteristics of their data. Typically, log_2 normalization is not needed for expression values that are smaller than 50.

Set a numeric cutoff for removing low expression features or not: This parameter specifies whether to filter out features with low expression frequency. The default choice is "No," indicating that features with low expression frequency will not be filtered out.

The cutoff for removing low expression features: If you select "Yes" for

the parameter "Set a numeric cutoff for removing low expression features or not," you will need to specify a threshold. Features with expression frequency (the number of samples expressing the feature divided by the total number of samples) below this threshold will be filtered out. The default value for this parameter is 0.1 (recommended).

After that, feature selection is performed：

Use survival information or not: Whether to use survival information is determined, and by default, it is set to "No." If the dataset contains survival information, users can choose to incorporate survival information and use the Cox method for feature selection. Features selected in this manner will have a higher correlation with the survival status of the samples.

If you choose to use survival information, further parameters will need to be configured：

A numeric cutoff for nominal p value derived from univariate Cox proportional hazards regression: If you choose to use survival information, you can set the threshold for the significance p-value. Features with p-values lower than this threshold are considered significantly associated with patient survival, and the default value is 0.05.If you do not use survival information, you will need to specify the feature selection method and threshold：

Choose a Get Elites method for mRNA dataset: There are three feature

selection methods available: MAD (Median Absolute Deviation), SD (Standard Deviation), and PCA (Principal Components Analysis). Users can choose the method that best suits their needs, with MAD being the default selection.

Choose a filtration method for mRNA dataset: If you choose "mad" or "sd" for the "Choose a Get Elites method for mRNA dataset" parameter, you will need to specify this parameter. This parameter provides two ways of feature selection: "elite.num": Select features based on a specific number. "elite.pct": Select features based on a specific proportion.The default method is "elite.num," where you specify the number of features you want to select.

An integer cutoff of exact number for selecting top elites: If you choose "elite.num" for the "Choose a filtration method for mRNA dataset" parameter, you will need to specify the number of features you want to select, denoted as "n." The platform will rank the results obtained from the "mad" or "sd" methods and select the top n features in descending order. The default value for this parameter is 1000.

A numeric cutoff of percentage for selecting top elites: If you choose "elite.pct" for the "Choose a filtration method for mRNA dataset" parameter, you will need to specify the feature selection proportion denoted as "n." The platform will rank the results obtained from the "mad" or "sd" methods and select the top features based on the specified

proportion n. The default value for this parameter is 0.1, which corresponds to 10% of the top features.

A numeric value which represents the ratio of principal components: If you choose "pca" for the "Choose a Get Elites method for mRNA dataset" parameter, you will need to specify the ratio used for selecting principal components. The default value is 0.95.Finally, the data is standardized, transforming it into a distribution with a mean of 0 and a variance of 1：

Centering the data after filtering or not: Whether to subtract the mean of their corresponding features is determined, and by default, it is set to "No," meaning that no subtraction of the mean is performed.

Scaling the data after filtering or not: Whether to divide by the standard deviation of their corresponding features is determined, and by default, it is set to "No," meaning that no division by the standard deviation is performed.

Once you have specified the preprocessing parameters for the mRNA data, you can proceed to set the parameters for other data types that require preprocessing. For detailed instructions, please refer to the guidance document in the " Get Elites settings for lncRNA dataset、Get Elites settings for DNA methylation dataset、Get Elites settings for copy number alterations dataset、Get Elites settings for binary somatic mutation dataset and Get Elites settings for radiomics dataset " parameter module.

iv.Get Elites settings for lncRNA dataset

The data can be categorized into two types, continuous and discrete. The lncRNA dataset falls under the category of continuous data, where rows represent features, and columns represent samples. To perform data preprocessing on the lncRNA dataset, relevant parameters need to be configured.

Firstly, filtering and log_2 normalization of samples and features are applied：

NA value action: The handling of NA (missing) values in the data can be done in three ways: Remove directly, KNN imputation, and No action.Remove directly: This option will delete samples that contain NA values. KNN imputation: It involves filling in NA values in samples using a K-nearest neighbors imputation method. No action: This option means no action will be taken to address NA values in the data.By default, the parameter uses "Remove directly," but users can choose the appropriate option based on the characteristics of their data.

Perform log2 transformation for data before calculating statistics or not: This parameter specifies whether log2 normalization of the data is required to constrain larger expression values within a smaller range. The default choice is "No," meaning that log2 normalization is not necessary. Users should make this determination based on the characteristics of their data. Typically, log2 normalization is not needed for expression values that

are smaller than 50.

Set a numeric cutoff for removing low expression features or not: This parameter specifies whether to filter out features with low expression frequency. The default choice is "No," indicating that features with low expression frequency will not be filtered out.

The cutoff for removing low expression features: If you select "Yes" for the parameter "Set a numeric cutoff for removing low expression features or not," you will need to specify a threshold. Features with expression frequency (the number of samples expressing the feature divided by the total number of samples) below this threshold will be filtered out. The default value for this parameter is 0.1 (recommended).

After that, feature selection is performed：

Use survival information or not: Whether to use survival information is determined, and by default, it is set to "No." If the dataset contains survival information, users can choose to incorporate survival information and use the Cox method for feature selection. Features selected in this manner will have a higher correlation with the survival status of the samples.

If you choose to use survival information, further parameters will need to be configured：

A numeric cutoff for nominal p value derived from univariate Cox proportional hazards regression: If you choose to use survival information,

you can set the threshold for the significance p-value. Features with p-values lower than this threshold are considered significantly associated with patient survival, and the default value is 0.05.If you do not use survival information, you will need to specify the feature selection method and threshold：

Choose a Get Elites method for lncRNA dataset: The available feature selection methods include MAD (Median Absolute Deviation), SD (Standard Deviation), and PCA (Principal Components Analysis). Users can choose the method that best suits their needs, with MAD being the default selection.

Choose a filtration method for lncRNA dataset: If you choose "mad" or "sd" for the "Choose a Get Elites method for lncRNA dataset" parameter, you will need to specify this parameter. This parameter provides two ways of feature selection: "elite.num": Select features based on a specific number. "elite.pct": Select features based on a specific proportion.The default method is "elite.num," where you specify the number of features you want to select.

An integer cutoff of exact number for selecting top elites: If you choose "elite.num" for the "Choose a filtration method for lncRNA dataset" parameter, you will need to specify the number of features you want to select, denoted as "n." The platform will rank the results obtained from the "mad" or "sd" methods and select the top n features in descending

order. The default value for this parameter is 1000.

A numeric cutoff of percentage for selecting top elites: If you choose "elite.pct" for the "Choose a filtration method for lncRNA dataset" parameter, you will need to specify the feature selection proportion, denoted as "n." The platform will rank the results obtained from the "mad" or "sd" methods and select the top features based on the specified proportion n. The default value for this parameter is 0.1, which corresponds to selecting the top 10% of features.

A numeric value which represents the ratio of principal components: If you choose "pca" for the "Choose a Get Elites method for lncRNA dataset" parameter, you will need to specify the ratio used for selecting principal components. The default value is 0.95, which means that enough principal components will be selected to explain 95% of the variance in the data.

Afterward, the data is standardized, transforming it into a distribution with a mean of 0 and a variance of 1：

Centering the data after filtering or not: Whether to subtract the mean of their corresponding features is an option. By default, it is set to "No," meaning that no subtraction of the mean is performed.

Scaling the data after filtering or not: Whether to divide by the standard deviation of their corresponding features is an option. By default, it is set to "No," meaning that no division by the standard deviation is performed.

After specifying the preprocessing parameters for the lncRNA dataset, you

should proceed to configure the preprocessing parameters for other data types as well. Refer to the documentation in the " Get Elites settings for mRNA dataset、Get Elites settings for DNA methylation dataset、Get Elites settings for copy number alterations dataset、Get Elites settings for binary somatic mutation dataset and Get Elites settings for radiomics dataset " parameter module for guidance on setting up preprocessing parameters for the radiomics dataset.

v.Get Elites settings for DNA methylation dataset

You can categorize different types of data into two classes: continuous and discrete. DNA methylation data falls under the category of continuous data, where rows represent features, and columns represent samples. To perform data preprocessing on the DNA methylation dataset, you need to set the appropriate parameters.

Firstly, filtering and $\log_2$ normalization of samples and features are applied.：

NA value action: The handling of NA (missing) values in the DNA methylation data can be done using three different methods: "Remove directly," "KNN imputation," or "No action." Remove directly: This method will remove samples that contain NA values. KNN imputation: It involves imputing (filling in) NA values in samples using a K-nearest neighbors imputation approach. No action: This option means no action will be taken to address NA values in the data.By default, the parameter uses "Remove

directly," but users can choose the appropriate option based on the characteristics of their data.

Perform log2 transformation for data before calculating statistics or not: This parameter specifies whether log_2 normalization of the data is needed to constrain larger expression values within a smaller range. The default choice is "No," meaning that log_2 normalization is not necessary. Users should make this determination based on the characteristics of their data. Typically, log_2 normalization is not needed for expression values that are smaller than 50.

Set a numeric cutoff for removing low expression features or not: This parameter specifies whether to filter out features with low expression frequency. The default choice is "No," indicating that features with low expression frequency will not be filtered out.

The cutoff for removing low expression features: If you select "Yes" for the parameter "Set a numeric cutoff for removing low expression features or not," you will need to specify a threshold. Features with expression frequency (the number of samples expressing the feature divided by the total number of samples) below this threshold will be filtered out. The default value for this parameter is 0.1, which is recommended.

After the preprocessing steps, the next step involves feature selection：

Use survival information or not: Whether to use survival information is an option, and by default, it is set to "No." If the dataset contains survival

information, users can choose to incorporate survival information and use the Cox method for feature selection. Features selected in this manner will have a higher correlation with the survival status of the samples.

If you choose to use survival information, you will need to further configure parameters related to survival analysis：

A numeric cutoff for nominal p value derived from univariate Cox proportional hazards regression: If you choose to use survival information, you can set the threshold for the significance p-value. Features with p-values lower than this threshold are considered significantly associated with patient survival, and the default value is 0.05.

If you do not use survival information, you will need to specify the feature selection method and threshold separately：

Choose a Get Elites method for DNA methylation dataset: The feature selection methods include MAD (Median Absolute Deviation), SD (Standard Deviation), and PCA (Principal Components Analysis). Users can choose the method that best suits their needs. By default, MAD is selected as the default method, but users can change it based on their specific requirements.

Choose a filtration method for DNA methylation dataset: If you choose "mad" or "sd" for the "Choose a Get Elites method for DNA methylation dataset" parameter, you will need to specify this parameter. This parameter offers two feature selection methods: "elite.num": Select

features based on a specific number. "elite.pct": Select features based on a specific proportion.The default method is "elite.num," where you specify the number of features you want to select.

An integer cutoff of exact number for selecting top elites: If you choose "elite.num" for the "Choose a filtration method for DNA methylation dataset" parameter, you will need to specify the number of features you want to select, denoted as "n." The platform will rank the results obtained from the "mad" or "sd" methods and select the top n features in descending order. The default value for this parameter is 1000.

A numeric cutoff of percentage for selecting top elites: If you choose "elite.pct" for the "Choose a filtration method for DNA methylation dataset" parameter, you will need to specify the feature selection proportion, denoted as "n." The platform will rank the results obtained from the "mad" or "sd" methods and select the top features based on the specified proportion n. The default value for this parameter is 0.1, which corresponds to selecting the top 10% of features.

A numeric value which represents the ratio of principal components: If you choose "pca" for the "Choose a Get Elites method for DNA methylation dataset" parameter, you will need to specify the ratio used for selecting principal components. The default value is 0.95, meaning that enough principal components will be selected to explain 95% of the variance in the data.

Finally, the data is standardized, transforming it into a distribution with a mean of 0 and a variance of 1. This standardization step ensures that the data has a common scale, making it suitable for various analytical techniques：

Centering the data after filtering or not: Whether to subtract the mean of their corresponding features is an option. By default, it is set to "No," meaning that no subtraction of the mean is performed.

Scaling the data after filtering or not: Whether to divide by the standard deviation of their corresponding features is an option. By default, it is set to "No," meaning that no division by the standard deviation is performed. This step is part of standardization, and if you choose "Yes," the data will be scaled to have a mean of 0 and a standard deviation of 1.

After specifying the preprocessing parameters for the DNA methylation dataset, you should proceed to configure the preprocessing parameters for other data types as needed. You can refer to the documentation in the "Get Elites settings for mRNA dataset、 Get Elites settings for lncRNA dataset、Get Elites settings for copy number alterations dataset、Get Elites settings for binary somatic mutation dataset and Get Elites settings for radiomics dataset " parameter module for guidance on setting up preprocessing parameters for the radiomics dataset.

vi.Get Elites settings for copy number alterations dataset

You can categorize different types of data into two classes: continuous and

discrete. Copy number alterations (CNA) data falls under the category of continuous data, where rows represent features, and columns represent samples. To perform data preprocessing on the CNA dataset, you need to set the appropriate parameters.

First, it's the filtering and log2 standardization of samples and features：

NA value action: The handling of NA (missing) values in the data can be done in three ways: Remove directly, KNN imputation, and No action. "Remove directly" involves deleting samples that contain NA values. "KNN imputation" means filling in NA values in samples using a K-nearest neighbors imputation method. "No action" indicates that no specific action will be taken with NA values, leaving them as they are By default, the parameter uses "Remove directly," but users can choose the appropriate method based on the actual data situation.

Perform log2 transformation for data before calculating statistics or not: This parameter specifies whether data should be log2 standardized to limit larger expression values within a smaller range. The default option is "No," which means no log2 standardization is required. Users should make this decision based on the actual data situation, typically not applying log2 standardization to expression values below 50.

Set a numeric cutoff for removing low expression features or not: This parameter specifies whether to filter out features with low expression frequency. The default option is "No," indicating that features with low

expression frequency will not be filtered out.

The cutoff for removing low expression features: If the "Set a numeric cutoff for removing low expression features or not" parameter is set to "Yes," then you need to specify a threshold. Features with expression frequencies (number of samples expressing the feature / total number of samples) below this threshold will be filtered out. The default value for this parameter is 0.1, which is recommended.

Next, feature selection is performed：

Use survival information or not: Whether to use survival information is set to "No" by default. If the dataset includes survival information, users can choose to incorporate this information and use the Cox method for feature selection. Features selected using this method will have a higher correlation with the survival outcomes of the samples.

If you choose to use survival information, you will need to further configure the parameters accordingly：

A numeric cutoff for nominal p value derived from univariate Cox proportional hazards regression: To set the significance p-value threshold, where values lower than this threshold indicate a significant correlation between the feature and patient survival status, the default value is 0.05.

If not using survival information, you need to specify the feature selection method and threshold：

Choose a Get Elites method for copy number alterations dataset: The

feature selection methods include MAD (Median Absolute Deviation), SD (Standard Deviation), and PCA (Principal Components Analysis). Users can choose the method based on their specific needs, with the default selection being MAD.

Choose a filtration method for copy number alterations dataset: If you choose either MAD or SD for the "Choose a Get Elites method for copy number alterations dataset" parameter, you will need to specify this parameter. This parameter offers two ways of feature selection: "elite.num": Select features based on the number of elites. "elite.pct": Select features based on the percentage of elites.The default selection is "elite.num."

An integer cutoff of exact number for selecting top elites: If you choose "elite.num" for the "Choose a filtration method for copy number alterations dataset" parameter, you will need to specify the number of features to select, denoted as "n." The platform will sort the results obtained from the MAD or SD method in descending order and select the top "n" features. The default value for this parameter is 1000.

A numeric cutoff of percentage for selecting top elites: If you choose "elite.pct" for the "Choose a filtration method for copy number alterations dataset" parameter, you will need to specify the proportion of features to select, denoted as "n." The platform will sort the results obtained from the MAD or SD method in descending order and select the top "n"

proportion of features. The default value for this parameter is 0.1.

A numeric value which represents the ratio of principal components: If you choose "pca" for the "Choose a Get Elites method for copy number alterations dataset" parameter, you will need to specify the ratio used for selecting principal components. The default value for this parameter is 0.95.

Finally, the data is standardized to achieve a distribution with a mean of 0 and a variance of 1：

Centering the data after filtering or not: Whether to subtract the mean of their respective features is set to "No" by default, meaning no subtraction of feature means will be performed.

Scaling the data after filtering or not: Whether to divide by their respective feature's standard deviation is set to "No" by default, meaning no division by feature standard deviations will be performed.

After specifying the preprocessing parameters for copy number alterations data, you can proceed to set other preprocessing parameters for different data types as needed. Please refer to the guidance documentation for the "Get Elites settings for mRNA dataset、Get Elites settings for lncRNA dataset、Get Elites settings for DNA methylation dataset、Get Elites settings for binary somatic mutation dataset and Get Elites settings for radiomics dataset " parameter module for further instructions on how to configure those settings.

vii.Get Elites settings for binary somatic mutation dataset

You can categorize different types of data into two classes: continuous and discrete. Binary somatic mutation datasets belong to the discrete data class, where rows represent features, and columns represent samples. To perform data preprocessing on the binary somatic mutation dataset, specific parameters need to be set.

First, perform filtering and $\log_2$ standardization on samples and features.：

NA value action: The handling of NA (missing) values in the data can be done in three ways: Remove directly, KNN imputation, and No action. "Remove directly" involves deleting samples that contain NA values. "KNN imputation" means filling in NA values in samples using a K-nearest neighbors imputation method. "No action" indicates that no specific action will be taken with NA values, leaving them as they are.By default, this parameter uses "Remove directly," but users can choose the appropriate method based on the actual data situation.

Perform log2 transformation for data before calculating statistics or not: This parameter specifies whether data should undergo $\log_2$ standardization to constrain larger expression values within a smaller range. The default option is "No," which means no $\log_2$ standardization is required. Users should make this decision based on the actual data situation, typically not applying $\log_2$ standardization to expression values below 50. For discrete variables, $\log_2$ standardization does not affect the

results.

Next, feature selection is performed：

Use survival information or not: Whether to use survival information is set to "No" by default. If the dataset includes survival information, users can choose to incorporate this information and use the Cox method for feature selection. Features selected using this method will have a higher correlation with the survival outcomes of the samples.

If you opt to utilize survival information, additional parameters need to be configured：

A numeric cutoff for nominal p value derived from univariate Cox proportional hazards regression: To set the significance p-value threshold, where values lower than this threshold indicate a significant correlation between the feature and patient survival status, the default value is 0.05.

If you are not using survival information, you will need to specify the feature selection method and threshold：

Choose a filtration method for binary somatic mutation dataset: This parameter offers two methods for feature selection: "elite.num" selects features based on the number of elites, while "elite.pct" selects features based on the proportion of elites. The default method is "elite.pct.".

An integer cutoff of mutation frequency for selecting elites: If you choose "elite.num" for the "Choose a filtration method for binary somatic mutation dataset" parameter, you will need to specify the mutation

frequency threshold "n" (where mutation indicates a feature having a value of 1 in a particular sample). The platform will select features with mutation frequencies greater than the threshold "n." The default threshold value is 100.

A numeric cutoff of 'mutation / sample' frequency for selecting elites: If you choose "elite.pct" for the "Choose a filtration method for binary somatic mutation dataset" parameter, you will need to specify the mutation frequency threshold "n" (where mutation frequency is calculated as the number of mutations divided by the total number of samples). The platform will select features with mutation frequencies greater than the threshold "n." The default threshold value is 0.1.

Finally, the data is standardized to achieve a distribution with a mean of 0 and a variance of 1：

Centering the data after filtering or not: Whether to subtract the mean of their corresponding features is set to "No" by default, meaning no subtraction of feature means will be performed.

Scaling the data after filtering or not: Whether to divide by their corresponding feature's standard deviation is set to "No" by default, meaning no division by feature standard deviations will be performed.

Once you have specified the preprocessing parameters for binary somatic mutation data, you can proceed to set other preprocessing parameters for different data types as needed. Please refer to the guidance

documentation for the "Get Elites settings for mRNA dataset、Get Elites settings for lncRNA dataset、Get Elites settings for DNA methylation dataset、Get Elites settings for copy number alterations dataset and Get Elites settings for radiomics dataset " parameter module for further instructions on how to configure those settings.

viii.Get Elites settings for radiomics dataset

To categorize different types of data, we can distinguish between continuous and discrete data. Radiomics datasets fall under the category of continuous data, where rows represent features, and columns represent samples. Preprocessing radiomics data involves configuring specific parameters.

Firstly, it involves filtering of samples and features, followed by $\log_2$ standardization.：

NA value action: The handling of NA (Not Available) values in the data involves three options: Remove directly, KNN imputation, and No action. The Remove directly option deletes samples that contain NA values, KNN imputation fills in NA values in the samples, and No action does not perform any operation on NA values. The default setting for this parameter is Remove directly, and users can choose the appropriate option based on the actual characteristics of their data.

Perform log2 transformation for data before calculating statistics or not: This parameter specifies whether $\log_2$ standardization is required for the

data, limiting larger expression values to a smaller range. The default option is No, indicating that log_2 standardization is not needed. Users should assess this based on the actual characteristics of their data, typically considering that expression values less than 50 may not require log_2 standardization.

Set a numeric cutoff for removing low expression features or not: This parameter determines whether to filter out features with low expression frequencies. The default choice is No, indicating that features with low expression frequencies will not be filtered out.

The cutoff for removing low expression features: If the "Set a numeric cutoff for removing low expression features or not" parameter is selected as Yes, then a threshold needs to be specified. Features with expression frequencies lower than this threshold (defined as the number of samples expressing the feature divided by the total number of samples) will be filtered out. The default value for this parameter is 0.1 (recommended).

Afterward, feature selection is performed：

Use survival information or not: Whether to use survival information is set to default as No. If the dataset includes survival information, users have the option to incorporate survival information and use the Cox method for feature selection. Features selected through this method will have a higher correlation with the survival status of the samples.

If choosing to use survival information, further parameters need to be

specified：

A numeric cutoff for nominal p value derived from univariate Cox proportional hazards regression: If using survival information, you need to set the threshold for the significance p-value. A p-value lower than this threshold indicates a significant association between the feature and the patient's survival status. The default value is 0.05.

If not using survival information, you will need to specify the feature selection method and threshold：

Choose a Get Elites method for radiomics dataset: It includes three feature selection methods: MAD (Median Absolute Deviation), SD (Standard Deviation), and PCA (Principal Components Analysis). Users can choose the appropriate method based on their specific situation. The default selection is MAD.

Choose a filtration method for radiomics dataset: If the "Choose a Get Elites method for radiomics dataset" parameter is set to MAD or SD, you need to specify this parameter. This parameter provides two ways of feature selection: elite.num selects based on the number dimension, while elite.pct selects based on the percentage dimension. The default is to use elite.num.

An integer cutoff of exact number for selecting top elites: If the "Choose a filtration method for radiomics dataset" parameter is set to elite.num, you need to specify the number of selected features, denoted as n. The

platform will rank the results from the MAD or SD method and select the top n features in descending order. The default value for this parameter is 1000.

A numeric cutoff of percentage for selecting top elites: If the "Choose a filtration method for radiomics dataset" parameter is set to elite.pct, you need to specify the selected feature percentage denoted as n. The platform will rank the results from the MAD or SD method and select the top features based on the specified percentage. The default value for this parameter is 0.1.

A numeric value which represents the ratio of principal components: If the "Choose a Get Elites method for radiomics dataset" parameter is set to PCA, you need to specify the ratio used for selecting principal components. The default value is 0.95.

Finally, standardization is applied to the data, transforming it into a distribution with a mean of 0 and a variance of 1：

Centering the data after filtering or not: Whether to subtract the mean of the corresponding feature is set to default as No, meaning no subtraction is applied.

Scaling the data after filtering or not: Whether to divide by the standard deviation of the corresponding feature is set to default as No, meaning no division is applied.

After specifying the preprocessing parameters for the radiomics data, you

can proceed to set other preprocessing parameters for different data types. Please refer to the guidance document for the " Get Elites settings for mRNA dataset、Get Elites settings for lncRNA dataset、Get Elites settings for DNA methylation dataset、Get Elites settings for copy number alterations dataset and Get Elites settings for binary somatic mutation dataset " parameter module for detailed instructions.

ix.Process Get Elites

Once you have specified all the preprocessing parameters for the different data types, click the "Process" button to initiate data preprocessing. After processing, feedback will be provided in the "Get Elites results" module on the right side of the webpage.

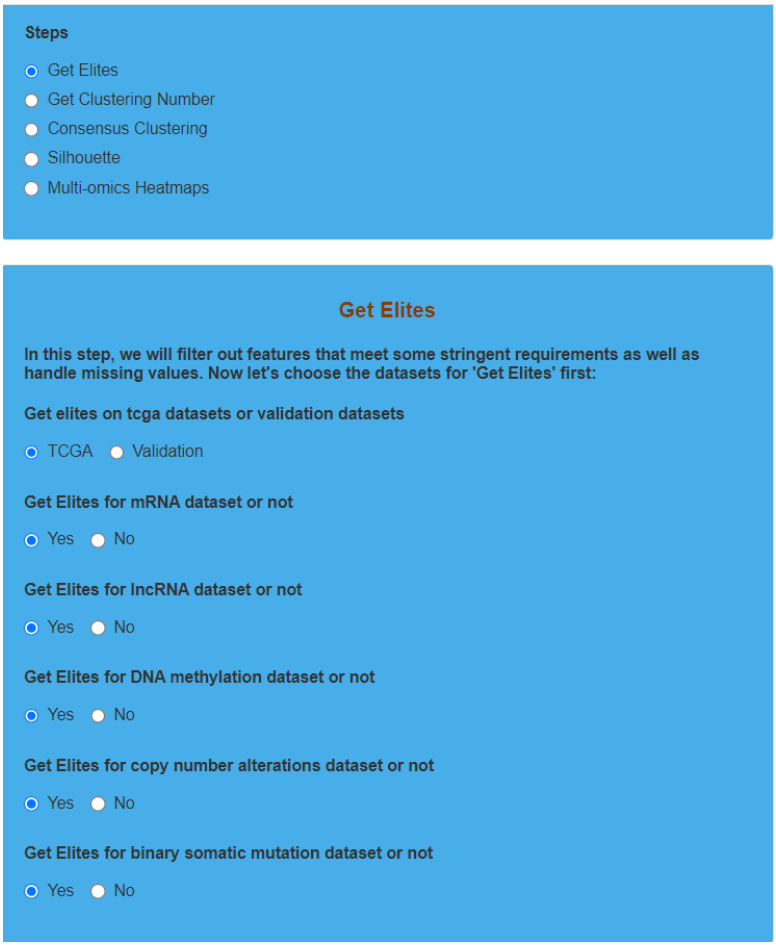
After completing the preprocessing of TCGA dataset, if there is an external validation set, you need to go back to the "Get Elites" parameter module and set the "Get elites on tcga datasets or validation datasets" parameter to "Validation." Then, repeat the steps to set the preprocessing parameters for each data type in the external validation set. Click the "Process" button to perform preprocessing.

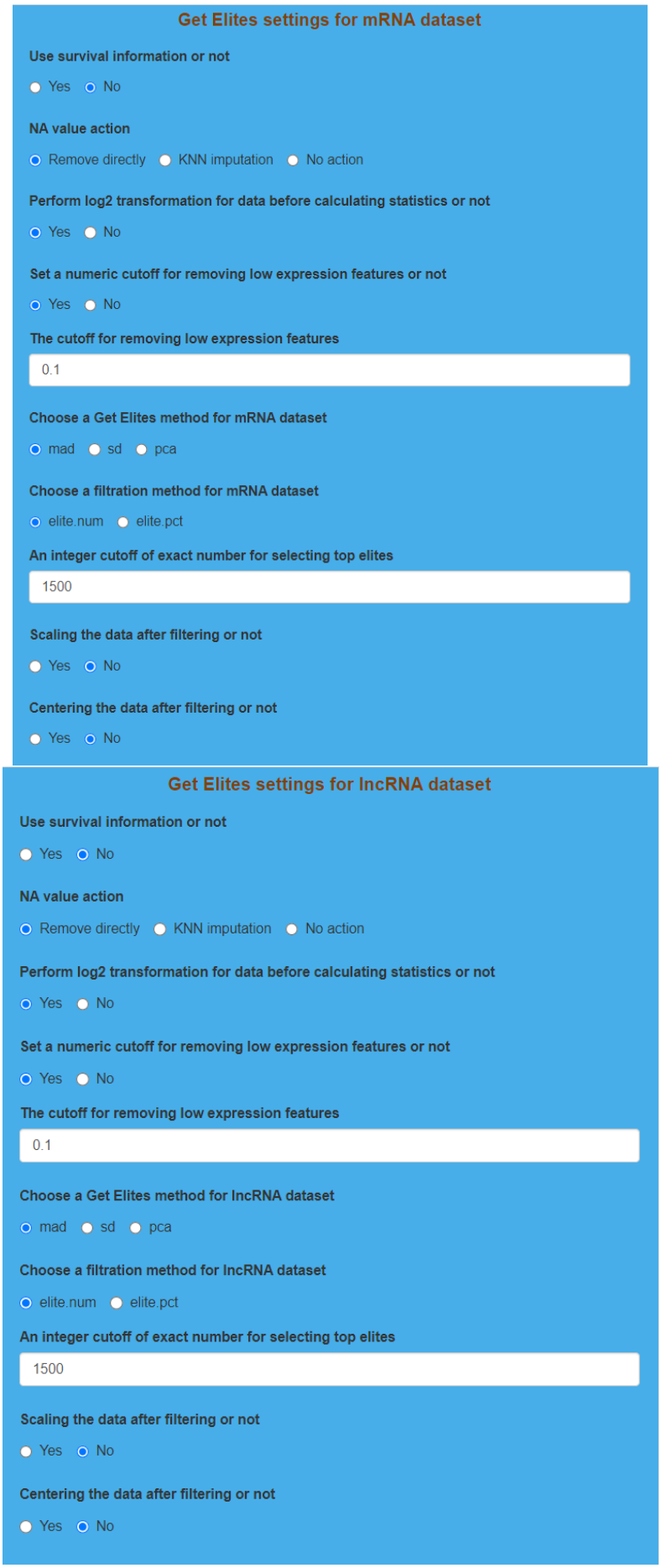
Once preprocessing is done for all datasets, go to the "GET Module" section under the "Steps" parameter module and select "Get Clustering Number" to determine the optimal clustering number. Refer to the guidance document for the "Get Clustering Number" parameter module for detailed instructions.

III.Examples with results interpretation

Parameter Settings: “Module switching options” on the upper left of the software chooses “GET Module”, and then the “Steps” chooses “Get Elites”. After that, we should indicate the omics data types for “Get Elites” (choose “Yes” under the corresponding parameters) and then set the parameters for each omics data in turn. For continuous omics data including mRNA, lncRNA, DNA methylation, copy number alterations and radiomics, we utilize “mad” method (“Choose a Get Elites method” chooses “mad”) to extract top 1500 features (“Choose a filtration method” chooses “elite.num”, and then “An integer cutoff of exact number for selecting top elites” indicates 1500) for each omics data according to mad values. In particular, for mRNA and lncRNA, we also perform “$\log_2$” transformation (“Perform log2 transformation for data before calculating statistics or not” chooses “Yes”) and set a numeric cutoff to remove features with low expression (“Set a numeric cutoff for removing low expression features or not” chooses “Yes”, and then “The cutoff for removing low expression features” indicates 0.1). For binary omics data containing somatic mutation, we extract the features whose mutation frequency divided by sample size is higher than 0.03 (“Choose a filtration method for binary somatic mutation dataset” chooses “elite.pct”, and then “A numeric cutoff of 'mutation / sample' frequency for selecting elites” indicates 0.03). Additionally, “Use survival information or not”

chooses “No”, “NA value action” chooses “Remove directly”, “Scaling the data after filtering or not” chooses “No”, and “Centering the data after filtering or not” chooses “No” for each omics data. Finally, we click the “Process” button to do “Get Elites”. If you want to do “Get Elites” on validation dataset, “Get elites on tcga datasets or validation datasets” should select “Validation”. Here we do not need to do “Get Elites” on validation dataset, thus “Get Elites for mRNA dataset or not” chooses “No”, and then we click the “Process” button.




Get Elites settings for mRNA dataset
Use survival information or not
Yes
No
NA value action
Remove directly
KNN imputation
No action
Perform log2 transformation for data before calculating statistics or not
Yes
No
Set a numeric cutoff for removing low expression features or not
Yes
No
The cutoff for removing low expression features
0.1
Choose a Get Elites method for mRNA dataset
mad
sd
pca
Choose a filtration method for mRNA dataset
elite.num
elite.pct
An integer cutoff of exact number for selecting top elites
1500
Scaling the data after filtering or not
Yes
No
Centering the data after filtering or not
Yes
No
Get Elites settings for lncRNA dataset
Use survival information or not
Yes
No
NA value action
Remove directly
KNN imputation
No action
Perform log2 transformation for data before calculating statistics or not
Yes
No
Set a numeric cutoff for removing low expression features or not
Yes
No
The cutoff for removing low expression features
0.1
Choose a Get Elites method for lncRNA dataset
mad
sd
pca
Choose a filtration method for lncRNA dataset
elite.num
elite.pct
An integer cutoff of exact number for selecting top elites
1500
Scaling the data after filtering or not
Yes
No
Centering the data after filtering or not
Yes
No

Get Elites settings for DNA methylation dataset

Use survival information or not

○ Yes ◉ No

NA value action

◉ Remove directly ○ KNN imputation ○ No action

Perform log2 transformation for data before calculating statistics or not

○ Yes ◉ No

Set a numeric cutoff for removing low expression features or not

○ Yes ◉ No

Choose a Get Elites method for DNA methylation dataset

◉ mad ○ sd ○ pca

Choose a filtration method for DNA methylation dataset

◉ elite.num ○ elite.pct

An integer cutoff of exact number for selecting top elites

1500

Scaling the data after filtering or not

○ Yes ◉ No

Centering the data after filtering or not

○ Yes ◉ No

Get Elites settings for copy number alterations dataset

Use survival information or not

○ Yes ◉ No

NA value action

◉ Remove directly ○ KNN imputation ○ No action

Perform log2 transformation for data before calculating statistics or not

○ Yes ◉ No

Set a numeric cutoff for removing low expression features or not

○ Yes ◉ No

Choose a Get Elites method for copy number alterations dataset

◉ mad ○ sd ○ pca

Choose a filtration method for copy number alterations dataset

◉ elite.num ○ elite.pct

An integer cutoff of exact number for selecting top elites

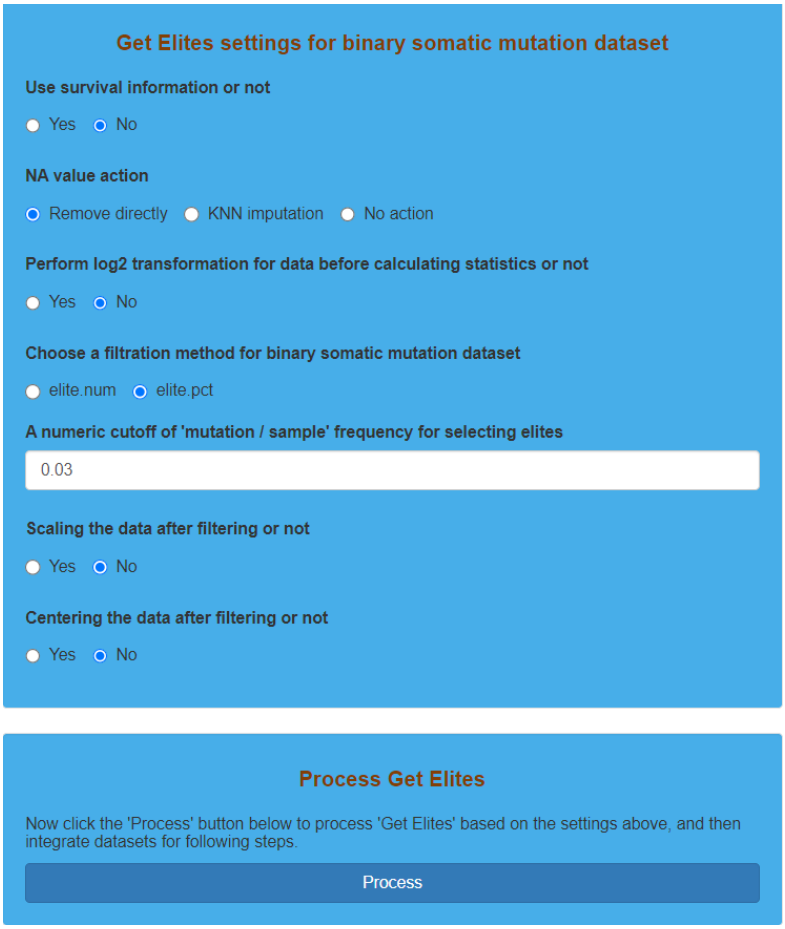
1500

Scaling the data after filtering or not

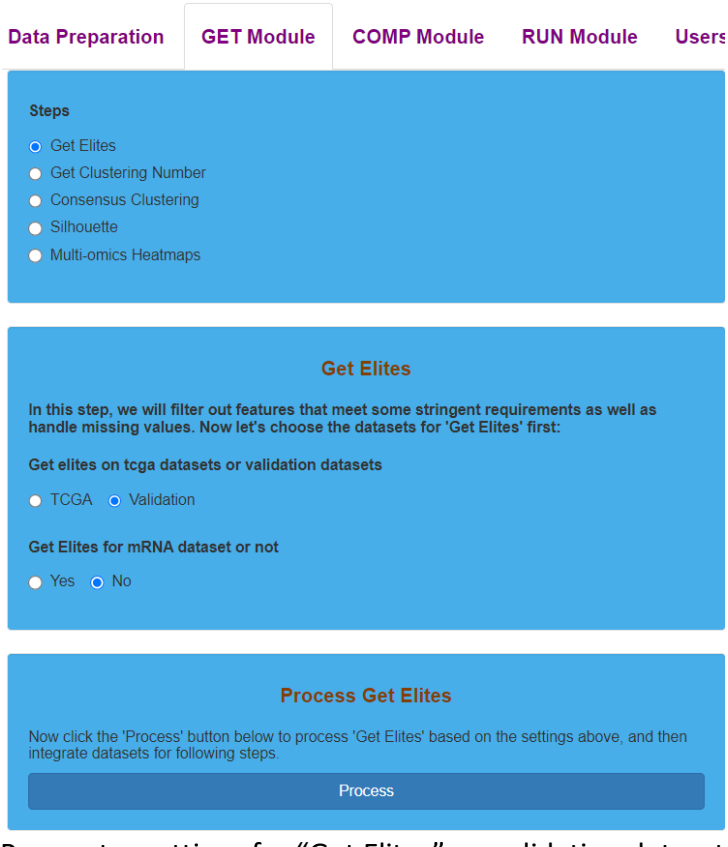
○ Yes ◉ No

Centering the data after filtering or not

○ Yes ◉ No



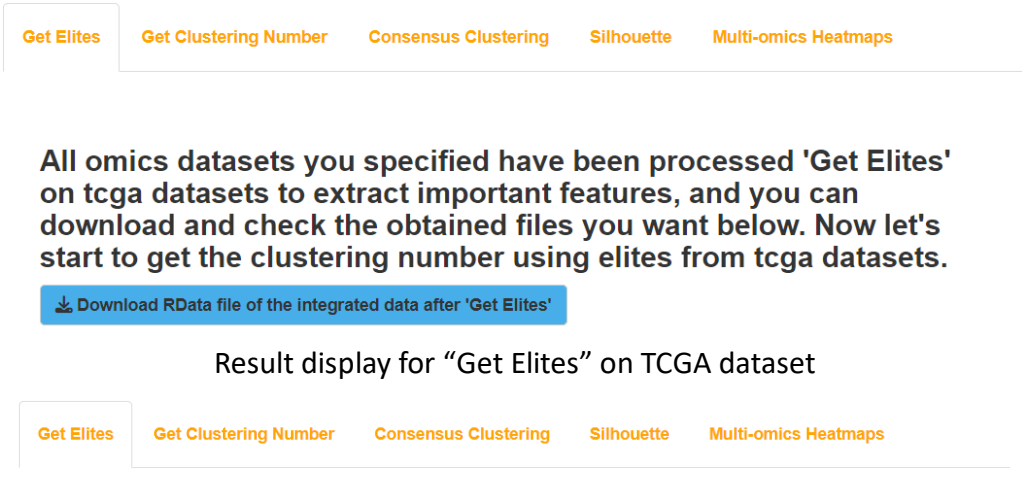
Parameter settings for “Get Elites” on TCGA dataset



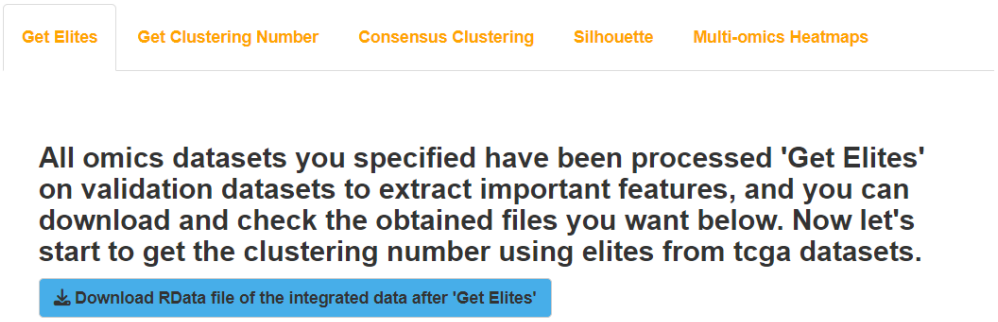
Parameter settings for “Get Elites” on validation dataset

Result Display:

The software will give feedback to the users after finishing the process of "Get Elites" on TCGA dataset or validation dataset.



Result display for "Get Elites" on TCGA dataset



Result display for "Get Elites" on validation dataset

(2). Get Clustering Number for TCGA dataset

I.Analysis introduction

"Get Optimal Clustering Number" combines two statistics of CPI and Gaps-statistics to plot and give the optimal clustering number.

II.Parameters setting guides

Get Clustering Number

After completing the preprocessing of the dataset, the next step is to determine the optimal number of clusters. Our platform uses two statistical measures, Clustering Prediction Index (CPI) and Gap-statistics

(Gapk), to jointly determine the optimal number of clusters. The parameters that the user needs to specify include:

The range of clustering number: Specify the possible range for the number of clusters by providing Minimum and Maximum values. The default values are 2 and 8, indicating that the possible range for the number of clusters is between 2 and 8.

Centering the data or not:Choose whether to center the data, i.e., subtract the mean of the corresponding feature. The default is Yes.

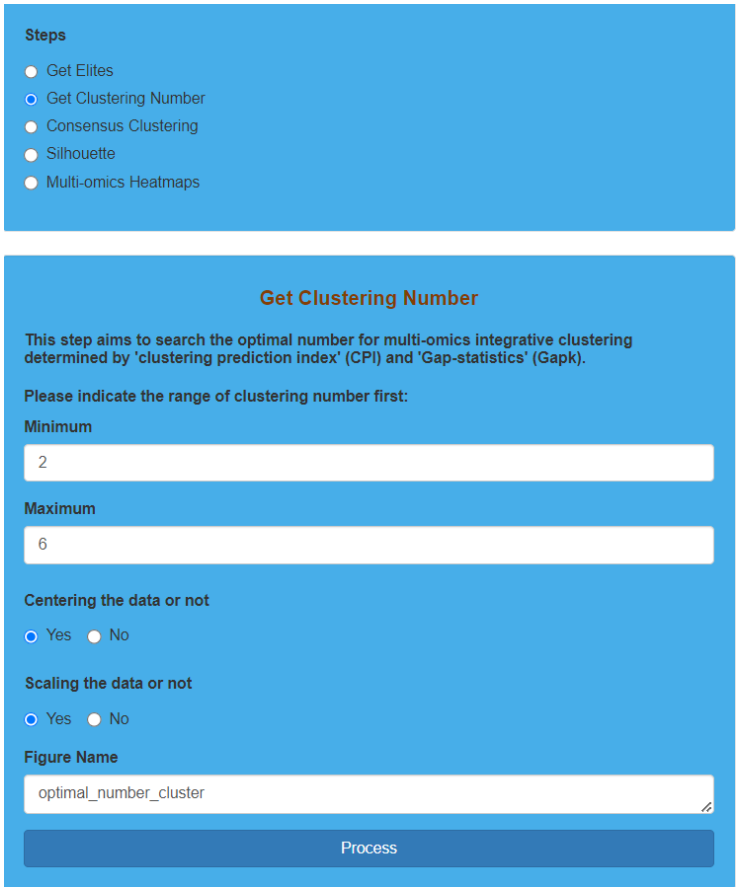
Scaling the data or not: Choose whether to scale the data, i.e., divide by the standard deviation of the corresponding feature. The default is Yes.

Figure Name: Specify the PDF name for the result images. The default is "optimal_number_cluster."

Once all the parameters for determining the optimal number of clusters are specified, click the "Process" button to run the program. After completion, feedback will be provided in the "Get Clustering Number results" module on the right side of the webpage. Users can choose to download the corresponding PDF images and RData result sets.Next, we will proceed with consensus clustering. Go to the "GET Module" section under the "Steps" parameter module and select "Consensus Clustering." Refer to the guidance document for the first "Consensus Clustering" parameter module for detailed instructions.

III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Get Clustering Number”, and then we indicate the range of clustering number (“Minimum” indicates 2 and “Maximum” indicates 6). After that, both “Centering the data or not” and “Scaling the data or not” choose “Yes”. Finally, we enter the “Figure Name”, and click the “Process” button to determine the optimal clustering number.

