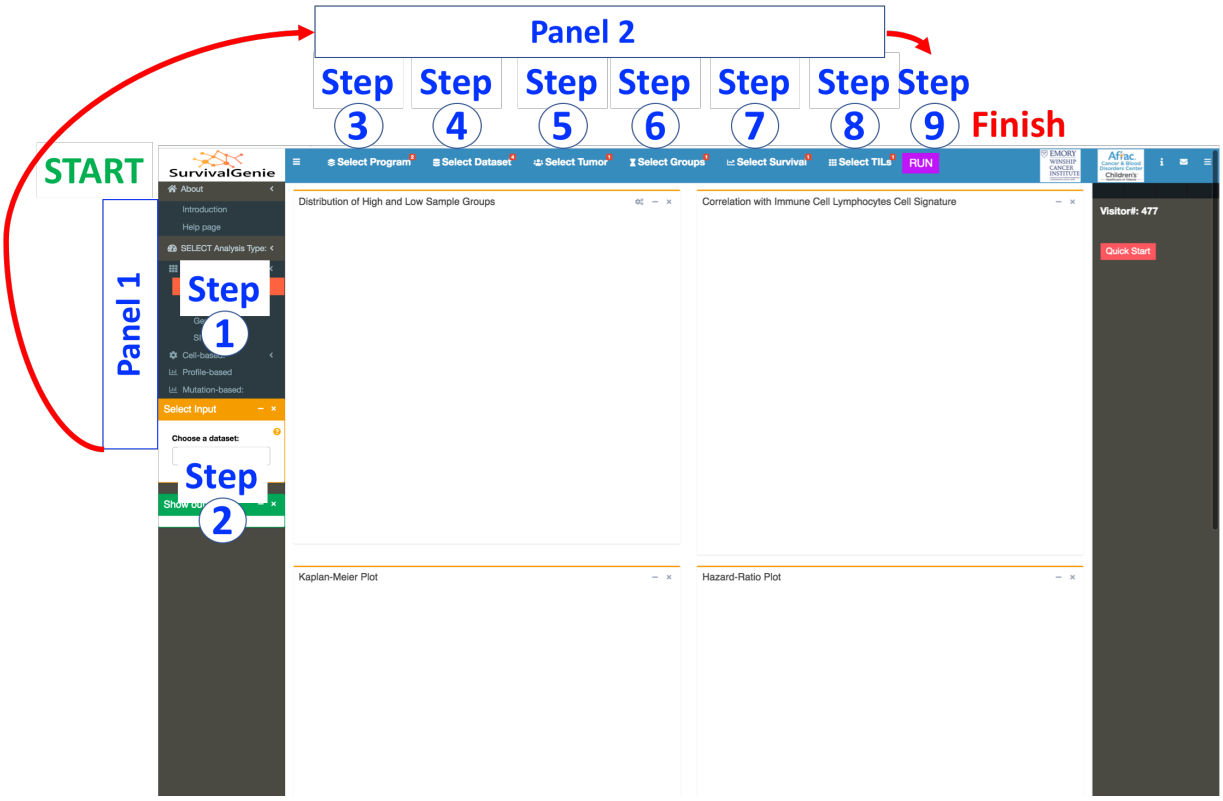
Step-by-Step-Guide

The Survival Genie web-interface is divided into **TWO INPUT PANELS:**

- Left-most Panel (Panel 1)
- Top-navigation Panel (Panel 2)



PANEL 1

This panel includes Step 1 & Step 2. User selects the molecular profile to be used for the survival analysis.

STEP 1: SELECT ANALYSIS TYPE

*Click on options under "**Select Analysis Type**" and select the analysis type to perform survival analysis on. The analyses can be performed using different genomics data types,*

i.e., single gene expression or gene set enrichment (gene-based), immune cell composition (cell-based), gene expression profiles (profile-based), and tumor mutation burden (mutation-based).

- **Gene-based** *is to query:*

    - **scRNA-seq clusters:** single-cell RNA-seq data consisting of gene markers from each cluster (e.g., scRNA-seq data analysis output). We have adapted the single-sample gene set enrichment analysis (ssGSEA[1]), to calculate the enrichment scores (ES) overall marker genes for each identified cluster for each tumor. Each cluster enrichment scores are then used to assist in assignment of risk groups to predict patient outcomes and of associations to tumors cell composition.

    - **Gene Sets:** list of genes as a set or signature (e.g., multiple-genes derived from pathways or biological processes). Here again, an aggregated enrichment score is computed for a gene set using ssGSEA method to assist in predicting survival outcomes.

    - **Gene Ratio:** input of two-genes. An expression ratio of two-genes is computed using bulk tumors normalized FPKM expression values to study associations with survival outcomes.