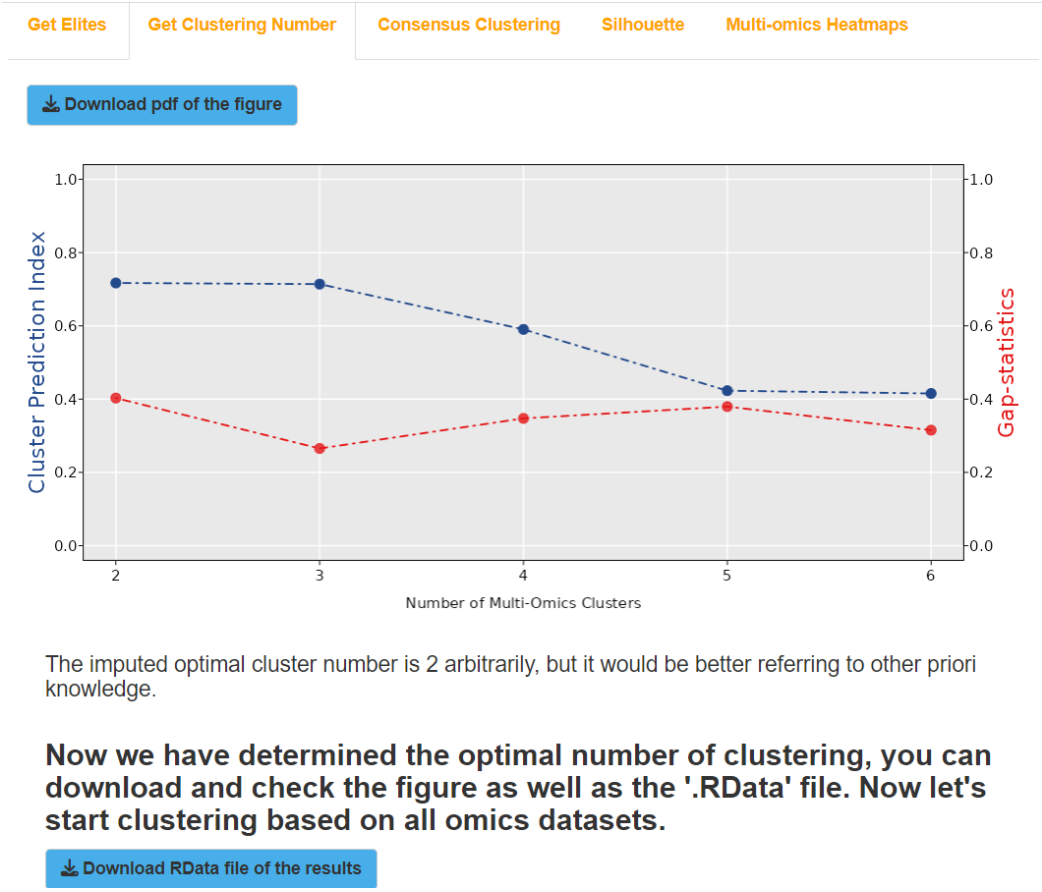


Parameter settings for “Get Clustering Number” on TCGA dataset

Result Display:

The software will plot according to the values of CPI and Gaps-statistics and suggest the optimal clustering number when the sum of CPI and Gaps-

statistics is the largest. However, we should also consider the prior information when determining the optimal clustering number, which will be used for the following “Consensus Clustering”. Here we combine prior information to determine “4” as the optimal clustering number.



Result display for “Get Clustering Number” on TCGA dataset

(3). Consensus Clustering for TCGA dataset

I.Analysis introduction

“Consensus Clustering” provides 10 clustering algorithms, including iClusterBayes, SNF, PINSPlus, NEMO, COCA, LRAcluster, ConsensusClustering, IntNMF, CIMLR, and MoCluster, from which users can choose one or more for clustering. If you want to run consensus

clustering, you should choose at least two algorithms to get the consensus clustering diagram.

II.Parameters setting guides

i.Consensus Clustering

The platform provides 10 clustering methods for users to choose from. If a user selects only one clustering method, the platform will perform clustering using that method alone. However, if consensus clustering is desired, the user needs to select two or more clustering methods.

Clustering algorithms (Choose at least two types if you want to get consensus results): The platform offers a total of 10 clustering methods, including iClusterBayes, SNF, PINSPlus, NEMO, COCA, LRAcluster, ConsensusClustering, IntNMF, CIMLR, and MoCluster. By default, all these methods are selected.

Once you have selected the clustering methods, if you choose only one method, proceed to the parameter module of that method for parameter settings. Refer to the documentation for the parameter modules of iClusterBayes, SNF, PINSPlus, NEMO, COCA, LRAcluster, ConsensusClustering, IntNMF, CIMLR, and MoCluster for detailed instructions.If you choose two or more methods, navigate to the next "Consensus Clustering" parameter module and follow the guidance document for parameter settings.

ii.iClusterBayes

If only the iClusterBayes (Integrative clustering by Bayesian latent variable model) clustering method is selected, you can specify the parameters for this clustering method.

The number of clusters: Users can utilize the optimal clustering number obtained in the "Get Clustering Number" step, with the parameter set to "System optimal" (default). Alternatively, users can define the clustering number based on prior knowledge by selecting "User defined."

User defined cluster number: If the "The number of clusters" parameter is set to "User defined," you need to input the user-defined number of clusters. The default value is 2.

Next, you will sequentially specify the prior probability for the indicator variable gamma for each omics data type.：

The prior probability for the indicator variable gamma of mRNA subdataset: If mRNA data is included, set the prior probability for its indicator variable gamma. The default value is 0.5.

The prior probability for the indicator variable gamma of lncRNA subdataset: If lncRNA data is included, set the prior probability for its indicator variable gamma. The default value is 0.5.

The prior probability for the indicator variable gamma of DNA methylation subdataset: If DNA methylation data is included, set the prior probability for its indicator variable gamma. The default value is 0.5.

The prior probability for the indicator variable gamma of copy number alterations subdataset: If copy number alterations (CNA) data is included, set the prior probability for its indicator variable gamma. The default value is 0.5.

The prior probability for the indicator variable gamma of binary somatic mutation subdataset: If binary somatic mutation data is included, set the prior probability for its indicator variable gamma. The default value is 0.5.

The prior probability for the indicator variable gamma of radiomics subdataset: If radiomics data is included, set the prior probability for its indicator variable gamma. The default value is 0.5.

The number of MCMC burnin: Specify the MCMC target number, with the default value set to 18000.

The number of MCMC draw: Specify the MCMC target number, with the default value set to 12000.

The standard deviation of random walk proposal for the latent variable: The standard deviation of the random walk proposal for latent variables, with the default value set to 0.05.

A numerical value to thin the MCMC chain in order to reduce autocorrelation: The thinning parameter used to refine the MCMC chain to reduce autocorrelation, with the default value set to 3.

After completing all parameter settings in the iClusterBayes parameter module, click the "Process" button to perform clustering. After

completion, feedback will be provided in the "Consensus Clustering results" module on the right side of the webpage. Users can choose to download the RData result set.Next, to evaluate the quality of the clustering results by calculating the sample-wise similarity of obtained subtypes, go to the "GET Module" section under the "Steps" parameter module and select "Silhouette." Refer to the guidance document for the "Silhouette" parameter module for detailed instructions.

iii. SNF

If only the SNF (Similarity Network Fusion) clustering method is selected, you can specify the parameters for this clustering method.

The number of clusters: Users can use the optimal clustering number obtained in the "Get Clustering Number" step, with the parameter set to "System optimal" (default). Alternatively, users can define the clustering number based on prior knowledge by selecting "User defined."

User defined cluster number: If the "The number of clusters" parameter is set to "User defined," you need to input the user-defined number of clusters. The default value is 2.

The number of neighbors in K-nearest neighbors part of the algorithm: Specify the number of neighbors in the K-nearest neighbors algorithm. The default value is 30.

The number of iterations for the diffusion process: Specify the number of iterations for the diffusion process. The default value is 20.

The variance for the local model: Specify the variance for the local model. The default value is 0.5.

After completing all parameter settings in the SNF parameter module, click the "Process" button to perform clustering. After completion, feedback will be provided in the "Consensus Clustering results" module on the right side of the webpage. Users can choose to download the RData result set.Next, to evaluate the quality of the clustering results by calculating the sample-wise similarity of obtained subtypes, go to the "GET Module" section under the "Steps" parameter module and select "Silhouette." Refer to the guidance document for the "Silhouette" parameter module for detailed instructions.

iv.PINSPlus

If only the PINSPlus (Perturbation Clustering for data INtegration and disease Subtyping) clustering method is selected, you can specify the parameters for this clustering method.

The number of clusters: Users can use the optimal clustering number obtained in the "Get Clustering Number" step, with the parameter set to "System optimal" (default). Alternatively, users can define the clustering number based on prior knowledge by selecting "User defined."

User defined cluster number: If the "The number of clusters" parameter is set to "User defined," you need to input the user-defined number of clusters. The default value is 2.

The normalization method for consensus clustering: Specify the normalization method for consensus clustering. The default is none, indicating no normalization.

Built-in clustering algorithm that PerturbationClustering will use: Specify the built-in clustering algorithm that PINSPlus will use. The default is kmeans.

The minimum number of iterations: Specify the minimum number of iterations. The default value is 50.

The maximum number of iterations: Specify the maximum number of iterations. The default value is 500.

After completing all parameter settings in the PINSPlus parameter module, click the "Process" button to perform clustering. After completion, feedback will be provided in the "Consensus Clustering results" module on the right side of the webpage. Users can choose to download the RData result set.Next, to evaluate the quality of the clustering results by calculating the sample-wise similarity of obtained subtypes, go to the "GET Module" section under the "Steps" parameter module and select "Silhouette." Refer to the guidance document for the "Silhouette" parameter module for detailed instructions.

v. NEMO

If you choose the NEMO (Neighborhood based multi-omics clustering) method, you can specify the parameters for this clustering method.

1. Number of Clusters: Users can use the optimal number of clusters obtained in the "Get Clustering Number" step. Choose "System optimal" for this parameter. Alternatively, users can define the number of clusters based on prior knowledge. Choose "User defined" for this parameter.

2. User Defined Cluster Number: If the "Number of Clusters" parameter is set to "User defined," input the desired number of clusters (default is 2).

3. Number of Neighbors to Use for Each Omic: Specify the number of neighbors for each type of omics data. The default is to use the system default parameters, which is calculated as "The number of samples divided by NUM.NEIGHBORS.RATIO."

4. Number of Neighbors Used for All Omics: If the "Number of Neighbors to Use for Each Omic" parameter is set to "The number of neighbors used for all omics," specify a common number of neighbors for all omics data (default is 30).

5. Input the Number of Neighbors for Each Omic: If the "Number of Neighbors to Use for Each Omic" parameter is set to "A list of numbers for each omic," input the number of neighbors for each type of omics data in sequence.

After configuring all parameters in the NEMO parameter module, click the "Process" button to perform clustering. Once completed, feedback on Consensus Clustering results will be available on the right side of the webpage. Users can choose to download the RData result set.

Subsequently, evaluate the quality of the clustering results by computing the sample-wise similarity of obtained subtypes. Proceed to the "Steps" parameter module under the "GET Module" section and choose "Silhouette." Refer to the guidance document for the Silhouette parameter module for further details.

vi. COCA

If you choose the COCA (Cluster-of-Clusters Analysis) clustering method, you can specify the parameters for this method.

1. Number of Clusters: Users can use the optimal number of clusters obtained in the "Get Clustering Number" step. Choose "System optimal" for this parameter. Alternatively, users can define the number of clusters based on prior knowledge. Choose "User defined" for this parameter.

2. User Defined Cluster Number: If the "Number of Clusters" parameter is set to "User defined," input the desired number of clusters (default is 2).

3. Clustering Methods for Subdataset Observations: Specify the clustering methods to be used for clustering the observations in each type of omics data. The default is "hclust."

4. Distances for Clustering Step: Specify the distances to be used in the clustering step for each subdataset. The default distance measure is "euclidean."

After configuring all parameters in the COCA parameter module, click the "Process" button to perform clustering. Once completed, feedback on

Consensus Clustering results will be available on the right side of the webpage. Users can choose to download the RData result set. Subsequently, evaluate the quality of the clustering results by computing the sample-wise similarity of obtained subtypes. Proceed to the "Steps" parameter module under the "GET Module" section and choose "Silhouette." Refer to the guidance document for the Silhouette parameter module for further details.

vii.LRAcluster

If you exclusively choose the LRAcluster (Integrated cancer omics data analysis by low rank approximation) clustering method, you can specify the parameters for this method.

1. Number of Clusters: Users can use the optimal number of clusters obtained in the "Get Clustering Number" step. Choose "System optimal" for this parameter. Alternatively, users can define the number of clusters based on prior knowledge. Choose "User defined" for this parameter.

2. User Defined Cluster Number: If the "Number of Clusters" parameter is set to "User defined," input the desired number of clusters (default is 2).

3. Cluster Algorithm for Similarity Matrix: Specify the cluster algorithm to be used for the similarity matrix. The default is "ward.D."

After configuring all parameters in the LRAcluster parameter module, click the "Process" button to perform clustering. Once completed, feedback on Consensus Clustering results will be available on the right side of the

webpage. Users can choose to download the RData result set. Subsequently, evaluate the quality of the clustering results by computing the sample-wise similarity of obtained subtypes. Proceed to the "Steps" parameter module under the "GET Module" section and choose "Silhouette." Refer to the guidance document for the Silhouette parameter module for further details.

viii.ConsensusClustering

If exclusively opting for the Consensus Clustering method, you can specify the parameters for this clustering approach.

1. Number of Clusters: Users can utilize the optimal number of clusters obtained in the "Get Clustering Number" step by choosing "System optimal" for this parameter. Alternatively, users can define the number of clusters based on prior knowledge by selecting "User defined" for this parameter.

2. User Defined Cluster Number: If the "Number of Clusters" parameter is set to "User defined," users should input the desired number of clusters (default is 2).

3. Normalization Method for Consensus Clustering: Specify the normalization method for consensus clustering. The default is "none," indicating no normalization.

4. Number of Subsamples: Specify the number of subsamples for secondary sampling. The default is 500.

5. Proportion of Items to Sample: Specify the proportion of items (samples) to sample in each subsample. The default is 0.8.

6. Proportion of Features to Sample: Specify the proportion of features to sample in each subsample. The default is 0.8.

7. Cluster Algorithm: Specify the cluster algorithm to be used. The default is "hc."

8. Hierarchical Linkage Method for Subsampling: Specify the hierarchical linkage method for subsampling. The default is "ward.D."

9. Hierarchical Method for Consensus Matrix: Specify the hierarchical method for the consensus matrix. The default is "ward.D."

10. Distance Function: Specify the distance function to be used in the clustering method. The default is "pearson."

11. Output Format for Heatmap: Choose the output format for the heatmap. The default is "none" (no heatmap output).

12. Write Output and Log to CSV or Not: Specify whether to save results and logs as CSV files. The default is "No."

13. Name of Output Directory: If selecting specific output formats or saving results as CSV, specify the directory name for saving output. The default is "consensuscluster."

14. Random Seed for Reproducible Results: Specify a random seed for reproducible results. The default is 123456.

After configuring all parameters in the Consensus Clustering module, click

the "Process" button to execute clustering. Upon completion, feedback on Consensus Clustering results will be provided on the right side of the webpage. Users can choose to download the RData result set. Subsequently, evaluate the quality of the clustering results as needed. If further analysis is required, refer to additional modules or steps provided in the tool interface.

ix.IntNMF

If exclusively choosing the IntNMF (Integrative Clustering via Non-negative Matrix Factorization) clustering method, you can specify the parameters for this method.

Number of Clusters: Users can utilize the optimal number of clusters obtained in the "Get Clustering Number" step by choosing "System optimal" for this parameter.

User defined cluster number: user define the number of clusters based on prior knowledge by selecting "User defined" for this parameter. The default value is 2.

No other parameters need to be set. Simply click the "Process" button to initiate clustering. Upon completion, feedback on Consensus Clustering results will be provided on the right side of the webpage. Users can choose to download the RData result set. Subsequently, evaluate the quality of the clustering results by computing the sample-wise similarity of obtained subtypes. Proceed to the "Steps" parameter module under the "GET

Module" section and choose "Silhouette." Refer to the guidance document for the Silhouette parameter module for further details.

x.CIMLR

If exclusively selecting the CIMLR (Cancer Integration via Multikernel Learning) clustering method, you can specify the parameters for this method.

1. Number of Clusters: Users can utilize the optimal number of clusters obtained in the "Get Clustering Number" step by choosing "System optimal" for this parameter. Alternatively, users can define the number of clusters based on prior knowledge by selecting "User defined" for this parameter. The default value is 2.

2. User Defined Cluster Number: If the "Number of Clusters" parameter is set to "User defined," users should input the desired number of clusters (default is 2).

3. Ratio of the Number of Cores for Multi-kernel Computation: Specify the ratio of the number of cores to be used when computing the multi-kernel. The default value is 0, indicating no specific ratio.

After configuring all parameters in the CIMLR parameter module, click the "Process" button to execute clustering. Upon completion, feedback on Consensus Clustering results will be provided on the right side of the webpage. Users can choose to download the RData result set. Subsequently, evaluate the quality of the clustering results by computing

the sample-wise similarity of obtained subtypes. Proceed to the "Steps" parameter module under the "GET Module" section and choose "Silhouette." Refer to the guidance document for the Silhouette parameter module for further details.

xi.MoCluster

If exclusively choosing the MoCluster (Multiple omics data integrative clustering) method, you can specify the parameters for this method.

The number of clusters: Users can utilize the optimal number of clusters obtained in the "Get Clustering Number" step by choosing "System optimal" for this parameter.Alternatively, users can define the number of clusters based on prior knowledge by selecting "User defined" for this parameter.

User defined cluster number: If the "Number of Clusters" parameter is set to "User defined," users should input the desired number of clusters (default is 2).

The number of components to calculate: Specify the number of components to calculate. The default value is 10.

Method: Specify the calculation method. The default value is "CPCA."

Normalization of different matrices: Specify the normalization method for different matrices. The default value is "lambda1."

The next step involves setting the indicator variable to load the absolute number (if k >= 1) or proportion (if 0 < k < 1) of non-zero coefficients of

the vector.

Format: Set the indicator variable to load the non-zero coefficients of the vector in the form of either an absolute number (default choice) or a proportion.

Setting: The setting for the indicator variable to load the non-zero coefficients of the vector can be configured in two ways: either use the same value for all omics data (default choice) - "For all omics," or set it individually for each type of omics data - "For each omic."

The absolute number: If the Format parameter is set to "Absolute number" and the Setting parameter is set to "For all omics," you need to specify the absolute number of non-zero coefficients for loading the vector at once. The default value is 10.

Input the absolute number for each omic: If the Format parameter is set to "Absolute number" and the Setting parameter is selected as "For each omic," you need to sequentially specify the absolute number of non-zero coefficients for loading the vector for each omics data.

The proportion: If the Format parameter is set to "The proportion" and the Setting parameter is set to "For all omics," you need to specify the proportion of non-zero coefficients for loading the vector at once. The default value is 0.1.

Input the proportion for each omic: If the Format parameter is set to "The proportion" and the Setting parameter is selected as "For each omic," you

need to sequentially specify the proportion of non-zero coefficients for loading the vector for each omics data.

The variables should be centered or not: Specify whether feature variables need to undergo centering processing, i.e., whether to subtract their corresponding feature means. The default selection is "Yes."

The variables should be scaled or not: Specify whether feature variables need to undergo scaling processing, i.e., whether to divide them by their corresponding feature standard deviations. The default selection is "Yes."

The cluster algorithm for distance: Specify the distance metric method for clustering, with the default value being "ward.D."

After completing all parameter settings in the MoCluster module, click the "Process" button to initiate clustering. Once completed, feedback on Consensus Clustering results will be provided in the module on the right side of the webpage. Users have the option to download the RData result set. Next, evaluate the quality of the clustering results by computing the sample-to-sample similarity of the obtained subtypes. Navigate to the "Steps" parameter module under the "GET Module" section and choose "Silhouette." For detailed instructions, refer to the guidance document for the Silhouette parameter module.

xii. Consensus Clustering

You have chosen two or more clustering methods, and you only need to specify the number of clusters to start consensus clustering. The selected

clustering method(s) will use default parameters.

1. The number of clusters: Users can use the optimal clustering number obtained in the "Get Clustering Number" step, with the parameter set to "System optimal" (default). Alternatively, users can define the cluster number based on prior knowledge, choosing "User defined" for this parameter.

2. User defined cluster number: If "The number of clusters" parameter is set to "User defined," this parameter requires input for the user-defined cluster number, with a default value of 2.

After configuring all parameters in this module, click the "Process" button to initiate consensus clustering. Once completed, feedback on Consensus Clustering results will be provided on the right side of the webpage, and users can choose to download the RData result set. Next, proceed to create a consensus clustering heatmap; for detailed instructions, refer to the guidance document for the "Consensus Heatmap" parameter module.

xiii.Consensus Heatmap

After completing the consensus clustering, you can create a consensus clustering heatmap. Before plotting, you need to specify the heatmap parameters:

1. Distance measurement for hierarchical clustering: Input the distance metric method for hierarchical clustering, with the default method being "euclidean."

2. Clustering method for hierarchical clustering: Input the clustering method for hierarchical clustering, with the default method being "ward.D."

3. Heatmap mapping color settings: Specify the heatmap colors. You can choose to use the system default colors ("System default" - default choice) or specify your own colors ("User defined").

4. Input heatmap mapping colors: If "Heatmap mapping color settings" is set to "User defined," input the specified colors (hexadecimal), for example, #000004FF.

5. Clustering subtypes color settings: Specify the colors for annotating the clusters obtained at the top of the heatmap. You can choose system default colors ("System default" - default choice) or specify your own colors ("User defined").

6. Input colors for clustering subtypes: If "Clustering subtypes color settings" is set to "User defined," input the specified colors (hexadecimal), for example, #2EC4B6.

7. Show sample ID or not: Set whether to display sample IDs on the heatmap. The default choice is "No," meaning sample IDs will not be displayed.

8. Figure Name: Specify the name of the output heatmap PDF, with the default being "consensusheatmap."

9. The width of the output figure: Specify the width of the output heatmap

PDF, with the default value being 5.5.

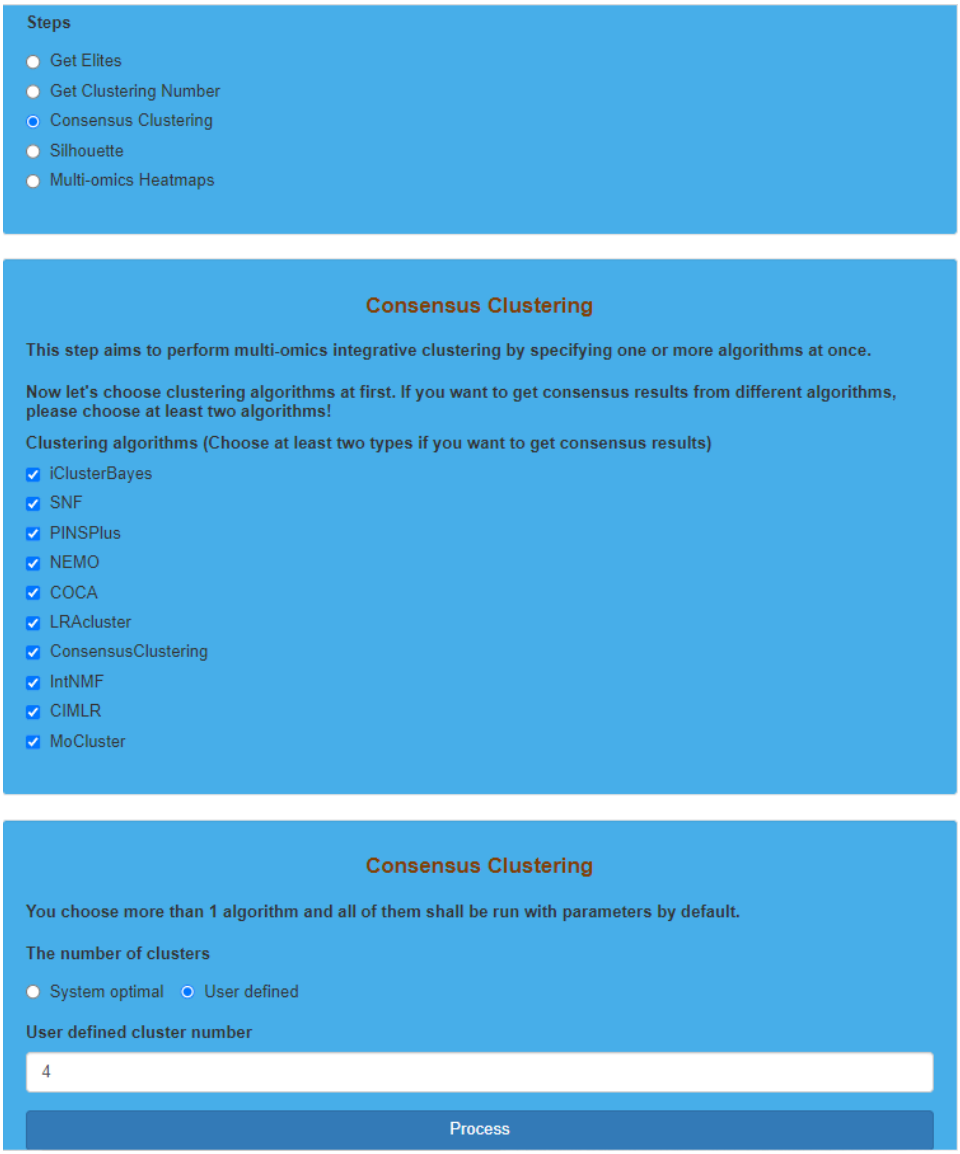
10. The height of the output figure: Specify the height of the output heatmap PDF, with the default value being 5.

After configuring the parameters for the consensus clustering heatmap, click the "Process" button to start creating the heatmap. Once completed, feedback will be provided in the "Consensus Clustering" results module on the right side of the webpage, and users can choose to download the RData dataset and the consensus clustering heatmap PDF. Next, evaluate the clustering results by calculating the sample-to-sample similarity of the obtained subtypes. Navigate to the "Steps" parameter module under the "GET Module" section and choose "Silhouette." Refer to the guidance document for the Silhouette parameter module for detailed instructions.

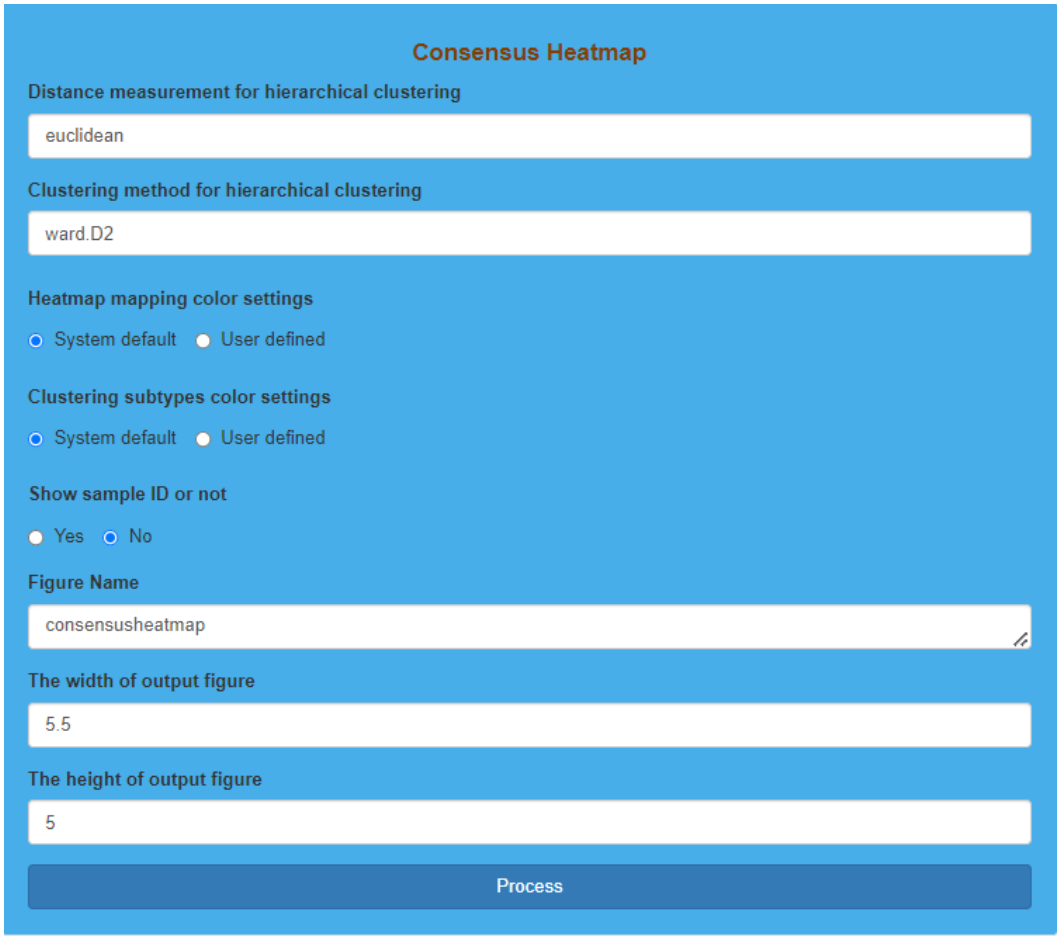
III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Consensus Clustering”, and then we choose the clustering algorithms (10 algorithms are chosen by default). The parameters for each clustering algorithm are set by default, but you can set the parameters concretely if you just indicate one clustering algorithm. Since we determine the optimal clustering number based on the prior information, therefore the “The number of clusters” chooses “User defined” and “User defined cluster number” indicates 4. After that, we click the “Process” button to do “Consensus Clustering”. After we get the results of “Consensus Clustering”, we set the

corresponding parameters to plot the consensus heatmap. “Distance measurement for hierarchical clustering” indicates “euclidean”, “Clustering method for hierarchical clustering” indicates “ward.D2”, both “Heatmap mapping color settings” and “Clustering subtypes color settings” choose “System default”, and “Show sample ID or not” chooses “No”. Besides, we indicate the “Figure Name”, “The width of output figure” and “The height of output figure”. Finally, we click the “Process” button to plot the consensus heatmap.



Parameter settings for “Consensus Clustering” on TCGA dataset



Parameter settings for consensus heatmap

Result Display:

The software will give feedback to the users if the “Consensus Clustering” is finished. Moreover, a consensus heatmap will be also displayed to reflect the quality of consensus clustering results.

Result display for “Consensus Clustering” on TCGA dataset

(4). Silhouette for TCGA dataset

I.Analysis introduction

“Silhouette” calculates and visualizes the similarity between samples in each subtype derived from clustering results using Silhouette Coefficient, which can be used to evaluate the clustering results.

II.Parameters setting guides

To evaluate the clustering results by calculating the sample-to-sample similarity of the obtained subtypes and create a similarity graph, follow these steps:

1. Colors for annotating each cluster: Specify the colors used for

annotating each cluster. The default choice is "System default," using the system's default colors. Alternatively, you can choose "User defined" and input your specified colors.

2. Input colors for annotating each cluster: If "Colors for annotating each cluster" is set to "User defined," input the specified colors (hexadecimal), for example, #2EC4B6.

3. Figure Name: Specify the name of the output similarity graph PDF, with the default being "silhouette."

4. The width of the output figure: Specify the width of the output similarity graph PDF, with the default value being 5.5.

5. The height of the output figure: Specify the height of the output similarity graph PDF, with the default value being 5.

After configuring the parameters for the similarity graph, click the "Process" button to calculate the sample-to-sample similarity of the obtained subtypes and create the similarity graph. Once completed, feedback will be provided in the "Silhouette" results module on the right side of the webpage, and users can choose to download the similarity graph PDF. Next, based on the clustering results and preprocessed multi-omics data, proceed to create multi-omics heatmaps.Navigate to the "Steps" parameter module under the "GET Module" section and choose "Multi-omics Heatmaps." Refer to the guidance document for the first "Multi-omics Heatmaps (multiple algorithms)" and the first "Multi-omics

Heatmaps (specified clustering method)" parameter modules for detailed instructions.

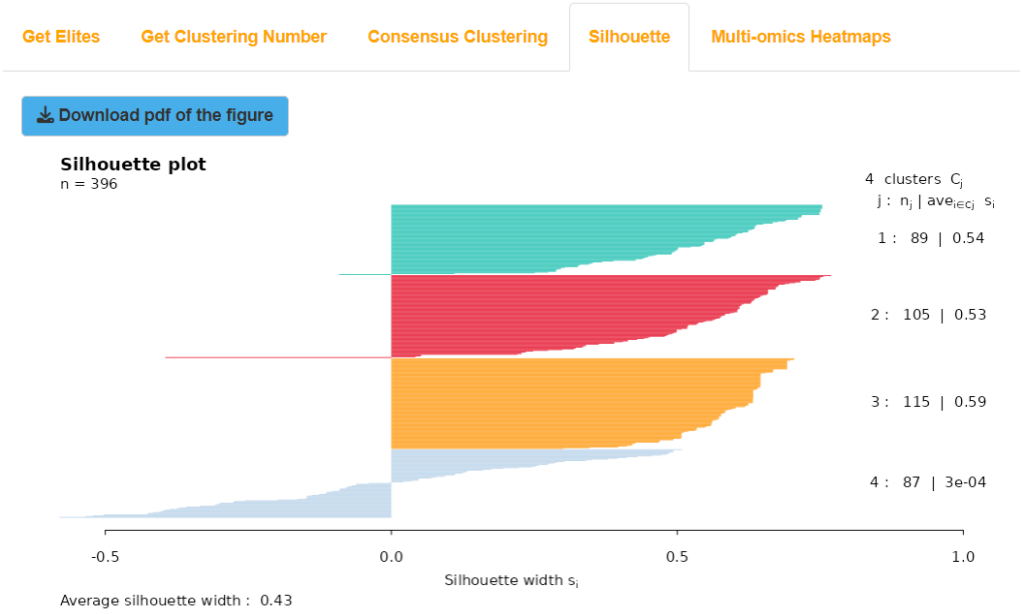
III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Silhouette”, and “Colors for annotating each cluster” chooses “System default”. Then, we indicate the “Figure Name”, “The width of output figure” and “The height of output figure”. Finally, we click the “Process” button to generate the Silhouette plot.

Parameter settings for “Silhouette” on TCGA dataset

Result Display:

The software will generate a Silhouette plot to evaluate the similarity between samples in each subtype derived from clustering results through Silhouette Coefficient, which can be used to evaluate the clustering results. Silhouette Coefficient ranges from -1 to 1, and the larger value indicates better clustering results.



The process of visualizing silhouette information from consensus clustering has been finished, you can download and check the figure. Now let's turn to the step of 'Multi-omics Heatmaps'.

Result display for “Silhouette” on TCGA dataset

(5). Multi-omics Heatmaps for TCGA dataset

I.Analysis introduction

“Multi-omics Heatmaps” combines multi-omics data and clustering results to generate multi-omics heatmaps, which can be utilized to evaluate the clustering results based on the expression differences of different subtypes in specific omics features.

II.Parameters setting guides

Multi-omics Heatmaps

Finally, based on the clustering results and preprocessed multi-omics data, create multi-omics heatmaps in three scenarios: (1) Use two or more clustering methods and generate heatmaps using the consensus

clustering results. (2) Use two or more clustering methods but create heatmaps using the results of a specific clustering method. (3) Use only one clustering method and generate heatmaps using the results of that clustering method.

Before creating a multi-omics heatmap, it is necessary to specify the processing parameters for each set of omics data：

Heatmaps for consensus clustering results of multiple clustering algorithms or not: Whether to use consensus clustering results for creating a multi-omics heatmap. This parameter is specified when employing two or more clustering methods, with the default choice being "Yes" to use consensus clustering results.

Indicate the clustering results derived from specified algorithm for multi-omics heatmaps: If the "Heatmaps for consensus clustering results of multiple clustering algorithms or not" parameter is set to "No," you need to input the specific clustering method to be used for creating a multi-omics heatmap.

Indicate if each omic data should be centered: Specify whether each type of omics data needs to undergo centering processing, i.e., subtracting their respective feature means. The default is selected for all.

Indicate if each omic data should be scaled: Specify whether each type of omics data needs to undergo scaling processing, i.e., dividing by their respective feature standard deviations. The default is checked for all.

Assign truncating values for extreme values in continuous normalized multi-omics data: Specify the truncation values for selected continuous type omics data. This truncation value is a positive number; if the data is greater than this value, it will be replaced by the truncation value, and if it is less than the negative of this value, it will be replaced by the negative of the truncation value.

Next, you will continue to specify the plotting parameters for multi-omics heatmaps, including:

1. Row Title Settings for Each Omic Data: Specify row titles for each omics data. Default is "System default," using system default row titles. If "User defined" is selected, users need to define row titles.

2. Input Row Title for Each Omic Data: If the "Row title settings for each omic data" parameter is set to "User defined," users need to input row titles for each omics data.

3. Legend Title Settings for Each Omic Data: Specify legend titles for each omics data. Default is "System default," using system default legend titles. If "User defined" is selected, users need to define legend titles.

4. Input Legend Title for Each Omic Data: If the "Legend title settings for each omic data" parameter is set to "User defined," users need to input legend titles for each omics data.

5. Show Dendrogram for Columns at the Top of Heatmap: Specify whether to draw sample clustering dendrogram at the top of the multi-omics

heatmap. Default is "Yes," otherwise "No."

6. Show Sample Names for Columns at the Bottom of Heatmap: Specify whether to display sample names at the bottom of the multi-omics heatmap. Default is "No," otherwise "Yes."

7. Show Dendrogram for Rows of Each Omic Data: Specify whether to draw feature clustering dendrogram for each omics data on the left side of the heatmap. Default is selected.

8. Colors for Annotating Each Cluster at the Top of Heatmap: Specify colors for annotating clusters at the top of the heatmap. Default is "System default," or choose "User defined" to input specified colors (hexadecimal format).

9. Colors for Heatmap of Each Omic Data: Specify colors for the heatmap of each omics data. Default is "System default" or choose "User defined" to input specified colors (hexadecimal format).

10. Annotation of Features in Dataset Heatmaps: Specify whether to annotate features in dataset-specific heatmaps. Default is "No," otherwise "Yes."

11. Feature Annotation Settings: If feature annotation is enabled, specify the number of features to be annotated, settings for annotation, and input features.

12. Sample Annotations from Survival Information: Specify whether to use clinical survival information for annotating the heatmap. Default is "No,"

otherwise "Yes."

13. Sample Annotation Settings: If sample annotation is enabled, specify the number of sample annotations, and input detailed information for each variable.

14. Heatmap Dimensions: Specify the width and height for each heatmap. Default values are 6 for width and 2 for height.

15. Figure Name: Specify the output PDF name for the multi-omics heatmap. Default is "moheatmap."

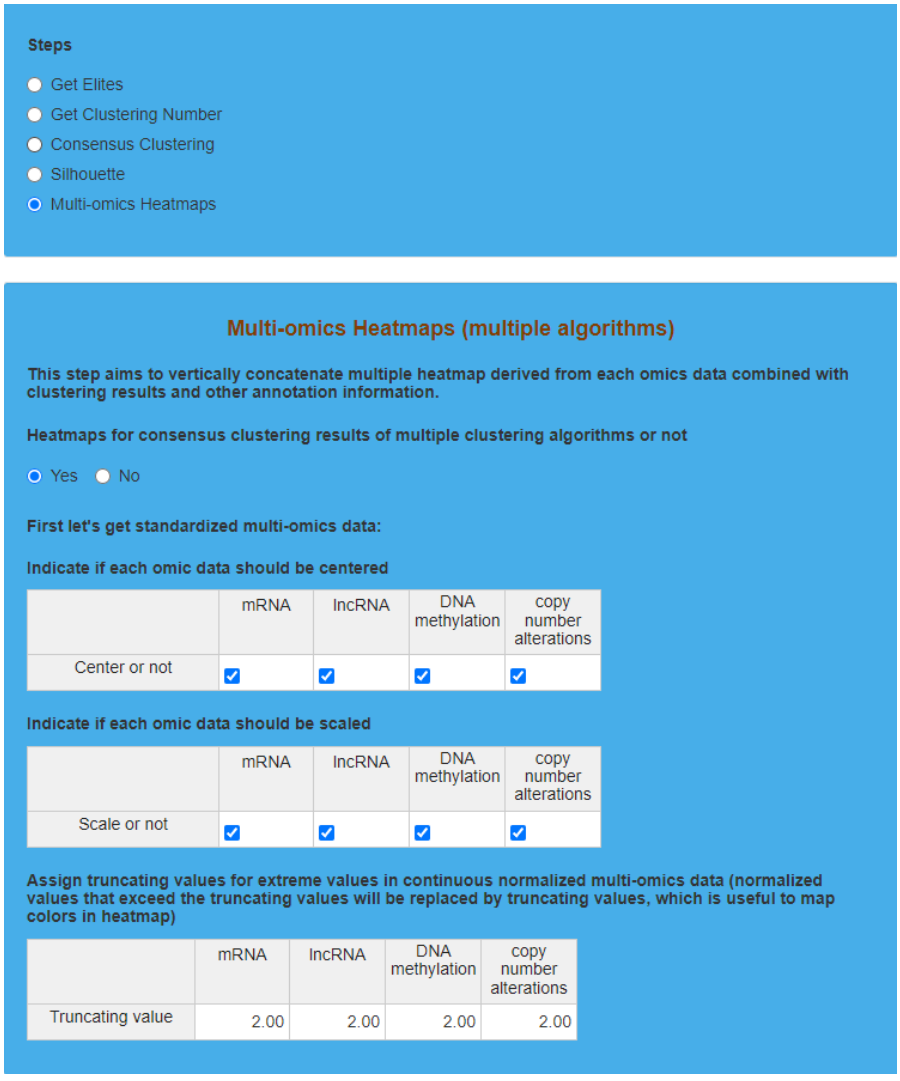
Once all processing and plotting parameters for each omics data are set, click the "Process" button to execute data processing and heatmap generation. Upon completion, feedback on Multi-omics Heatmaps results will be provided on the right side of the webpage. Users can choose to download the RData result dataset and the multi-omics heatmap in PDF format. This concludes all analyses in the GET Module. Next, click the "COMP Module" button at the top of the webpage to perform comparative analysis of the obtained subtypes (refer to the guidance document under the "Steps" parameter module in the COMP Module for details).

III.Examples with results interpretation

Parameter Settings: There are many parameters that need to be set in this step. The “Steps” chooses “Multi-omics Heatmaps”, and we should specify whether to use the results of consensus clustering (“Heatmaps for

consensus clustering results of multiple clustering algorithms or not" chooses "Yes" by default. Otherwise, we choose "No" and then indicate the clustering results derived from specified algorithm). Then, we need to standardize each omics data ("Indicate if each omic data should be centered", "Indicate if each omic data should be scaled" and "Assign truncating values for extreme values in continuous normalized multi-omics data" use default settings). After that, we start to set the parameters of the multi-omics heatmap such as "Row title settings for each omic data", "Legend title settings for each omic data", "Show the sample names for columns at the bottom of heatmap or not", "Colors for annotating each cluster at the top of heatmap", "Colors for heatmap of each omic data", "The width for each heatmap", "The height for each heatmap" and "Figure Name" (All these parameters use default settings). In addition, we should indicate the parameters of "Click the boxes below to indicate whether show the feature names for rows of each omic data", "Click the boxes below to indicate whether show dendrogram for rows of each omic data", "Input distance method for clustering each omic data at feature dimension" and "Input clustering method for each omic data at feature dimension" for each omics data (All these parameters use default settings). In particular, clustering algorithms of iClusterBayes, CIMLR, MoCluster contain the results of feature screening, and the software provides the option to display the results. Additionally, clustering

algorithms of COCA, LRAcluster, ConsensusClustering, and MoCluster, as well as consensus clustering will generate a hierarchical clustering diagram of samples, and the users can decide whether to display it or not. Here we choose “No” for “Show the dendrogram for columns at the top of heatmap or not”. In addition to annotation for sample through clustering results, “Multi-omics Heatmaps” also allows users to select specific clinical features for annotation, and we select “PAM_Subtype”, “oneNN_Subtype”, “Lund_Subtype” and “TCGA_Subtype”. Finally, we click the “Process” button to plot the multi-omics heatmap.



Multi-omics Heatmaps (multiple algorithms)

Then let's set the parameters of the heatmap:

Row title settings for each omic data

(•) System default ( ) User defined

Legend title settings for each omic data

(•) System default ( ) User defined

Show the dendrogram for columns at the top of heatmap or not

( ) Yes (•) No

Show the sample names for columns at the bottom of heatmap or not

( ) Yes (•) No

Click the boxes below to indicate whether show the feature names for rows of each omic data

	mRNA	lncRNA	DNA methylation	copy number alterations	binary somatic mutation
Show rownames	☐	☐	☐	☐	☐

Input distance method for clustering each omic data at feature dimension

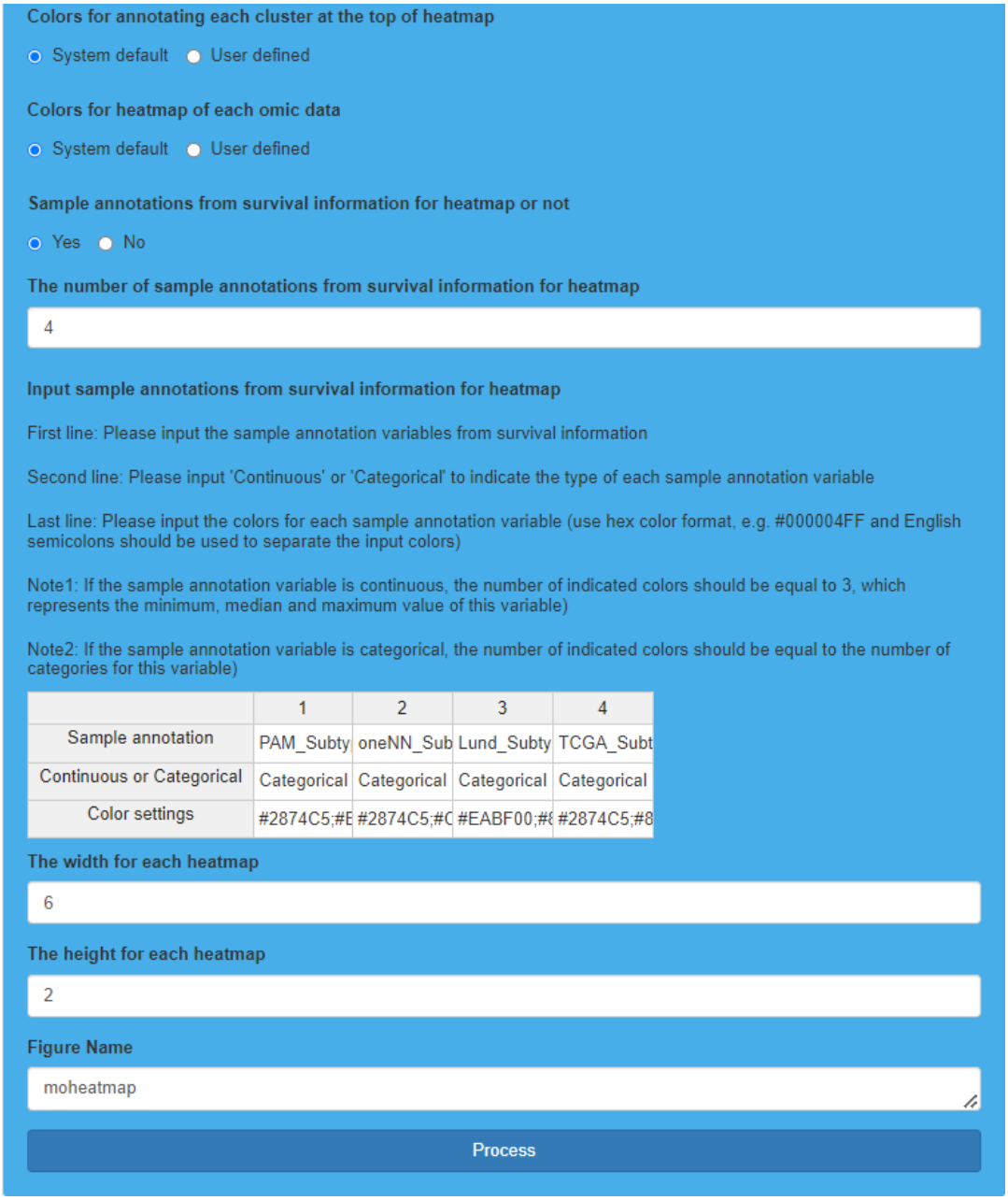
	mRNA	lncRNA	DNA methylation	copy number alterations	binary somatic mutation
Distance method	pearson	pearson	pearson	pearson	pearson

Input clustering method for each omic data at feature dimension

	mRNA	lncRNA	DNA methylation	copy number alterations	binary somatic mutation
Clustering method	ward.D	ward.D	ward.D	ward.D	ward.D

Click the boxes below to indicate whether show dendrogram for rows of each omic data

	mRNA	lncRNA	DNA methylation	copy number alterations	binary somatic mutation
Show row dendrogram	☑	☑	☑	☑	☑

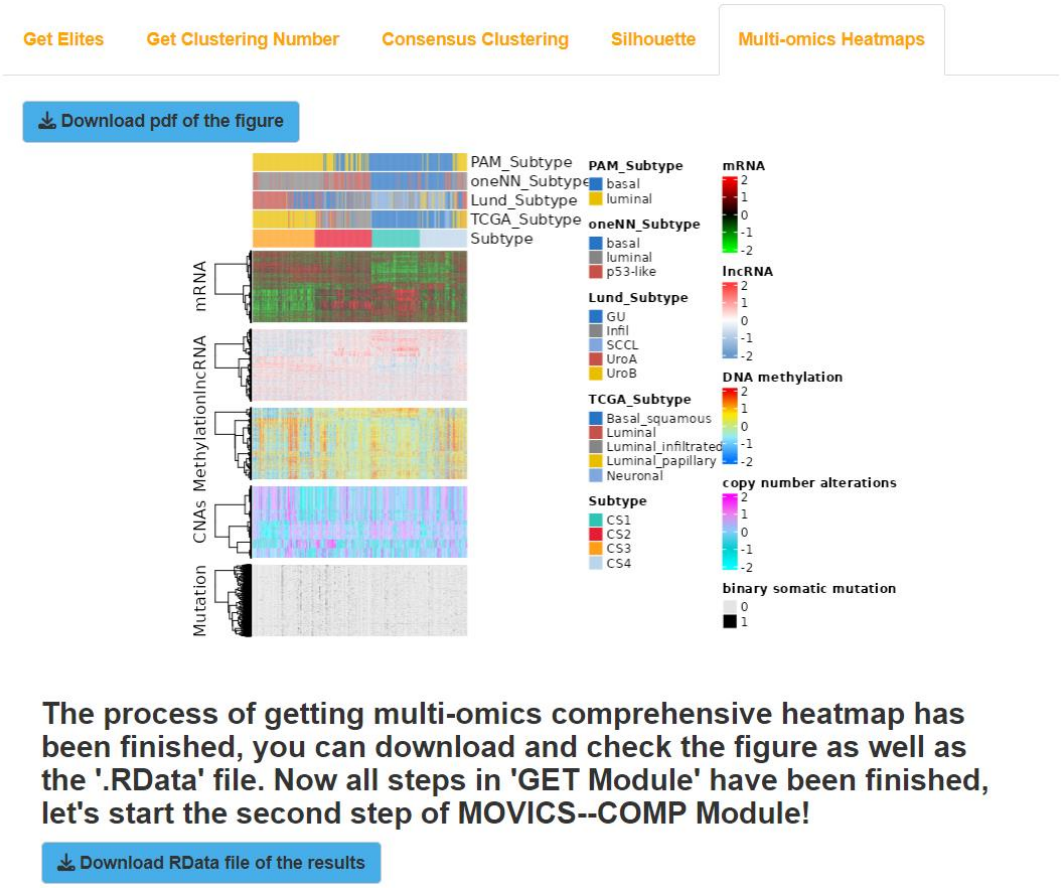


Parameter settings for “Multi-omics Heatmaps” on TCGA dataset

Result Display:

The software will generate a multi-omics heatmap, and users can observe the expression differences of features in each omics data for each subtype. Because the generated multi-omics heatmap is relatively large, it may not be displayed clearly enough on the website. So, users can download the

heatmap, and then enlarge it for observation.



Result display for “Multi-omics Heatmaps” on TCGA dataset

The third step: COMP Module (TCGA dataset)

“COMP Module” compares characteristics of different subtypes obtained from clustering results, which is divided into seven sub-modules, containing “Compare survival outcome”, “Compare clinical features”, “Compare mutational frequency”, “Compare total mutation burden”, “Compare fraction genome altered”, “Compare drug sensitivity” and “Compare agreement with other subtypes”. The TCGA datasets should contain corresponding types of data for specified analysis in “COMP

Module”.

(1). Compare survival outcome for TCGA dataset

I.Analysis introduction

“Compare survival outcome” generates KAPLAN-MEIER curves to show the significance of survival differences among different subtypes.

II.Parameters setting guides

i.COMP Module

After completing the clustering of multi-omics data, we will proceed with comparative analysis between the subtypes obtained. Firstly, we will compare survival outcome between different subtypes, which requires the prior preparation of clinical survival data. Next, we will compare clinical features of each subtype on specified clinical variables, and clinical survival data should be prepared in advance for this analysis as well. Following that, we will compare mutational frequency of genes between subtypes, necessitating the preparation of binary somatic mutation data. The subsequent step involves comparing the differences in total mutation burden among subtypes, and again, binary somatic mutation data needs to be prepared beforehand. Additionally, we will compare fraction of the genome altered by copy number alterations in genes affected by copy number increase or decrease in each subtype, requiring prepared copy number alterations data. Moreover, we will use IC50 values to compare the response levels of each subtype to drug treatments in the GDSC

database, and mRNA or lncRNA data needs to be prepared in advance for this analysis. Finally, we will assess the consistency between clustering results and traditional classification results, evaluating the quality of the clustering results, and clinical survival data should be prepared beforehand for this evaluation.

The first comparison involves the survival outcomes between different subtypes. Go to the Steps parameter module under the COMP Module and select "Compare survival outcome". For specific parameter settings, please refer to the guidance document for the "Compare survival outcome" parameter module.

ii.Compare survival outcome

The first choice is to compare the survival outcomes among subtypes obtained through clustering. It is necessary to prepare clinical survival data first, and the parameters that need to be specified include:

Compare survival outcome on TCGA datasets or Validation datasets: Specify the dataset for survival outcome comparison, with TCGA as the default selection. If choosing Validation, it is necessary to first complete the analysis of "Run nearest template prediction" (select "Validation" for the parameter "Run nearest template prediction on TCGA or validation cohort") or "Run partition around medoids classifier" (select "Validation" for the parameter "Run partition around medoids classifier on TCGA or validation cohort") under the RUN Module.

Model-free approaches for subtype prediction in validation cohort: If the "Compare survival outcome on TCGA datasets or Validation datasets" parameter is set to "Validation", it is necessary to specify the model-free method for subtype prediction in the external validation cohort. The default selection is "NTP" (Run nearest template prediction analysis), which uses the subtype prediction results obtained from the analysis of the external validation cohort using Run nearest template prediction. Alternatively, you can choose "PAM", which uses the subtype prediction results obtained from the analysis of the external validation cohort using Run partition around medoids classifier.

Format conversion of the survival time: The unit for survival time can be set, with the default selection being "No conversion", which uses the original unit (days) from the survival data and plots Kaplan-Meier curves with days as the x-axis unit. Additionally, you can choose to convert the unit to "Years" or "Months" for survival time.

Setting for the x-axis cutoff for showing the maximal survival time: Set the cutoff value for the x-axis of the Kaplan-Meier curve, which represents the maximum survival time. The default is to use the system-defined maximum survival time, selected as "System default". If you choose "User defined", you need to specify the maximum survival time manually.

The x-axis cutoff for showing the maximal survival time (Unit: day): If the "Format conversion of the survival time" parameter is set to "No

conversion" (using days as the unit for survival time) and the "Setting for the x-axis cutoff for showing the maximal survival time" parameter is set to "User defined", then the user needs to specify the maximum survival time. The default value is 3650 days.

The x-axis cutoff for showing the maximal survival time (Unit: year): If the "Format conversion of the survival time" parameter is set to "Years" and the "Setting for the x-axis cutoff for showing the maximal survival time" parameter is set to "User defined", then the user needs to specify the maximum survival time. The default value is 10 years.

The x-axis cutoff for showing the maximal survival time (Unit: month): If the "Format conversion of the survival time" parameter is set to "Months" and the "Setting for the x-axis cutoff for showing the maximal survival time" parameter is set to "User defined", then the user needs to specify the maximum survival time. The default value is 120 months.

Estimate probability of surviving beyond a certain number of years: Whether to estimate the n-year survival probability. The default selection is "No" (not estimated), otherwise choose "Yes".

The number of years for surviving probability estimation: If the "Estimate probability of surviving beyond a certain number of years" parameter is set to "Yes", then you need to specify the number of n-year survival probabilities to be estimated. For example, if you want to estimate survival probabilities for 1 year, 3 years, and 5 years, this parameter would be set

to 3.

Input the year: After specifying "The number of years for surviving probability estimation", you need to enter the years for survival probability estimation one by one. For example, if you enter 1, 3, 5, it means you want to estimate the survival probabilities for 1 year, 3 years, and 5 years, respectively.

Setting for colors of clustering subtypes: Specify the colors for annotating the clusters obtained for each subtype. You can choose the system default colors (default selection) or specify your own colors (User defined).

Input color for each subtype: If the "Setting for colors of clustering subtypes" parameter is set to "User defined", you need to input the specified colors (in hexadecimal format) one by one. For example #2EC4B6.

Method for adjusting p values: Specify the p-value adjustment method for comparing the survival outcomes of each subtype. There are eight options: holm, hochberg, hommel, bonferroni, BH, BY, fdr, and none (no p-value adjustment). The default selection is BH.

The way for drawing a horizontal/vertical line at median survival: Specify the drawing method for the median survival time line on the Kaplan-Meier curve. You can draw the line in the horizontal direction (Horizontal line), vertical direction (Vertical line), both horizontal and vertical directions (Horizontal & Vertical), or choose not to draw any lines (No lines - default selection).

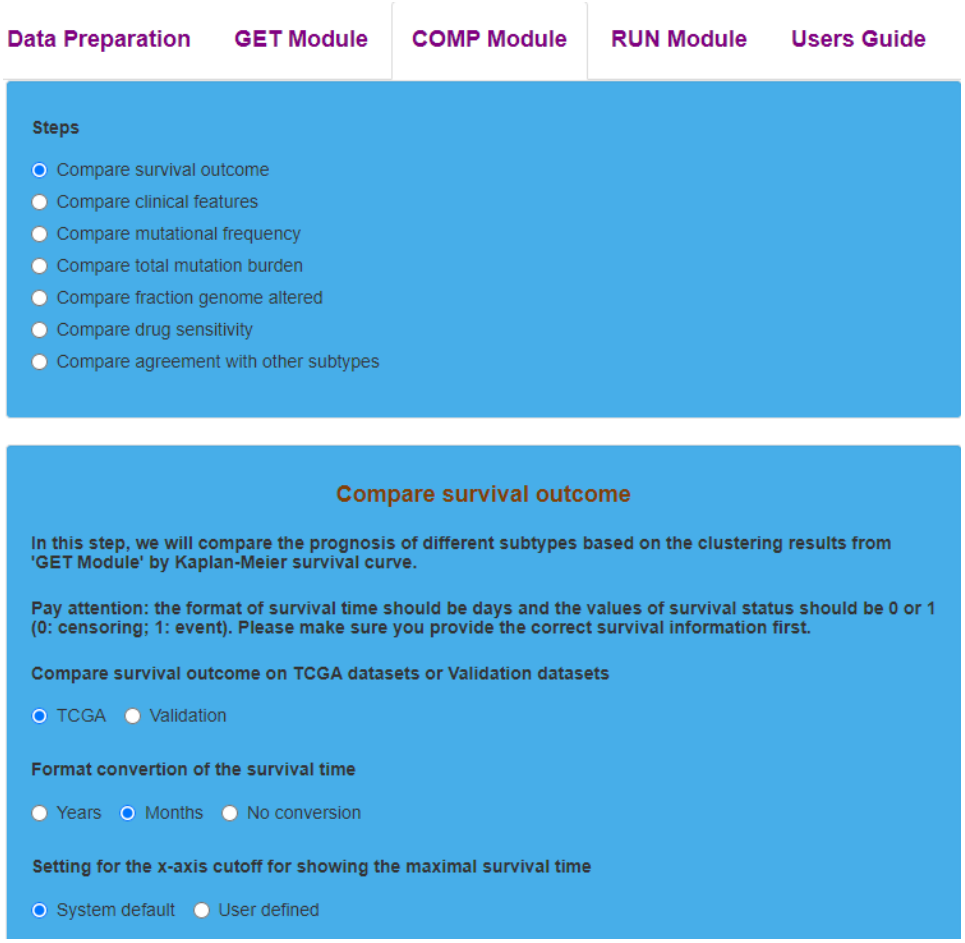
Figure Name: Specify the name for the output PDF file of the Kaplan-Meier curve. If the "Compare survival outcome on TCGA datasets or Validation datasets" parameter is set to TCGA, the default setting for this parameter is KAPLAN-MEIER_CURVE(TCGA). If the parameter is set to Validation, the default setting is KAPLAN-MEIER_CURVE(Validation).

Once all parameters are set, click the "Process" button to compare the survival outcomes among the subtypes obtained from clustering. After completion, feedback will be provided in the "Survival" module on the right side of the webpage. Users can choose to download the RData result set and Kaplan-Meier curve PDF. Next, we will compare the subtypes on specified clinical variables. Navigate to the "Steps" parameter module under the "COMP Module" and select "Compare clinical features". For detailed instructions, refer to the documentation for the "Compare clinical features" parameter module.

III.Examples with results interpretation

Parameter Settings: “Module switching options” on the upper left of the software chooses “COMP Module”, and then the “Steps” chooses “Compare survival outcome”, and “Compare survival outcome on TCGA datasets or Validation datasets” chooses “TCGA”. After that, “Format convertion of the survival time” chooses “Months” to convert data unit to month. Besides, we indicate the maximum value of the survival time as 10 years (“Setting for the x-axis cutoff for showing the maximal survival time”

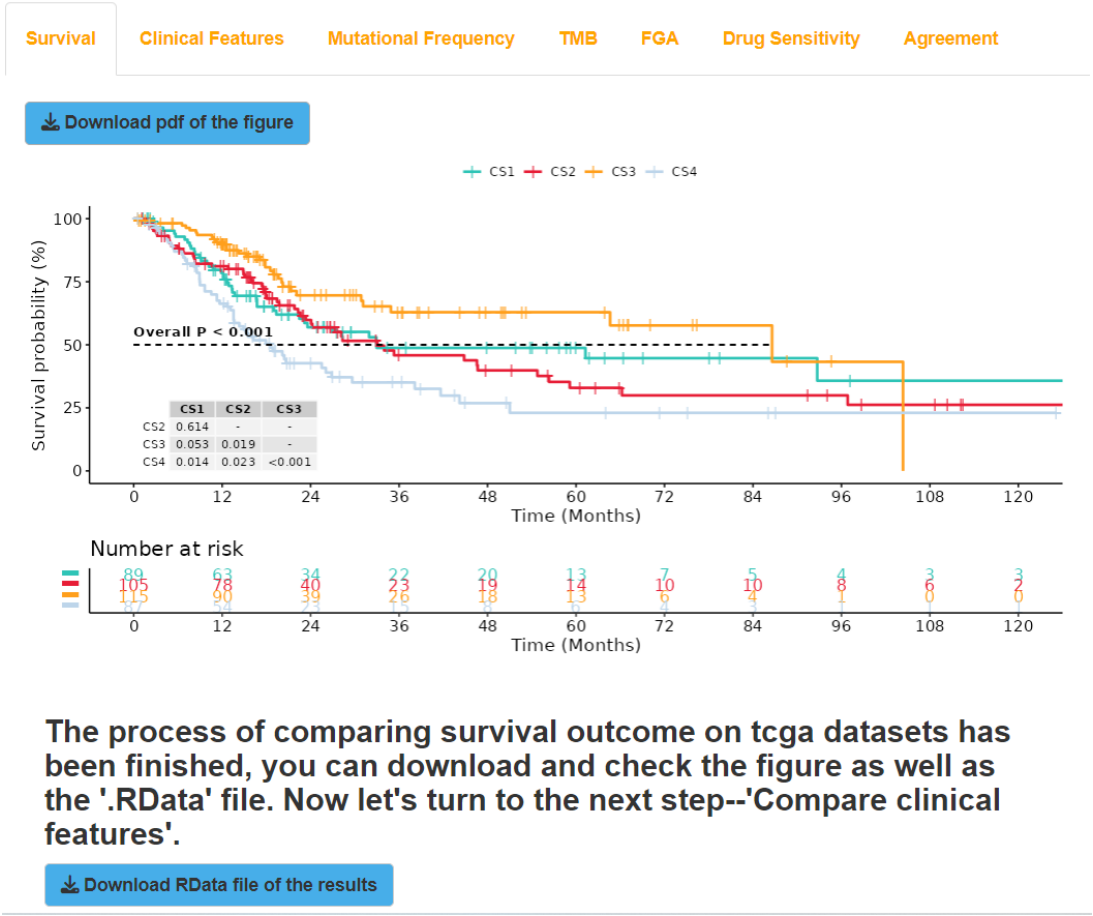
chooses “System default”). Then, we will calculate the survival rate at 5-year and 10-year. “Estimate probability of surviving beyond a certain number of years” chooses “Yes”, and then we indicate 2 as “The number of years for surviving probability estimation” and enter 5 and 10 for “Input the year”. For the adjustment of the *p* value, the software provides 8 methods and we choose “none” (Method for adjusting p values). For the median survival time, we choose “Horizontal line” (The way for drawing a horizontal/vertical line at median survival). Additionally, “Setting for colors of clustering subtypes” and “Figure Name” use default settings. Finally, we click the “Process” button to compare survival outcomes among subtypes.



Parameter settings for “Compare survival outcome” on TCGA dataset

Result Display:

The software will generate Kaplan-Meier curve for each subtype derived from clustering results, and compare the survival differences among each subtype at the same time. “Overall P<0.05” reveals significant survival differences among each subtype.



Result display for “Compare survival outcome” on TCGA dataset

(2). Compare clinical features for TCGA dataset

I.Analysis introduction

“Compare clinical features” generates a table to screen out clinical variables which are significantly associated with subtypes.

II.Parameters setting guides

Compare clinical features

Then, compare the subtypes on specified clinical variables, and you need to prepare clinical survival data in advance. The parameters to specify include:

Compare clinical features on tcga datasets or validation datasets: Specify the dataset for clinical variable comparison, with the default choice being TCGA. If you choose Validation, you need to first complete the analysis under the RUN Module, either Run nearest template prediction (choose Validation for the parameter of Run nearest template prediction on tcga or validation cohort) or Run partition around medoids classifier (choose Validation for the parameter of Run partition around medoids classifier on tcga or validation cohort).

Model-free approaches for subtype prediction in validation cohort: If the parameter "Compare clinical features on tcga datasets or validation datasets" is set to "Validation", it is necessary to specify the model-free method for subtype prediction in the external validation queue. The default choice is NTP, which utilizes the subtype predictions obtained from the Run nearest template prediction analysis on the external validation dataset. Alternatively, selecting PAM uses the subtype predictions obtained from the Run partition around medoids classifier analysis on the external validation dataset.

The number of variables in survival and clinical information chosen for summary and statistical tests: Specify the number of clinical variables for comparison; the default value is 5.

Input the variables: Enter the clinical variables for comparison one by one. The names entered must match the names of the clinical variables in the

dataset, such as OS, OS.time, age_at_initial_pathologic_diagnosis, gender, ajcc_pathologic_tumor_stage.

User defined stratifying variable or not: Whether to customize the dimensions for comparison. The default is No, using the subtypes obtained from clustering as the comparison dimensions, comparing the quantities (for discrete variables) or values (for continuous variables) of clinical variables across different subtypes.

Input the stratifying variable: If the "User defined stratifying variable or not" parameter is set to Yes, the user needs to input the stratifying variable for clinical variable comparison. This variable must come from the clinical variables in the dataset.

The number of the categorical variables in chosen variables: Specify the number of discrete variables among the clinical variables for comparison. The default value is 3.

Input the categorical variables: Enter the discrete clinical variables for comparison one by one, such as OS, gender, ajcc_pathologic_tumor_stage.

The number of variables for which the p-values should be those of nonparametric tests: Specify the number of clinical variables for which non-parametric tests will be conducted, with a default value of 0.

Input variables for which the p-values should be those of nonparametric tests: If the parameter "The number of variables for which the p-values should be those of nonparametric tests" is set to a value greater than 0,

you need to input the clinical variables for which non-parametric tests will be conducted, such as OS.time.

The number of variables for which the p-values should be those of exact tests: Specify the number of clinical variables for which exact tests will be conducted, with a default value of 0.

Input variables for which the p-values should be those of exact tests: If the parameter "The number of variables for which the p-values should be those of exact tests" is set to a value greater than 0, you need to input the clinical variables for which exact tests will be conducted. The input should include variables like ajcc_pathologic_tumor_stage.

Whether NA should be handled as a regular factor level rather than missing value: The handling of NA values can be set to treat them as regular values or as missing values. The default selection is "No", treating them as missing values.

Transform the '.txt' output file to a '.docx' WORD file or not: Specify whether to output results in Word format. The default selection is "Yes".

The name of the output table: Specify the name for the output result table. If the "Compare clinical features on tcga datasets or validation datasets" parameter is set to TCGA, the default value for this parameter is "Summarization_of_clinical_variables_stratified_by_current_subtypes(TCGA)"; if the parameter is set to Validation, the default value is "Summarization_of_clinical_variables_stratified_by_current_subtypes(V

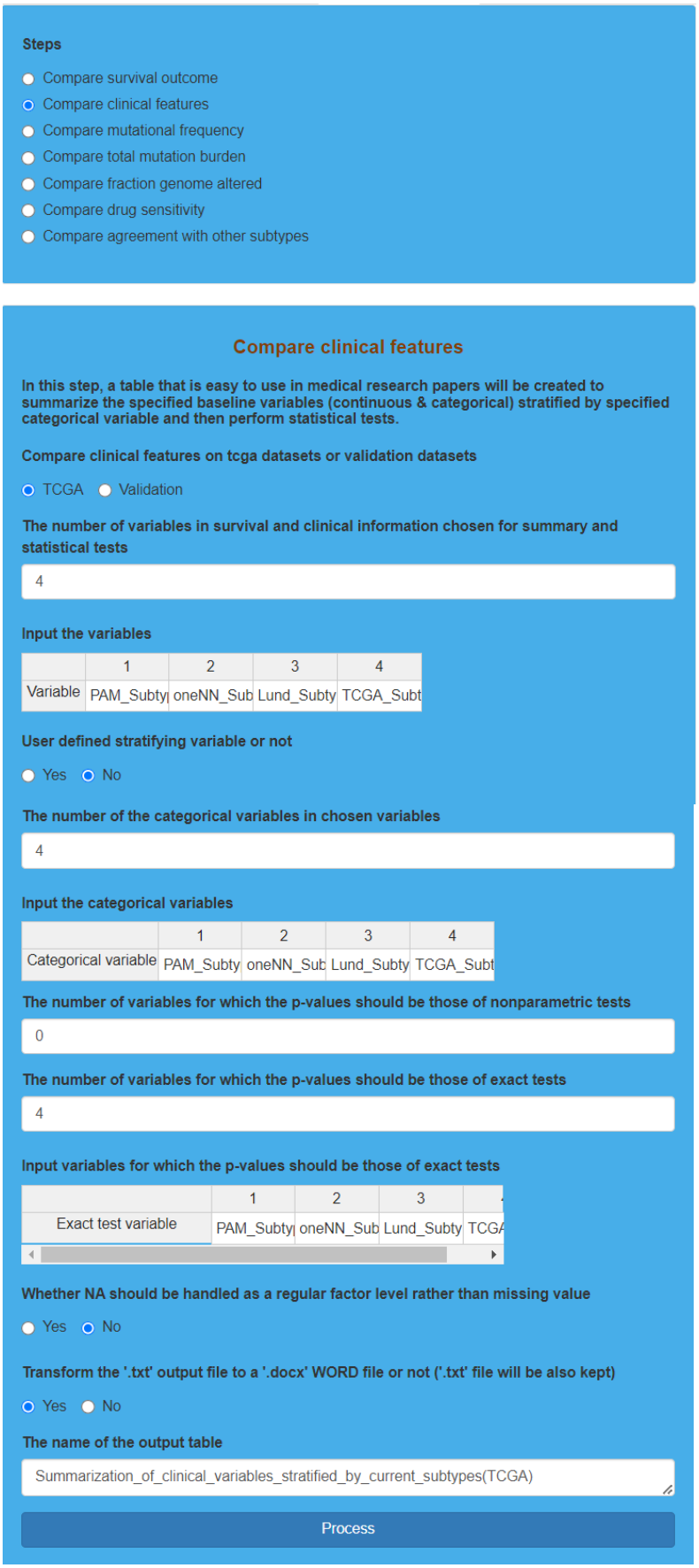
alidation)".

After setting all parameters, click the "Process" button to compare the specified clinical variables among the subtypes obtained by clustering. Once completed, feedback will be provided in the "Clinical Features" results module on the right side of the webpage. Users can choose to download the RData result set and result table in four formats (Copy: click to copy results; CSV: click to download a CSV result file; Excel: click to download an Excel result file; Print: click to go to the PDF viewing page, where a PDF result file can be downloaded). Next, we will proceed to compare the mutational frequency among subtypes. Go to the "COMP Module" under the "Steps" parameter module and select "Compare mutational frequency". Refer to the documentation for the "Compare mutational frequency" parameter module for detailed instructions.

III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Compare clinical features”, and “Compare clinical features on tcga datasets or validation datasets” chooses “TCGA”. First, we indicate 4 as the number of clinical features for comparison among subtypes derived from clustering results (“The number of variables in survival and clinical information chosen for summary and statistical tests”). After that, we enter the features containing “PAM_Subtype”, “oneNN_Subtype”, “Lund_Subtype”, and “TCGA_Subtype” for “Input the variables” and then define “Subtype” as

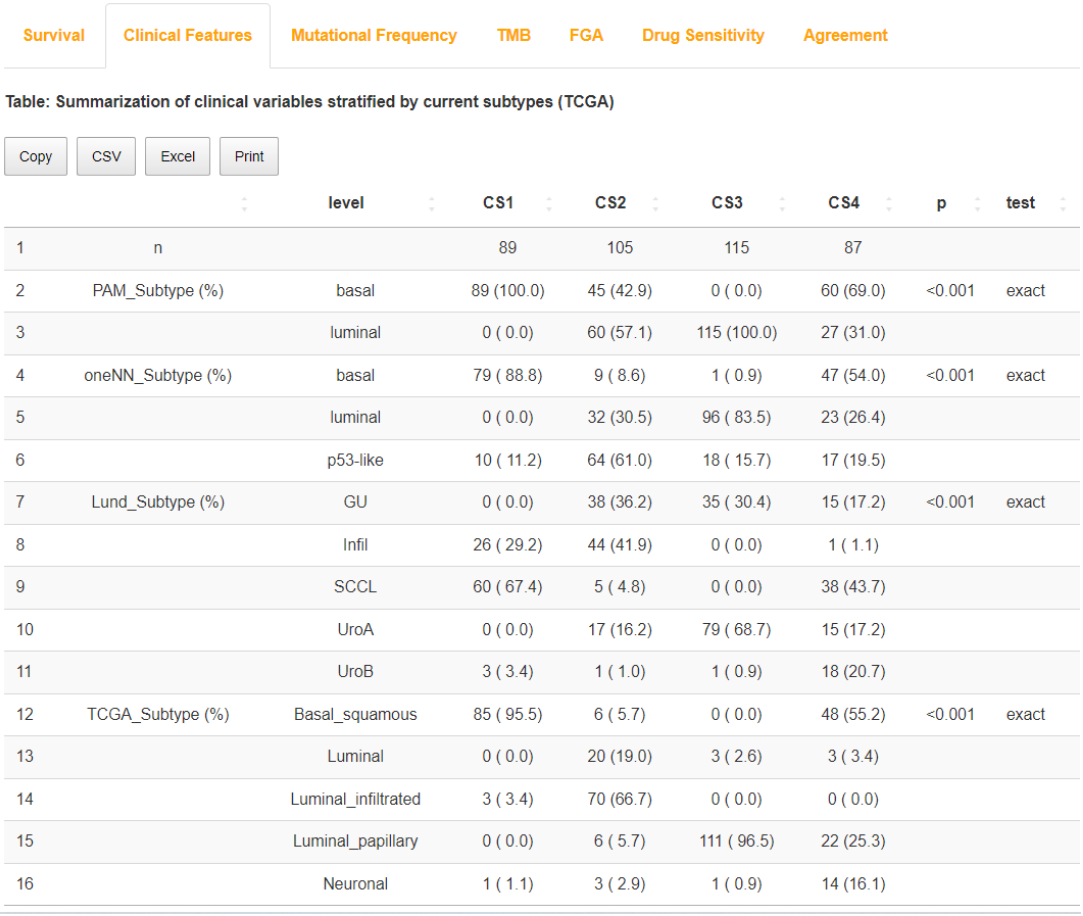
stratifying variable (“User defined stratifying variable or not” chooses “No” by default). Besides, we also indicate 4 as the number of categorical clinical features (“The number of the categorical variables in chosen variables”), and then enter all features (“Input the categorical variables”). Moreover, we indicate 0 as the number of features for nonparametric test (“The number of variables for which the p-values should be those of nonparametric tests”) and 4 as the number of features for exact test (“The number of variables for which the p-values should be those of exact tests”) respectively, and then enter the corresponding features (all features for “Input variables for which the p-values should be those of exact tests”). Additionally, “Whether NA should be handled as a regular factor level rather than missing value” chooses “No” and “Transform the '.txt' output file to a '.docx' WORD file or not” chooses “Yes” by default. Finally, we indicate the “The name of the output table” using default setting, and then click the “Process” button to compare clinical features among subtypes.



Parameter settings for “Compare clinical features” on TCGA dataset

Result Display:

The software will generate a table to display the comparison results of clinical features among subtypes derived from clustering results. "*p*<0.05" reveals the clinical feature that is significantly correlated with subtypes.



Survival | Clinical Features | Mutational Frequency | TMB | FGA | Drug Sensitivity | Agreement

Table: Summarization of clinical variables stratified by current subtypes (TCGA)

Copy CSV Excel Print

		level	CS1	CS2	CS3	CS4	p	test
1	n		89	105	115	87		
2	PAM_Subtype (%)	basal	89 (100.0)	45 (42.9)	0 (0.0)	60 (69.0)	<0.001	exact
3		luminal	0 (0.0)	60 (57.1)	115 (100.0)	27 (31.0)		
4	oneNN_Subtype (%)	basal	79 (88.8)	9 (8.6)	1 (0.9)	47 (54.0)	<0.001	exact
5		luminal	0 (0.0)	32 (30.5)	96 (83.5)	23 (26.4)		
6		p53-like	10 (11.2)	64 (61.0)	18 (15.7)	17 (19.5)		
7	Lund_Subtype (%)	GU	0 (0.0)	38 (36.2)	35 (30.4)	15 (17.2)	<0.001	exact
8		Infil	26 (29.2)	44 (41.9)	0 (0.0)	1 (1.1)		
9		SCCL	60 (67.4)	5 (4.8)	0 (0.0)	38 (43.7)		
10		UroA	0 (0.0)	17 (16.2)	79 (68.7)	15 (17.2)		
11		UroB	3 (3.4)	1 (1.0)	1 (0.9)	18 (20.7)		
12	TCGA_Subtype (%)	Basal_squamous	85 (95.5)	6 (5.7)	0 (0.0)	48 (55.2)	<0.001	exact
13		Luminal	0 (0.0)	20 (19.0)	3 (2.6)	3 (3.4)		
14		Luminal_infiltrated	3 (3.4)	70 (66.7)	0 (0.0)	0 (0.0)		
15		Luminal_papillary	0 (0.0)	6 (5.7)	111 (96.5)	22 (25.3)		
16		Neuronal	1 (1.1)	3 (2.9)	1 (0.9)	14 (16.1)		

Result display for "Compare clinical features" on TCGA dataset

(3). Compare mutational frequency for TCGA dataset

I.Analysis introduction

"Compare mutational frequency" utilizes a table to display the mutation frequency of genes that meet certain conditions, and then draws a waterfall chart to display the genes whose mutation frequency is

significantly different in each subtype.

II.Parameters setting guides

Compare mutational frequency

Next, we will compare the mutational frequency among the subtypes obtained through clustering. To do this, you need to prepare binary somatic mutation data, and the parameters to specify include:

Compare mutational frequency on tcga datasets or validation datasets: Specify the dataset for comparing the mutational frequency, with the default selection being TCGA. If you choose Validation, you need to first complete the analysis under the RUN Module, either Run nearest template prediction on tcga or validation cohort (selecting Validation) or Run partition around medoids classifier on tcga or validation cohort (selecting Validation).

Model-free approaches for subtype prediction in validation cohort: If the parameter "Compare mutational frequency on tcga datasets or validation datasets" is set to Validation, you need to specify the model-free method for subtype prediction in the external validation queue. The default choice is NTP, which utilizes the subtype prediction results obtained from the Run nearest template prediction analysis on the external validation set. Alternatively, choose PAM if using the results from the Run partition around medoids classifier analysis on the external validation set.

The frequency cutoff for mutation data: Enter the threshold for

mutational frequency. Only genes with a mutational frequency (number of samples with mutations/total number of samples) greater than this threshold will be selected for comparison. The default threshold value is 0.05.

Statistical method for independence testing: Specify the statistical method for the independence test. The default choice is Fisher's exact test.

The correction method for multiple comparison: Specify the correction method for p-values in multiple comparisons. The available options include Holm, Hochberg, Hommel, Bonferroni, BH, BY, and FDR. The default choice is BH.

Transform the '.txt' output file to a '.docx' WORD file or not: Specify whether to output the results in Word format. The default choice is Yes.

The name of the output table: Specify the name of the output result table. If the parameter "Compare mutational frequency on TCGA datasets or Validation datasets" is selected as TCGA, this parameter is set by default to "Independent_test_between_subtype_and_mutation(TCGA)"; if Validation is selected, this parameter is set by default to "Independent_test_between_subtype_and_mutation(Validation)".

Perform clustering within each subtype for oncoprint or not: Whether to cluster within each subtype in the oncoprint, the default is set to "Yes" for clustering.

The nominal p value cutoff for significant mutations shown in oncoprint:

Enter the nominal p-value threshold. Genes with p-values less than this threshold from the comparison of mutation frequencies between subtypes will be selected.

The adjusted p value cutoff for significant mutations shown in oncoprint: Enter the adjusted p-value threshold. Genes with adjusted p-values less than this threshold from the comparison of mutation frequencies between subtypes will be selected.

Genes that are selected by both "The nominal p value cutoff for significant mutations shown in oncoprint" and "The adjusted p value cutoff for significant mutations shown in oncoprint" parameters will be displayed on the oncoprint.

User defined mutation color for oncoprint or not: If the user selects "Yes" for the parameter "Whether to customize the colors used to annotate mutations in the oncoprint", it means the user wants to customize the colors for mutations in the oncoprint.

Input the mutation color for oncoprint: If the user chooses "Yes" for the parameter "User defined mutation color for oncoprint or not", they need to specify the color (in hexadecimal format, e.g., #21498D) used for annotating mutations in the oncoprint.

User defined background color for oncoprint or not: Whether the user customizes the background color in oncoprint, with the default option being “No” to use the system's default colors.

Input the background color for oncoprint: If the "User defined background color for oncoprint or not" parameter is set to "Yes", the user needs to specify the background color for oncoprint (in hexadecimal format), for example, #dcddde.

Setting for colors to annotate each subtype in oncoprint: If "Setting for colors of clustering subtypes in oncoprint" is chosen as "User defined", the user needs to input the specified colors in hexadecimal format, such as #2EC4B6.

Input color for each subtype: If "Setting for colors to annotate each subtype in oncoprint" is chosen as "User defined", the user needs to input the specified colors in hexadecimal format, such as #2EC4B6.

Sample annotations from survival information for oncoprint or not: Specify whether to use clinical survival information for annotation at the top of oncoprint. Choose "No" for not using or "Yes" for using.

The number of sample annotations from survival information for oncoprint: If the parameter "Sample annotations from survival information for oncoprint or not" is set to "Yes", specify the number of clinical survival variables to be annotated. The default value is 3.

Input sample annotations from survival information for oncoprint: Next, you need to specify detailed information for clinical survival variables. In the first line, enter the variable names to be annotated accurately from the clinical survival dataset. In the second line, enter the data types

corresponding to each clinical survival variable. In the third line, enter the annotation colors (in hexadecimal), separated by semicolons, for each clinical survival variable, for example, #000004FF. If it is a continuous variable, the number of colors should be 3, corresponding to the minimum value, median, and maximum value of that variable. If it is a discrete variable, the number of colors should match the number of categories for that variable.

The width of output figure: Specify the width of the oncoprint output in PDF, with a default value of 8.

The height of output figure: Specify the width of the oncoprint output in PDF, with a default value of 4.

Figure Name: Specify the name of the oncoprint output in PDF. If the "Compare mutational frequency on TCGA datasets or Validation datasets" parameter is set to TCGA, the default value for this parameter is "oncoprint_for_mutations_with_frequency_over_than_the_setting_cutoff (TCGA)"; if set to Validation, the default value is "oncoprint_for_mutations_with_frequency_over_than_the_setting_cutoff (Validation)".

After setting all the parameters, click the "Process" button to compare the gene mutation frequency among the subtypes obtained from clustering. After completion, feedback will be provided in the "Mutational Frequency" module on the right side of the webpage. Users can choose to

download the RData result set, oncoprint output in PDF format, and the table comparing the mutation frequencies of selected genes across subtypes (available in four formats: Copy, CSV, Excel, Print). Next, we will compare the differences in the total mutation burden among subtypes. Go to the "COMP Module" and select the "Steps" parameter module, then choose "Compare total mutation burden". For detailed instructions, refer to the documentation for the "Compare total mutation burden" parameter module.

III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Compare mutational frequency”, and “Compare mutational frequency on tcga datasets or validation datasets” chooses “TCGA”. First, we indicate 0.03 as the frequency cutoff for mutation data. Only features mutated in over than such proportion would be included in testing. Then, we choose “fisher” for “Statistical method for independence testing”. In addition, “The correction method for multiple comparison” chooses “BH”, “Transform the '.txt' output file to a '.docx' WORD file or not” chooses “Yes”, and we also use the default name for “The name of the output table”. After that, we set the parameters for oncoprint. We first choose “Yes” for “Perform clustering within each subtype for oncoprint or not”, and indicate 0.05 for “The nominal p value cutoff for significant mutations shown in oncoprint” and 0.25 for “The adjusted p value cutoff for significant mutations shown

in oncoprint". Besides, we use the default settings for the mutation and background color for oncoprint (Corresponding parameters choose "No"), as well as the annotation color for each subtype in oncoprint ("Setting for colors to annotate each subtype in oncoprint" chooses "System default"). In addition to annotation for sample through clustering results, this step also allows users to select specific clinical features for annotation, and we select "PAM_Subtype", "oneNN_Subtype", "Lund_Subtype" and "TCGA_Subtype". Moreover, we utilize the default settings for "The width of output figure" and "Figure Name", and then indicate 5 as "The height of output figure". Finally, we click the "Process" button to compare mutational frequency among each subtype derived from clustering results.

Steps

- ○ Compare survival outcome
- ○ Compare clinical features
- ◉ Compare mutational frequency
- ○ Compare total mutation burden
- ○ Compare fraction genome altered
- ○ Compare drug sensitivity
- ○ Compare agreement with other subtypes

Compare mutational frequency

In this step, a table and an oncoprint will be generated to compare mutational frequency among different multi-omics integerative clusters to test the independency between subtypes and mutational status.

Compare mutational frequency on tcga datasets or validation datasets

◉ TCGA ○ Validation

The frequency cutoff for mutation data (Only features that mutated in over than such proportion would be included in testing)

0.03

Statistical method for independence testing

◉ fisher ○ chisq

The correction method for multiple comparison

- ○ holm
- ○ hochberg
- ○ hommel
- ○ bonferroni
- ◉ BH
- ○ BY
- ○ fdr

Transform the '.txt' output file to a '.docx' WORD file or not ('.txt' file will be also kept)

◉ Yes ○ No

The name of the output table

Independent_test_between_subtype_and_mutation(TCGA)

Perform clustering within each subtype for oncoprint or not

◉ Yes ○ No

The nominal p value cutoff for significant mutations shown in oncoprint

0.05

The adjusted p value cutoff for significant mutations shown in oncoprint

0.25

User defined mutation color for oncoprint or not

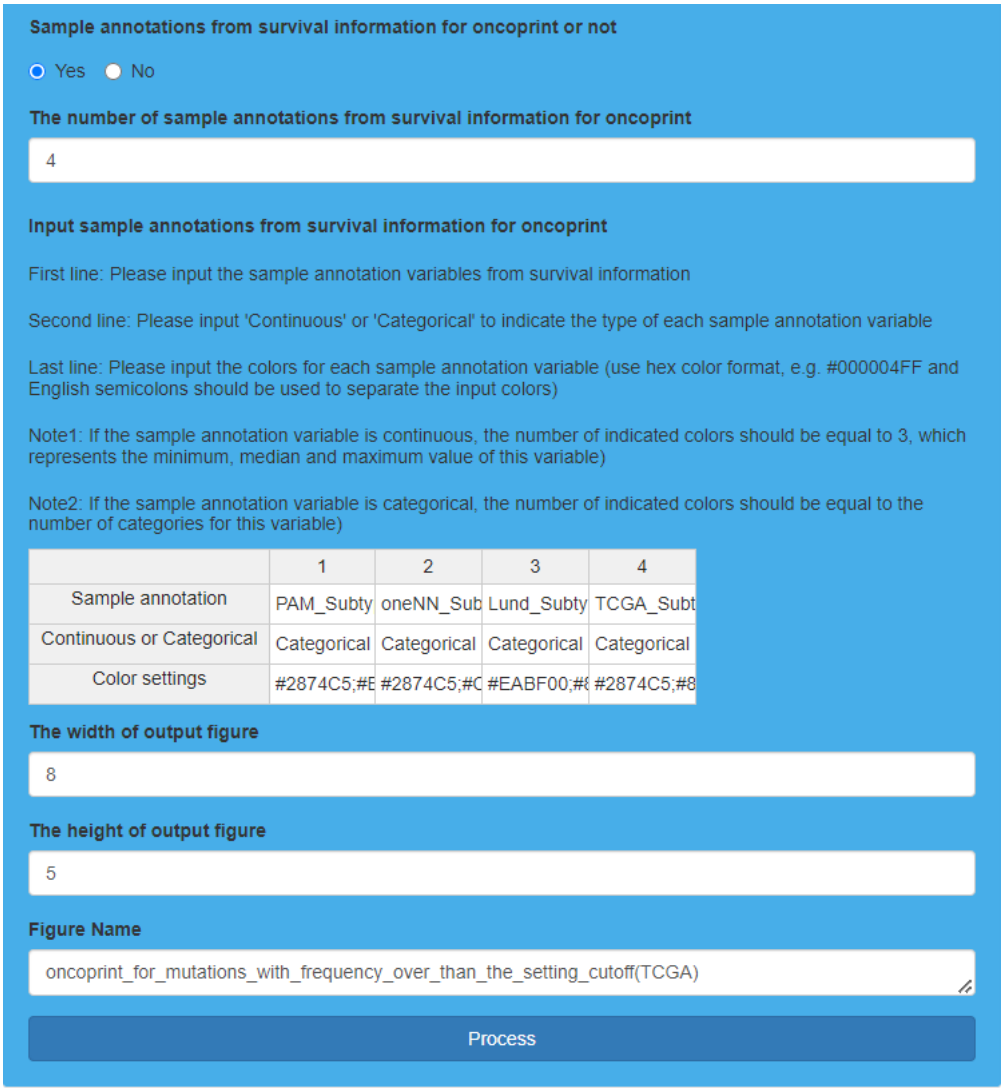
○ Yes ◉ No

User defined background color for oncoprint or not

○ Yes ◉ No

Setting for colors to annotate each subtype in oncoprint

◉ System default ○ User defined



Parameter settings for “Compare mutational frequency” on TCGA dataset

Result Display:

The software will generate a table to show the comparison of mutations among subtypes for specific genes with mutation frequency that is higher than the setting cutoff. Additionally, the software also generates an oncoprint to display the mutations of the genes whose mutation frequency is significantly different among subtypes on the basis of the table.

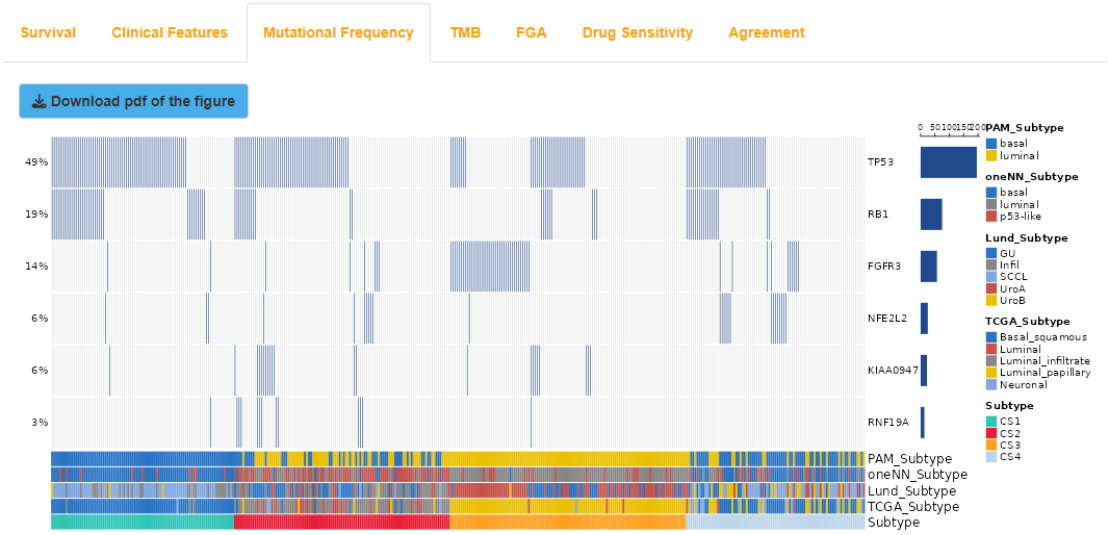


Table: Independent test between subtype and mutation (TCGA)

Copy CSV Excel Print

Gene (Mutated)	TMB	CS1	CS2	CS3	CS4	pvalue	padj
FGFR3	57 (14%)	2 (2.2%)	6 (5.7%)	39 (33.9%)	10 (11.5%)	3.44e-11	4.67e-08
TP53	196 (50%)	66 (74.2%)	56 (53.3%)	35 (30.4%)	39 (44.8%)	7.01e-09	4.76e-06
RB1	75 (19%)	35 (39.3%)	13 (12.4%)	9 (7.8%)	18 (20.7%)	1.52e-07	6.88e-05
RNF19A	13 (3%)	1 (1.1%)	11 (10.5%)	1 (0.9%)	0 (0.0%)	5.86e-05	1.99e-02
NFE2L2	25 (6%)	3 (3.4%)	7 (6.7%)	1 (0.9%)	14 (16.1%)	1.13e-04	3.07e-02
KIAA0947	22 (6%)	1 (1.1%)	12 (11.4%)	9 (7.8%)	0 (0.0%)	2.41e-04	5.45e-02

Result display for “Compare mutational frequency” on TCGA dataset

(4). Compare total mutation burden for TCGA dataset

I.Analysis introduction

“Compare total mutation burden” compares the total mutation burden (TMB) among subtypes by drawing a box-violin plot, and then uses a table to show the TMB of each sample.