    - **Single Gene:** single-gene normalized FPKM expression data is used to study associations with survival outcomes.

- **Cell-based** *is to query:*

    - **CIBERSOFT TILs:** Proportion estimates of tumor-infiltrating lymphocytes (TILs) by LM6 and LM22 cell signature matrix for tumors using CIBERSOFT method[2] (https://cibersort.stanford.edu/). The estimates were obtained using the bulk tumors normalized FPKM expression data and are available for all cancer datasets.
    - **Digital TIL%:** Percentage of tumor-infiltrating lymphocytes (TILs) based on H&E images from cancer imaging archive (https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=33948919). This analysis can only be done for 13 TCGA cancer types.

- **Profile-based** *is to query:*

    - **Weighted gene profile** (e.g., derived from weighted gene co-expression network analysis such as 18-gene T-cell inflamed gene expression profile, GEP scores). The GEP score is calculated as weighted sum of the normalized FPKM expression values of the gene signature for each tumor. The weightings for

each gene in the signature is provided by the user in the input data file (see below Step (2) for input data format requirements). Here the example data file are the weightings of each 18-gene signature from (claim # 21c) as published in the Patent filed under "WO201609437" https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2016094377&tab=PCTCLAIMS.

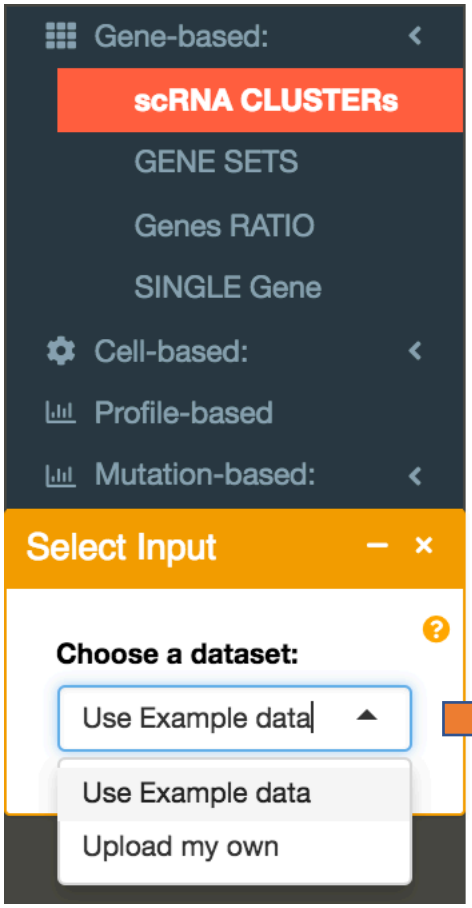
- **Mutation-based** *is to query:*

    - Effect of Tumor Mutation Burden (TMB) within a cancer type by the total number of non-synonymous somatic mutations, exonic mutations, and all somatic mutations per Mb. This analysis type can be selected to study effect of high TMB tumors on patient survival. The somatic mutations reported by Mutect2 from GDC were downloaded and used to estimate TMB of tumor samples.

STEP 2: SELECT INPUT

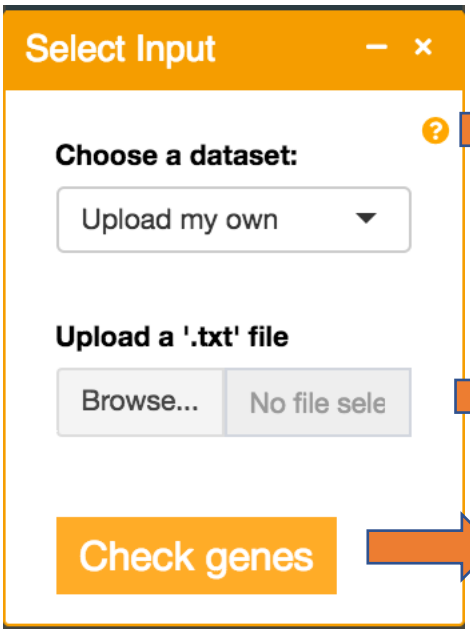
After selecting for the analysis type, a box panel will appear underneath. This is the step to specify inputs within the selected analysis type.

- **<u>scRNA-seq clusters:</u>**
    - The input requires a tab-delimited text (.txt) data file containing list of all gene markers for each identified cluster obtained from any single-cell RNA-seq data analysis tool.



Users have two "**File Input**" options: they can use an example data file and test the analysis OR upload their own data file.

    - When users select "**Upload my own**" option for "File Input", they must "Upload" an input data file from their local computer.



The format requirements of the user-uploaded data file can be viewed by clicking on the orange-colored **question mark ("?")** icon.