II.Parameters setting guides

Compare total mutation burden

The next step involves comparing the differences among subtypes in terms of Total Mutation Burden (TMB). Before proceeding, you need to prepare binary somatic mutation data, specifying the following

parameters:

Compare total mutation burden on tcga datasets or validation datasets: Specify the dataset for comparing the Total Mutation Burden (TMB), with the default selection being TCGA. If choosing Validation, you need to first complete the analysis under the RUN Module, selecting either Run nearest template prediction on tcga or validation cohort (choose Validation for the parameter) or Run partition around medoids classifier on tcga or validation cohort (choose Validation for the parameter).

Model-free approaches for subtype prediction in validation cohort: If the parameter "Compare total mutation burden on tcga datasets or validation datasets" is set to Validation, you need to specify the model-free method for subtype prediction in the external validation queue. The default selection is NTP, which uses the subtype prediction results obtained from the analysis of the external validation set using Run nearest template prediction. Alternatively, choose PAM to use the subtype prediction results obtained from the analysis of the external validation set using Run partition around medoids classifier.

Whether remove repeated variants in a particuar sample, mapped to multiple transcripts of same gene: Specify whether to remove duplicate mutations mapped to multiple transcripts of the same gene in specific samples. The default option is Yes for removal.

Remove possible FLAGS: Specify whether to remove potential FLAGS

genes. The default option is No for no removal. FLAGS genes (such as TTN, MUC16, etc.) are often non-pathogenic and passenger genes, but they frequently mutate in most publicly available exome studies, and some of them are suspicious.

Whether user defined a list of variant classifications that should be considered as non-synonymous and the rest will be considered synonymous: Specify whether the user wants to define a list of non-synonymous mutations, and the rest will be considered synonymous (silent) mutations. The default option for this parameter is No, which means using the mutation categories with high/moderate impact, including "Frame_Shift_Del", "Frame_Shift_Ins", "Splice_Site", "Translation_Start_Site", "Nonsense_Mutation", "Nonstop_Mutation", "In_Frame_Del", "In_Frame_Ins" and "Missense_Mutation". For more details, please refer to the website http://asia.ensembl.org/Help/Glossary?id=535.

The number of variant classifications that should be considered as non-synonymous: If the user chooses "Yes" for the parameter "Whether user defined a list of variant classifications that should be considered as non-synonymous and the rest will be considered synonymous", the number of non-synonymous variant classifications needs to be specified.

Input variant classification: After specifying the number of non-synonymous variant classifications, you need to input each non-

synonymous variant classification one by one.。

The estimation of exome size: The estimate for the exonome size is 38 by default. For detailed information, please refer to the website https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0424-2.

The method for statistical testing: The default statistical test method for comparing total mutation burden between subtypes is nonparametric.

Show the sample size within each subtype at the top of the figure or not: Whether to display the sample size of each subtype in the output boxplot-violin plot, the default is to display, choose Yes.

Setting for colors to annotate each subtype: Specify the colors for annotating the subtypes obtained from clustering. You can choose the system default colors (System default, selected by default) or choose to define your own colors (User defined).

Input color for each subtype: If the "Setting for colors to annotate each subtype" parameter is set to "User defined", users need to input the specified colors (hexadecimal) one by one, for example, #2EC4B6.

The width of boxviolin plot: Specify the width of the output boxplot-violin plot in PDF, with a default value of 6.

The height of boxviolin plot: Specify the height of the output boxplot-violin plot in PDF, with a default value of 6.

Figure Name: Specify the name of the output boxplot-violin plot in PDF. If

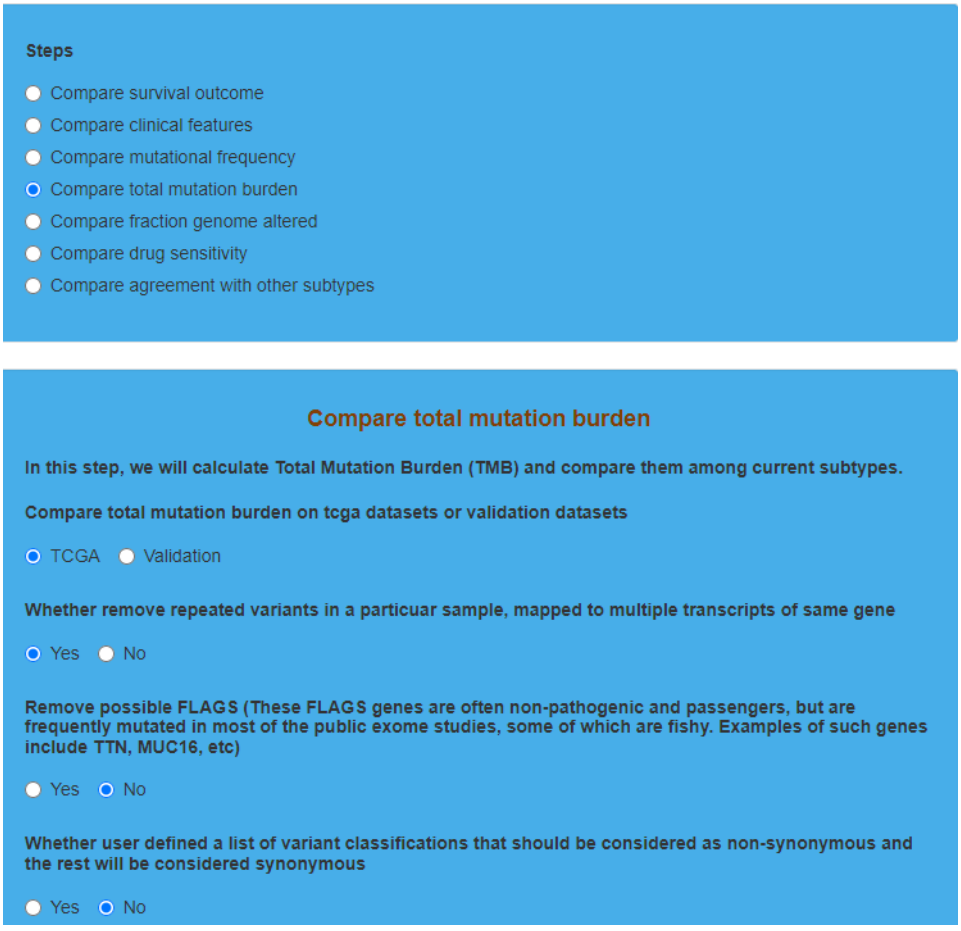
the "Compare total mutation burden on TCGA datasets or Validation datasets" parameter is set to TCGA, the default setting for this parameter is "distribution_of_TMB_and_titv(TCGA)"; if it is set to Validation, the default setting is "distribution_of_TMB_and_titv(Validation)".

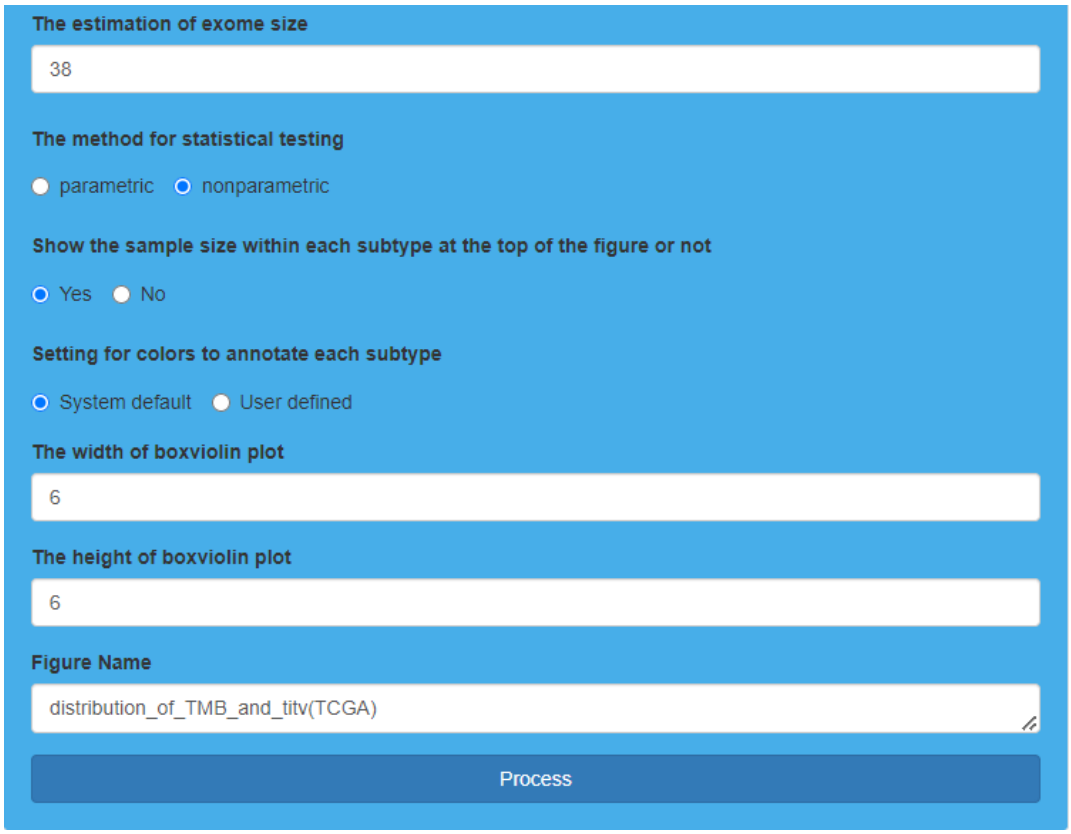
After setting all parameters, click the "Process" button to compare the total mutation burden among subtypes obtained from clustering. Once completed, feedback will be provided in the TMB (Total Mutation Burden) results module on the right side of the webpage. Please note that the generation of the boxplot-violin plot is a complex process and may take several seconds to display, so please be patient. Users can choose to download the RData result set, the boxplot-violin plot in PDF format, and the table containing the calculated total mutation burden for each sample in each subtype (available in four formats: Copy, CSV, Excel, Print). Next, the comparison will focus on the percentage of the genome altered by copy number variations in each subtype. To proceed, go to the "Steps" parameter module under the "COMP Module" section and select "Compare fraction genome altered." For detailed instructions, refer to the documentation for the "Compare fraction genome altered" parameter module.

III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Compare total mutation burden”, and “Compare total mutation burden on tcga datasets or

validation datasets" chooses "TCGA". First, we choose "Yes" for "Whether remove repeated variants in a particular sample", and then we choose "No" for "Remove possible FLAGS". After that, we choose "No" for "Whether user defined a list of variant classifications" and indicate 38 as "The estimation of exome size". Moreover, "The method for statistical testing" chooses "nonparametric" and "Show the sample size within each subtype at the top of the figure or not" chooses "Yes". Finally, we use default settings for "Setting for colors to annotate each subtype" (System default), "The width of boxviolin plot", "The height of boxviolin plot" as well as "Figure Name", and then click the "Process" button to compare total mutation burden (TMB) among subtypes derived from clustering results.





Parameter settings for “Compare total mutation burden” on TCGA dataset

Result Display:

The software will generate a box-violin plot to show the comparison results of total mutation burden (TMB) among subtypes, and the total mutation burden (TMB) of each sample will be also displayed through a table. “p<0.05” shows significantly different on total mutation burden (TMB) among subtypes.

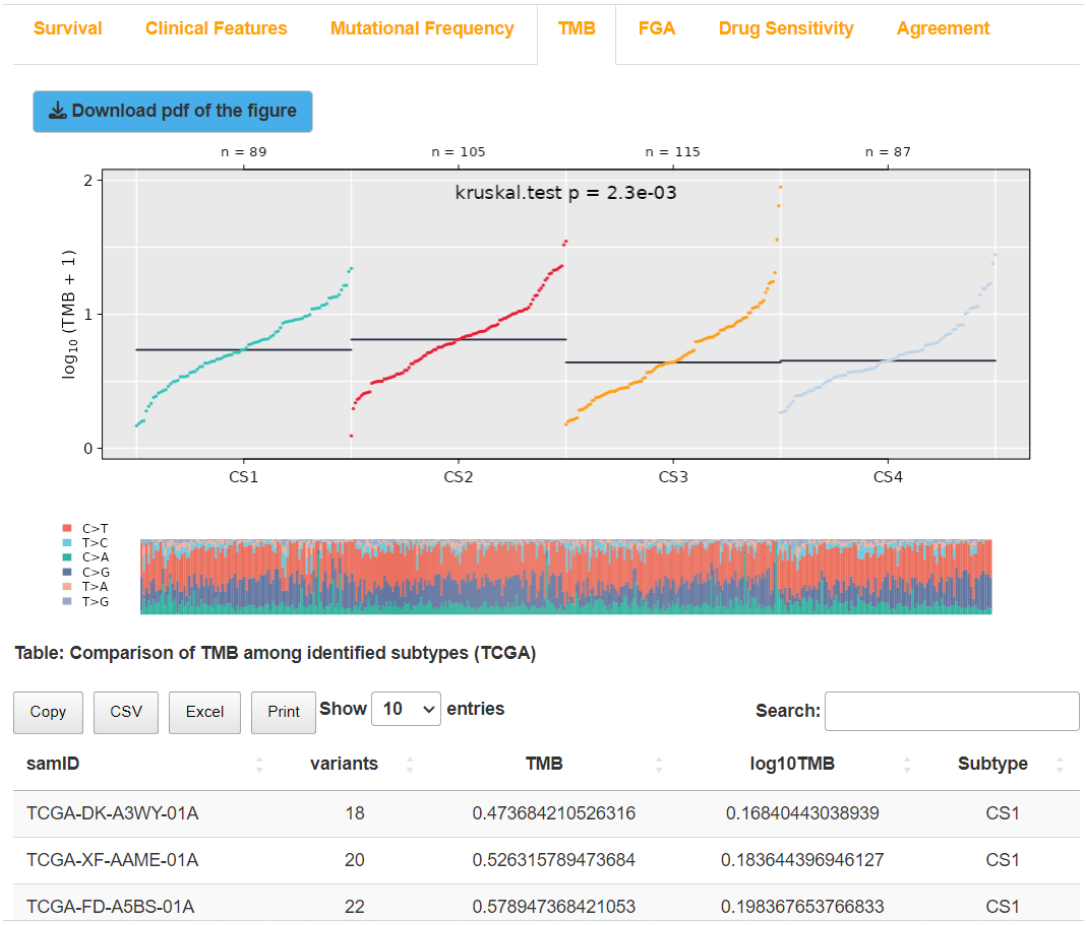


Table: Comparison of TMB among identified subtypes (TCGA)

Copy CSV Excel Print Show 10 entries Search:

samID	variants	TMB	log10TMB	Subtype
TCGA-DK-A3WY-01A	18	0.473684210526316	0.16840443038939	CS1
TCGA-XF-AAME-01A	20	0.526315789473684	0.183644396946127	CS1
TCGA-FD-A5BS-01A	22	0.578947368421053	0.198367653766833	CS1

Result display for “Compare total mutation burden” on TCGA dataset

(5). Compare fraction genome altered for TCGA dataset

I.Analysis introduction

“Compare fraction genome altered” compares the fraction of genome altered by copy number gain or loss among subtypes through bar charts, and tabulates according to the specifics of each sample at the same time. Fraction genome altered (FGA) represents the fraction of the genome altered by copy number gain or loss. Specifically, FGA can be divided into FGG and FGL, which represent the fraction of genome gained and the fraction of genome lost respectively caused by copy number gain or loss.

II.Parameters setting guides

Compare fraction genome altered

Comparing the percentage of the genome affected by copy number alterations (FGA), specifically including the percentage of the genome increased due to copy number gains (FGG) and the percentage decreased due to copy number losses (FGL) in each subtype. Prior to this analysis, preparation of copy number alterations (CNA) data is required, with the following specified parameters:

Compare fraction genome altered on tcga datasets or validation datasets: Specify the dataset for comparing the percentage of genome alterations, with the default choice being TCGA. If Validation is selected, you need to first complete the Run nearest template prediction analysis (selecting Validation for the "Run nearest template prediction on tcga or validation cohort" parameter) or Run partition around medoids classifier analysis (selecting Validation for the "Run partition around medoids classifier on tcga or validation cohort" parameter) in the RUN Module section.

Model-free approaches for subtype prediction in validation cohort: If the parameter "Compare fraction genome altered on tcga datasets or validation datasets" is set to "Validation", you need to specify the non-model method for subtype prediction in the external validation queue. The default choice is NTP (Run nearest template prediction analysis results from the external validation set), or alternatively, you can choose PAM

(using Run partition around medoids classifier analysis results from the external validation set).

The type of the 'value' column in segmented copy number dataset: Specify whether the data type in the copy number variation dataset is "segment value" or "copy number". The default choice is "segment value" (segment_mean). If you choose "copy-number value" (original), the platform will convert the copy number to segment value using the formula: segment value = $\log_2$(copy number/2).

The cutoff for identifying copy-number gain or loss: Specify the threshold for determining copy-number changes (increases or decreases). The default value is 0.2.

The method for statistical testing: Specify the statistical test method used for comparing subtypes. The default selection is nonparametric.

Setting for mapping colors for bars of FGA, FGG and FGL: Specify the colors for FGA, FGG, and FGL when drawing the comparison bar plot. You can choose the system default colors (System default, selected by default) or specify your own colors (User defined).

The mapping colors for bars of FGA, FGG and FGL: If you have chosen "User defined" for the "Setting for mapping colors for bars of FGA, FGG, and FGL" parameter, you need to input the specified colors in hexadecimal format, for example, #008B8A.

Setting for colors to annotate each subtype: Specify the colors for

annotating each subtype obtained from clustering. You can choose the system's default color (System default, selected by default) or specify your own colors (User defined).

Input color for each subtype: If the "Setting for colors to annotate each subtype" parameter is selected as "User defined", users need to input the specified colors in hexadecimal format, such as #2EC4B6.

The width of barplot: Specify the width of the output comparison bar plot in the PDF, with a default value of 8.

The height of barplot: Specify the height of the output comparison bar plot in the PDF, with a default value of 4.

Figure Name: Specify the name of the output PDF for the comparison bar plot. If the "Compare fraction genome altered on TCGA datasets or Validation datasets" parameter is set to TCGA, the default setting for this parameter is "barplot_of_fraction_genome_altered(TCGA)"; if Validation is chosen, the default setting is "barplot_of_fraction_genome_altered(Validation)".

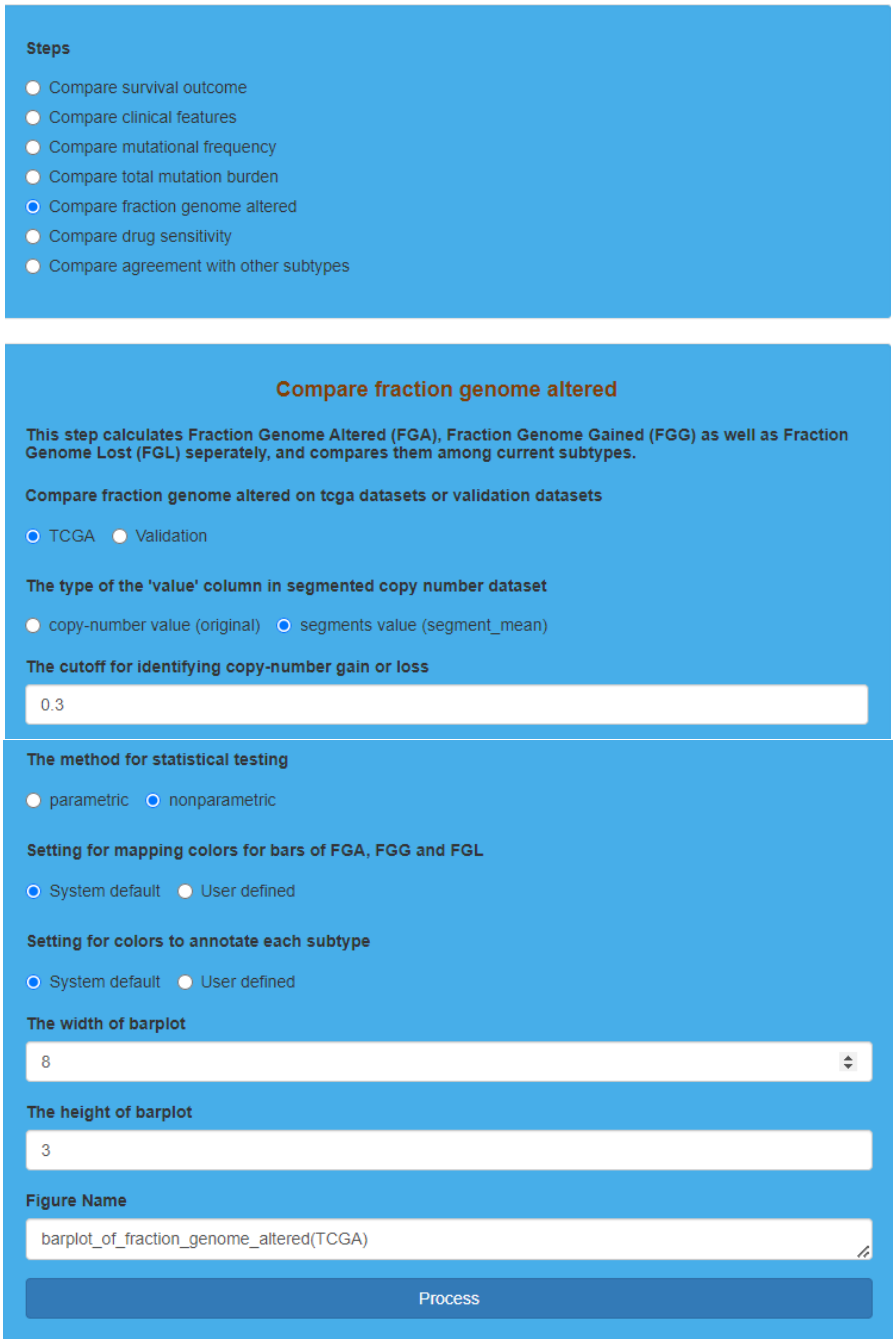
After setting all parameters, click the "Process" button to compare FGA, FGG, and FGL among the subtypes obtained from clustering. Once completed, feedback will be provided in the "FGA Results" module on the right side of the webpage. Users can choose to download the RData result set, the comparison bar plot in PDF format, and the table of FGA, FGG, and FGL calculations for each sample in each subtype (available in four formats:

Copy, CSV, Excel, and Print). Next, compare the response of each subtype to drug treatment in the GDSC database using IC_{50} values. Go to the "Steps" parameter module under the "COMP Module" section and select "Compare drug sensitivity." Refer to the guidance document for the "Compare drug sensitivity" parameter module for more details.

III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Compare fraction genome altered”, and “Compare fraction genome altered on tcga datasets or validation datasets” chooses “TCGA”. First, we choose “segments value (segment_mean)” for “The type of the 'value' column in segmented copy number dataset”, and then indicate 0.3 as “The cutoff for identifying copy-number gain or loss”. After that, we choose “nonparametric” for “The method for statistical testing” and we utilize default settings for “Setting for mapping colors for bars of FGA, FGG and FGL” (System default) and “Setting for colors to annotate each subtype” (System default). Finally, we indicate “The width of barplot” as well as “Figure Name” using default settings, and indicate 3 as “The height of barplot”. Then, we click the “Process” button to compare the fraction of genome altered (FGA), the fraction of genome gained (FGG) and the fraction of genome lost (FGL)

among each subtype derived from clustering results.



Parameter settings for “Compare fraction genome altered” on TCGA dataset

Result Display:

The software will generate bar plots to show the comparison results of FGA, FGG and FGL among subtypes, and the values of FGA, FGG, and FGL of each sample will be also displayed through a table. The sign of “*”

reveals significantly different on FGA, or FGG, or FGL among subtypes.

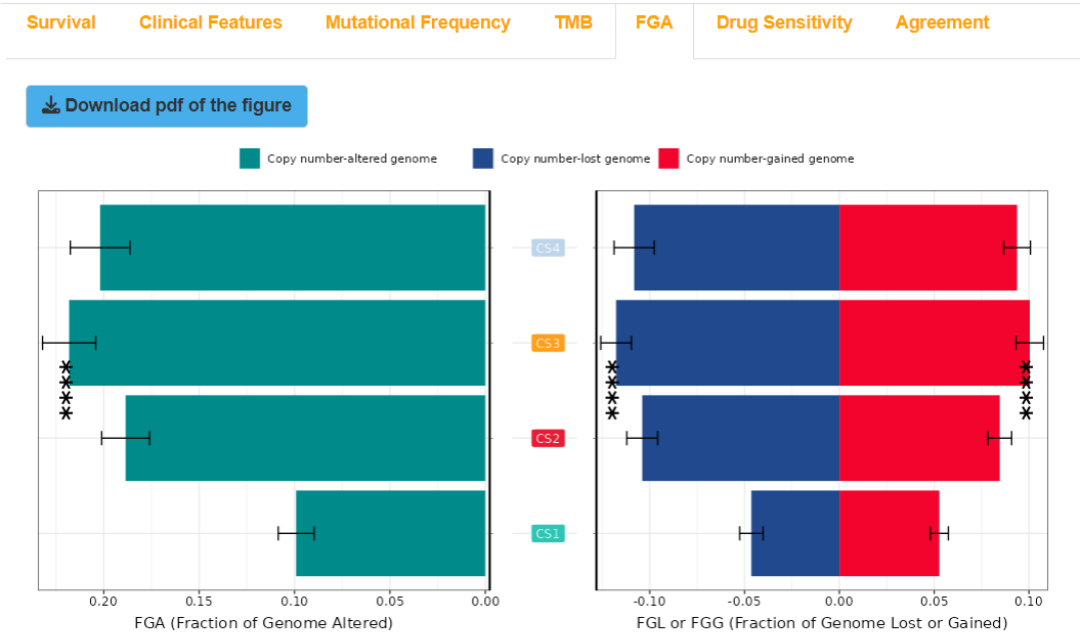


Table: Comparison of fraction genome altered among identified subtypes (TCGA)

Copy CSV Excel Print Show 10 entries Search:

samID	FGA	FGG	FGL	Subtype
TCGA-2F-A9KO-01A	0.0760313477721086	0.0629308952614183	0.0131004525106903	CS2
TCGA-2F-A9KP-01A	0.135278005558389	0.0631249554639437	0.0721530500944457	CS3
TCGA-2F-A9KQ-01A	0.183251543840721	0.039102419536416	0.144149124304305	CS3

Result display for “Compare fraction genome altered” on TCGA dataset

(6). Compare drug sensitivity for TCGA dataset

I.Analysis introduction

“Compare drug sensitivity” uses IC_{50} to compare the responses to drugs in GDSC database among subtypes, using box-violin plots for visualization. Besides, we listed estimated IC_{50} for each sample in the form of a table.

II.Parameters setting guides

Compare drug sensitivity

In addition, the platform also compares the response of each subtype to drug treatment in the GDSC database using IC_{50} values, i.e., conducting drug sensitivity comparison analysis. To perform this analysis, mRNA or lncRNA data needs to be prepared, and the specified parameters include:

Compare drug sensitivity on tcga datasets or validation datasets: Specify the dataset for drug sensitivity comparison, with the default choice being TCGA. If you choose Validation, you need to first complete the Run nearest template prediction analysis under the RUN Module section (select Validation for the parameter Run nearest template prediction on tcga or validation cohort) or the Run partition around medoids classifier analysis (select Validation for the parameter Run partition around medoids classifier on tcga or validation cohort) in the RUN Module.

Model-free approaches for subtype prediction in validation cohort: If the parameter "Compare drug sensitivity on tcga datasets or validation datasets" is set to "Validation"， users need to specify the non-model method for subtype prediction in the external validation queue. The default choice is NTP, which utilizes the subtype prediction results from the external validation set obtained through Run nearest template prediction analysis. Otherwise, you can choose PAM, which uses the subtype prediction results from the external validation set obtained

through Run partition around medoids classifier analysis.

The name of the drug from GDSC for which you would like to predict sensitivity: Enter the drugs from the GDSC database for which users want to perform drug sensitivity comparison. The default value is Cisplatin.

Train the models on only a subset of the CGP cell lines or not: Specify whether to train the model only on a subset of CGP cell lines. The default option is "No", indicating that training will be performed on the entire CGP cell line dataset.

The tissue type from which the cell lines originated: If the parameter "Train the models on only a subset of the CGP cell lines or not" is set to "Yes", the default value is "lung". The user needs to choose the appropriate tissue type based on the selected cancer type.

The method for statistical testing: Choose the statistical test method for drug sensitivity comparison. The default selection is nonparametric.

Setting for colors to annotate each subtype: Specify the colors for annotating the clusters obtained for each subtype. You can either choose the system's default colors (System default) or specify your own colors (User defined).

Input color for each subtype: If you choose "User defined" for the "Setting for colors to annotate each subtype" parameter, you need to input the specified colors in hexadecimal format, such as #2EC4B6.

The seed for reproducing the result of comparing drug sensitivity: Enter

the random seed used to reproduce the results of drug sensitivity comparison. The default value is 123456.

The width of boxviolin plot: Specify the width of the output boxplot-violin plot.

The height of boxviolin plot: Specify the height of the output boxplot-violin plot.

The prefix for the name of output boxviolin plot: Specify the prefix for the output boxplot-violin plot PDF name. If the "Compare drug sensitivity on TCGA datasets or Validation datasets" parameter is set to TCGA, the default value for this parameter is "boxviolin_of_estimated_IC50(TCGA)"; if the parameter is set to Validation, the default value is "boxviolin_of_estimated_IC50(Validation)".

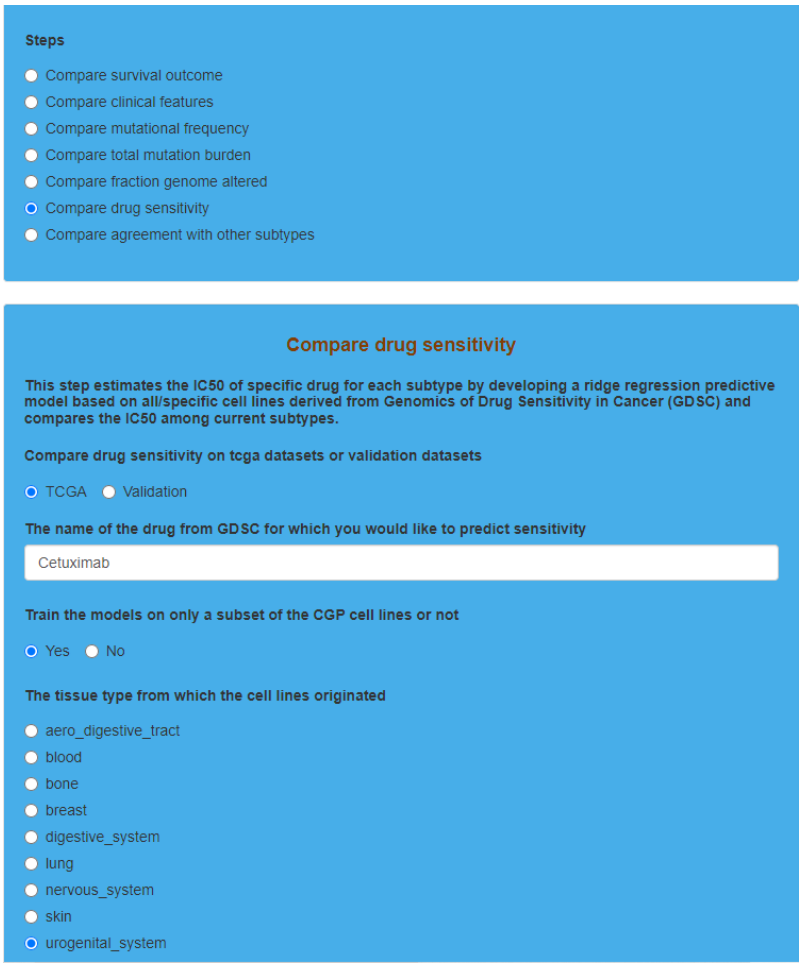
After setting all the parameters, click the "Process" button to perform the drug sensitivity comparison analysis among the subtypes obtained through clustering. Once completed, feedback will be provided in the "Drug Sensitivity" module on the right side of the webpage. Users can choose to download the RData result set, drug sensitivity comparison boxplot-violin plot PDF, and the IC50 calculation result table for each sample in each subtype (four formats available: Copy, CSV, Excel, Print). Finally, to assess the consistency between clustering results and traditional classification results, go to the "COMP Module" under the "Steps" parameter module and select "Compare agreement with other

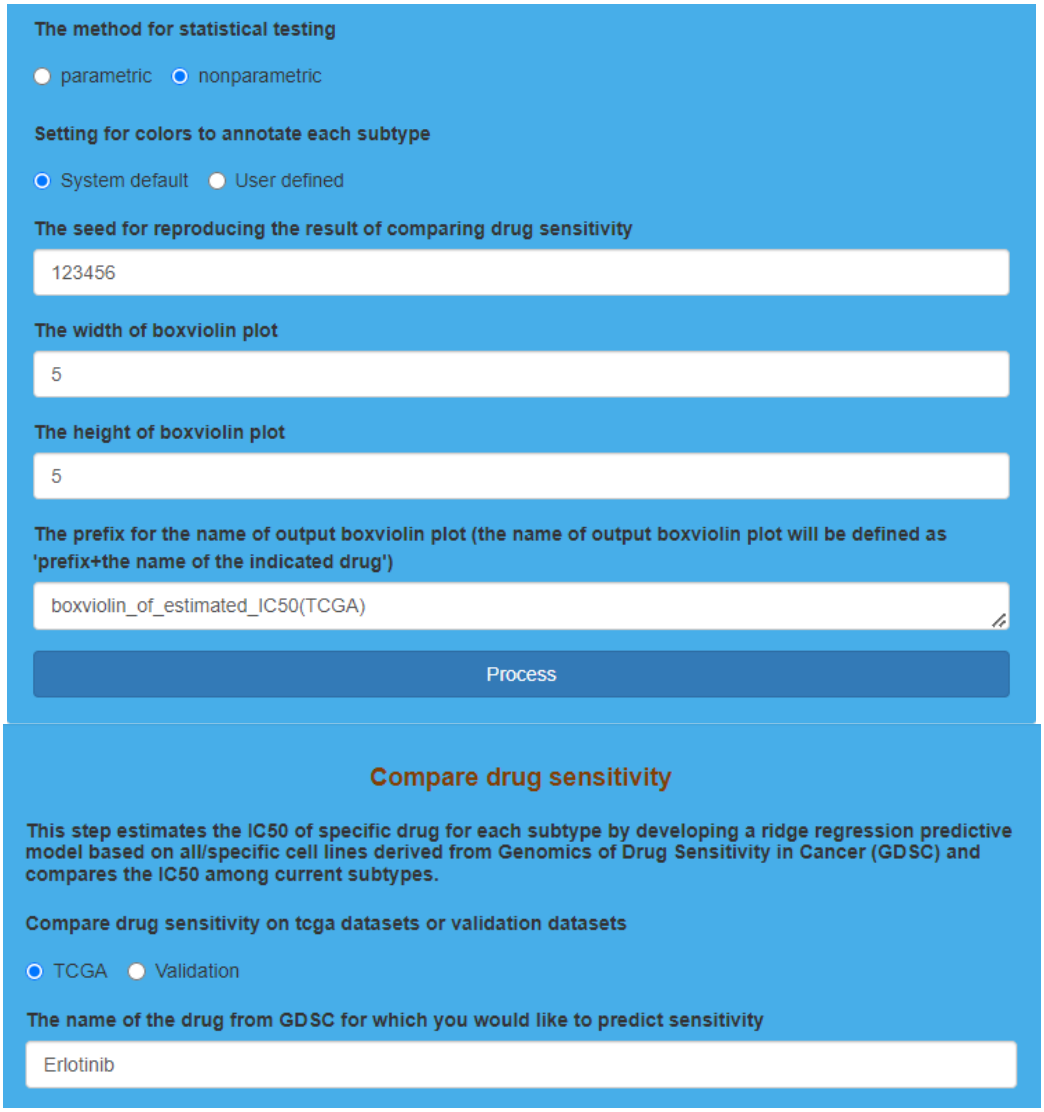
subtypes." Refer to the documentation for the "Compare agreement with other subtypes" parameter module for detailed instructions.

III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Compare drug sensitivity”, and “Compare drug sensitivity on tcga datasets or validation datasets” chooses “TCGA”. First, we indicate the name of the drugs from GDSC, here we indicate “Cetuximab” and “Erlotinib” separately. Then, we choose “Yes” for “Train the models on only a subset of the CGP cell lines or not”, and then choose “urogenital_system” for “The tissue type from which the cell lines originated”. After that, we choose “nonparametric” for “The method for statistical testing”. In addition, we use default settings for “Setting for colors to annotate each subtype” (System default), “The seed for reproducing the result of comparing drug sensitivity” (123456), “The width of boxviolin plot”, “The height of boxviolin plot” and “The prefix for the name of output boxviolin plot”. Finally, we click the “Process” button to compare drug sensitivity among subtypes derived from clustering

results.


Steps
Compare survival outcome
Compare clinical features
Compare mutational frequency
Compare total mutation burden
Compare fraction genome altered
Compare drug sensitivity
Compare agreement with other subtypes
Compare drug sensitivity
This step estimates the IC50 of specific drug for each subtype by developing a ridge regression predictive model based on all/specific cell lines derived from Genomics of Drug Sensitivity in Cancer (GDSC) and compares the IC50 among current subtypes.
Compare drug sensitivity on tcga datasets or validation datasets
TCGA
Validation
The name of the drug from GDSC for which you would like to predict sensitivity
Cetuximab
Train the models on only a subset of the CGP cell lines or not
Yes
No
The tissue type from which the cell lines originated
aero_digestive_tract
blood
bone
breast
digestive_system
lung
nervous_system
skin
urogenital_system



Parameter settings for “Compare drug sensitivity” on TCGA dataset

Result Display:

The software will generate box-violin plots to show the IC_{50} comparison results among subtypes for drugs, and the estimated IC_{50} of each sample for drugs will be also displayed through tables. “p<0.05” reveals significance of differences on the response to drugs among subtypes.

Survival | Clinical Features | Mutational Frequency | TMB | FGA | Drug Sensitivity | Agreement

Download pdf of the figure

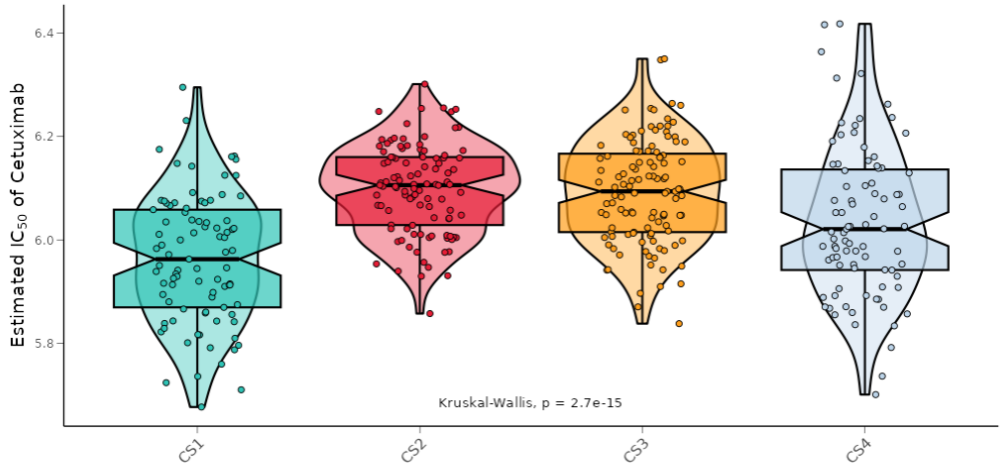


Table: Comparison of estimated IC50 (TCGA) for Cetuximab among identified subtypes

Copy | CSV | Excel | Print | Show 10 entries | Search:

	Est.IC50	Subtype
TCGA-HQ-A5NE-01A	5.73612825749444	CS1
TCGA-FD-A5BT-01A	6.23068306420836	CS1
TCGA-XF-AAMW-01A	5.79149025378978	CS1

Survival | Clinical Features | Mutational Frequency | TMB | FGA | Drug Sensitivity | Agreement

Download pdf of the figure

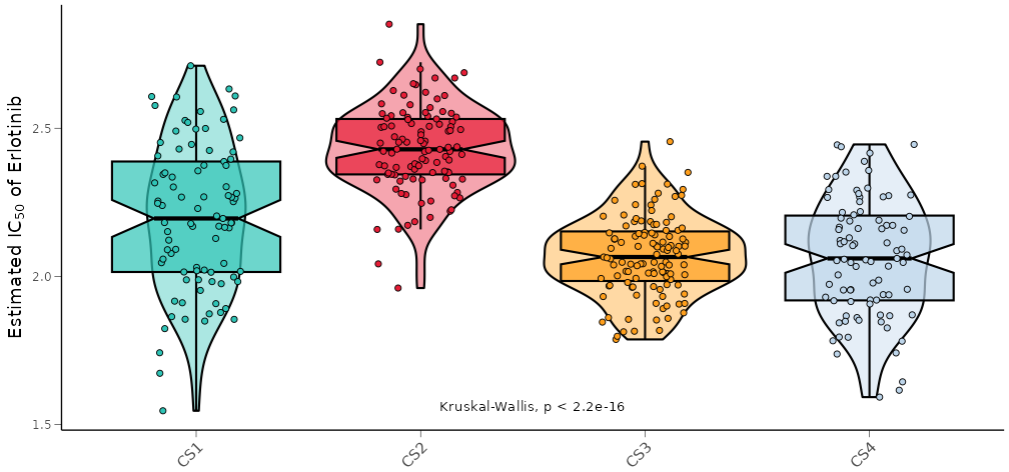


Table: Comparison of estimated IC50 (TCGA) for Erlotinib among identified subtypes

Copy | CSV | Excel | Print | Show 10 entries | Search:

	Est.IC50	Subtype
TCGA-HQ-A5NE-01A	2.03047625576851	CS1
TCGA-FD-A5BT-01A	2.52634061080072	CS1
TCGA-XF-AAMW-01A	1.87375678066546	CS1

Result display for “Compare drug sensitivity” on TCGA dataset

(7). Compare agreement with other subtypes for TCGA dataset

I.Analysis introduction

"Compare agreement with other subtypes" compares the consistency of the clustering results with the current classification results (e.g., PAM50, pstage), which can evaluate the clustering results. A bar chart of four evaluation indicators, containing Rand Index (RI), Adjusted Mutual Information (AMI), Jaccard Index (JI), and Fowlkes-Mallows (FM) as well as an alluvial diagram are utilized for visualization. In addition, the software also lists the values of four evaluation indicators above in the form of a table.

II.Parameters setting guides

Compare agreement with other subtypes

Finally, to assess the consistency between clustering results and traditional classification results, four statistical metrics—Rand Index (RI), Adjusted Mutual Information (AMI), Jaccard Index (JI), and Fowlkes-Mallows (FM)—are used to evaluate the quality of the clustering results. Clinical survival data must be prepared beforehand, and the specified parameters include:

Compare agreement with other subtypes on tcga datasets or validation datasets: Specify the dataset for consistency comparison with traditional classification results, with TCGA being the default selection. If choosing

Validation, it is necessary to first complete the analysis under the RUN Module module, selecting either Run nearest template prediction on tcga or validation cohort (choose Validation) or Run partition around medoids classifier on tcga or validation cohort (choose Validation).

Model-free approaches for subtype prediction in validation cohort: If the parameter "Compare agreement with other subtypes on TCGA datasets or Validation datasets" is set to "Validation", you need to specify the model-free method for subtype prediction in the external validation queue. The default choice is NTP, which uses the subtype prediction results obtained from the Run nearest template prediction analysis on the external validation cohort. Alternatively, you can choose PAM, which uses the subtype prediction results obtained from the Run partition around medoids classifier analysis on the external validation cohort.

The number of the traditional subtypes for comparison: Specify the number of traditional classification variables for consistency comparison, ranging from 1 to 6, with a default value of 1.

Input the variable name of traditional subtypes in survival and clinical information for comparison: Enter the traditional classification variables for consistency comparison in order. The names entered must come from clinical variables in the dataset, such as ajcc_pathologic_tumor_stage.

Setting for colors to annotate each subtype: Specify the colors for annotating each subtype obtained from clustering. You can choose the

system's default colors (System default, the default choice) or define your own colors (User defined).

Input color for each subtype: If the parameter "Setting for colors to annotate each subtype" is chosen as "User defined", the user needs to input the specified colors one by one in hexadecimal format, for example, #2EC4B6.

The width for box in alluvial diagram: Specify the width of the boxes in the impact plot; the default value is 0.1.

The width of the figure: Specify the width of the output image in PDF format; the default value is 6.

The height of the figure: Specify the height of the output image in PDF format; the default value is 5.

Figure Name: Specify the name of the output PDF image. If the "Compare agreement with other subtypes on TCGA datasets or validation datasets" parameter is set to TCGA, the default value for this parameter is "agreement_between_current_subtype_and_other_classifications (TCGA)"; if it is set to Validation, the default value is "agreement_between_current_subtype_and_other_classifications (Validation)".

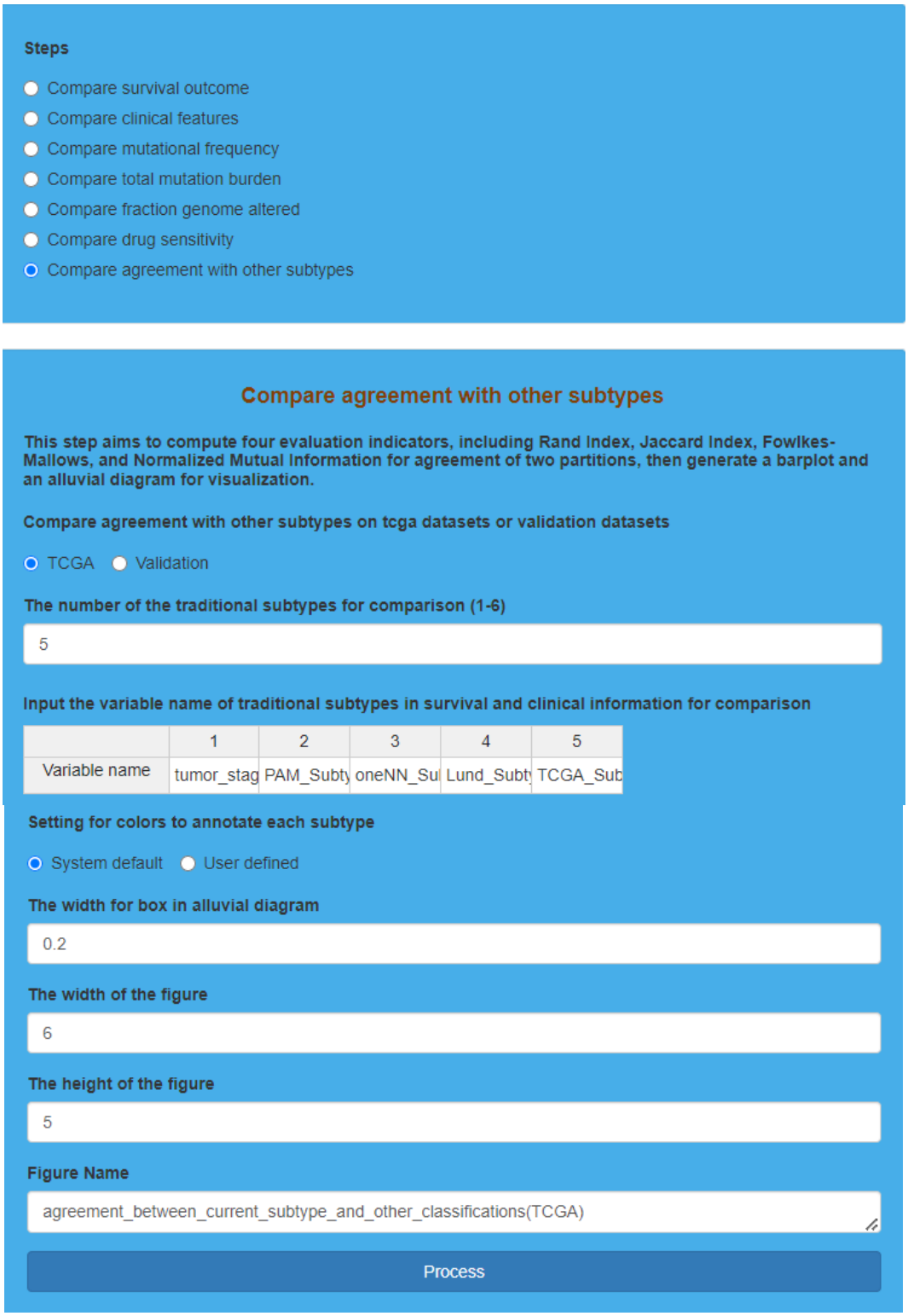
After setting all the parameters, click the "Process" button to compare the consistency between the clustering results and traditional classification

results. Once completed, the feedback will be provided in the "Agreement" results module on the right side of the webpage. Users can choose to download the RData result set, the PDF image (consisting of bar charts and impact plots for the four consistency metrics), and the table of results for the four consistency metrics (in four formats: Copy, CSV, Excel, Print). At this point, all analyses in the COMP Module have been completed. Next, click on the "RUN Module" button at the top of the webpage to proceed with downstream analyses for the clustered subtypes (refer to the guidance document for the "Steps" parameter module under the RUN Module).

III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Compare agreement with other subtypes”, and “Compare agreement with other subtypes on tcga datasets or validation datasets” chooses “TCGA”. We first indicate 5 as “The number of the traditional subtypes for comparison”, and then enter the names of the traditional subtypes containing “tumor_stage”, ” PAM_Subtype”, “oneNN_Subtype”, “Lund_Subtype”, and “TCGA_Subtype” (“Input the variable name of traditional subtypes in survival and clinical information for comparison”). Besides, we indicate 0.2 as “The width for box in alluvial diagram”. Additionally, “Setting for colors to annotate each subtype”, “The width of the figure”, “The height of the figure”, and “Figure Name” use default settings. Finally, we click the “Process” button to

compare agreement between subtypes derived from clustering results and other traditional subtypes.



Parameter settings for “Compare agreement with other subtypes” on TCGA dataset

Result Display:

The software will generate an alluvial diagram to compare agreement

between subtypes derived from clustering results and other traditional subtypes. In addition, a bar plot and a table are also generated to display the values of four statistical indicators which are utilized to evaluate the agreement, including Rand Index (RI), Adjusted Mutual Information (AMI), Jaccard Index (JI), and Fowlkes-Mallows (FM). All these indicators range from 0 to 1, and the larger the values are, the clustering results are more similar to the current classification results.

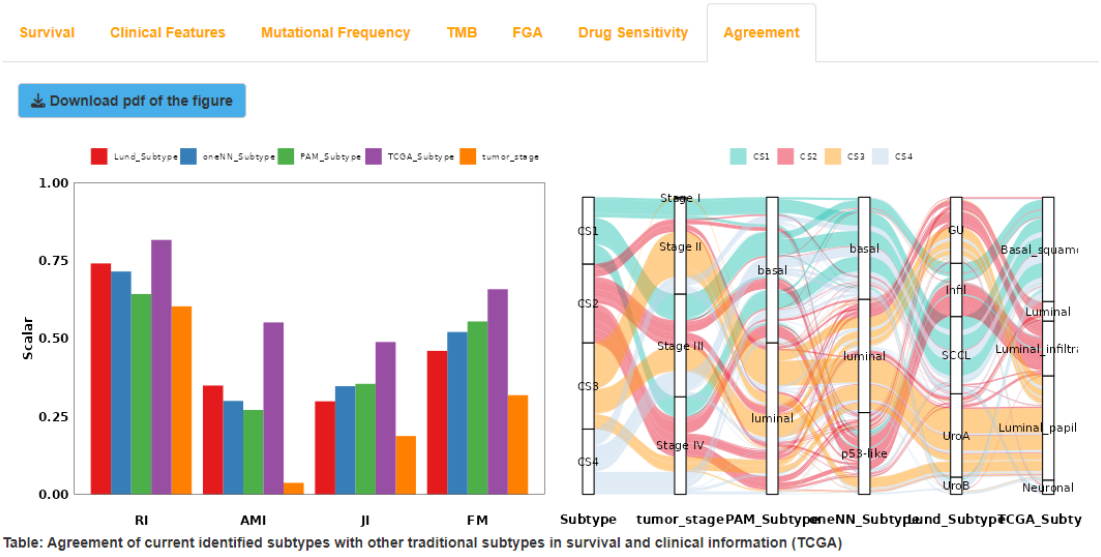


Table: Agreement of current identified subtypes with other traditional subtypes in survival and clinical information (TCGA)

Copy CSV Excel Print

current.subtype	other.subtype	RI	AMI	JI	FM
Subtype	tumor_stage	0.602518859480885	0.036837126675318	0.18686406319479	0.317705899909264
Subtype	PAM_Subtype	0.642117376294592	0.270395597187127	0.354206081860551	0.55409718215014
Subtype	oneNN_Subtype	0.714243702851298	0.299437633971342	0.346635093258493	0.520382446628201
Subtype	Lund_Subtype	0.740109960363125	0.34869068756393	0.298159593936673	0.459854686039391
Subtype	TCGA_Subtype	0.815445595192431	0.551017844716095	0.488264908175566	0.657448754910777

Result display for “Compare agreement with other subtypes” on TCGA dataset

The fourth step: RUN Module

“RUN Module” performs downstream analyses among different subtypes obtained from clustering results and predicts subtype for external

validation datasets through model-free methods, which is also divided into seven sub-modules, including “Run differential expression analysis”, “Run biomarker identification procedure”, “Run gene set enrichment analysis”, “Run gene set variation analysis”, “Run nearest template prediction”, “Run partition around medoids classifier” and “Run consistency evaluation using Kappa statistics”.

(1). Run differential expression analysis for TCGA dataset

I.Analysis introduction

“Run differential expression analysis” provides three types of algorithms including DESeq2, edgeR, and limma to find out differentially expressed genes for each subtype, which are displayed in the form of a table.

II.Parameters setting guides

i.RUN Module

After completing the comparative analysis of the clustered subtypes on the TCGA dataset, we will proceed with downstream analyses for these subtypes on the same dataset. Additionally, we will use a model-free method to predict the subtypes of an external validation set and perform Kappa consistency comparison. To do this, mRNA or lncRNA data must be prepared. The following steps will be performed:

1. Differential Expression Analysis: Run a differential expression analysis among the clustered subtypes.

2. Biomarker Identification: Based on the results of the differential

expression analysis, run a procedure to identify biomarkers, filtering for upregulated or downregulated genes in each subtype.

3. Gene Set Enrichment Analysis (GSEA): Perform GSEA to analyze the enrichment of gene sets, identifying upregulated or downregulated pathways in each subtype.

4. Gene Set Variation Analysis (GSVA): Conduct GSVA to analyze gene set variation, calculating enrichment scores for each sample within each subtype on specified pathways.

5. External Validation Set Prediction: Use the model-free method NTP to predict the subtypes of samples in an external validation set based on the identified biomarker genes.

6. Alternative Model-Free Prediction: Additionally, use another model-free method, PAM, to predict the subtypes of samples in the external validation set.

7. Evaluation of Consistency: Evaluate the consistency between the predicted subtypes and the clustered subtypes using similarity and reproducibility metrics like IGP.

8. Kappa Consistency Evaluation: Finally, assess the consistency between the clustered subtypes and the predicted subtypes or between predictions using both NTP and PAM, utilizing the Kappa statistic.

To start, perform differential expression analysis among the subtypes. Navigate to the "RUN Module" and choose "Run differential expression

analysis" under the "Steps" parameter module. Refer to the documentation for the specific parameter settings in the "Run differential expression analysis" parameter module for detailed instructions on configuration.

ii.Run differential expression analysis

First, conduct differential expression analysis among subtypes in the TCGA dataset, specifying the following parameters:

Choose the algorithm for differential expression analysis: Specify the algorithm for differential expression analysis. The platform provides three differential expression analysis algorithms: DESeq2, edgeR, and limma. Users can choose the appropriate algorithm based on their specific situation. The default selection is DESeq2.。

Indicate the prefix of output file: Specify the prefix for the output file of the differential expression analysis. The default is empty, meaning no specified prefix.

The result already exists, overwrite it or skip this step directly: If the differential expression analysis result file already exists, specify whether to overwrite it. The default is to choose "Overwrite" to redo the differential expression analysis and replace the existing results. Alternatively, you can choose "Skip" to skip the differential expression analysis if the result file already exists.

Whether sort adjusted p value in ascending order for output table:

Whether to arrange the differential expression analysis results in ascending order based on adjusted p-values. The default option is "Yes".

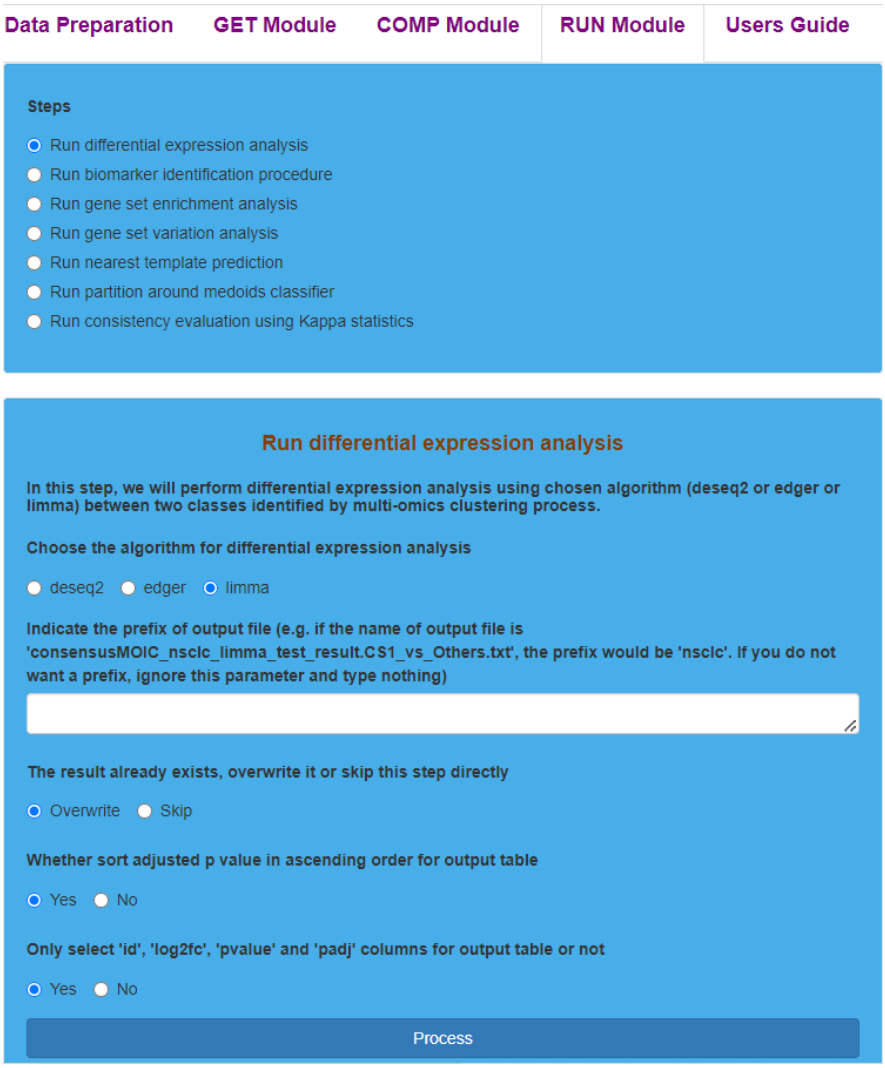
Only select 'id', 'log2fc', 'pvalue' and 'padj' columns for output table or not: Whether to display only the 'id', 'log2fc', 'pvalue', and 'padj' columns in the differential expression analysis results. The default option is "Yes."

After setting all the parameters, click the "Process" button to perform differential expression analysis among the subtypes obtained through clustering. After completion, feedback will be provided in the DEA (Differential Expression Analysis) results module on the right side of the webpage. Users can choose to download the RData result set and the table of differential expression analysis results in four formats (Copy: click to copy results; CSV: click to download CSV result file; Excel: click to download Excel result file; Print: click to go to the PDF browsing page, where you can download the PDF result file). Next, further filtering of upregulated or downregulated marker genes for each subtype will be set based on the results of the differential expression analysis. Go to the "Steps" parameter module under the "Run biomarker identification procedure" in the "RUN Module" section, as detailed in the guidance document for the "Run biomarker identification procedure" parameter module.

III.Examples with results interpretation

Parameter Settings: “Module switching options” on the upper left of the

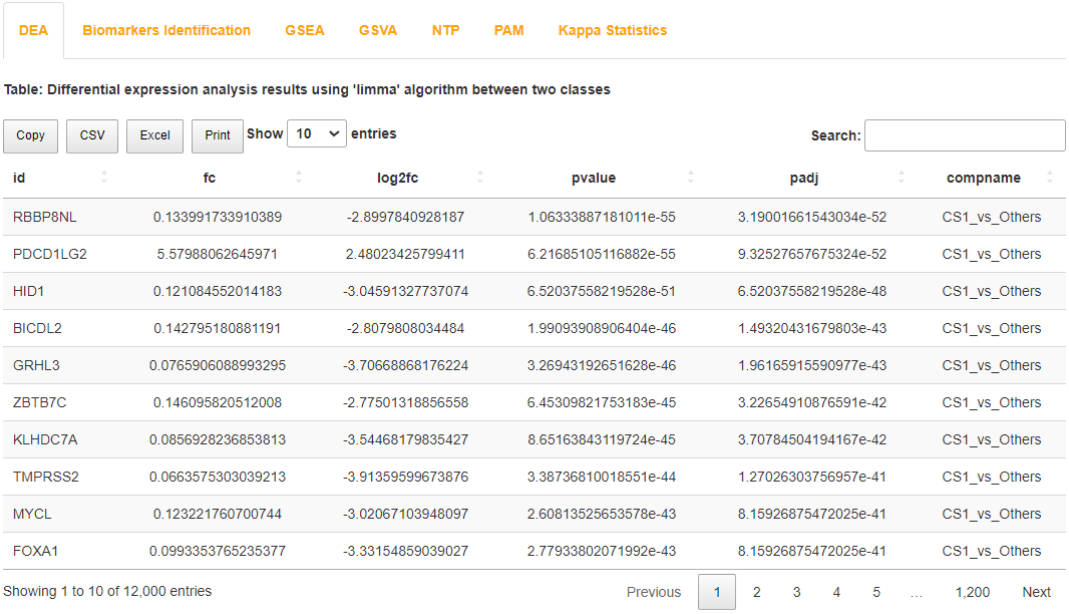
software chooses “RUN Module”, and then the “Steps” chooses “Run differential expression analysis”. First, we choose “limma” as “Choose the algorithm for differential expression analysis”, and then we indicate the prefix of the output “.txt” file (we ignore “Indicate the prefix of output file” and type nothing). Additionally, “overwrite it or skip this step directly” chooses “Overwrite”, “Whether sort adjusted p value in ascending order for output table” chooses “Yes” and “Only select 'id', 'log2fc', 'pvalue' and 'padj' columns for output table or not” chooses “Yes”. Finally, we click the “Process” button to run differential expression analysis.



Parameter settings for “Run differential expression analysis” on TCGA dataset

Result Display:

The software will generate a table to display the results of differential expression analysis.



DEA | Biomarkers Identification | GSEA | GSVA | NTP | PAM | Kappa Statistics

Table: Differential expression analysis results using 'limma' algorithm between two classes

Copy | CSV | Excel | Print | Show 10 entries | Search:

id	fc	log2fc	pvalue	padj	compname
RBBP8NL	0.133991733910389	-2.8997840928187	1.06333887181011e-55	3.19001661543034e-52	CS1_vs_Others
PDCD1LG2	5.57988062645971	2.48023425799411	6.21685105116882e-55	9.32527657675324e-52	CS1_vs_Others
HID1	0.121084552014183	-3.04591327737074	6.52037558219528e-51	6.52037558219528e-48	CS1_vs_Others
BICDL2	0.142795180881191	-2.8079808034484	1.99093908906404e-46	1.49320431679803e-43	CS1_vs_Others
GRHL3	0.0765906088993295	-3.70668868176224	3.26943192651628e-46	1.96165915590977e-43	CS1_vs_Others
ZBTB7C	0.146095820512008	-2.77501318856558	6.45309821753183e-45	3.22654910876591e-42	CS1_vs_Others
KLHDC7A	0.0856928236853813	-3.54468179835427	8.65163843119724e-45	3.70784504194167e-42	CS1_vs_Others
TMPRSS2	0.0663575303039213	-3.91359599673876	3.38736810018551e-44	1.27026303756957e-41	CS1_vs_Others
MYCL	0.123221760700744	-3.02067103948097	2.60813525653578e-43	8.15926875472025e-41	CS1_vs_Others
FOXA1	0.0993353765235377	-3.33154859039027	2.77933802071992e-43	8.15926875472025e-41	CS1_vs_Others

Showing 1 to 10 of 12,000 entries | Previous | 1 | 2 | 3 | 4 | 5 | ... | 1,200 | Next

Result display for “Run differential expression analysis” on TCGA dataset

(2). Run biomarker identification procedure for TCGA dataset

I.Analysis introduction

“Run biomarker identification procedure” sets conditions to further screen out up-regulated and down-regulated marker genes separately for each subtype, which are displayed in the form of heatmaps and tables.

II.Parameters setting guides

Run biomarker identification procedure

On the basis of the differential expression analysis, further filtering is performed to select upregulated or downregulated marker genes for each subtype. These filtered upregulated or downregulated marker genes will

serve as prediction templates for subtype prediction using the Nearest Template Prediction (NTP) method in the subsequent steps. The parameters to be specified include:

Specify the criteria for filtering marker genes based on the results of the differential expression analysis:

Indicate the algorithm for completed differential expression analysis: Specify the algorithm used for the previous differential expression analysis, including three options: DESeq2, edgeR, and limma. Note that the results of the corresponding differential expression analysis for this algorithm must exist.

Indicate the nominal p value for identifying significant markers: Specify the nominal p-value threshold for filtering marker genes from the differential expression analysis results, with a default value of 0.05.

Indicate the adjusted p value for identifying significant markers: Specify the adjusted p-value threshold for filtering marker genes from the differential expression analysis results, with a default value of 0.05.

Indicate the direction of identifying significant marker: Specify whether to filter up-regulated or down-regulated marker genes, with the default choice being up-regulated marker genes. Based on the previous step's differential expression analysis results, genes with $\log_2FC>1$ are considered up-regulated, and those with $\log_2FC<1$ are considered down-regulated.

Indicate the number of top markers sorted by log2fc should be identified for each subtype: For the differential expression analysis results of each subtype, sort them in descending order based on the absolute values of log2FC, and select the top n results for each subtype. The default value for n is 200.

Using the specified differential expression analysis results, select the top n up-regulated or down-regulated marker genes with larger absolute log2FC based on the nominal p-value threshold and adjusted p-value threshold.

Then, specify the parameters for drawing the expression heatmap of marker genes:

Sample annotations from survival information for heatmap or not: Specify whether to annotate the heatmap at the top with clinical survival information. Choose "No" for not using it by default or "Yes" to use it.

The number of sample annotations from survival information for heatmap: If the "Sample annotations from survival information for heatmap or not" parameter is set to "Yes", specify the number of clinical survival variables to annotate. The default value is 3.

Input sample annotations from survival information for heatmap: Then, you need to specify the detailed information for clinical survival variables. In the first row, accurately enter the variable names that need annotation from the clinical survival dataset. In the second row, enter the data types

corresponding to each clinical survival variable. In the third row, enter the annotation colors (in hexadecimal format) for each clinical survival variable, for example, #000004FF, using semicolons to separate multiple colors. If it is a continuous variable, the number of colors should be 3, corresponding to the minimum value, median, and maximum value of that variable. If it is a discrete variable, the number of colors should match the number of categories for that variable.

Colors for annotating each cluster at the top of heatmap: Specify the colors for annotating the clusters obtained by clustering at the top of the heatmap. You can choose the system's default colors (System default, selected by default) or specify your own colors (User defined).

Input colors for annotating each cluster: If you choose "User defined" for the "Colors for annotating each cluster at the top of the heatmap" parameter, you need to enter the specified colors in hexadecimal format, such as #2EC4B6.

Indicate if expression data should be centered: Whether to perform centering on mRNA or lncRNA expression data, i.e., subtracting the mean of their respective features. The default option is Yes to perform centering.

Indicate if expression data should be scaled: Whether to perform scaling on mRNA or lncRNA expression data, i.e., dividing by the standard deviation of their respective features. The default option is Yes to perform scaling.

Assign marginal cutoff for truncating values in data or not: Whether to specify the plotting cutoff value for mRNA or lncRNA expression data. The default option is Yes to specify a cutoff value.

Marginal cutoff for truncating values in data: If the parameter "Assign marginal cutoff for truncating values in data" is set to Yes, you need to specify the plotting cutoff value for mRNA or lncRNA expression data. The cutoff value is a positive number, with a default of 3. If the data is greater than this cutoff value, it will be replaced by the cutoff value; if it is less than the negative of the cutoff value, it will be replaced by the negative of the cutoff value.

Show rownames (feature names) in heatmap or not: Whether to display the names of the marker genes for each subtype in the rows of the heatmap, default choice is No for not displaying.

Show colnames (sample ID) in heatmap or not: Whether to display the IDs of each sample in the columns of the heatmap, default choice is No for not displaying.

Colors for heatmap: Specify the colors for drawing the heatmap; you can choose the system's default colors (System default, the default choice) or specify your own colors (User defined).

The number of colors for heatmap: If the "Colors for heatmap" parameter is set to "User defined", the user needs to input the number of colors for drawing the heatmap.

Input colors for heatmap: If the "Colors for heatmap" parameter is set to "User defined", the user needs to enter the specified colors in hexadecimal format, such as #00FF00.

The width of output figure: Specify the width for the output PDF of the heatmap of marker gene expression, with a default value of 8.

The height of output figure: Specify the height for the output PDF of the heatmap of marker gene expression, with a default value of 8.

Indicate the prefix of output figure: Specify the prefix for the output PDF of the heatmap of marker gene expression, with a default value of an empty string (no prefix) and a system default prefix of "markerheatmap".

After setting all the parameters, click the "Process" button to filter upregulated or downregulated marker genes for each subtype. Once completed, feedback will be provided in the Biomarkers Identification results module on the right side of the webpage. Users can choose to download the RData result set, the PDF of the heatmap of marker gene expression, and the table of selected marker genes for each subtype in various formats (Copy, CSV, Excel, Print). The next step involves Gene Set Enrichment Analysis (GSEA) to identify enriched pathways for upregulated or downregulated genes in each subtype. Navigate to the "RUN Module" and select the "Steps" parameter module. Refer to the documentation for "Prepare a gene set background file" and "Run gene set enrichment

analysis" for detailed instructions.

III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Run biomarker identification procedure”. First, we choose the algorithm (limma) that we use for differential expression analysis. Then, we indicate 0.05 for both the “nominal *p* value” and “adjusted *p* value” to identify significant markers for each subtype derived from clustering results. After that, we choose “up-regulated” and “down-regulated” for “Indicate the direction of identifying significant marker” in turn to screen significant up-regulated and down-regulated markers for each subtype respectively. Furthermore, we indicate 30 and 50 as “the number of top markers sorted by log2fc should be identified for each subtype” for up-regulated and down-regulated markers respectively. For the processing of expression data, we normalize the data by default (both “Indicate if expression data should be centered” and “Indicate if expression data should be scaled” choose “Yes”). After that, we choose “Yes” for “Assign marginal cutoff for truncating values in data or not” and indicate 3 as “Marginal cutoff for truncating values in data”. For the settings of the heatmap, “Colors for annotating each cluster at the top of heatmap” (System default), “Show rownames (feature names) in heatmap or not” (No), “Show colnames (sample ID) in heatmap or not” (No), and “Colors for heatmap” (System default) use default settings. We indicate 12 for “The width of output figure” and “The

height of output figure". In addition to annotation for sample through clustering results, this step also allows users to select specific clinical features for annotation, and we select "PAM_Subtype", "oneNN_Subtype", "Lund_Subtype" and "TCGA_Subtype". Finally, we indicate the prefix of the output figure (we ignore "Indicate the prefix of output figure" and type nothing), and then click the "Process" button to run biomarker identification procedure.

Steps

- ○ Run differential expression analysis
- ◉ Run biomarker identification procedure
- ○ Run gene set enrichment analysis
- ○ Run gene set variation analysis
- ○ Run nearest template prediction
- ○ Run partition around medoids classifier
- ○ Run consistency evaluation using Kappa statistics

Run biomarker identification procedure

This step aims to identify uniquely and significantly expressed (overexpressed or downexpressed) biomarkers for each subtype identified by multi-omics clustering process. A template including top markers will be genearated for subtype external verification and a heatmap will also be generated.

Indicate the algorithm for completed differential expression analysis

○ deseq2 ○ edger ◉ limma

Indicate the nominal p value for identifying significant markers

0.05

Indicate the adjusted p value for identifying significant markers

0.05

Indicate the direction of identifying significant marker

◉ up-regulated ○ down-regulated

Indicate the number of top markers sorted by log2fc should be identified for each subtype

30

Sample annotations from survival information for heatmap or not

◉ Yes ○ No

The number of sample annotations from survival information for heatmap

4

Input sample annotations from survival information for heatmap

First line: Please input the sample annotation variables from survival information

Second line: Please input 'Continuous' or 'Categorical' to indicate the type of each sample annotation variable

Last line: Please input the colors for each sample annotation variable (use hex color format, e.g. #000004FF and English semicolons should be used to separate the input colors)

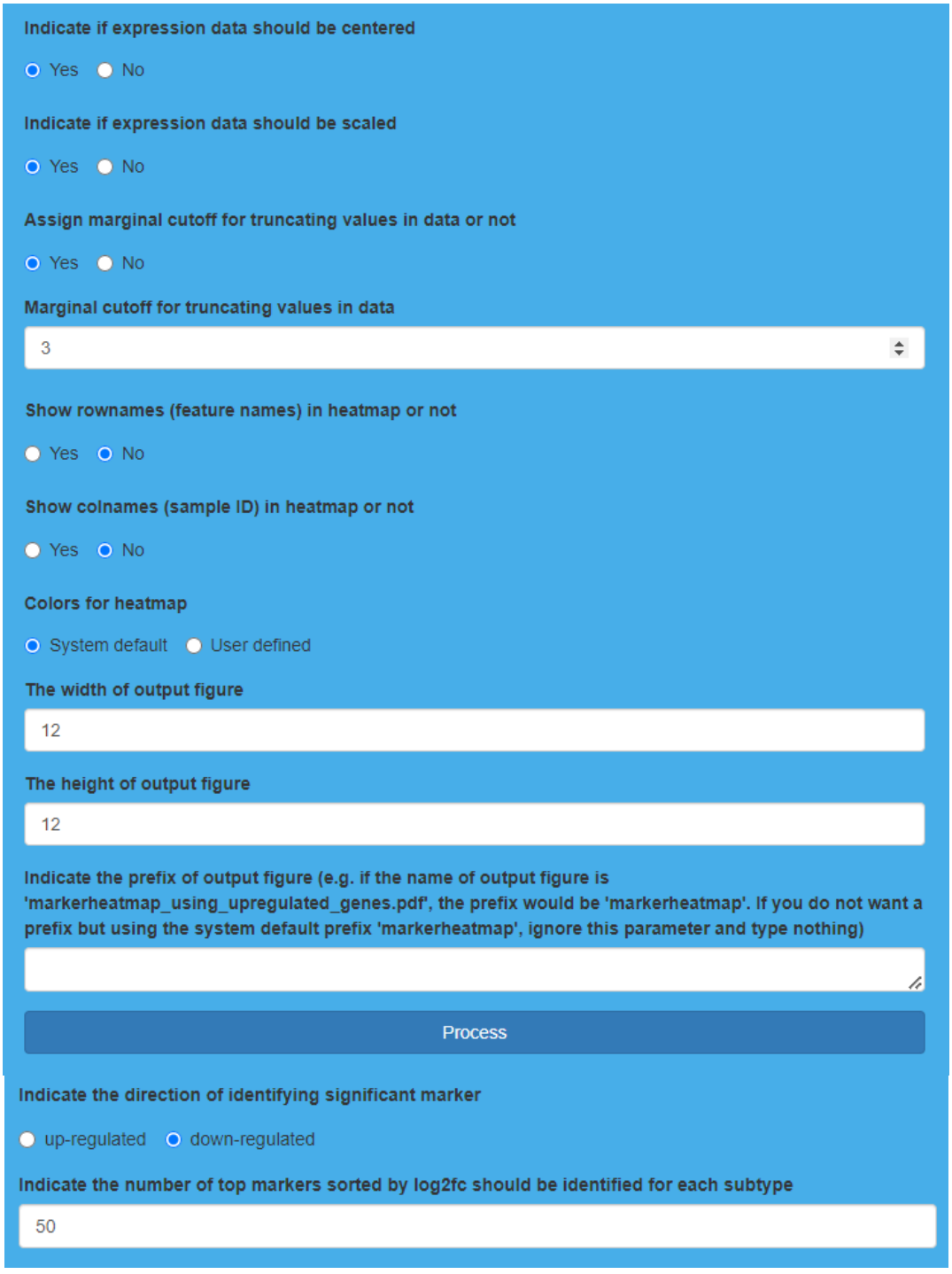
Note1: If the sample annotation variable is continuous, the number of indicated colors should be equal to 3, which represents the minimum, median and maximum value of this variable)

Note2: If the sample annotation variable is categorical, the number of indicated colors should be equal to the number of categories for this variable)

	1	2	3	4
Sample annotation	PAM_Subty	oneNN_Sub	Lund_Subty	TCGA_Subt
Continuous or Categorical	Categorical	Categorical	Categorical	Categorical
Color settings	#2874C5;#E	#2874C5;#C	#EABF00;#8	#2874C5;#8

Colors for annotating each cluster at the top of heatmap

◉ System default ○ User defined



Parameter settings for “Run biomarker identification procedure” on TCGA dataset

Result Display:

The software will generate heatmaps to show the expression of the screening up-regulated and down-regulated markers in each subtype (nominal *p* value < 0.05 & adjusted *p* value < 0.05). In addition, the

screening markers for each subtype are also displayed through tables.

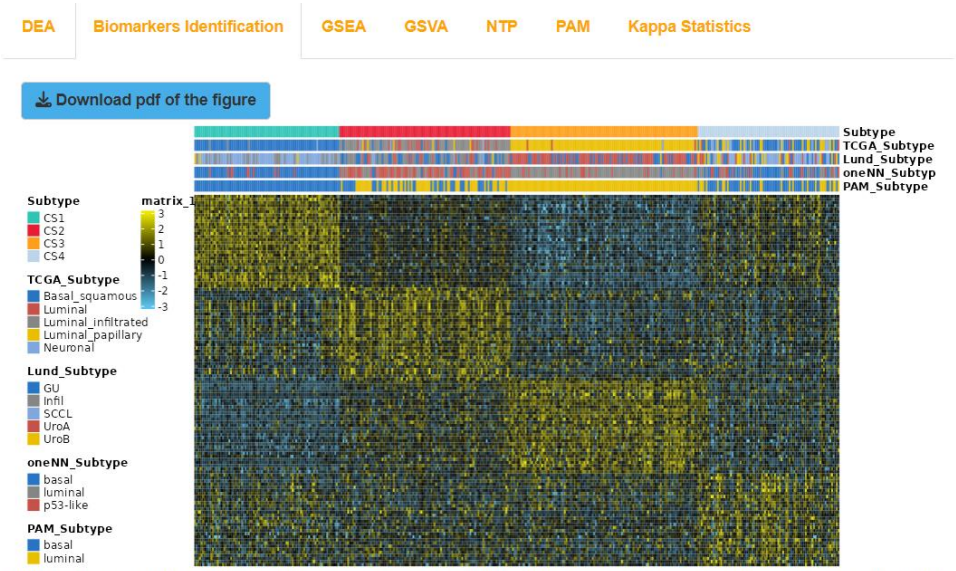


Table: Subtype-specific upregulated biomarkers for each identified subtype based on differential expression analysis results using 'limma' algorithm

Copy CSV Excel Print Show 10 entries Search:

probe	class	dirct
SAA1	CS1	up
MYOSLID	CS1	up

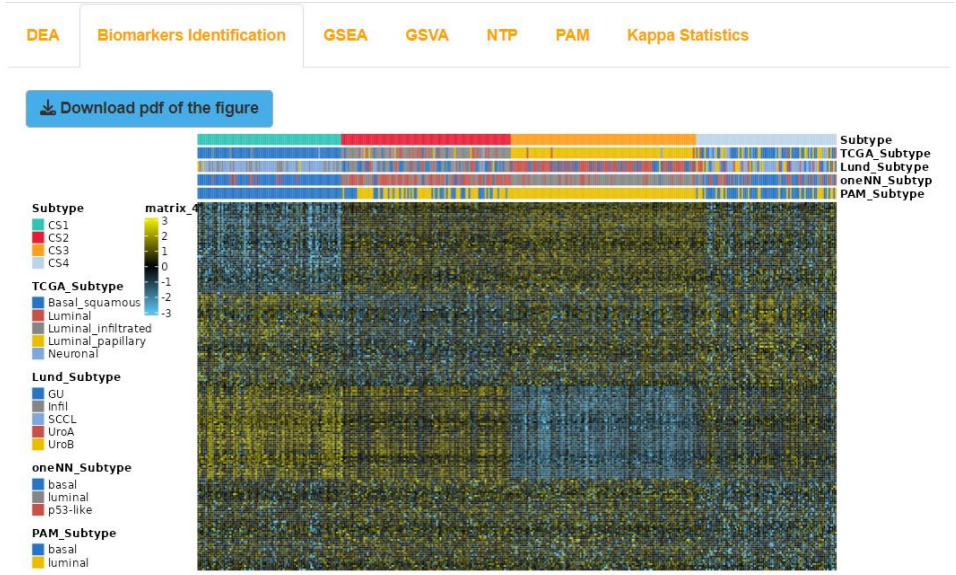


Table: Subtype-specific downregulated biomarkers for each identified subtype based on differential expression analysis results using 'limma' algorithm

Copy CSV Excel Print Show 10 entries Search:

probe	class	dirct
GPX2	CS1	down
UPK1B	CS1	down

Result display for “Run biomarker identification procedure” on TCGA dataset

(3). Run gene set enrichment analysis for TCGA dataset

I.Analysis introduction

“Run gene set enrichment analysis” screens out up-regulated and down-regulated pathways respectively for each subtype based on the given gene set background files, and then shows the information of these pathways through a table. Additionally, the software also calculates enrichment scores of the screened pathways in each subtype, which are displayed in the form of heatmaps and tables.

II.Parameters setting guides

i.Prepare a gene set background file

The next step is to perform Gene Set Enrichment Analysis (GSEA) to identify enriched pathways for each subtype. To begin, you need to prepare a gene set background file:

The manner for gene set background file preparation: Specify the method for preparing the gene set background file. The platform offers three preparation methods: System default (default choice) using the built-in gene set background file, Download from specified URL to obtain the gene set background file from a specified link, and Upload manually to upload the gene set background file from your local device.

Download url for gene set background file: If the parameter "The manner for gene set background file preparation" is set to "Download from

specified URL", the user needs to specify the download link for the gene set background file.

Upload the gene set background file: If the parameter "The manner for gene set background file preparation" is set to "Upload manually", the user needs to click the "Browse" button to select the corresponding gene set background file from their local device.

After specifying the gene set background file, click the "Process" button to prepare and process the gene set background file. Once completed, feedback will be provided in the GSEA Results module on the right side of the webpage. Next, continue to specify the other parameters for gene set enrichment analysis (GSEA). For detailed guidance, refer to the documentation for the "Run gene set enrichment analysis" parameter module.

ii.Run gene set enrichment analysis

The next step involves specifying additional parameters for gene set enrichment analysis (GSEA), including:

First, specify the parameters and result filtering criteria for gene set enrichment analysis:

Indicate the algorithm for completed differential expression analysis: Specify the algorithm used in the previous differential expression analysis, including deseq2, edger, and limma. Note that the results of the differential expression analysis corresponding to this algorithm must exist.

Indicate the direction of identifying significant pathway: Filter for significantly up-regulated or down-regulated pathways in each subtype, with the default choice being up-regulated pathways. Based on the gene set enrichment analysis results, pathways are considered significantly up-regulated if the Normalized Enrichment Score (NES) > 0, and significantly down-regulated if NES < 0.

Indicate the number of top pathways sorted by NES should be identified for each subtype: Sort the gene set enrichment analysis results for each subtype based on the absolute value of the Normalized Enrichment Score (NES) in descending order. Select the top n pathways for each subtype, with the default value of n being 10.

Indicate the number of permutations for gene set enrichment analysis: Specify the number of permutations for gene set enrichment analysis, with the default value being 1000. Setting this parameter to 10000 may enhance result reproducibility.

Indicate minimal size of each gene set for analysis: Specify the minimum gene set size for gene set enrichment analysis, with the default value being 10. This means that pathways with at least 10 genes will be selected for gene set enrichment analysis.

Indicate maximal size of each gene set for analysis: Specify the maximum gene set size for gene set enrichment analysis, with the default value being 500. This means that pathways with up to 500 genes will be selected for

gene set enrichment analysis.

Indicate the nominal p value for identifying significant pathways: Specify the nominal p-value threshold for filtering significant pathways from the gene set enrichment analysis results, with the default value being 0.05.

Indicate the adjusted p value for identifying significant markers: Specify the adjusted p-value threshold for filtering significant pathways from the gene set enrichment analysis results, with the default value being 0.25.

Specify the method for calculating enrichment scores:

Indicate the method to employ in the estimation of gene set enrichment scores per sample: Specify the method used for estimating the gene set enrichment scores for the selected pathways in each sample. The options include gsva, ssgsea, zscore, and plage, with the default selection being gsva.

Indicate the method to calculate subtype-specific pathway enrichment scores: Specify the method for calculating the enrichment score of the selected pathways across different subtypes. The options include mean and median, with the default selection being mean for calculating the average value.

Specify the parameters for drawing the enrichment score heatmap of the selected pathways across different subtypes:

Colors for annotating each cluster at the top of heatmap: Specify the color for annotating the clusters obtained from clustering at the top of the

heatmap. You can choose the system's default color (System default) or define your own colors (User defined).

Input colors for annotating each cluster: If the "Colors for annotating each cluster at the top of heatmap" parameter is set to "User defined", the user needs to enter the specified colors (hexadecimal) one by one. For example, #2EC4B6.

Colors for heatmap: Specify the colors for drawing the heatmap. You can choose the system default colors (System default) or specify your own colors (User defined).

Input colors for heatmap: If the "Colors for heatmap" parameter is set to "User defined", users need to enter the specified colors in hexadecimal format one by one, for example, #0000FF.

The width of output figure: Specify the width for the output heatmap PDF, with a default value of 15.

The height of output figure: Specify the height for the output heatmap PDF, with a default value of 10.

Indicate the prefix of output figure: Specify the prefix for the output heatmap PDF, with a default value of an empty string (no prefix), using the system default prefix "gseaheatmap".

After setting all parameters, click the "Process" button to perform gene set enrichment analysis. After completion, feedback will be provided in the GSEA Results module on the right side of the webpage. Users can

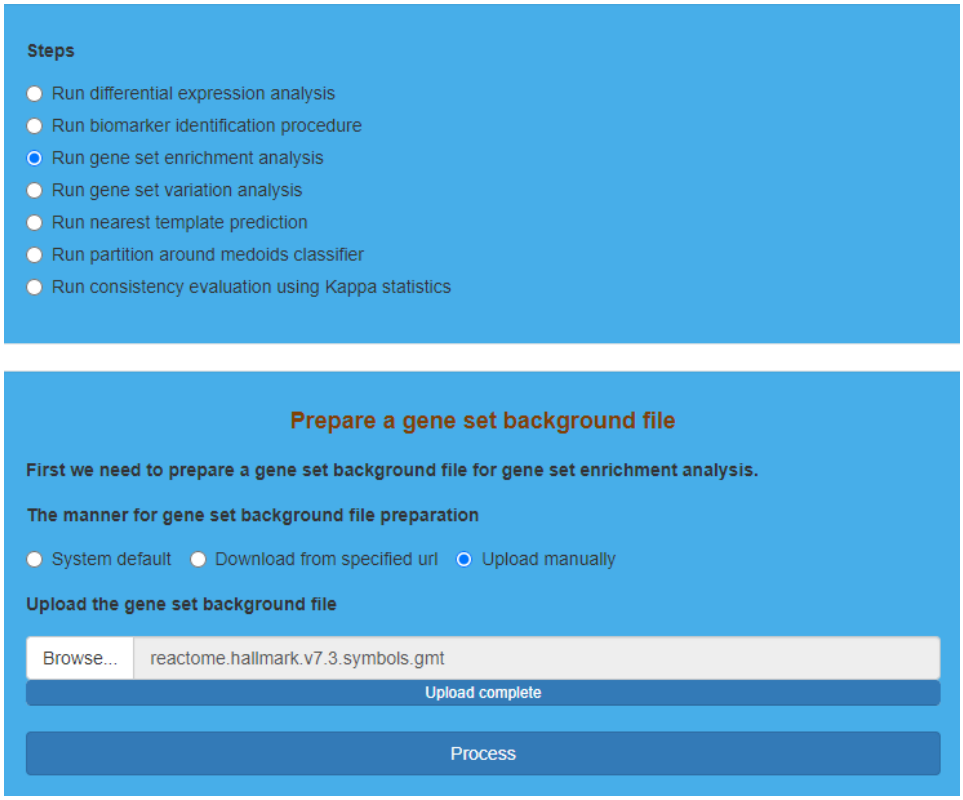
choose to download the RData result set, the PDF of the heatmap showing the enrichment scores for selected pathways in each subtype, the table of gene set enrichment analysis results (available in four formats: Copy, CSV, Excel, Print), and the table of enrichment scores for selected pathways in each subtype (also available in four formats: Copy, CSV, Excel, Print). Next, gene set variation analysis (GSVA) will be conducted to calculate the enrichment scores for each sample in a given pathway within each subtype. To proceed, go to the "RUN Module" and select "Run gene set variation analysis" in the Steps parameter module. Refer to the documentation for guidance on the Run gene set variation analysis parameters.

III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Run gene set enrichment analysis”. First, we upload the gene set background file (reactome.hallmark.v7.3.symbols.gmt), and choose “Upload manually” for “The manner for gene set background file preparation”. Then, we click the “Browse” button to upload the gene set background file and click the “Process” button to finish the preparation of the gene set background file. After that, we choose the algorithm (limma) that we use for differential expression analysis, and then choose “up-regulated” and “down-regulated” for “Indicate the direction of identifying significant pathway” in turn to screen significant up-regulated and down-regulated pathways

for each subtype derived from clustering results respectively. After that, we indicate 10 as “the number of top pathways sorted by NES should be identified for each subtype”. Besides, we indicate 1000 as “the number of permutations for gene set enrichment analysis”, 10 as “minimal size of each gene set for analysis” and 500 as “maximal size of each gene set for analysis”. Furthermore, we indicate 0.05 for “nominal *p* value” and 0.25 for “adjusted *p* value” to identify significant pathways for each subtype derived from clustering results. Additionally, “Indicate the method to employ in the estimation of gene set enrichment scores per sample” chooses “gsva” and “Indicate the method to calculate subtype-specific pathway enrichment scores” chooses “mean”. For the settings of the heatmap, we utilize default settings for “Colors for annotating each cluster at the top of heatmap” (System default), “Colors for heatmap” (System default). Moreover, we indicate 10 for “The width of output figure” and 20 for “The height of output figure”. Finally, we indicate the prefix of the output figure (we ignore “Indicate the prefix of output figure” and type nothing), and then click the “Process” button to run gene set enrichment

analysis.