Browser and select your input.

Users can check the input genes by clicking on the orange-filled action button **"Check genes"**. The genes missing in the processed cancer datasets will not be used towards the analysis. Users have the option to leave the missing (or un-identified) genes in the data file and proceed with the identified list of genes for the analysis or correct for the missing genes (see requirements for details) and try it again.

File format requirements:

- Maximum file size limit of up to 500 MB.
- ASCII formatted tab-delimited .txt file
- Here is a screenshot of a user-uploaded 'File Input' for scRNA-seq cluster analysis:

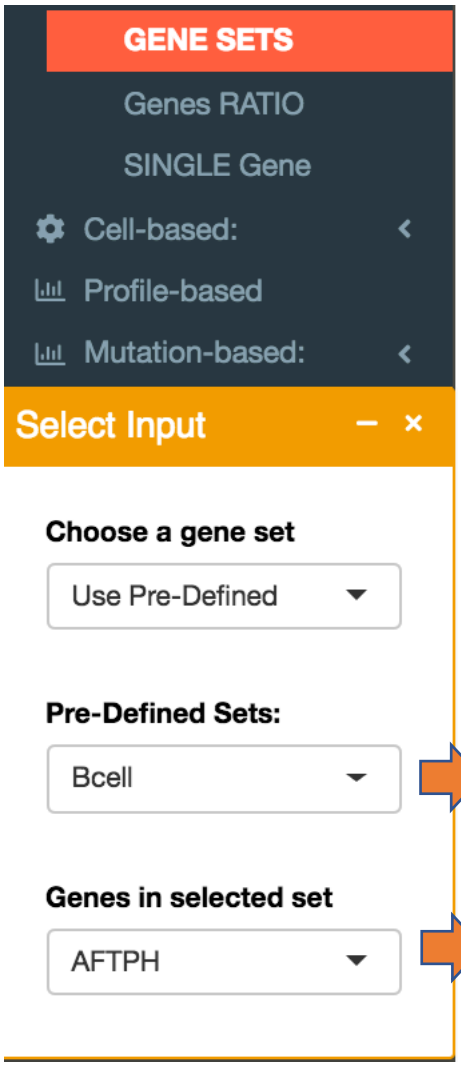
	p_val	avg_logFC	pct.1	pct.2	p_val_adj	cluster	gene
IL32	0	1.68902314	0.778	0.029	0	T cells	IL32
CD3E	0	1.59622586	0.904	0.01	0	T cells	CD3E
LTB	0	1.31772138	0.783	0.127	0	T cells	LTB
IFITM1	0	1.16484289	0.961	0.414	0	T cells	IFITM1
CD7	0	1.15653804	0.756	0.055	0	T cells	CD7
PTPRCAP	0	1.05105594	0.805	0.117	0	T cells	PTPRCAP
SARAF	0	1.04499308	0.972	0.686	0	T cells	SARAF
RPS26	0	0.96444842	0.987	0.914	0	T cells	RPS26
CD3D	0	0.9562295	0.679	0.015	0	T cells	CD3D
LEPROTL1	0	0.90061834	0.754	0.193	0	T cells	LEPROTL1
LDHB	0	0.89581084	0.825	0.31	0	T cells	LDHB
CD69	0	0.8809193	0.688	0.128	0	T cells	CD69
RPS27	0	0.87364867	1	0.934	0	T cells	RPS27
RPL3	0	0.87091851	1	0.93	0	T cells	RPL3
IL7R	0	0.85774719	0.55	0.004	0	T cells	IL7R
EEF1A1	0	0.85479086	1	0.969	0	T cells	EEF1A1
NPM1	0	0.84670327	0.935	0.581	0	T cells	NPM1
RPSA	0	0.82688569	0.999	0.904	0	T cells	RPSA
RPL5	0	0.82221242	0.999	0.93	0	T cells	RPL5
RPS12	0	0.78560196	1	0.987	0	T cells	RPS12
RPS3	0	0.77765597	1	0.959	0	T cells	RPS3
RPS27A	0	0.765089	1	0.951	0	T cells	RPS27A
RPS5	0	0.76436185	0.999	0.923	0	T cells	RPS5

- The file must have the below columns for:

- Cluster #s or names, labeled as **"cluster"**
- Human HGNC gene symbols, labeled as **"gene"**
- FDR, column labeled as **"p_val_adj"**. *Note that gene markers are filtered for FDR<=0.05 by cluster for the analysis by default.*
- Any other columns are optional
- The order of the columns does not matter
- Only unique gene symbols per cluster are counted.