Steps
Run differential expression analysis
Run biomarker identification procedure
Run gene set enrichment analysis
Run gene set variation analysis
Run nearest template prediction
Run partition around medoids classifier
Run consistency evaluation using Kappa statistics
Prepare a gene set background file
First we need to prepare a gene set background file for gene set enrichment analysis.
The manner for gene set background file preparation
System default
Download from specified url
Upload manually
Upload the gene set background file
Browse...
reactome.hallmark.v7.3.symbols.gmt
Upload complete
Process

Run gene set enrichment analysis

This step aims to perform gene set enrichment analysis using a background file to identify subtype-specific (overexpressed or downexpressed) functional pathways for each subtype.

Indicate the algorithm for completed differential expression analysis

○ deseq2 ○ edger ◉ limma

Indicate the direction of identifying significant pathway

◉ up-regulated ○ down-regulated

Indicate the number of top pathways sorted by NES should be identified for each subtype

10

Indicate the number of permutations for gene set enrichment analysis (1000 by default and 10000 will be better for reproducibility)

1000

Indicate minimal size of each gene set for analysis

10

Indicate maximal size of each gene set for analysis

500

Indicate the nominal p value for identifying significant pathways

0.05

Indicate the adjusted p value for identifying significant pathways

0.25

Indicate the method to employ in the estimation of gene set enrichment scores per sample

◉ gsva ○ ssgsea ○ zscore ○ plage

Indicate the method to calculate subtype-specific pathway enrichment scores

◉ mean ○ median

Colors for annotating each cluster at the top of heatmap

◉ System default ○ User defined

Colors for heatmap

◉ System default ○ User defined

The width of output figure

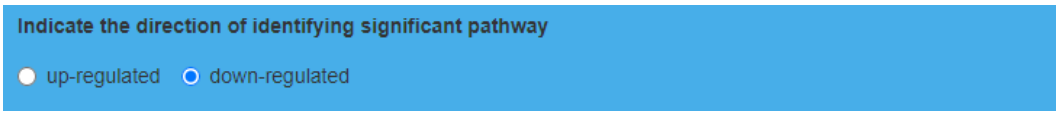
10

The height of output figure

20

Indicate the prefix of output figure (e.g. if the name of output figure is 'gseaheatmap_using_upregulated_pathways.pdf', the prefix would be 'gseaheatmap'. If you do not want a prefix but using the system default prefix 'gseaheatmap', ignore this parameter and type nothing)

Process



Parameter settings for “Run gene set enrichment analysis” on TCGA dataset

Result Display:

The software will generate heatmaps to show the enrichment scores of the screening up-regulated and down-regulated pathways in each subtype (nominal *p* value < 0.05 & adjusted *p* value < 0.25). In addition, the gene set enrichment analysis results as well as enrichment scores of the screening up-regulated and down-regulated pathways in each subtype are also displayed through tables.

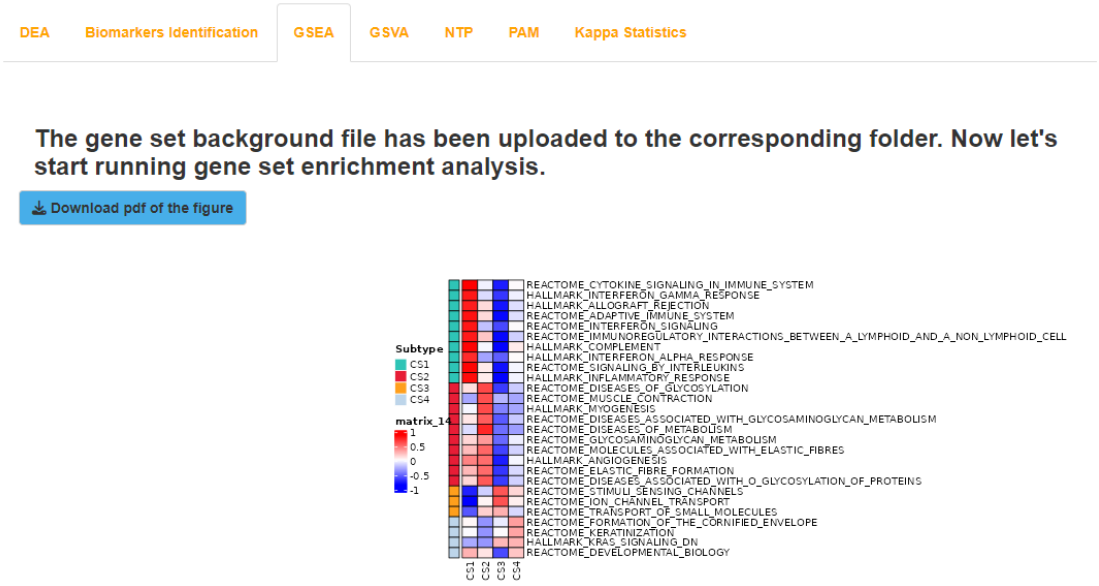


Table 1: GSEA results with upregulated pathways for each identified subtype based on differential expression analysis results using 'limma' algorithm

Copy CSV Excel Print Show 10 entries Search:

	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalue	Subtype
	127	0.65455508768311	3.10065410909388	0.00198807157057654	0.00628930817610063	0.0028649357034869	CS1
	70	0.730702519862898	3.08165589822424	0.00204081632653061	0.00628930817610063	0.0028649357034869	CS1
	56	0.742781213281168	3.0209333303717	0.00205761316872428	0.00628930817610063	0.0028649357034869	CS1
	88	0.673704881640801	2.98102220743303	0.00204081632653061	0.00628930817610063	0.0028649357034869	CS1
	44	0.725415383190205	2.7792422795105	0.00209643605870021	0.00628930817610063	0.0028649357034869	CS1
MPHOID_CELL	36	0.743003578658666	2.70561420645678	0.0020746887966805	0.00628930817610063	0.0028649357034869	CS1
	53	0.653268909894806	2.64155332490265	0.00200400801603206	0.00628930817610063	0.0028649357034869	CS1
	26	0.774246038549765	2.63409256567362	0.00195694716242661	0.00628930817610063	0.0028649357034869	CS1
	78	0.600584493808231	2.60216617829169	0.00196463654223969	0.00628930817610063	0.0028649357034869	CS1
	67	0.620408260146148	2.60202928159608	0.00204918032786885	0.00628930817610063	0.0028649357034869	CS1

Showing 1 to 10 of 75 entries Previous 1 2 3 4 5 … 8 Next

Table 2: subtype-specific enrichment scores on upregulated pathways among identified subtypes based on differential expression analysis results using 'limma' algorithm

Copy CSV Excel Print Show 10 entries Search:

	CS1	CS2	CS3	CS4
_SYSTEM	0.908587480157412	-0.00703097246683867	-0.731196664584179	0.0415393887437293
	0.827772425351802	-0.0829658839413201	-0.645342447420137	-0.0139278750780818
	0.808304282740688	0.164237954622694	-0.761357249150002	-0.0709937319385595
	0.851339340824841	0.166880881588706	-0.807107961642894	-0.0664391382709674
	0.829096184105986	-0.170831669511776	-0.58425531884131	0.0361532811033746
ONS_BETWEEN_A_LYMPHOID_AND_A_NON_LYMPHOID_CELL	0.823931335409712	0.234259403977811	-0.805856843564454	-0.118173362312457
	0.896054194449996	0.0123733956545938	-0.815744426580805	0.105893684035972
	0.761569274386063	-0.270164805705296	-0.507306660782453	0.0467695684322682
	0.892778683272245	0.109958331913571	-0.77140562167291	0.00970130141281623
	0.873311796778305	0.104031583633586	-0.818095354801434	0.00250461236052602

Showing 1 to 10 of 27 entries Previous 1 2 3 Next

DEA Biomarkers Identification GSEA GSVA NTP PAM Kappa Statistics

The gene set background file has been uploaded to the corresponding folder. Now let's start running gene set enrichment analysis.

Download pdf of the figure

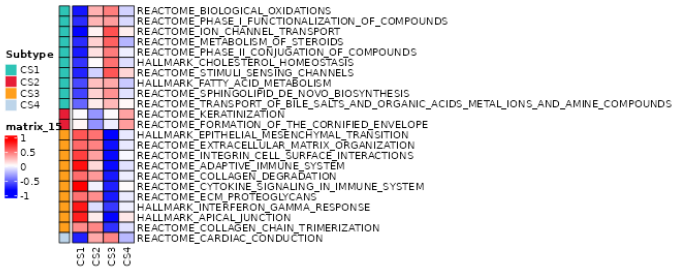


Table 1: GSEA results with downregulated pathways for each identified subtype based on differential expression analysis results using 'limma' algorithm

Copy CSV Excel Print Show 10 entries Search:

	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalue	Subtype
	33	-0.632603318795763	-2.26590982745526	0.00197628458498024	0.00628930817610063	0.0028649357034869	CS1
	23	-0.621685385219035	-2.01957180484275	0.00199203187250996	0.00628930817610063	0.0028649357034869	CS1
	22	-0.603304886642277	-1.93222015885375	0.00200803212851406	0.00628930817610063	0.0028649357034869	CS1
	14	-0.65048218418364	-1.85006402143245	0.00823045267489712	0.0178152836380684	0.00811530310382927	CS1
	10	-0.702378459319984	-1.80063327244324	0.00628930817610063	0.0151474887058198	0.00690005641262337	CS1
	11	-0.679493282592888	-1.80042966122803	0.00814663951120163	0.0178152836380684	0.00811530310382927	CS1
	14	-0.620989130910023	-1.76618157534796	0.00823045267489712	0.0178152836380684	0.00811530310382927	CS1
	18	-0.563601457413615	-1.74973682801151	0.0115606936416185	0.0232573954437266	0.0105943198697185	CS1
	10	-0.635059117971616	-1.62805188942618	0.0188679245283019	0.0350697292863002	0.0159751306074867	CS1
OMPOUNDS	11	-0.582961427097482	-1.5446525691807	0.0427698574338086	0.0769857433808554	0.0350689135745355	CS1

Showing 1 to 10 of 105 entries Previous 1 2 3 4 5 ... 11 Next

Table 2: subtype-specific enrichment scores on downregulated pathways among identified subtypes based on differential expression analysis results using 'limma' algorithm

Copy CSV Excel Print Show 10 entries Search:

	CS1	CS2	CS3	CS4
	-0.770627938433205	0.30078898877846	0.478862151853589	-0.117353320001267
OF_COMPOUNDS	-0.690282557518981	0.2902722223263	0.371670638297782	-0.0934623463135406
	-0.82968840107253	0.0759021720188665	0.630546992600368	0.0919210701592323
	-0.694107591835339	0.191254563852626	0.577466370026564	-0.216799817204467
OMPOUNDS	-0.752876549162572	0.135499724942267	0.483673390997174	-0.0169441639514944
	-0.661275657421243	0.0435428334732109	0.518727959054105	-0.0774361504756129
	-0.720677139563765	-0.122166013194812	0.6146682723452	0.177473197275284
	-0.587882839698883	0.252331456205827	0.329353092081203	-0.149628316911315
YNTHESIS	-0.614055252758763	0.192551894749849	0.408282999947314	-0.0631777007983775
ND_ORGANIC_ACIDS_METAL_IONS_AND_AMINE_COMPOUNDS	-0.491767197799781	0.101597406419201	0.277918102843314	0.0609146468705173

Showing 1 to 10 of 23 entries Previous 1 2 3 Next

Result display for “Run gene set enrichment analysis” on TCGA dataset

(4). Run gene set variation analysis for TCGA dataset

I.Analysis introduction

“Run gene set variation analysis” calculates enrichment scores of each sample in each subtype according to the given gene set background files, and then displays the results using tables and a corresponding heatmap.

II.Parameters setting guides

i.Prepare a gene set list of interest

Following this, gene set variation analysis (GSVA) will be conducted to calculate the enrichment scores for each sample in a given pathway within each subtype. Similarly, you will need to prepare a file containing a list of gene sets of interest：

The manner for the gene set list of interest preparation: Specify the method for preparing the list of gene sets of interest. The platform provides three preparation methods: System default (default choice), Download from specified URL, and Upload manually.

Download url for the gene set list of interest: If the "The manner for the gene set list of interest preparation" parameter is set to "Download from specified URL"， the user needs to specify the download link for the file containing the list of gene sets of interest.

Upload the gene set list of interest: If the "The manner for the gene set list of interest preparation" parameter is set to "Upload manually"， the user needs to click the "Browse" button to select the relevant file containing the list of gene sets of interest from their local device.

After specifying the file for the list of gene sets of interest, click the "Process" button to prepare and process the file. Once completed, the GSVA results module on the right side of the webpage will provide feedback. Next, you need to continue specifying the additional

parameters for GSVA gene set variation analysis, as detailed in the guidance document for the "Run gene set variation analysis" parameter module.

ii.Run gene set variation analysis

Next, we will continue to specify additional parameters for gene set variation analysis (GSVA), including:

First, specify the calculation method for enrichment scores and the normalization parameters for the calculated results:

Indicate the method to employ in the estimation of gene set enrichment scores per sample: Specify the method used to estimate the gene set enrichment scores for the selected pathways in each sample. There are four available methods: gsva, ssgsea, zscore, and plage. The default selection is gsva.

Indicate if enrichment scores should be centered or not: Specify whether to perform Centering on the enrichment score calculation results, i.e., subtracting the mean enrichment score of each pathway. The default choice is Yes for processing.

Indicate if enrichment scores should be scaled or not: Specify whether to perform Scaling on the enrichment score calculation results, i.e., dividing by the standard deviation of the enrichment score of each pathway. The default choice is Yes for processing.

Assign marginal cutoff for truncating enrichment scores or not: Specify

whether to assign a marginal cutoff for truncating the results of the enrichment score calculation. The default choice is Yes for specifying a cutoff value.

Marginal cutoff for truncating enrichment scores: If the parameter "Assign marginal cutoff for truncating enrichment scores or not" is set to Yes, you need to specify the cutoff value for the enrichment score calculation results. The default cutoff value is a positive number, typically set to 1. If the data is greater than this cutoff value, it will be replaced by the cutoff value. If the data is less than the negative of the cutoff value, it will be replaced by the negative of the cutoff value.

Next, specify the parameters for drawing the heatmap of enrichment scores for the selected pathways across samples:

Sample annotations from survival information for heatmap or not: Specify whether to use clinical survival information for annotation at the top of the heatmap. Choose "No" if you do not want to use clinical survival information; otherwise, choose "Yes".

The number of sample annotations from survival information for heatmap: If the "Sample annotations from survival information for heatmap or not" parameter is set to "Yes", specify the number of clinical survival variables to annotate. The default value is 3.

Input sample annotations from survival information for heatmap: Then, you need to specify detailed information about the clinical survival

variables. In the first line, accurately enter the variable names that need to be annotated in the clinical survival dataset. In the second line, enter the data types corresponding to each clinical survival variable. In the third line, enter the annotation colors for each clinical survival variable in hexadecimal format, for example, #000004FF, using semicolons to separate multiple colors. If it is a continuous variable, the number of colors entered should be 3, corresponding to the minimum value, median, and maximum value of that variable. If it is a discrete variable, the number of colors entered should match the number of categories for that variable.

Colors for annotating each cluster at the top of heatmap: Specify the color for annotating the clusters obtained from clustering at the top of the heatmap. You can choose the system's default color (System default, selected by default) or specify your own color (User defined).

Input colors for annotating each cluster: If you have chosen "User defined" for the "Colors for annotating each cluster at the top of the heatmap" parameter, you need to enter the specified colors (in hexadecimal format) one by one. For example, #2EC4B6.

Colors for heatmap: Specify the colors for drawing the heatmap; you can choose the system's default colors (System default, selected by default) or specify your own colors (User defined).

The number of colors for heatmap: If the "Colors for heatmap" parameter is set to "User defined", you need to specify the number of colors for

drawing the heatmap.

Input colors for heatmap: If the "Colors for heatmap" parameter is set to "User defined", you need to enter the specified colors (in hexadecimal format) one by one. For example, #00FF00.

Distance measurement for hierarchical clustering: Specify the method for hierarchical clustering distance measurement for input data. The default value is "euclidean".

Clustering method for hierarchical clustering: Enter the clustering method for hierarchical clustering. The default value is "ward.D".

Show rownames (feature names) in heatmap or not: Specify whether to display the names of the selected pathways in the rows of the heatmap. The default option is Yes for display.

Show colnames (sample ID) in heatmap or not: Specify whether to display the IDs of each sample in the columns of the heatmap. The default option is No for no display.

The width of output figure: Specify the width of the output heatmap PDF. The default value is 8.

The height of output figure: Specify the height of the output heatmap PDF. The default value is 8.

Indicate the prefix of output figure: Specify the prefix for the output heatmap PDF. The default is empty, meaning no prefix is specified, and the system default prefix "enrichment_heatmap_using" will be used.

After setting all the parameters, click the "Process" button to perform gene set variation analysis. Once completed, feedback will be provided in the GSVA results module on the right side of the webpage. Users can choose to download the RData result set, the heatmap PDF showing the enrichment scores of selected pathways in each sample, the raw enrichment score results table (in four formats: Copy, CSV, Excel, Print), and the standardized enrichment score results table (in four formats: Copy, CSV, Excel, Print). Additionally, to validate the clustering results, the platform uses the non-model method NTP to predict the subtypes of external validation set samples based on the obtained subtype marker genes. Go to the "RUN Module" module, then under the "Steps" parameter module, select "Run nearest template prediction." For detailed instructions, refer to the documentation for the "Run nearest template prediction" parameter module.

III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Run gene set variation analysis”. First, we upload the gene set list of interest (h.all.v7.2.symbols.gmt), and choose “Upload manually” for “The manner for the gene set list of interest preparation”. Then, we click the “Browse” button to upload the gene set list of interest and click the “Process” button to finish the preparation of the gene set list of interest. After that, “Indicate the method to employ in the estimation of gene set enrichment scores per sample” chooses “gsva”

by default. Besides, we choose “Yes” for both “Indicate if enrichment scores should be centered or not” and “Indicate if enrichment scores should be scaled or not”. Furthermore, we choose “Yes” for “Assign marginal cutoff for truncating enrichment scores or not”, and then indicate 1 as “Marginal cutoff for truncating enrichment scores”. For the settings of the heatmap, “Colors for annotating each cluster at the top of heatmap” (System default), “Colors for heatmap” (System default), “Distance measurement for hierarchical clustering” (euclidean), “Clustering method for hierarchical clustering” (ward.D), “Show rownames (feature names) in heatmap or not” (Yes), “Show colnames (sample ID) in heatmap or not” (No) use default settings. Moreover, we indicate 12 for “The width of output figure” and 10 for “The height of output figure”. In addition to annotation for sample through clustering results, this step also allows users to select specific clinical features for annotation, and we select “PAM_Subtype”, “oneNN_Subtype”, “Lund_Subtype” and “TCGA_Subtype”. Finally, we indicate the prefix of the output figure (we ignore “Indicate the prefix of output figure” and type nothing), and then click the “Process” button to run gene set variation analysis.

Steps

- ○ Run differential expression analysis
- ○ Run biomarker identification procedure
- ○ Run gene set enrichment analysis
- ◉ Run gene set variation analysis
- ○ Run nearest template prediction
- ○ Run partition around medoids classifier
- ○ Run consistency evaluation using Kappa statistics

Prepare a gene set list of interest

First we need to prepare a gene set list of interest for gene set variation analysis.

The manner for the gene set list of interest preparation

○ System default ○ Download from specified url ◉ Upload manually

Upload the gene set list of interest

Browse... h.all.v7.2.symbols.gmt

Upload complete

Process

Run gene set variation analysis

This step aims to use gene set variation analysis to calculate enrichment score of each sample in each subtype based on a given gene set list of interest.

Indicate the method to employ in the estimation of gene set enrichment scores per sample

◉ gsva ○ ssgsea ○ zscore ○ plage

Indicate if enrichment scores should be centered or not

◉ Yes ○ No

Indicate if enrichment scores should be scaled or not

◉ Yes ○ No

Assign marginal cutoff for truncating enrichment scores or not

◉ Yes ○ No

Marginal cutoff for truncating enrichment scores

1

Sample annotations from survival information for heatmap or not

(•) Yes ( ) No

The number of sample annotations from survival information for heatmap

4

Input sample annotations from survival information for heatmap

First line: Please input the sample annotation variables from survival information

Second line: Please input 'Continuous' or 'Categorical' to indicate the type of each sample annotation variable

Last line: Please input the colors for each sample annotation variable (use hex color format, e.g. #000004FF and English semicolons should be used to separate the input colors)

Note1: If the sample annotation variable is continuous, the number of indicated colors should be equal to 3, which represents the minimum, median and maximum value of this variable)

Note2: If the sample annotation variable is categorical, the number of indicated colors should be equal to the number of categories for this variable)

	1	2	3	4
Sample annotation	PAM_Subty	oneNN_Sub	Lund_Subty	TCGA_Subt
Continuous or Categorical	Categorical	Categorical	Categorical	Categorical
Color settings	#2874C5;#E	#2874C5;#C	#EABF00;#8	#2874C5;#8

Colors for annotating each cluster at the top of heatmap

(•) System default ( ) User defined

Colors for heatmap

(•) System default ( ) User defined

Distance measurement for hierarchical clustering

euclidean

Clustering method for hierarchical clustering

ward.D

Show rownames (feature names) in heatmap or not

(•) Yes ( ) No

Show colnames (sample ID) in heatmap or not

( ) Yes (•) No

The width of output figure

12

The height of output figure

10

Indicate the prefix of output figure (e.g. if the name of output figure is 'enrichment_heatmap_using_gsva.pdf', the prefix would be 'enrichment_heatmap_using'. If you do not want a prefix but using the system default prefix 'enrichment_heatmap_using', ignore this parameter and type nothing)

Process

Parameter settings for “Run gene set variation analysis” on TCGA dataset

Result Display:

The software will generate a heatmap to show the enrichment scores of the pathways of interest in each sample. In addition, the raw enrichment scores as well as the z-scored enrichment scores of the pathways of interest in each sample are also displayed through tables.

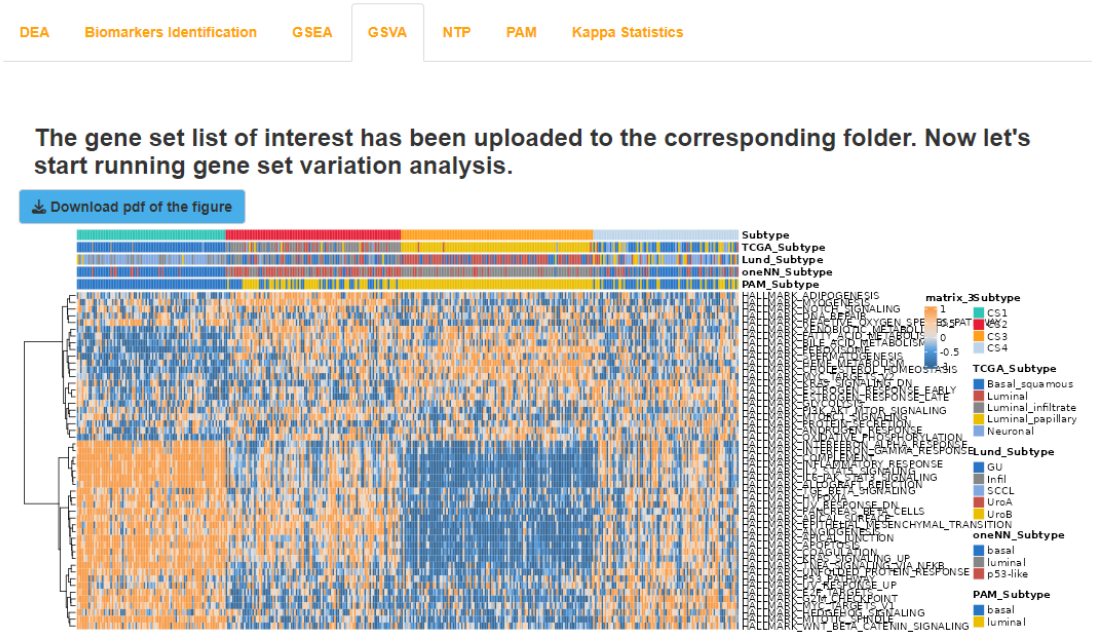


Table 1: Raw enrichment score based on the given gene set list of interest by using gsva method

Copy CSV Excel Print Show 10 entries Search:

	TCGA-HQ-A5NE-01A	TCGA-FD-A5BT-01A	TCGA-DK-A3IQ-01A	TCGA-ZF-AA4X-01A	TCGA-GD-A3 01A
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.609218342433021	0.342134708499665	-0.378235223995979	-0.415108127429381	0.449228402
HALLMARK_HYPOXIA	0.358997941976495	0.47732064628403	-0.0368311341472628	-0.453998566810624	0.461055813
HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.0323779511502909	0.00213634297680262	-0.533605330050768	0.235882964475868	-0.442244592
HALLMARK_MITOTIC_SPINDLE	0.597602119146635	0.278140073467567	0.559150124604204	-0.582666033541275	0.552710255
HALLMARK_WNT_BETA_CATENIN_SIGNALING	0.418776091290295	0.327385512194761	0.0453084471775966	-0.828641349786719	0.452445899
HALLMARK_TGF_BETA_SIGNALING	0.646759875102749	0.335207757642067	0.482956598665959	-0.27711834193699	0.529793887
HALLMARK_IL6_JAK_STAT3_SIGNALING	0.191113230567938	0.541986286185345	-0.595254171404916	-0.375533077776739	-0.149116146
HALLMARK_DNA_REPAIR	-0.141512544244856	0.300803310529454	0.179262027213133	-0.673758559800182	-0.473760174
HALLMARK_G2M_CHECKPOINT	0.394217500582665	0.625125334451975	-0.67334000667331	-0.798083798083763	0.107774441
HALLMARK_APOPTOSIS	0.398781383159376	0.391502718290118	0.0193253323332012	-0.448959828281674	0.192061819

Showing 1 to 10 of 50 entries Previous 1 2 3 4 5 Next

Table 2: z-scored enrichment score based on the given gene set list of interest by using gsva method

Copy CSV Excel Print Show 10 entries Search:

	TCGA-HQ-A5NE-01A	TCGA-FD-A5BT-01A	TCGA-DK-A3IQ-01A	TCGA-ZF-AA4X-01A	TCGA-GD-A3O 01A
HALLMARK_TNFA_SIGNALING_VIA_NFKB	1	1	-1	-1	1
HALLMARK_HYPOXIA	1	1	-0.160501448405905	-1	1
HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.1318780794035	0.0283053056664498	-1	0.828850891782026	-1
HALLMARK_MITOTIC_SPINDLE	1	0.612744740885331	1	-1	1
HALLMARK_WNT_BETA_CATENIN_SIGNALING	1	0.788397855210029	0.130803092463057	-1	1
HALLMARK_TGF_BETA_SIGNALING	1	0.865828527380933	1	-0.735957675060966	1
HALLMARK_IL6_JAK_STAT3_SIGNALING	0.470033109513886	1	-1	-1	-0.41352557800
HALLMARK_DNA_REPAIR	-0.409676246660787	0.810892719475831	0.475499997467221	-1	-1
HALLMARK_G2M_CHECKPOINT	0.793470604401491	1	-1	-1	0.26131038782
HALLMARK_APOPTOSIS	1	1	0.0186317285522308	-1	0.60534965416

Showing 1 to 10 of 50 entries Previous 1 2 3 4 5 Next

Result display for “Run gene set variation analysis” on TCGA dataset

(5). Run nearest template prediction for TCGA dataset and validation datasets

I.Analysis introduction

Nearest Template Prediction (NTP) is a model-free method, and “Run nearest template prediction” utilizes NTP to predict the subtype of each sample in external validation dataset based on marker genes for each

subtype obtained from TCGA dataset, which are displayed through a table. Then, a heatmap is drawn to show the consistency between prediction results and clustering results.

II.Parameters setting guides

Run nearest template prediction

To validate the quality of the clustering results, the platform employs the non-model method Nearest Template Prediction (NTP) to predict the subtypes of samples in an external validation set based on the obtained subtype marker genes. The parameters that need to be specified include:

Run nearest template prediction on tcga or validation cohort: Select the dataset for subtype prediction using the NTP method, with the default choice being the external validation set. After completing the subtype prediction on the external validation set, you also have the option to reselect TCGA and use NTP to predict the subtypes.

To perform subtype prediction using the NTP method, you need to specify the template information. The prediction template must be obtained from previous analyses:

Indicate the algorithm: Specify the algorithm used in the previous differential expression analysis, including DESeq2, edgeR, and limma.

Indicate the direction: Specify the use of previously obtained up-regulated or down-regulated marker genes for each subtype, with the default selection being up-regulated marker genes.

Then specify the other parameters for NTP subtype prediction:

Indicate if the expression data should be further scaled: Whether to perform additional scaling on mRNA or lncRNA expression data, i.e., dividing by the corresponding feature's standard deviation. The default selection is Yes for processing.。

Indicate if the expression data should be further centered: 是 Whether to perform additional centering on mRNA or lncRNA expression data, i.e., subtracting the corresponding feature's mean. The default selection is Yes for processing.

Indicate the permutations for p-value estimation: Specify the number of permutations for p-value estimation. The default value is 1000.

Indicate the distance measurement: Specify the distance metric, which includes cosine, pearson, spearman, and kendall. The default choice is cosine.

Input an integer value for p-value reproducibility: Specify the random seed for reproducibility of NTP prediction results. The default value is 123456.

The width of output figure: Specify the width for the output of the consistency heatmap PDF, with a default value of 5.

The height of output figure: Specify the height for the output of the consistency heatmap PDF, with a default value of 5.。

The name of the nearest template prediction heatmap: Specify the name

for the output of the consistency heatmap PDF. If the "Run nearest template prediction on TCGA or validation cohort" parameter is set to TCGA, the default value for this parameter is set to "ntpheatmap(TCGA)"; if the parameter is set to Validation, the default value is "ntpheatmap(Validation)".

Once all the parameters are set, click the "Process" button to perform subtype prediction using the NTP method for the specified dataset. After completion, feedback will be provided in the NTP results module on the right side of the webpage. Users can choose to download the RData result set, the consistency heatmap PDF, and the subtype prediction result table (in four formats: Copy, CSV, Excel, Print). In addition, the platform also offers another model-free method, PAM, for predicting the subtypes of samples in an external validation set. To use this method, navigate to the "RUN Module" module, then to the "Steps" parameter module, and choose "Run partition around medoids classifier." Refer to the documentation for guidance on the parameters of the "Run partition around medoids classifier" module.

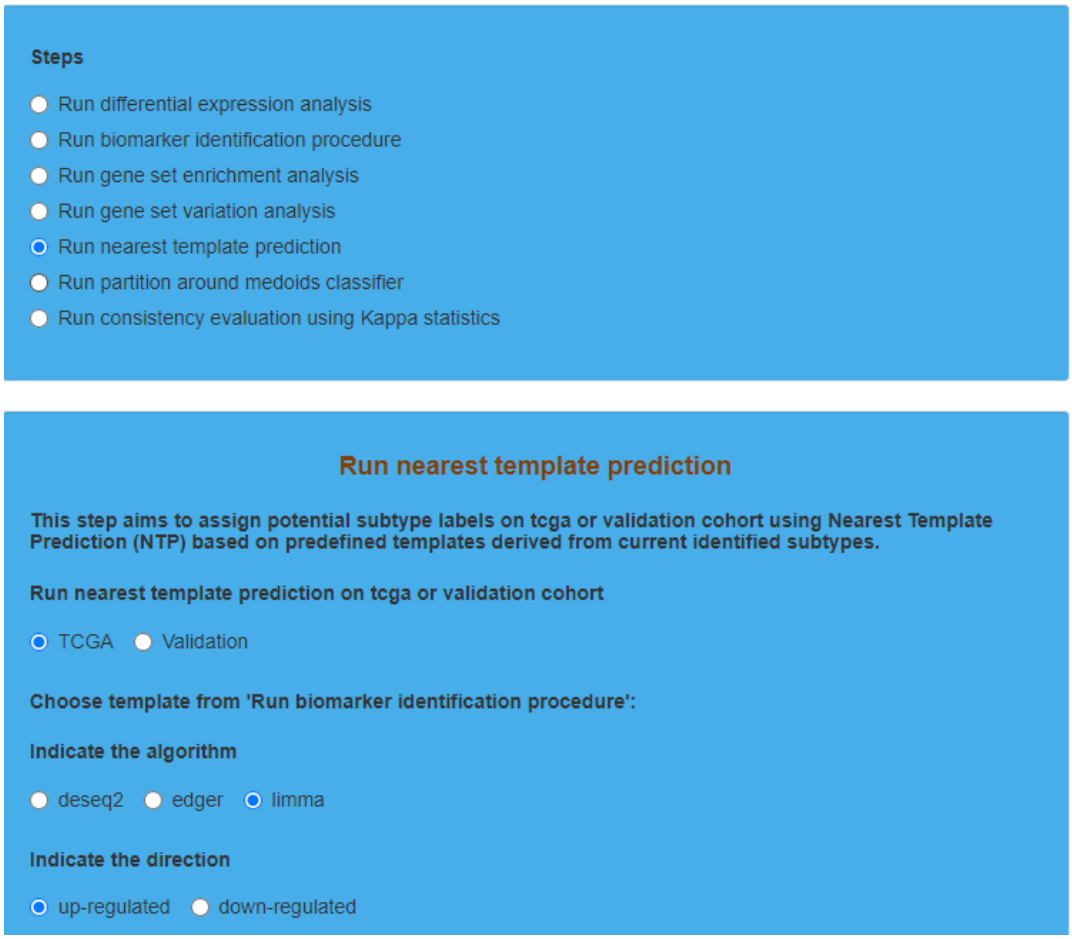
In addition, if you have obtained the NTP prediction results for an external validation set, you can go back to the "COMP Module" module and choose "Validation" to perform comparative analysis between predicted subtypes on the external validation set. It's important to note that users need to select the appropriate analysis within the "COMP Module" based on the

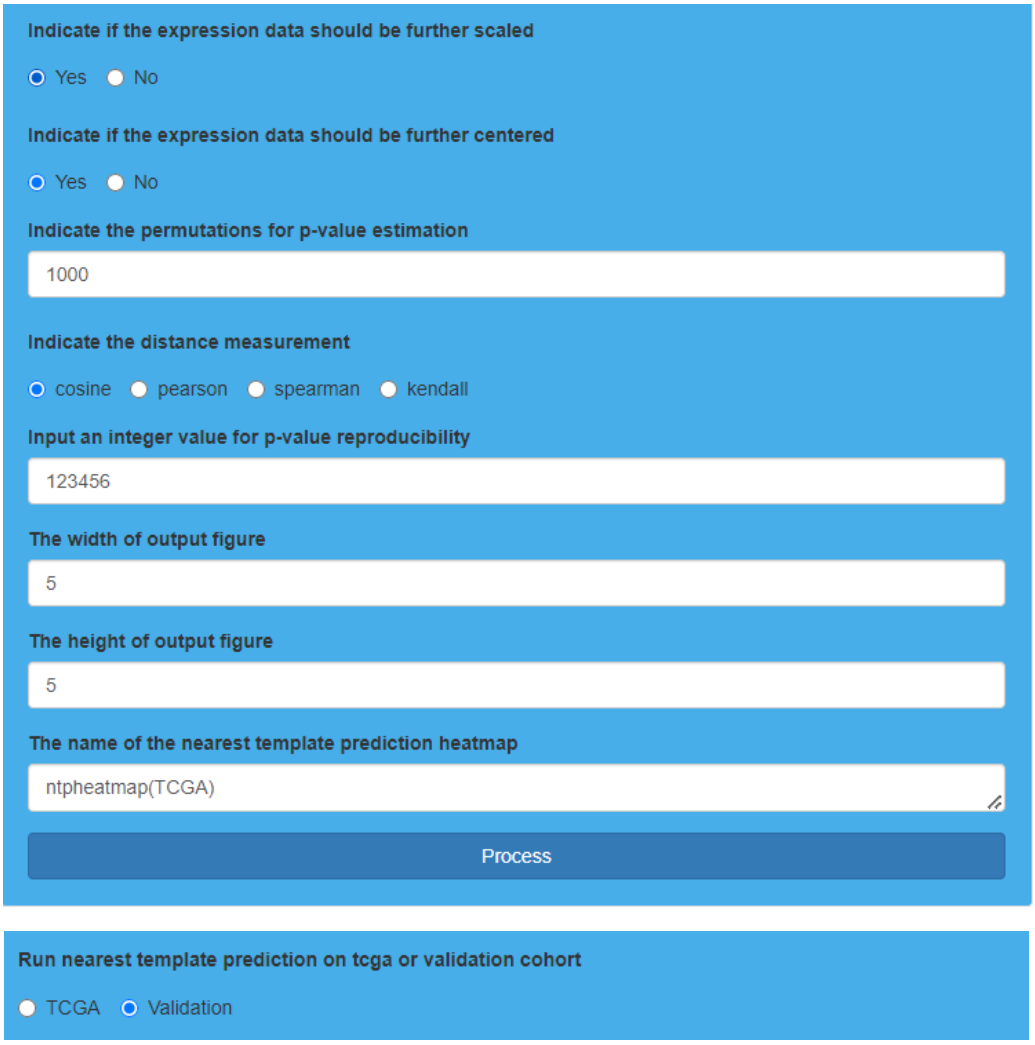
omics data types available in the external validation set. For example: Compare survival outcome, Compare clinical features, and Compare agreement with other subtypes analyses require clinical survival data. Compare drug sensitivity analysis requires mRNA or lncRNA data. Compare mutational frequency and Compare total mutation burden analyses require binary somatic mutation data. Compare fraction genome altered analysis requires copy number alterations data. For specific instructions, please refer to the documentation for each parameter in the "COMP Module" module.

III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Run nearest template prediction”, and “Run nearest template prediction on tcga or validation cohort” chooses “TCGA”. First, we should indicate the template from the results of differential expression analysis, here “Indicate the algorithm” chooses “limma” and “Indicate the direction” chooses “up-regulated”. For the processing of expression data, we normalize the data by default (both “Indicate if the expression data should be further scaled” and “Indicate if the expression data should be further centered” choose “Yes”). After that, we indicate 1000 as “the permutations for p-value estimation”, and then choose “cosine” for “the distance measurement”. Additionally, we indicate 123456 as “an integer value for p-value reproducibility” and utilize default settings for “The width of output figure”, “The height of

output figure" and "The name of the nearest template prediction heatmap". Finally, we click the "Process" button to run nearest template prediction. For two validation datasets, "Run nearest template prediction on tcga or validation cohort" chooses "Validation", keep other parameters unchanged, and then click the "Process" button to run nearest template prediction in turn.





Parameter settings for “Run nearest template prediction”

Result Display:

The software will generate tables to show the subtype prediction results of each sample in TCGA dataset as well as two validation datasets using “Nearest Template Prediction” (NTP). Besides, consistency between prediction results and templates in TCGA dataset as well as two validation

datasets will be evaluated through heatmaps, and higher degree of consistency indicates better prediction performance.

DEA | Biomarkers Identification | GSEA | GSVA | NTP | PAM | Kappa Statistics

Download pdf of the figure

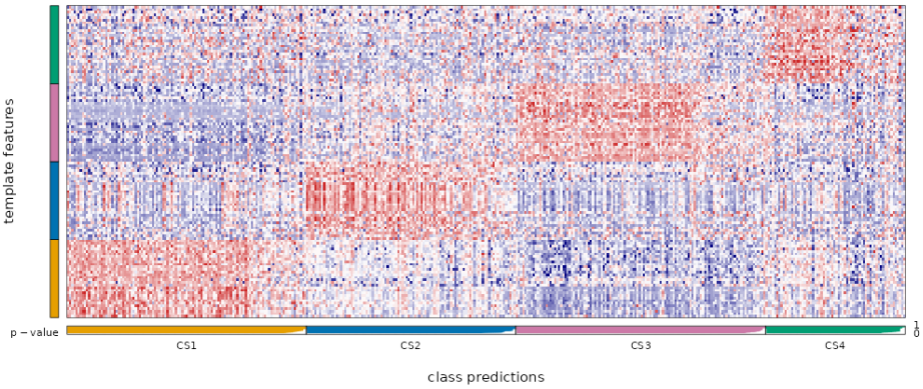


Table: the results of nearest template prediction (TCGA)

Copy | CSV | Excel | Print | Show 10 entries | Search:

	prediction	d.CS1	d.CS2	d.CS3	d.CS4	p.value	FDR
TCGA-HQ-A5NE-01A	CS1	0.5624	0.6853	0.5774	0.5985	0.2527	0.2613
TCGA-FD-A5BT-01A	CS1	0.5187	0.6355	0.8644	0.8229	0.001	0.0015
TCGA-DK-A3IQ-01A	CS2	0.7343	0.3889	0.7571	0.7477	0.001	0.0015

Result display for “Run nearest template prediction” on TCGA dataset

DEA | Biomarkers Identification | GSEA | GSVA | NTP | PAM | Kappa Statistics

Download pdf of the figure

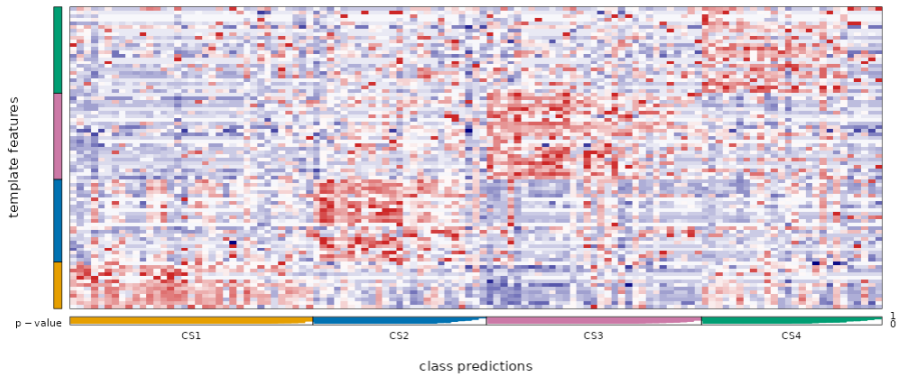


Table: the results of nearest template prediction (Validation)

Copy | CSV | Excel | Print | Show 10 entries | Search:

	prediction	d.CS1	d.CS2	d.CS3	d.CS4	p.value	FDR
FR_1_U133_2.CEL	CS3	0.7556	0.7056	0.5346	0.692	0.002	0.0045
FR_103_U133_2.CEL	CS3	0.8266	0.7992	0.4677	0.6782	0.001	0.0027
FR_106_U133_2.CEL	CS2	0.7434	0.572	0.8391	0.8381	0.001	0.0027

Result display for “Run nearest template prediction” on validation dataset (affy)

DEA | Biomarkers Identification | GSEA | GSVA | NTP | PAM | Kappa Statistics

Download pdf of the figure

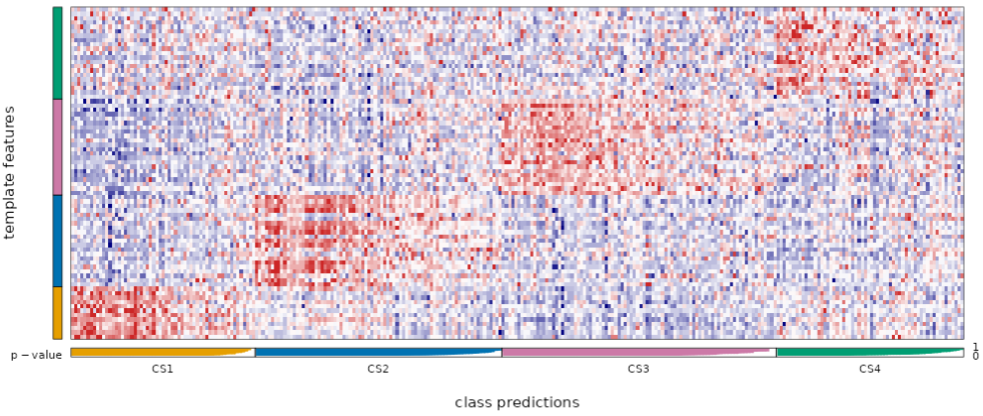


Table: the results of nearest template prediction (Validation)

Copy | CSV | Excel | Print | Show 10 entries | Search:

	prediction	d.CS1	d.CS2	d.CS3	d.CS4	p.value	FDR
GSM806803	CS3	0.8511	0.7957	0.4802	0.7163	0.001	0.0029
GSM806804	CS2	0.7392	0.3537	0.6758	0.7841	0.001	0.0029
GSM806805	CS3	0.8066	0.7669	0.7664	0.7794	0.8312	0.837

Result display for “Run nearest template prediction” on validation dataset (illumina)

(6). Run partition around medoids classifier for TCGA dataset and validation datasets

I.Analysis introduction

Partition around Medoids (PAM) is also a model-free prediction method. “Run partition around medoids classifier” uses PAM to predict the subtype of each sample in external validation dataset and evaluates the consistency between prediction results and clustering results through a similarity and reproducibility indicator named IGP. The predicted subtype of each sample and IGP value of each subtype are displayed in the form of tables respectively.

II.Parameters setting guides

Run partition around medoids classifier

In addition, the platform also offers another model-free method, Partition Around Medoids (PAM), for predicting subtypes of samples in an external validation set. The consistency of the prediction results with the clustering results is evaluated using similarity and reproducibility metrics, such as the In-Group Proportion (IGP). The parameters that need to be specified include:

Run partition around medoids classifier on tcga or validation cohort: Select the dataset for subtype prediction using the PAM method, with the default choice being the external validation set (Validation). After

completing subtype prediction on the external validation set, you can also choose to reselect TCGA and use PAM to predict subtypes.

Indicate action for NA values in normalized expression training data: Specify the handling method for missing values (NA) in the training dataset (TCGA). The default choice is to "Remove directly," which means directly removing NA. If you choose "KNN imputation," the KNN method will be used to impute NA.

Indicate action for NA values in normalized expression testing data: Specify the handling method for missing values (NA) in the predicted dataset. The default choice is to "Remove directly", which means directly removing NA. If you choose "KNN imputation", the KNN method will be used to impute NA. If the "Run partition around medoids classifier on tcga or validation cohort" parameter is set to "Validation", the predicted dataset is the external validation set. If it is set to "TCGA", the predicted dataset is the TCGA dataset.

Whether indicate a subset of genes to be used: Specify whether to use specific genes for training and prediction. The default is No, which means not specifying.

Indicate a subset of genes to be used: If the parameter "Whether indicate a subset of genes to be used" is set to Yes, you need to enter the specified gene names one by one, separated by commas.

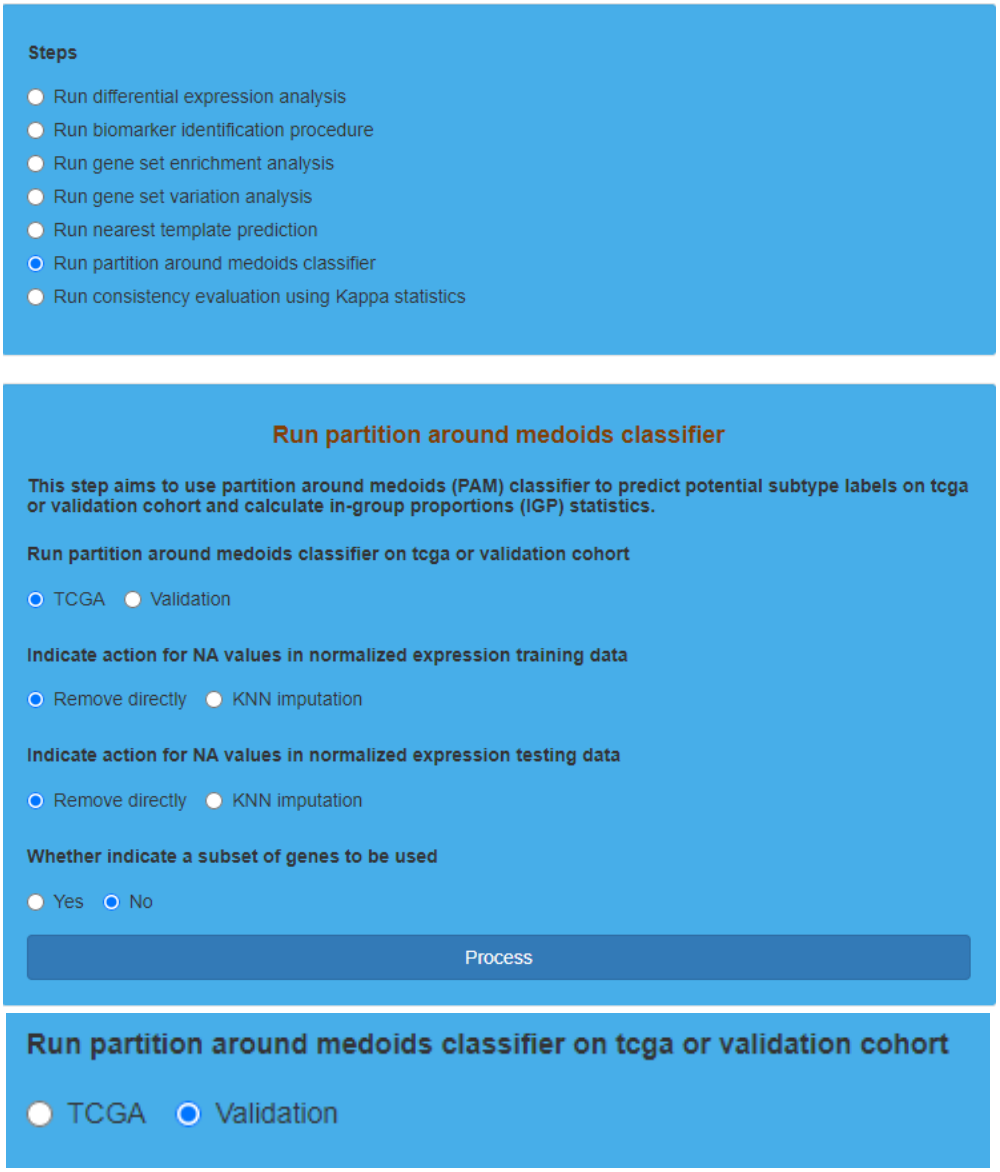
After setting all the parameters, click the "Process" button to perform subtype prediction using the PAM method. Once completed, the platform will provide feedback in the PAM Results module on the right side of the webpage. You can choose to download the RData result set, the IGP calculation result table (in four formats: Copy, CSV, Excel, Print), and the subtype prediction result table (in four formats: Copy, CSV, Excel, Print). Finally, evaluate the consistency between clustering results and prediction results, or between NTP and PAM prediction results, using the Kappa statistic. Go to the "RUN Module" and select "Steps" in the parameter module, then choose "Run consistency evaluation using Kappa statistics." Refer to the guidance document for the parameters in the Run consistency evaluation using Kappa statistics module.

Additionally, if you have obtained the PAM prediction results for an external validation set, you can go back to the COMP Module module and choose Validation for comparing subtypes' prediction analysis on the external validation set. It's important to note that users should select the appropriate analysis in the COMP Module based on the omics data type in the external validation set. For example, analyses such as Compare survival outcome, Compare clinical features, and Compare agreement with other subtypes require clinical survival data. Compare drug sensitivity analysis requires mRNA or lncRNA data, and Compare mutational frequency and Compare total mutation burden analyses

require binary somatic mutation data. Similarly, Compare fraction genome altered analysis requires copy number alterations data. For detailed instructions, please refer to the respective parameter documents in the COMP Module module.

III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Run partition around medoids classifier”, and “Run partition around medoids classifier on tcga or validation cohort” chooses “TCGA”. For the processing of the “NA” values in “normalized expression training data” and “normalized expression testing data”, we choose “Remove directly”. Here, the “training data” refers to TCGA dataset. If we want to predict the subtype of each sample in TCGA dataset, the “testing data” refers to TCGA dataset. If we want to predict the subtype of each sample in validation dataset, the “testing data” refers to validation dataset. After that, we choose “No” for “Whether indicate a subset of genes to be used”. Finally, we click the “Process” button to run partition around medoids classifier. For two validation datasets, “Run partition around medoids classifier on tcga or validation cohort” chooses “Validation”, keep other parameters unchanged, and then click the “Process” button to run partition around medoids classifier in turn.



Parameter settings for “Run partition around medoids classifier”

Result Display:

The software will generate tables to show the subtype prediction results of each sample in TCGA dataset as well as two validation datasets using “Partition around Medoids” (PAM) classifier. Besides, the values of IGP statistics for each subtype will be also displayed through tables, which

reflects the similarity and reproducibility between the training data and testing data. Larger value of IGP indicates higher consistency.

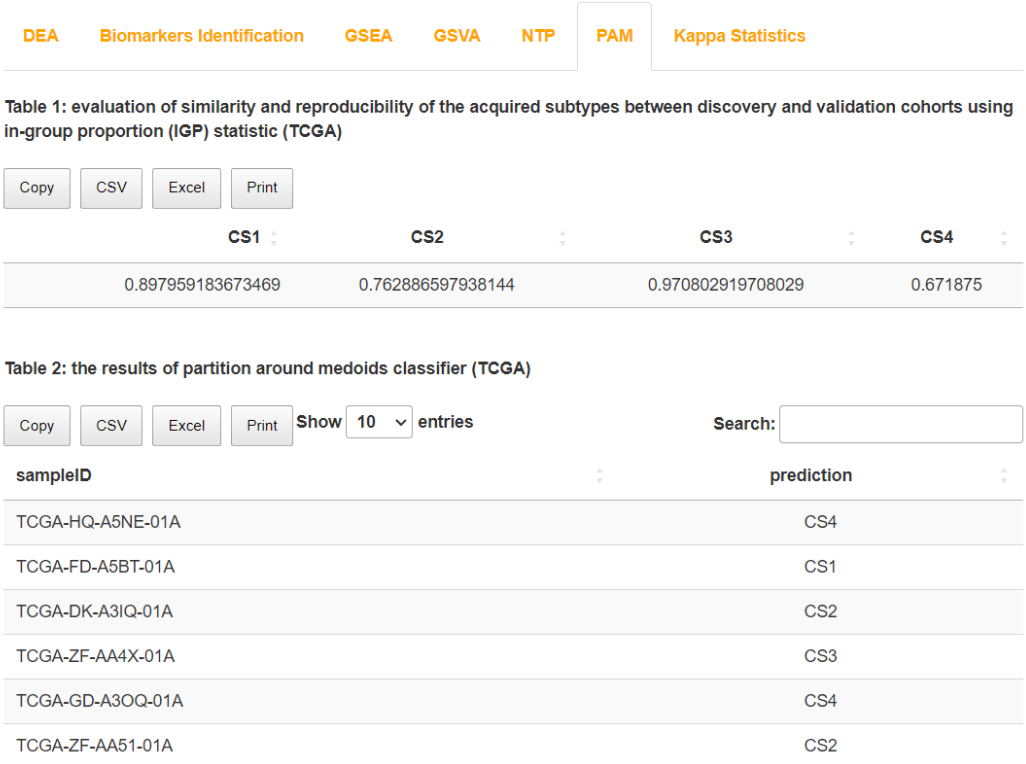

DEA Biomarkers Identification GSEA GSVA NTP PAM Kappa Statistics

Table 1: evaluation of similarity and reproducibility of the acquired subtypes between discovery and validation cohorts using in-group proportion (IGP) statistic (TCGA)

Copy CSV Excel Print

CS1	CS2	CS3	CS4
0.897959183673469	0.762886597938144	0.970802919708029	0.671875

Table 2: the results of partition around medoids classifier (TCGA)

Copy CSV Excel Print Show 10 entries Search:

sampleID	prediction
TCGA-HQ-A5NE-01A	CS4
TCGA-FD-A5BT-01A	CS1
TCGA-DK-A3IQ-01A	CS2
TCGA-ZF-AA4X-01A	CS3
TCGA-GD-A3OQ-01A	CS4
TCGA-ZF-AA51-01A	CS2

Result display for “Run partition around medoids classifier” on TCGA dataset

DEA | Biomarkers Identification | GSEA | GSVA | NTP | PAM | Kappa Statistics

Table 1: evaluation of similarity and reproducibility of the acquired subtypes between discovery and validation cohorts using in-group proportion (IGP) statistic (Validation)

Copy CSV Excel Print

	CS1	CS2	CS3	CS4
	0.941176470588235	0.727272727272727	0.846153846153846	0.818181818181818

Table 2: the results of partition around medoids classifier (Validation)

Copy CSV Excel Print Show 10 entries Search:

sampleID	prediction
FR_1_U133_2.CEL	CS3
FR_103_U133_2.CEL	CS3
FR_106_U133_2.CEL	CS2
FR_12_U133_2.CEL	CS4
FR_120_U133_2.CEL	CS1
FR_127_U133_2_2.CEL	CS3

Result display for “Run partition around medoids classifier” on validation dataset (affy)

DEA | Biomarkers Identification | GSEA | GSVA | NTP | PAM | Kappa Statistics

Table 1: evaluation of similarity and reproducibility of the acquired subtypes between discovery and validation cohorts using in-group proportion (IGP) statistic (Validation)

Copy CSV Excel Print

	CS1	CS2	CS3	CS4
	0.757142857142857	0.671875	0.905982905982906	0.628571428571429

Table 2: the results of partition around medoids classifier (Validation)

Copy CSV Excel Print Show 10 entries Search:

sampleID	prediction
GSM806803	CS3
GSM806804	CS2
GSM806805	CS3
GSM806810	CS1
GSM806814	CS1
GSM806822	CS2

Result display for “Run partition around medoids classifier” on validation dataset (illumina)

(7). Run consistency evaluation using Kappa statistics for TCGA dataset and validation datasets

I.Analysis introduction

“Run consistency evaluation using Kappa statistics” calculates Kappa statistics, and then generates heatmaps to evaluate the consistency between clustering results and prediction results or the consistency between prediction results derived from NTP and PAM.

II.Parameters setting guides

i.Run consistency evaluation using Kappa statistics

Finally, to evaluate the consistency between clustering results and prediction results or between NTP and PAM prediction results, you need to specify the prediction result sets obtained from NTP and PAM analyses:

Choose the results you have obtained in 'Run nearest template prediction' and 'Run partition around medoids classifier' procedures: Specify the obtained prediction result sets in NTP and PAM analyses based on the actual situation, which can be divided into four scenarios:

(1) If you choose to run nearest template prediction on the TCGA cohort using the NTP method to predict the subtypes of the TCGA dataset, a Kappa consistency comparison between the TCGA clustering results and the NTP-predicted TCGA results will be conducted. Refer to the documentation for the Run Kappa on TCGA cohort (CMOIC vs NTP)

parameter module for details.

(2) If you choose to run partition around medoids classifier on the TCGA cohort using the PAM method to predict the subtypes of the TCGA dataset, a Kappa consistency comparison between the TCGA clustering results and the PAM-predicted TCGA results will be conducted. Refer to the documentation for the Run Kappa on TCGA cohort (CMOIC vs PAM) parameter module for details.

(3) If you simultaneously choose to run nearest template prediction on the TCGA cohort and run partition around medoids classifier on the TCGA cohort, using both the NTP and PAM methods to predict the subtypes of the TCGA dataset, a Kappa consistency comparison between the NTP-predicted TCGA results and the PAM-predicted TCGA results will be conducted. Refer to the documentation for the Run Kappa on TCGA cohort (NTP vs PAM) parameter module for details.

(4) If you simultaneously choose to run nearest template prediction on the validation cohort and run partition around medoids classifier on the validation cohort, using both the NTP and PAM methods to predict the subtypes of the external validation dataset, a Kappa consistency comparison between the NTP-predicted external validation results and the PAM-predicted external validation results will be conducted. Refer to the documentation for the Run Kappa on validation cohort (NTP vs PAM) parameter module for details.

After specifying the prediction result sets obtained from NTP and PAM analyses, Kappa consistency comparisons will be conducted based on different scenarios. Please refer to the documentation for the following parameter modules for details: Run Kappa on tcga cohort (CMOIC vs NTP), Run Kappa on tcga cohort (CMOIC vs PAM), Run Kappa on tcga cohort (NTP vs PAM), Run Kappa on validation cohort (NTP vs PAM).

ii. Run Kappa on tcga cohort (CMOIC vs NTP)

Calculate the Kappa consistency between TCGA clustering results and NTP-predicted TCGA subtype results, and generate a consistency heatmap. The required parameters include:

Indicate the label of the first subtype: Specify a plotting abbreviation label for TCGA clustering results, which will be used for the y-axis and legend in the heatmap. The default label is CMOIC_TCGA.

Indicate the label of the second subtype: Specify a plotting abbreviation label for NTP-predicted TCGA subtype results, which will be used for the x-axis and legend in the heatmap. The default label is NTP_TCGA.

The width of output figure: Specify the width for the output heatmap PDF. The default value is 5.

The height of output figure: Specify the width for the output heatmap PDF. The default height is 5.

The name of the consistency heatmap: Specify the name for the output heatmap PDF, with the default value being constheatmap

(CMOIC_VS_NTP_TCGA).

After specifying all the parameters, click the "Process" button to perform the Kappa consistency comparison between TCGA clustering results and NTP-predicted TCGA subtypes. Once completed, feedback will be provided in the Kappa Statistics results module on the right side of the webpage. Users can choose to download the Kappa consistency heatmap PDF. Additionally, users need to perform Kappa consistency comparisons for other scenarios based on the actual situation. Please refer to the guidance document for the parameters in "Run Kappa on tcga cohort (CMOIC vs PAM)", "Run Kappa on tcga cohort (NTP vs PAM)", and "Run Kappa on validation cohort (NTP vs PAM)" for further details.

If you have completed the necessary Kappa consistency comparisons, go to the "Finish" parameter module and click the "Finish" button. Refer to the guidance document for the "Finish" parameter module for detailed instructions.

iii.Run Kappa on tcga cohort (CMOIC vs PAM)

Calculate the Kappa consistency between TCGA clustering results and PAM-predicted TCGA subtypes, and generate a consistency heatmap. The required parameters include:

Indicate the label of the first subtype: Specify a plotting abbreviation label for TCGA clustering results. This label will be used for the y-axis and legend of the heatmap. The default is set to CMOIC_TCGA.

Indicate the label of the second subtype: Specify a plotting abbreviation label for PAM-predicted TCGA subtype results. This label will be used for the x-axis and legend of the heatmap. The default is set to PAM_TCGA.

The width of output figure: Specify the width of the output heatmap PDF, with a default value of 5.

The height of output figure: Specify the height of the output heatmap PDF, with a default value of 5.

The name of the consistency heatmap: Specify the name of the output heatmap PDF, with a default value of "constheatmap(CMOIC_VS_PAM_TCGA)".

After specifying all the parameters, click the "Process" button to perform the Kappa consistency comparison between TCGA clustering results and PAM-predicted TCGA subtypes. Once completed, the feedback will be provided in the Kappa Statistics module on the right side of the webpage, where users can choose to download the Kappa consistency heatmap PDF. Additionally, users may need to conduct Kappa consistency comparisons for other scenarios based on the actual situation, as outlined in the documentation for the "Run Kappa on tcga cohort (CMOIC vs NTP)", "Run Kappa on tcga cohort (NTP vs PAM)", and "Run Kappa on validation cohort (NTP vs PAM)" parameter modules.

If the necessary Kappa consistency comparisons are completed, proceed to the "Finish" parameter module, and click the "Finish" button as

outlined in the Finish parameter module documentation.

iv.Run Kappa on tcga cohort (NTP vs PAM)

Calculate the Kappa consistency between NTP predicted TCGA subtype results and PAM predicted TCGA subtype results, and generate a consistency heatmap. The parameters to specify include:

Indicate the label of the first subtype: Specify a plotting abbreviation label for the NTP-predicted TCGA subtypes, which will be used for the vertical axis and legend in the heatmap. The default is set to NTP_TCGA.

Indicate the label of the second subtype: Specify a plotting abbreviation label for the PAM-predicted TCGA subtypes, which will be used for the horizontal axis and legend in the heatmap. The default is set to PAM_TCGA.

The width of output figure: Specify the width of the output heatmap PDF, with a default value of 5.

The height of output figure: Specify the height of the output heatmap PDF, with a default value of 5.

The name of the consistency heatmap: Specify the name of the output heatmap PDF, with a default value of constheatmap (NTP_VS_PAM_TCGA).

Once all the parameters are specified, click the "Process" button to perform the Kappa consistency comparison between NTP-predicted TCGA subtypes and PAM-predicted TCGA subtypes. After completion, feedback will be provided in the Kappa Statistics results module on the right side of the webpage. Users can choose to download the Kappa consistency

heatmap PDF. Additionally, users may need to perform Kappa consistency comparisons for other scenarios based on the actual situation. Refer to the documentation for the parameters in the "Run Kappa on tcga cohort (CMOIC vs NTP)", "Run Kappa on tcga cohort (CMOIC vs PAM)", and "Run Kappa on validation cohort (NTP vs PAM)" modules for more details.

If you have completed the necessary Kappa consistency comparisons, go to the "Finish" module, and click the "Finish" button. Refer to the documentation for the "Finish" module for detailed instructions.

v.Run Kappa on validation cohort (NTP vs PAM)

To calculate the Kappa consistency between NTP-predicted external validation set subtypes and PAM-predicted external validation set subtypes and generate a consistency heatmap, you need to specify the following parameters:

Indicate the label of the first subtype: Specify a plotting abbreviation label for NTP-predicted external validation set subtypes, which will be used for the y-axis and legend in the heatmap. The default label is "NTP_Validation".

Indicate the label of the second subtype: Specify a plotting abbreviation label for PAM-predicted external validation set subtypes, which will be used for the x-axis and legend in the heatmap. The default label is "PAM_Validation".

The width of output figure: Specify the width of the output heatmap PDF.

The default value is 5.

The height of output figure: Specify the height of the output heatmap PDF. The default value is 5.

The name of the consistency heatmap: Specify the name of the output heatmap PDF. The default value is constheatmap (NTP_VS_PAM_Validation).

After setting all the parameters, click the "Process" button to perform the Kappa consistency comparison between NTP-predicted external validation set subtypes and PAM-predicted external validation set subtypes. Once completed, feedback will be provided in the Kappa Statistics results module on the right side of the webpage. Users can choose to download the Kappa consistency heatmap PDF. Additionally, users may need to perform Kappa consistency comparisons for other scenarios based on actual situations. For detailed instructions, refer to the documentation for the "Run Kappa on tcga cohort (CMOIC vs NTP)," "Run Kappa on tcga cohort (CMOIC vs PAM)," and "Run Kappa on tcga cohort (NTP vs PAM)" parameters in the module.

If you have completed the necessary Kappa consistency comparisons, go to the "Finish" parameter module, and click the "Finish" button. For detailed instructions, refer to the documentation for the "Finish" parameter module.

vi.Finish

Click the "Finish" button, and at this point, you have completed the entire analysis workflow on the MOVICShiny platform. The Kappa Statistics results module on the right side of the webpage will provide feedback to the user. At this stage, you can download and organize the corresponding result files, including the RData result set, plot PDFs, and result tables, for further analysis based on the generated output.

Additionally, if you have obtained NTP or PAM prediction results for an external validation set, you can return to the COMP Module module to select "Validation" for comparing subtypes in the external validation set. It's important to note that users should choose the appropriate analysis in the COMP Module based on the type of omics data available in the external validation set. For example, analyses like "Compare survival outcome", "Compare clinical features" and "Compare agreement with other subtypes" require clinical survival data, while "Compare drug sensitivity" analysis requires mRNA or lncRNA data. Analyses like "Compare mutational frequency" and "Compare total mutation burden" need binary somatic mutation data, and "Compare fraction genome altered" analysis requires copy number alterations data. For detailed instructions, please refer to the guidance documentation for each parameter in the COMP Module.

III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Run consistency evaluation using Kappa statistics”, and “Choose the results you have obtained in 'Run nearest template prediction' and 'Run partition around medoids classifier' procedures” chooses all the options. Then, we set the parameters containing “Indicate the label of the first subtype”, “Indicate the label of the second subtype”, “The width of output figure”, “The height of output figure” and “The name of the consistency heatmap” for “Run Kappa on tcga cohort (CMOIC vs NTP)”, “Run Kappa on tcga cohort (CMOIC vs PAM)”, “Run Kappa on tcga cohort (NTP vs PAM)” and “Run Kappa on validation cohort (NTP vs PAM)” in turn. Here, “CMOIC” represents the clustering results, “NTP” represents the prediction results using NTP method, and “PAM” represents the prediction results using PAM method. After that, we click the “Process” buttons to run consistency evaluation using Kappa statistics sequentially. Finally, we click the “Finish” button to finish the whole multi-omics analysis procedure.

Run Kappa on tcga cohort (CMOIC vs NTP)

Indicate the label of the first subtype

CMOIC_TCGA

Indicate the label of the second subtype

NTP_TCGA

The width of output figure

5

The height of output figure

5

The name of the consistency heatmap

constheatmap(CMOIC_VS_NTP_TCGA)

Process

Run Kappa on tcga cohort (CMOIC vs PAM)

Indicate the label of the first subtype

CMOIC_TCGA

Indicate the label of the second subtype

PAM_TCGA

The width of output figure

5

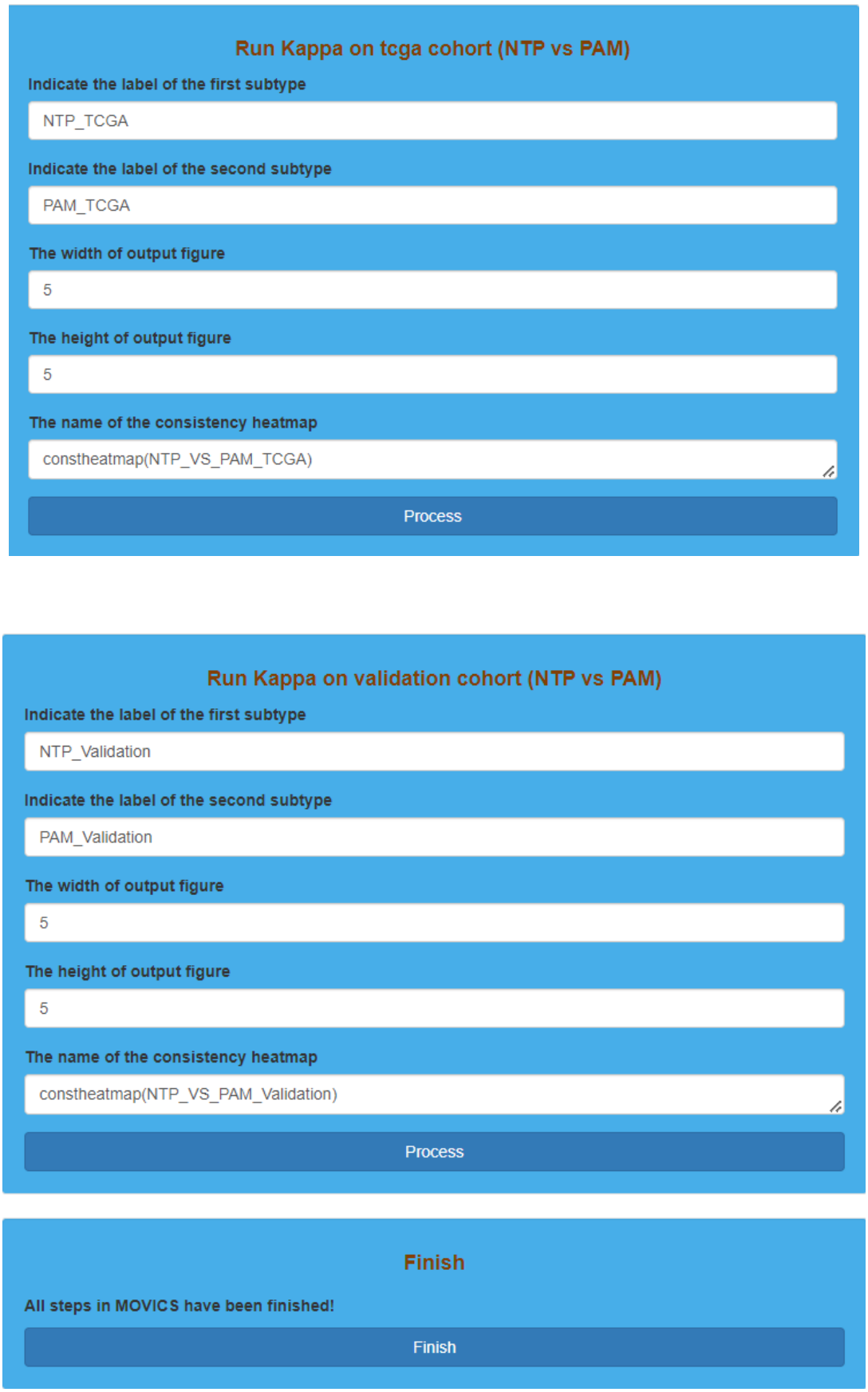
The height of output figure

5

The name of the consistency heatmap

constheatmap(CMOIC_VS_PAM_TCGA)

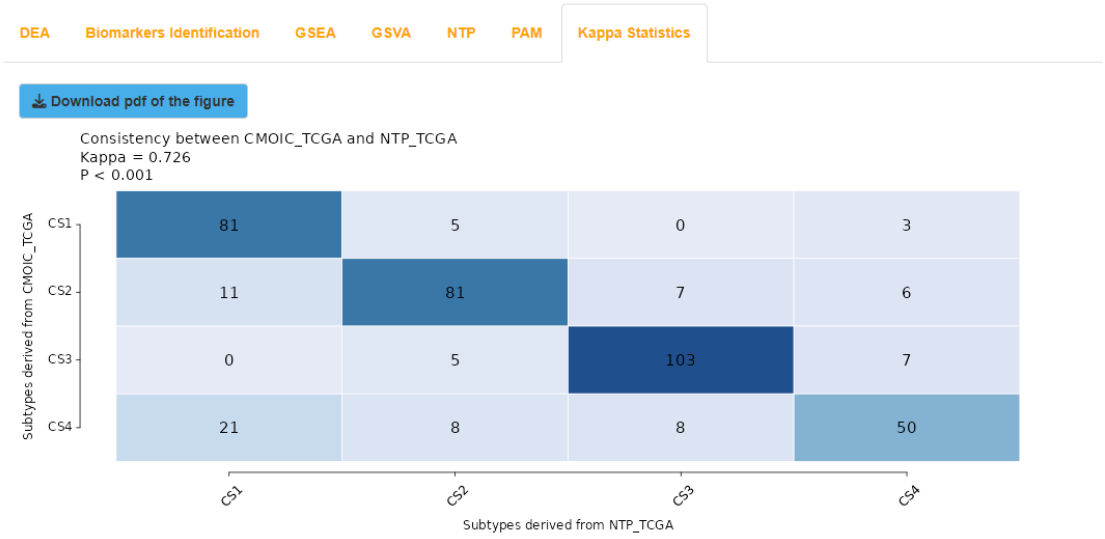
Process



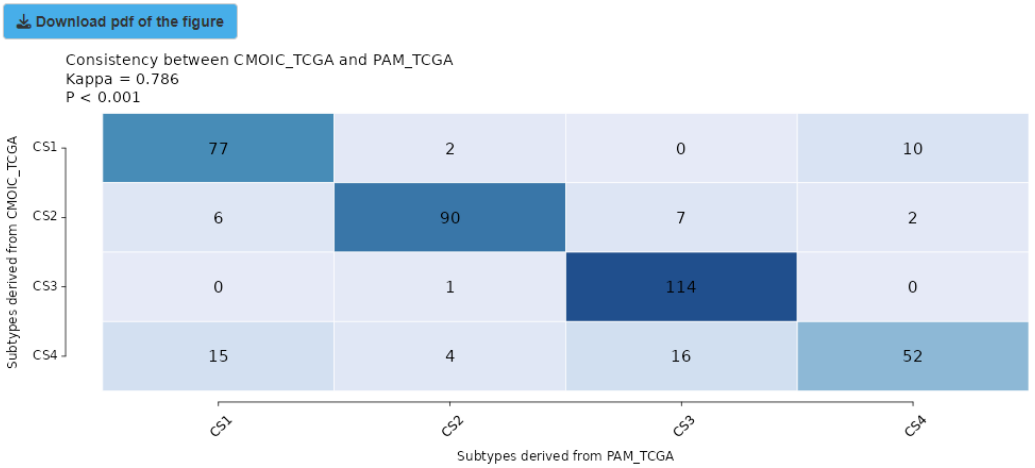
Parameter settings for “Run consistency evaluation using Kappa statistics”

Result Display:

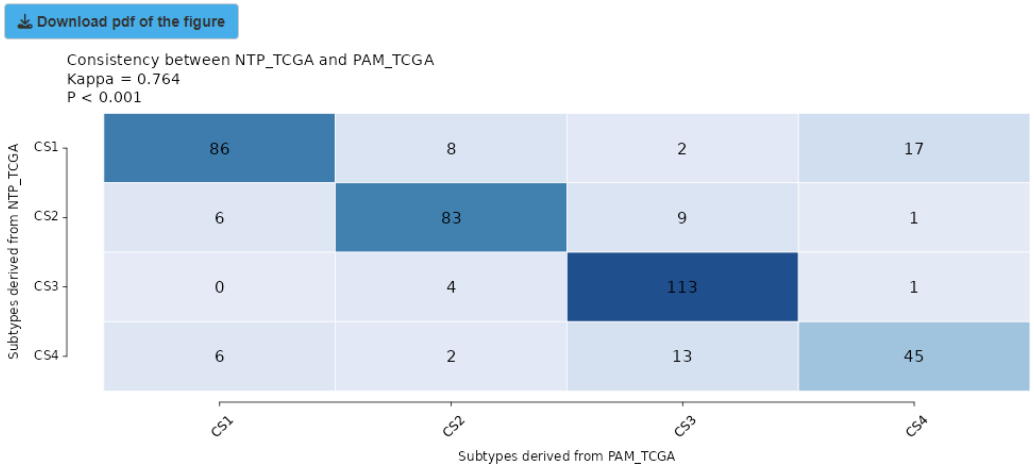
The software will calculate Kappa statistics, and then generate heatmaps to evaluate the consistency between clustering results and prediction results or the consistency between prediction results derived from NTP and PAM method. The greater the Kappa statistic is, the higher the degree of consistency is. Generally speaking, Kappa statistic that is greater than 0.4 indicates a high degree of consistency, and greater than 0.7 indicates an extremely high degree of consistency.



The process of running consistency evaluation using Kappa statistics between current subtypes derived from multi-omics clustering and NTP-predicted subtypes on tcga cohort has been finished, you can download and check the figure. Now keep on the next part of 'Run consistency evaluation using Kappa statistics'.

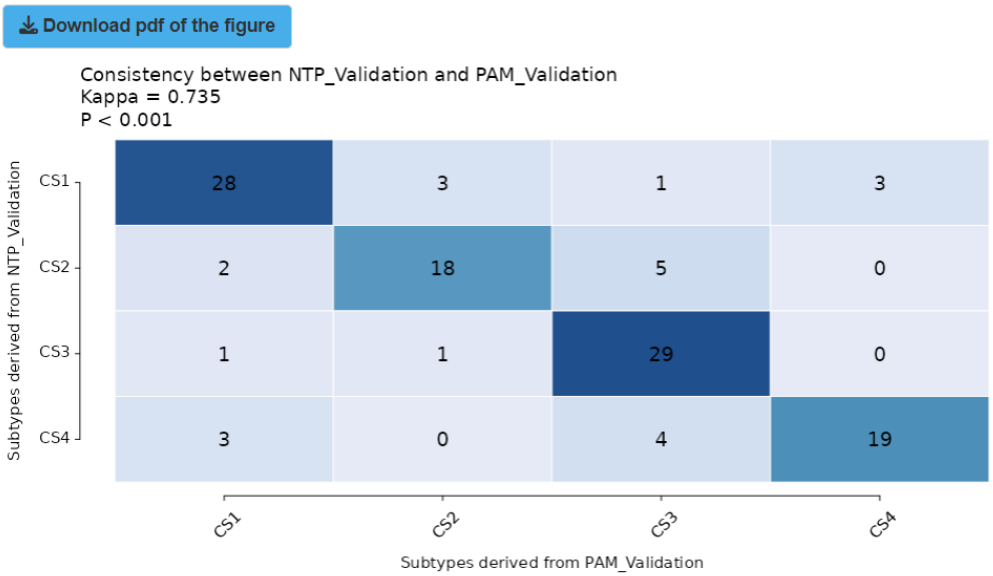


The process of running consistency evaluation using Kappa statistics between current subtypes derived from multi-omics clustering and PAM-predicted subtypes on tcga cohort has been finished, you can download and check the figure. Now keep on the next part of 'Run consistency evaluation using Kappa statistics'.



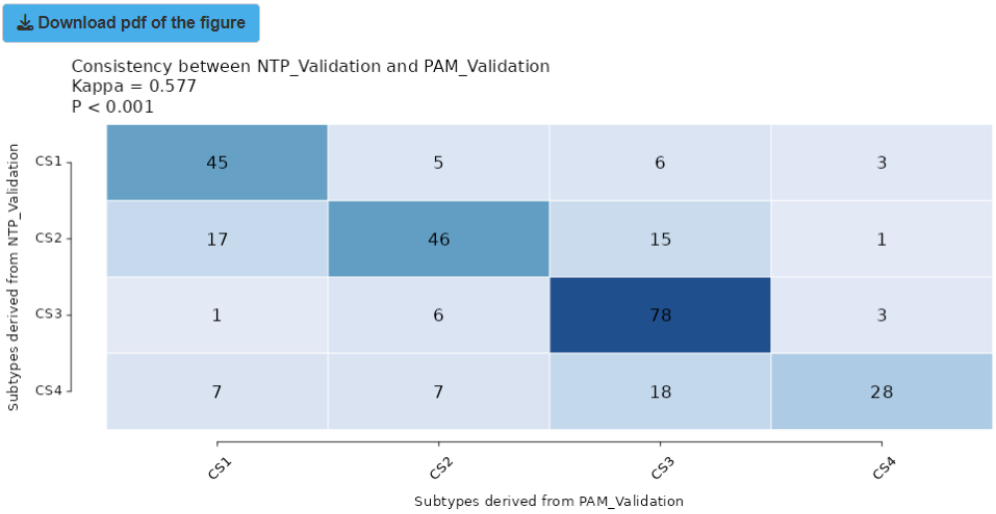
The process of running consistency evaluation using Kappa statistics between NTP-predicted subtypes and PAM-predicted subtypes on tcga cohort has been finished, you can download and check the figure. Now keep on the next part of 'Run consistency evaluation using Kappa statistics'.

Result display for “Run consistency evaluation using Kappa statistics” on TCGA dataset



The process of running consistency evaluation using Kappa statistics between NTP-predicted subtypes and PAM-predicted subtypes on validation cohort has been finished, you can download and check the figure. Now all steps in 'RUN Module' have been finished!

Result display for “Run consistency evaluation using Kappa statistics” on validation dataset (affy)



The process of running consistency evaluation using Kappa statistics between NTP-predicted subtypes and PAM-predicted subtypes on validation cohort has been finished, you can download and check the figure. Now all steps in 'RUN Module' have been finished!

Now all steps in MOVICS have been finished, you can download and check all the tables and figures, as well as the '.RData' files. Then you can further process the obtained results, thanks for using MOVICS RShiny and we are looking forward to your valuable feedback and another visit!

Result display for “Run consistency evaluation using Kappa statistics” on validation

dataset (illumina)

The fifth step: COMP Module (Validation dataset)

The prediction results of external validation dataset from both NTP and PAM methods can be used to carry out analyses in “COMP Module”, which can further validate the accuracy and reliability of the clustering results. The external validation datasets should contain corresponding types of data for specified analysis in “COMP Module”. In our given examples, since the two validation datasets only include mRNA as well as clinical and survival data, we only finish “Compare survival outcome”, “Compare clinical features”, “Compare drug sensitivity” and “Compare agreement with other subtypes”.

(1). Compare survival outcome for validation dataset

I.Analysis introduction

Details can be found in “The third step: COMP Module (TCGA dataset)”.

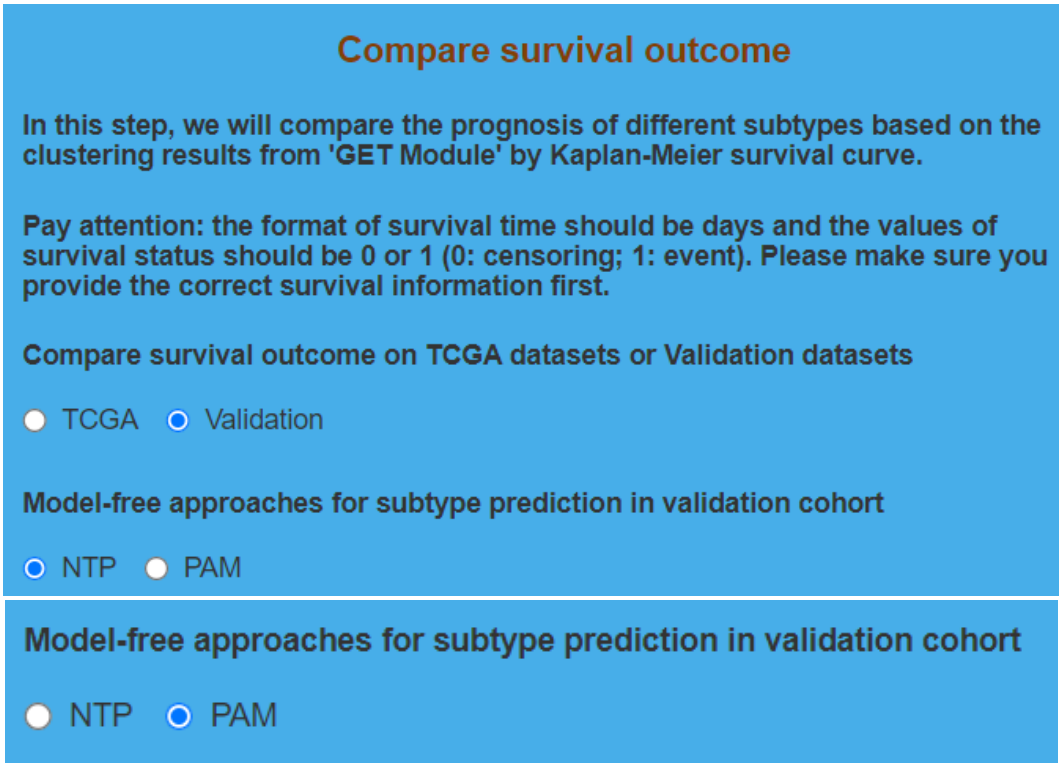
II.Parameters setting guides

Details can be found in “The third step: COMP Module (TCGA dataset)”.

III.Examples with results interpretation

Parameter Settings: “Module switching options” on the upper left of the software chooses “COMP Module”, and then the “Steps” chooses “Compare survival outcome”, and “Compare survival outcome on TCGA datasets or Validation datasets” chooses “Validation”. After that, “Model-free approaches for subtype prediction in validation cohort” chooses “NTP”

and “PAM” in turn, and other parameters keep unchanged. Finally, we click the “Process” button to compare survival outcomes among subtypes derived from NTP and PAM methods for two validation datasets in turn.

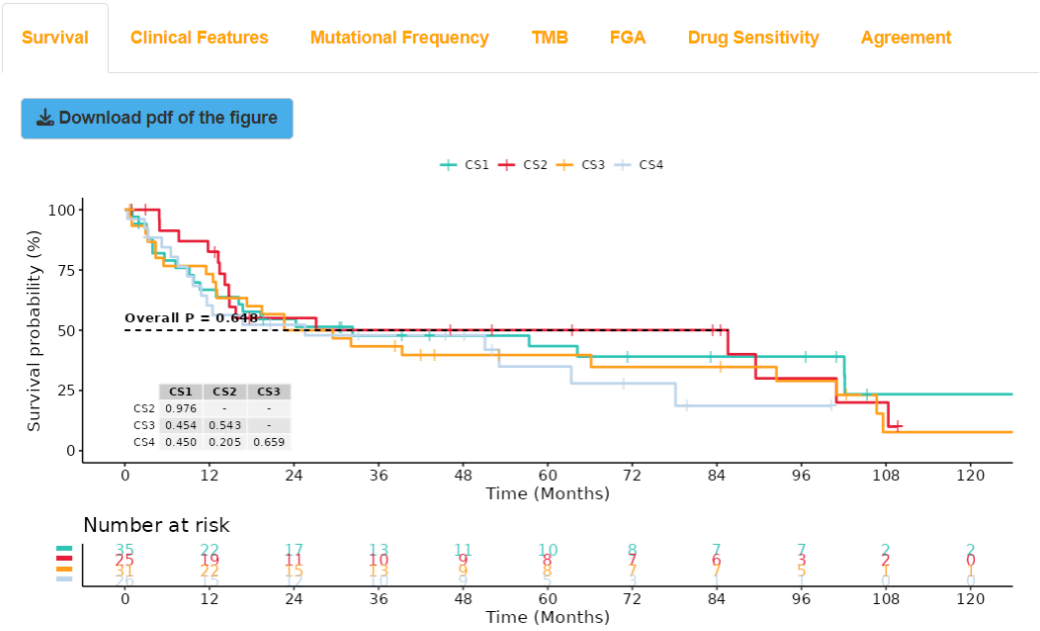


Parameter settings for “Compare survival outcome” on validation dataset

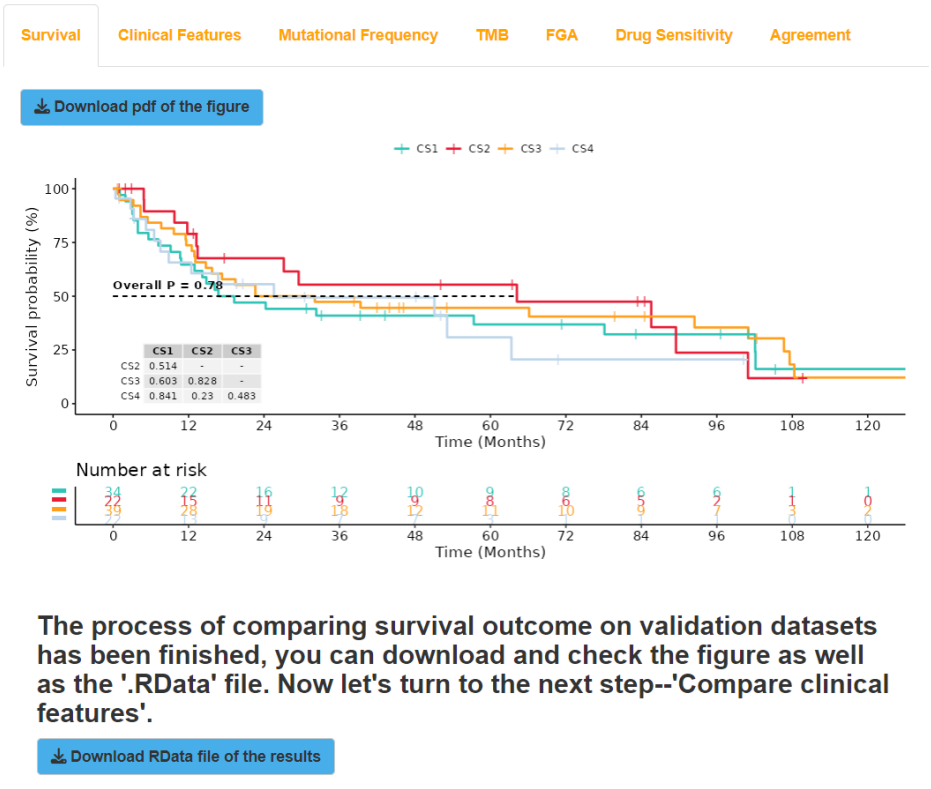
Result Display:

Like the TCGA dataset, the software will also generate Kaplan-Meier curve for each subtype derived from NTP and PAM methods for two validation datasets, and compare the survival differences among each subtype at the

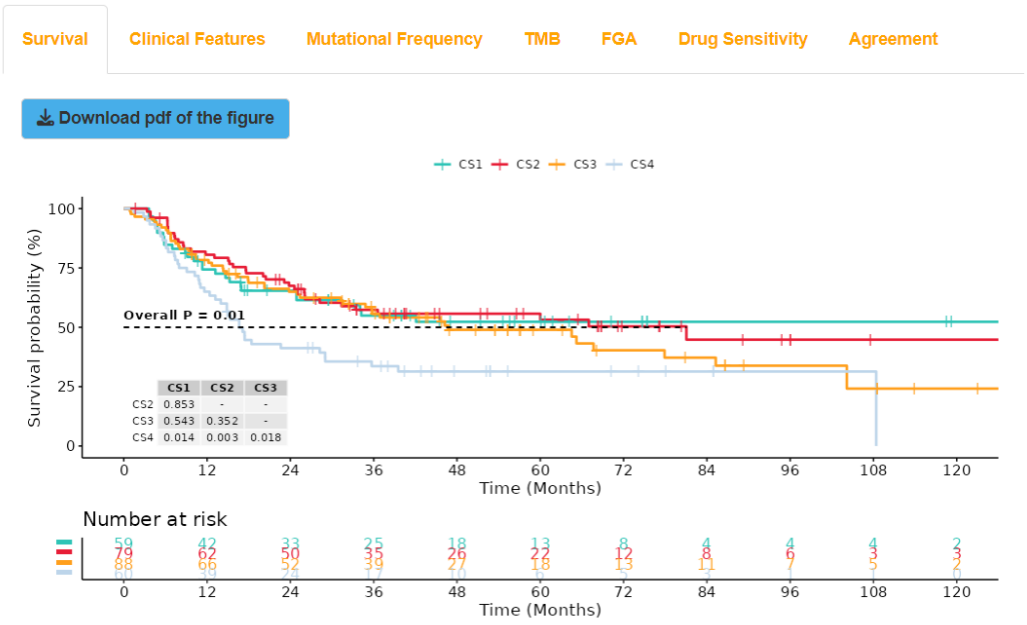
same time.



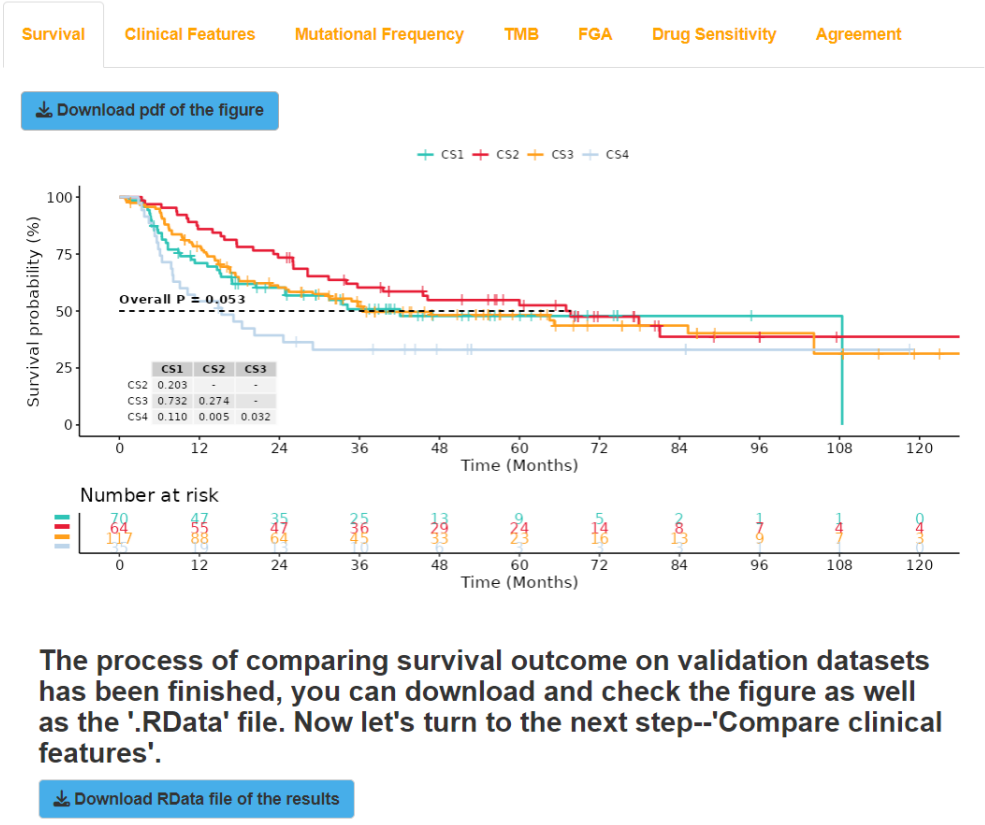
Result display for “Compare survival outcome” on validation dataset (affy+NTP)



Result display for “Compare survival outcome” on validation dataset (affy+PAM)



Result display for “Compare survival outcome” on validation dataset (illumina+NTP)



Result display for “Compare survival outcome” on validation dataset (illumina+PAM)

(2). Compare clinical features for validation dataset

I.Analysis introduction

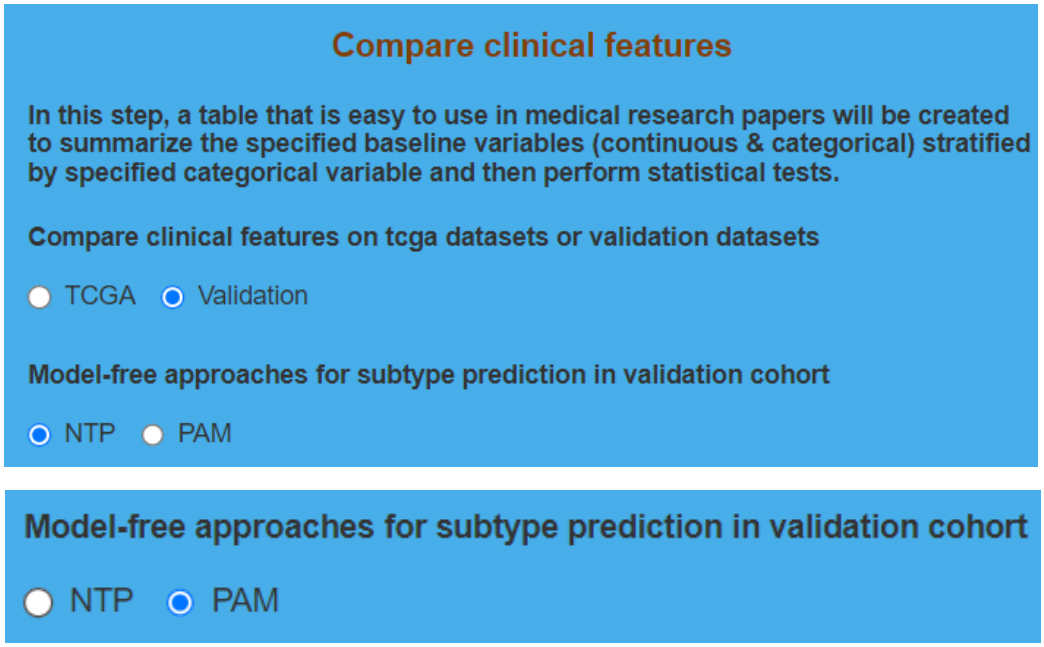
Details can be found in “The third step: COMP Module (TCGA dataset)”.

II.Parameters setting guides

Details can be found in “The third step: COMP Module (TCGA dataset)”.

III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Compare clinical features”, and “Compare clinical features on tcga datasets or validation datasets” chooses “Validation”. Other parameters keep unchanged, and then we click the “Process” button to compare clinical features among subtypes derived from NTP and PAM methods for two validation datasets in turn.



Parameter settings for “Compare clinical features” on validation dataset

Result Display:

Like the TCGA dataset, the software will also generate tables to display the

comparison results of clinical features among subtypes derived from NTP and PAM methods for two validation datasets.

Survival | Clinical Features | Mutational Frequency | TMB | FGA | Drug Sensitivity | Agreement

Table: Summarization of clinical variables stratified by current subtypes (Validation)

Copy CSV Excel Print

		level	CS1	CS2	CS3	CS4	p	test
1	n		35	25	31	26		
2	Age (%)	<=70	21 (60.0)	11 (44.0)	23 (74.2)	10 (38.5)	0.028	exact
3		>70	14 (40.0)	14 (56.0)	8 (25.8)	16 (61.5)		
4	Gender (%)	FEMALE	11 (31.4)	7 (28.0)	6 (19.4)	3 (11.5)	0.272	exact
5		MALE	24 (68.6)	18 (72.0)	25 (80.6)	23 (88.5)		
6	T_Stage (%)	T2	8 (22.9)	7 (28.0)	12 (38.7)	5 (19.2)	0.739	exact
7		T3	19 (54.3)	13 (52.0)	12 (38.7)	15 (57.7)		
8		T4	8 (22.9)	5 (20.0)	7 (22.6)	6 (23.1)		
9	futime (median [IQR])		600.00 [195.65, 2067.35]	480.00 [390.00, 2541.93]	690.00 [259.71, 1818.79]	570.00 [239.74, 1582.07]	0.879	nonnorm
10	fustat (%)	0	13 (37.1)	10 (40.0)	7 (22.6)	9 (34.6)	0.507	exact
11		1	22 (62.9)	15 (60.0)	24 (77.4)	17 (65.4)		

Result display for “Compare clinical features” on validation dataset (affy+NTP)

Survival | Clinical Features | Mutational Frequency | TMB | FGA | Drug Sensitivity | Agreement

Table: Summarization of clinical variables stratified by current subtypes (Validation)

Copy CSV Excel Print

		level	CS1	CS2	CS3	CS4	p	test
1	n		34	22	39	22		
2	Age (%)	<=70	21 (61.8)	11 (50.0)	24 (61.5)	9 (40.9)	0.367	exact
3		>70	13 (38.2)	11 (50.0)	15 (38.5)	13 (59.1)		
4	Gender (%)	FEMALE	10 (29.4)	8 (36.4)	6 (15.4)	3 (13.6)	0.150	exact
5		MALE	24 (70.6)	14 (63.6)	33 (84.6)	19 (86.4)		
6	T_Stage (%)	T2	6 (17.6)	6 (27.3)	16 (41.0)	4 (18.2)	0.136	exact
7		T3	17 (50.0)	14 (63.6)	15 (38.5)	13 (59.1)		
8		T4	11 (32.4)	2 (9.1)	8 (20.5)	5 (22.7)		
9	futime (median [IQR])		548.99 [234.58, 2069.57]	683.96 [313.24, 2396.06]	690.00 [353.35, 2225.54]	555.00 [169.53, 1536.21]	0.565	nonnorm
10	fustat (%)	0	9 (26.5)	10 (45.5)	11 (28.2)	9 (40.9)	0.371	exact
11		1	25 (73.5)	12 (54.5)	28 (71.8)	13 (59.1)		

Result display for “Compare clinical features” on validation dataset (affy+PAM)

Survival | Clinical Features | Mutational Frequency | TMB | FGA | Drug Sensitivity | Agreement

Table: Summarization of clinical variables stratified by current subtypes (Validation)

Copy CSV Excel Print

		level	CS1	CS2	CS3	CS4	p	test
1	n		59	79	88	60		
2	Age (%)	<=70	42 (71.2)	52 (65.8)	57 (64.8)	30 (50.0)	0.087	exact
3		>70	17 (28.8)	27 (34.2)	31 (35.2)	30 (50.0)		
4	Gender (%)	FEMALE	10 (23.3)	13 (20.6)	12 (17.6)	12 (30.0)	0.490	exact
5		MALE	33 (76.7)	50 (79.4)	56 (82.4)	28 (70.0)		
6	T_Stage (%)	T2	38 (64.4)	52 (65.8)	45 (51.1)	31 (51.7)	0.236	exact
7		T3	17 (28.8)	20 (25.3)	29 (33.0)	24 (40.0)		
8		T4	4 (6.8)	7 (8.9)	14 (15.9)	5 (8.3)		
9	futime (median [IQR])		900.00 [335.55, 1706.55]	992.10 [472.05, 1995.00]	987.14 [367.50, 1686.14]	507.00 [268.42, 1172.25]	0.043	nonnorm
10	fustat (%)	0	33 (55.9)	43 (54.4)	40 (45.5)	19 (31.7)	0.020	exact
11		1	26 (44.1)	36 (45.6)	48 (54.5)	41 (68.3)		

Result display for “Compare clinical features” on validation dataset (illumina+NTP)

Survival | Clinical Features | Mutational Frequency | TMB | FGA | Drug Sensitivity | Agreement

Table: Summarization of clinical variables stratified by current subtypes (Validation)

Copy CSV Excel Print

		level	CS1	CS2	CS3	CS4	p	test
1	n		70	64	117	35		
2	Age (%)	<=70	42 (60.0)	42 (65.6)	75 (64.1)	22 (62.9)	0.921	exact
3		>70	28 (40.0)	22 (34.4)	42 (35.9)	13 (37.1)		
4	Gender (%)	FEMALE	12 (23.1)	11 (23.9)	13 (14.4)	11 (42.3)	0.031	exact
5		MALE	40 (76.9)	35 (76.1)	77 (85.6)	15 (57.7)		
6	T_Stage (%)	T2	41 (58.6)	34 (53.1)	70 (59.8)	21 (60.0)	0.325	exact
7		T3	24 (34.3)	26 (40.6)	31 (26.5)	9 (25.7)		
8		T4	5 (7.1)	4 (6.2)	16 (13.7)	5 (14.3)		
9	futime (median [IQR])		713.55 [270.00, 1255.50]	1309.50 [720.69, 2116.40]	912.00 [372.00, 1632.00]	453.00 [190.50, 1233.00]	<0.001	nonnorm
10	fustat (%)	0	36 (51.4)	31 (48.4)	56 (47.9)	12 (34.3)	0.411	exact
11		1	34 (48.6)	33 (51.6)	61 (52.1)	23 (65.7)		

Result display for “Compare clinical features” on validation dataset (illumina+PAM)

(3). Compare drug sensitivity for validation dataset

I.Analysis introduction

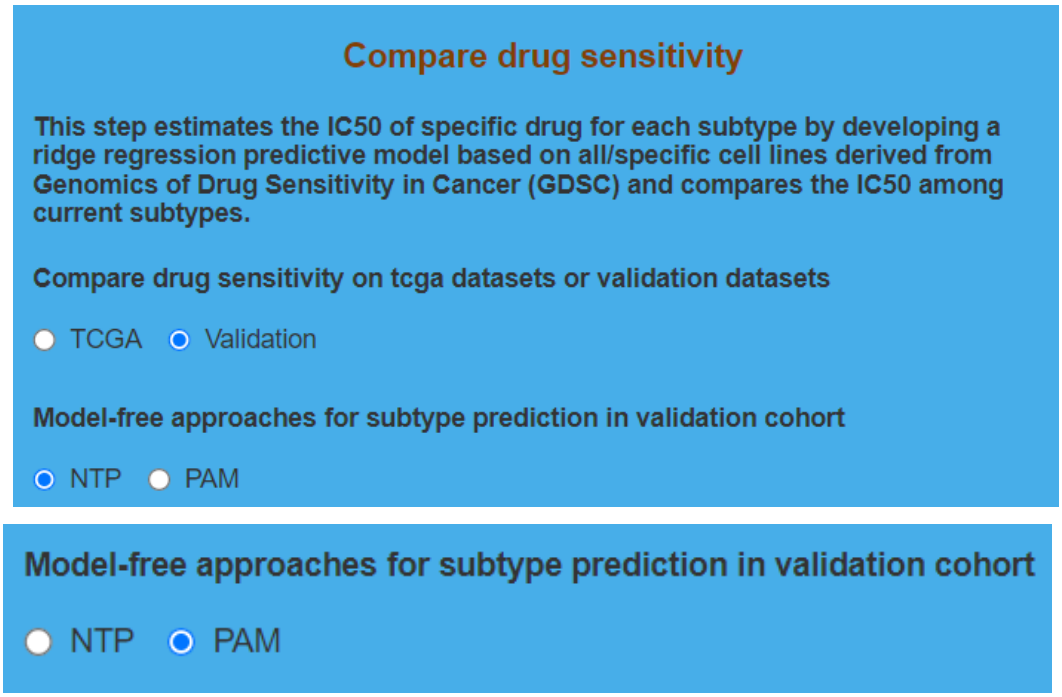
Details can be found in “The third step: COMP Module (TCGA dataset)”.

II.Parameters setting guides

Details can be found in “The third step: COMP Module (TCGA dataset)”.

III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Compare drug sensitivity”, and “Compare drug sensitivity on tcga datasets or validation datasets” chooses “Validation”. Other parameters keep unchanged, and then we click the “Process” button to compare drug sensitivity among subtypes derived from NTP and PAM methods for two validation datasets in turn.



Parameter settings for “Compare drug sensitivity” on validation dataset

Result Display:

Like the TCGA dataset, the software will also generate box-violin plots to

show the IC_{50} comparison results for drugs among subtypes derived from NTP and PAM methods for two validation datasets, and the estimated IC_{50} of each sample in two validation datasets for drugs will be also displayed through tables.

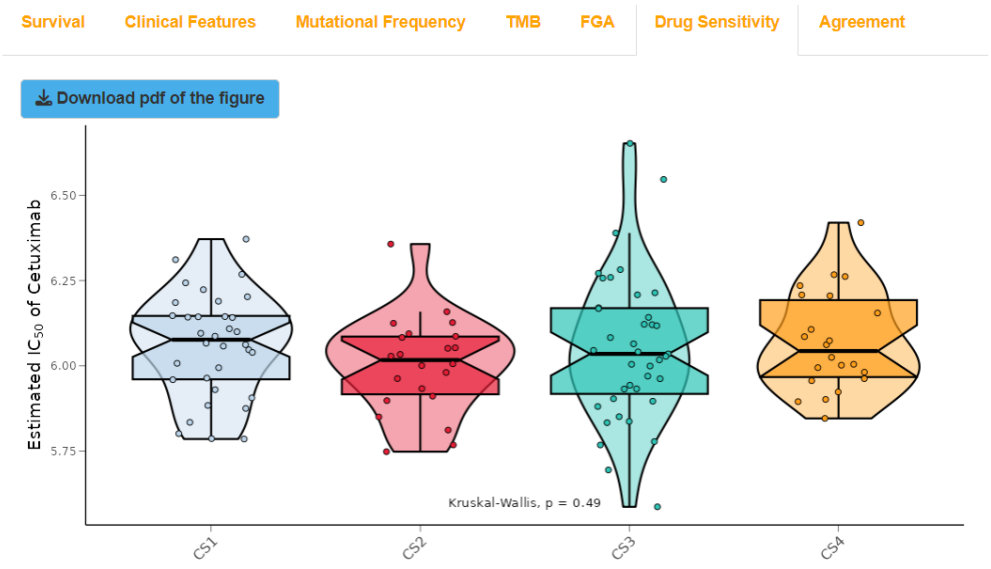


Table: Comparison of estimated IC50 (Validation) for Cetuximab among identified subtypes

Copy CSV Excel Print Show 10 entries Search:

	Est.IC50	Subtype
FR_120_U133_2.CEL	1.87491573880593	CS1
FR_14_U133_2.CEL	2.02567634852706	CS1
FR_181_U133_2.CEL	1.99193585977775	CS1

Result display for “Compare drug sensitivity” on validation dataset (affy+NTP+ Cetuximab)

Survival | Clinical Features | Mutational Frequency | TMB | FGA | Drug Sensitivity | Agreement

Download pdf of the figure

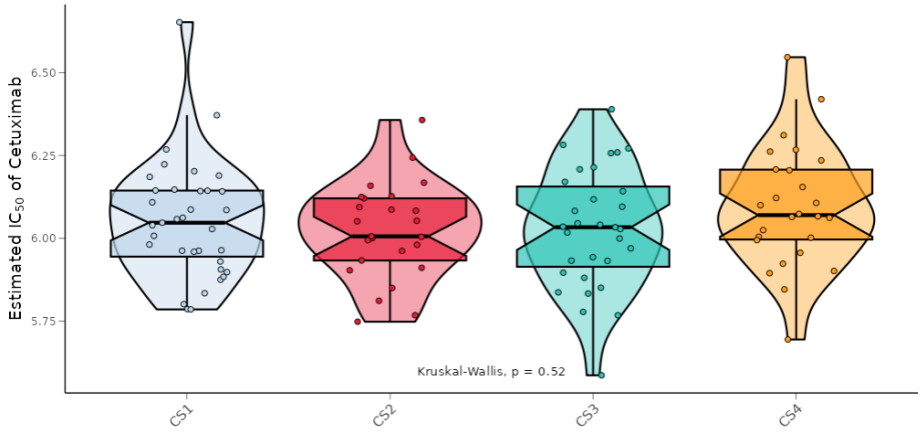


Table: Comparison of estimated IC50 (Validation) for Cetuximab among identified subtypes

Copy | CSV | Excel | Print | Show 10 entries | Search:

	Est.IC50	Subtype
FR_120_U133_2.CEL	5.78698094205484	CS1
FR_14_U133_2.CEL	6.05795247622992	CS1
FR_181_U133_2.CEL	5.96264550272376	CS1

Result display for “Compare drug sensitivity” on validation dataset (affy+PAM+ Cetuximab)

Survival | Clinical Features | Mutational Frequency | TMB | FGA | Drug Sensitivity | Agreement

Download pdf of the figure

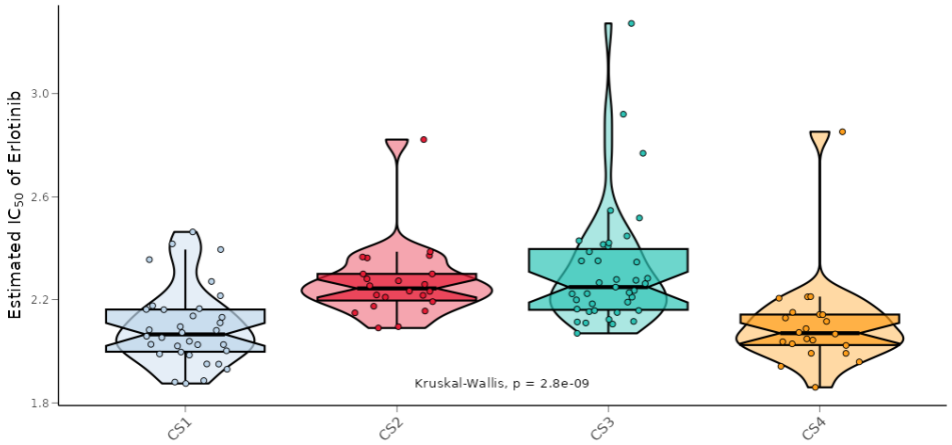


Table: Comparison of estimated IC50 (Validation) for Erlotinib among identified subtypes

Copy | CSV | Excel | Print | Show 10 entries | Search:

	Est.IC50	Subtype
FR_120_U133_2.CEL	1.87491573880593	CS1
FR_14_U133_2.CEL	2.02567634852706	CS1
FR_181_U133_2.CEL	1.99193585977775	CS1

Result display for “Compare drug sensitivity” on validation dataset (affy+NTP+ Erlotinib)

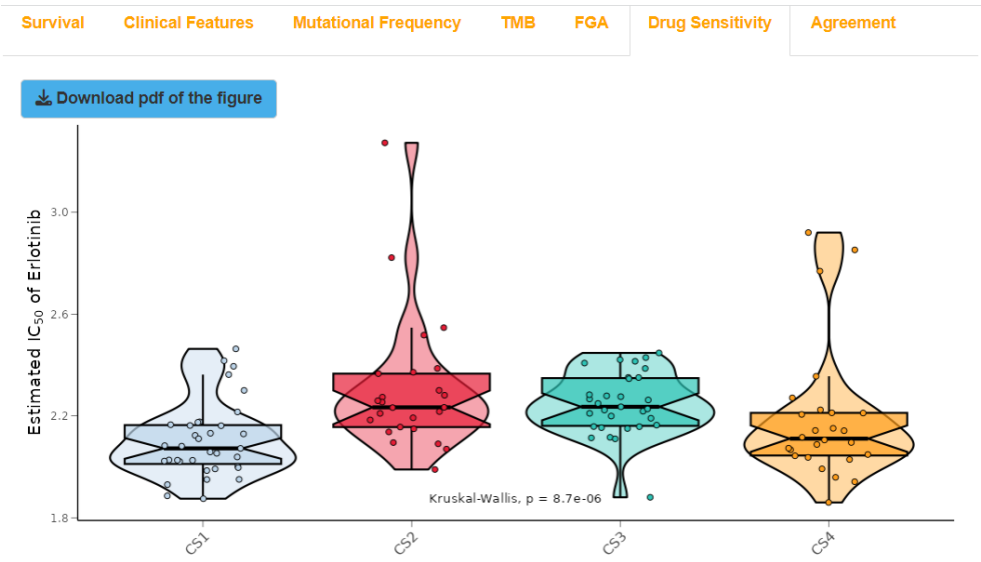


Table: Comparison of estimated IC50 (Validation) for Erlotinib among identified subtypes

Copy CSV Excel Print Show 10 entries Search:

	Est.IC50	Subtype
FR_120_U133_2.CEL	1.87491573880593	CS1
FR_14_U133_2.CEL	2.02567634852706	CS1
FR_181_U133_2.CEL	1.99193585977775	CS1

Result display for “Compare drug sensitivity” on validation dataset (affy+PAM+ Erlotinib)

Survival | Clinical Features | Mutational Frequency | TMB | FGA | Drug Sensitivity | Agreement

Download pdf of the figure

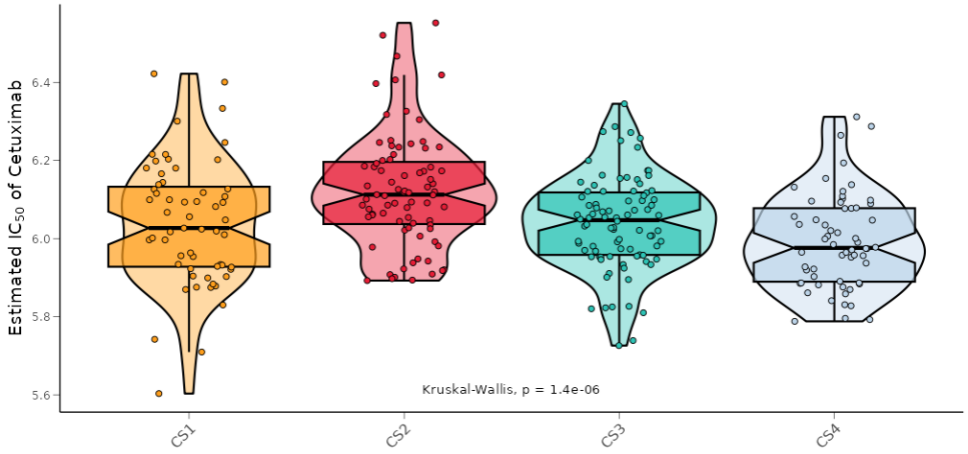


Table: Comparison of estimated IC50 (Validation) for Cetuximab among identified subtypes

Copy | CSV | Excel | Print | Show 10 entries | Search:

	Est.IC50	Subtype
GSM806810	6.05597392190824	CS1
GSM806864	5.9341718070258	CS1
GSM806868	6.02420552926831	CS1

Result display for “Compare drug sensitivity” on validation dataset (illumina+NTP+ Cetuximab)

Survival | Clinical Features | Mutational Frequency | TMB | FGA | Drug Sensitivity | Agreement

Download pdf of the figure

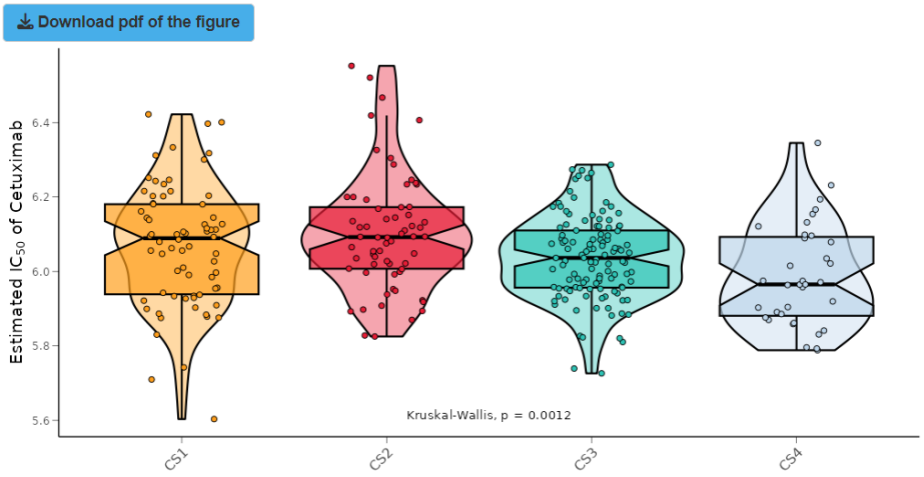


Table: Comparison of estimated IC50 (Validation) for Cetuximab among identified subtypes

Copy | CSV | Excel | Print | Show 10 entries | Search:

	Est.IC50	Subtype
GSM806810	6.05597392190822	CS1
GSM806814	6.06247363049068	CS1
GSM806864	5.9341718070258	CS1

Result display for “Compare drug sensitivity” on validation dataset (illumina+PAM+ Cetuximab)

Survival | Clinical Features | Mutational Frequency | TMB | FGA | Drug Sensitivity | Agreement

Download pdf of the figure

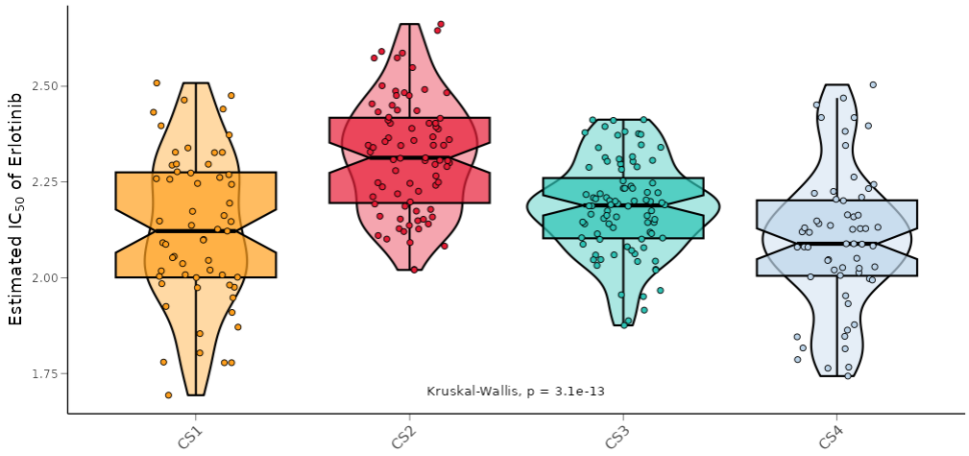


Table: Comparison of estimated IC50 (Validation) for Erlotinib among identified subtypes

Copy | CSV | Excel | Print | Show 10 entries | Search:

	Est.IC50	Subtype
GSM806810	2.00030403109302	CS1
GSM806864	2.007819835814	CS1
GSM806868	2.29641779049512	CS1

Result display for “Compare drug sensitivity” on validation dataset (illumina+NTP+ Erlotinib)

Survival | Clinical Features | Mutational Frequency | TMB | FGA | Drug Sensitivity | Agreement

Download pdf of the figure

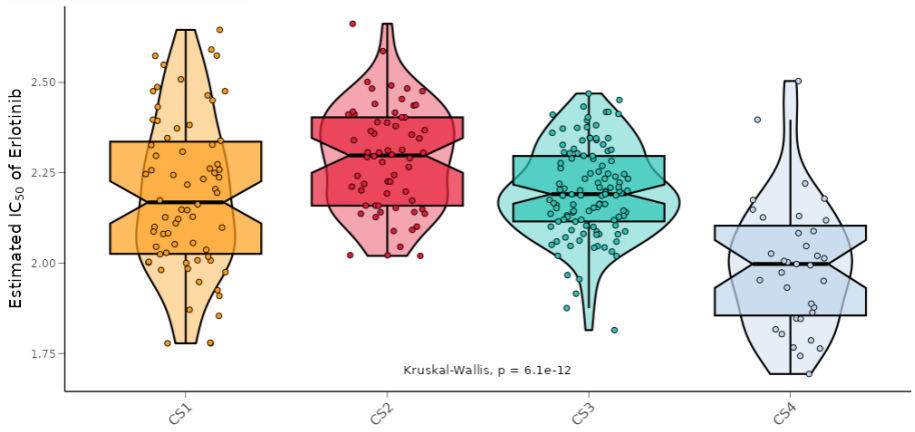


Table: Comparison of estimated IC50 (Validation) for Erlotinib among identified subtypes

Copy | CSV | Excel | Print | Show 10 entries | Search:

	Est.IC50	Subtype
GSM806810	2.00030403109301	CS1
GSM806814	2.11024789954329	CS1
GSM806864	2.00781983581398	CS1

Result display for “Compare drug sensitivity” on validation dataset (illumina+PAM+ Erlotinib)

(4). Compare agreement with other subtypes for validation dataset

I.Analysis introduction

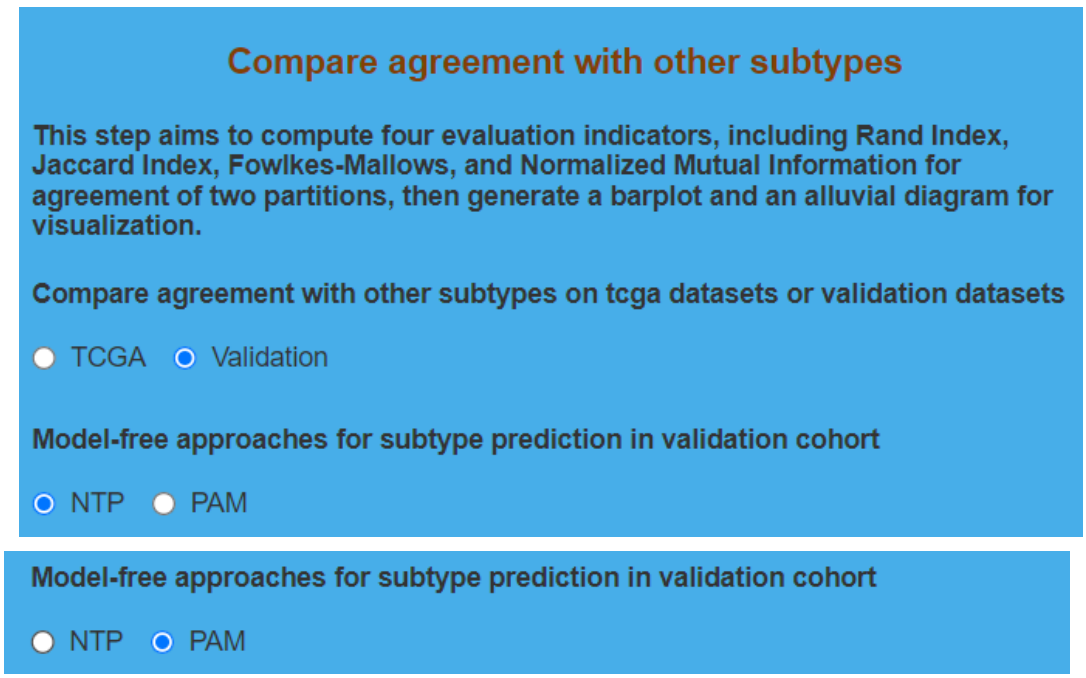
Details can be found in “The third step: COMP Module (TCGA dataset)”.

II.Parameters setting guides

Details can be found in “The third step: COMP Module (TCGA dataset)”.

III.Examples with results interpretation

Parameter Settings: The “Steps” chooses “Compare agreement with other subtypes”, and “Compare agreement with other subtypes on tcga datasets or validation datasets” chooses “Validation”. Other parameters keep unchanged, and then we click the “Process” button to compare agreement between subtypes derived from NTP as well as PAM methods and other traditional subtypes for two validation datasets in turn.



Parameter settings for “Compare agreement with other subtypes” on validation dataset

Result Display:

Like the TCGA dataset, the software will also generate alluvial diagrams to compare agreement between subtypes derived from NTP as well as PAM methods and other traditional subtypes for two validation datasets. Additionally, bar plots and tables are also generated for two validation datasets to display the values of four statistical indicators which are utilized to evaluate the agreement, including Rand Index (RI), Adjusted Mutual Information (AMI), Jaccard Index (JI), and Fowlkes-Mallows (FM).

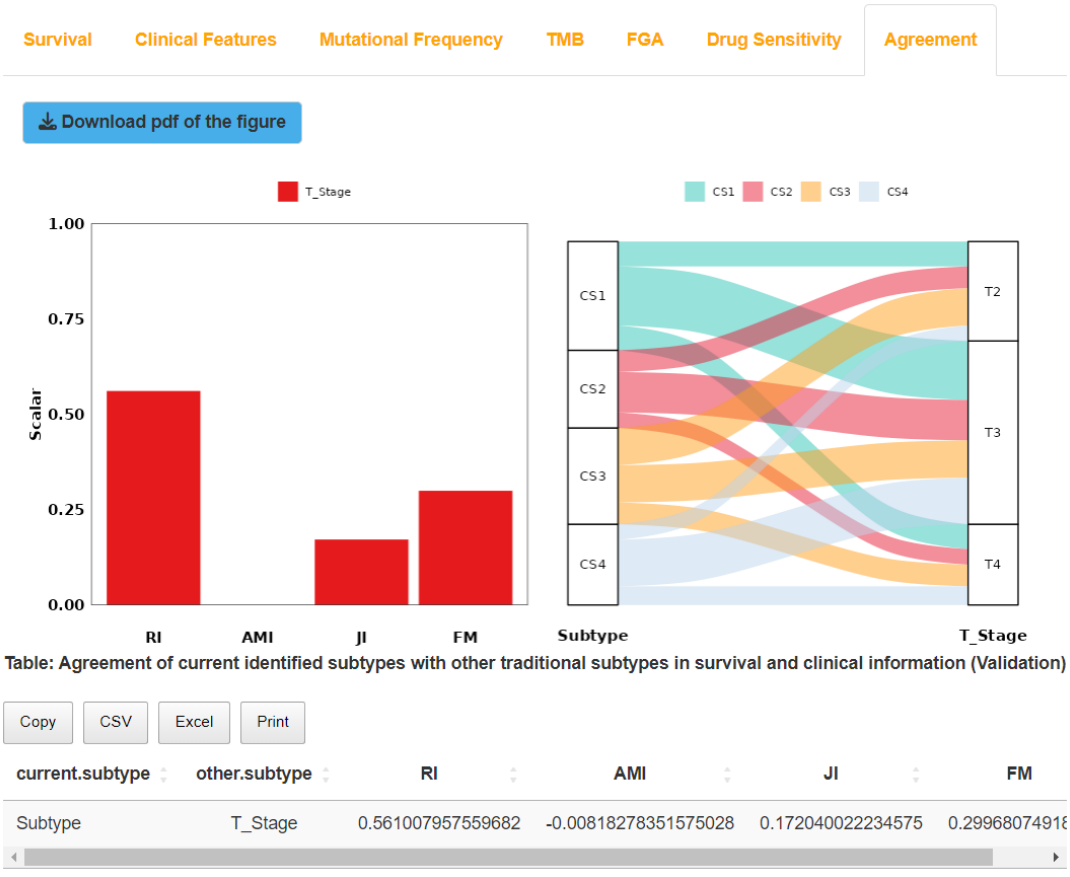
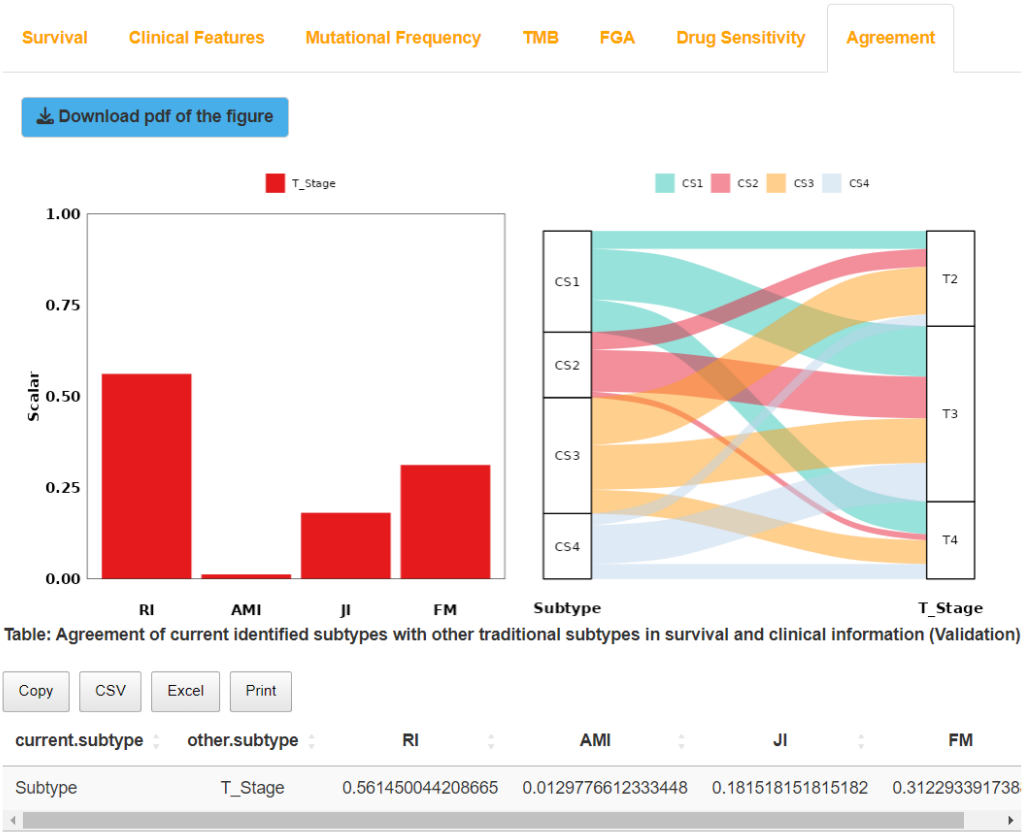


Table: Agreement of current identified subtypes with other traditional subtypes in survival and clinical information (Validation)

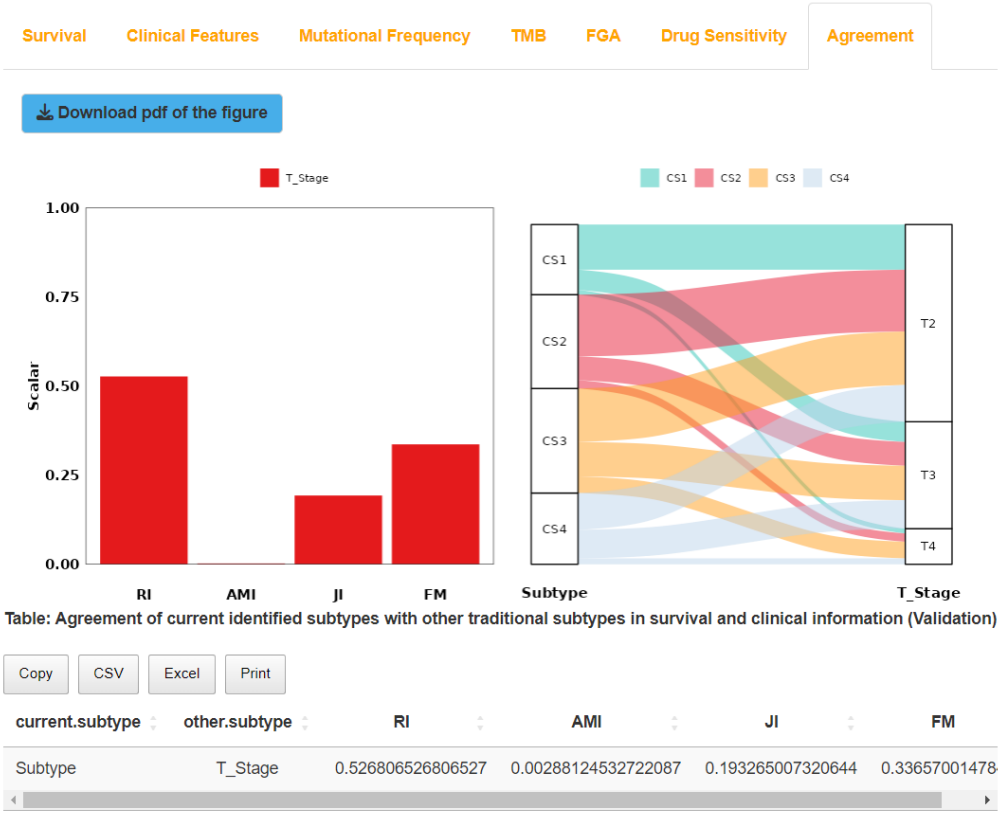
Copy CSV Excel Print

current.subtype	other.subtype	RI	AMI	JI	FM
Subtype	T_Stage	0.561007957559682	-0.00818278351575028	0.172040022234575	0.2996807491

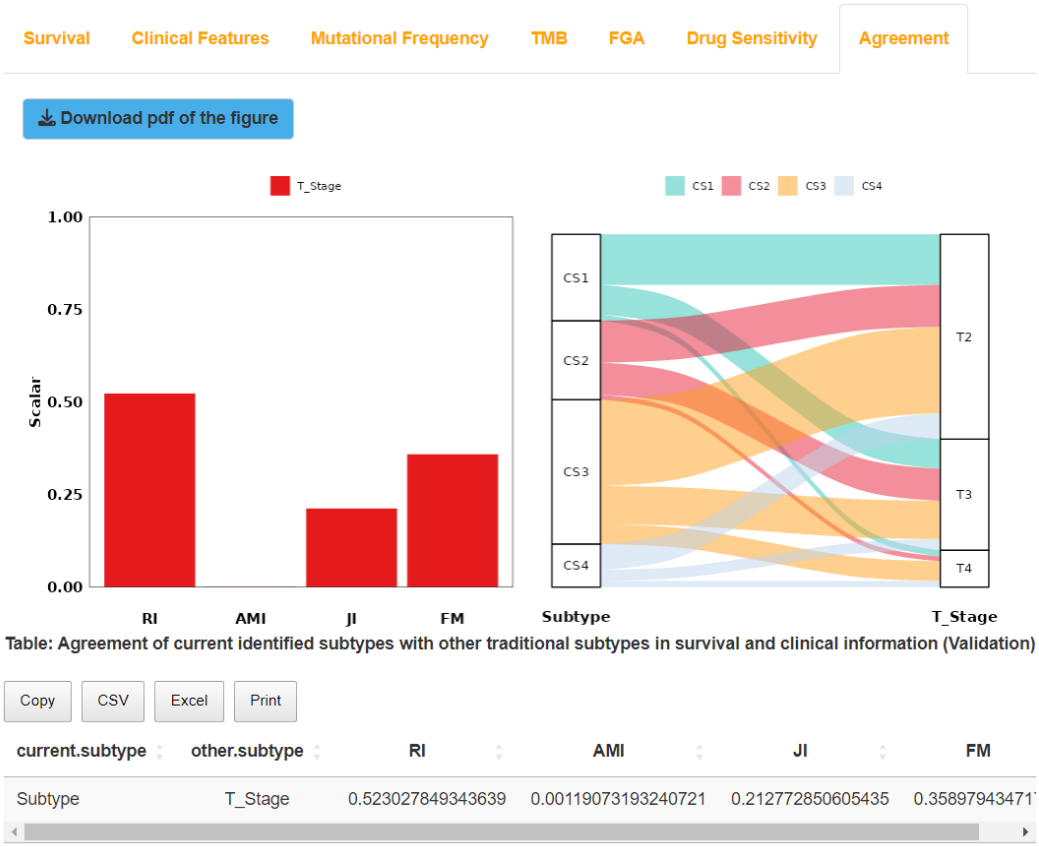
Result display for “Compare agreement with other subtypes” on validation dataset (affy+NTP)



Result display for “Compare agreement with other subtypes” on validation dataset (affy+PAM)



Result display for “Compare agreement with other subtypes” on validation dataset (illumina+NTP)



Result display for “Compare agreement with other subtypes” on validation dataset (illumina+PAM)

So far, the use of the website has been all introduced. Now, let’s try to perform multi-omics analysis on specified cancers through this website, and we welcome you to point out errors or give valuable advice on our website. If you have any questions or meet any difficulties during the use of the website, please do not hesitate to contact us by writing e-mails to:

Xiaofan Lu: ***xlu.cpu@foxmail.com***