Gene Sets:

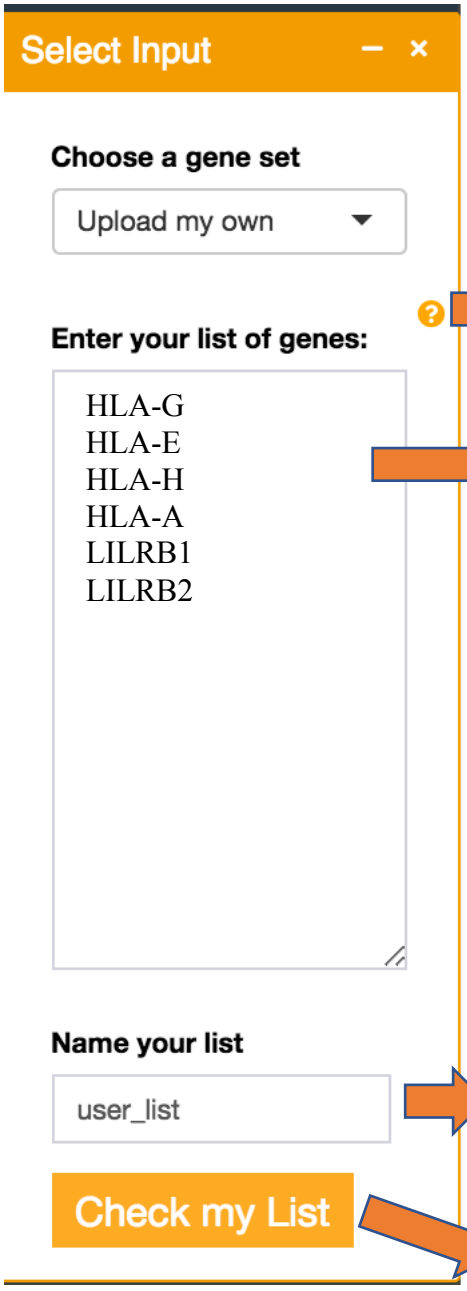
- For Gene Set analysis, users can **"Choose a gene set"** either from our list of pre-defined gene sets or upload their own defined list of genes.



The **Pre-defined Sets** include 24 cell subsets collected[6], and 26 signature gene sets from cbioportal[7]. Users can select a **Pre-defined set** from the drop-down-menu.

Users can take a look on all markers genes in the selected pre-defined gene set under the "**Genes in selected set**" option. This option is just to view if you are curious to know what genes are part of the gene set.

- When user select "**Upload my own**" gene set, they must "**Enter list of genes"** in the text box provided.



The format requirements can also be viewed by clicking on the blue question-mark icon("?") next to **"Enter your list of genes"** option.

User enters the list of genes, It must be,

- **HUMAN HGNC GENE SYMBOLS**
- Either **UPPERCASE** or **LOWERCASE**
- **ONE** GENE **PER LINE**
- **LESS** than **50** Genes
- **NO WHITE SPACES**
- **NO EMPTY LINES**

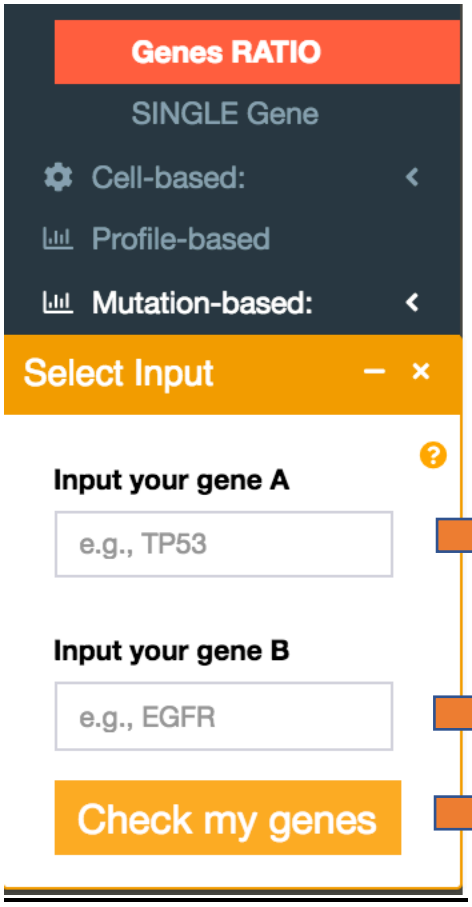
User must also name their input list under "Name your list" box. The default is "user_list".

Check the input genes by clicking on **"Check my List"**

Gene Ratio:

- User inputs two-genes (e.g., numerator gene A and denominator gene B) for gene ratio survival analysis. An expression ratio is estimated using

normalized FPKM expression value of gene A divided by normalized FPKM expression value of gene B.



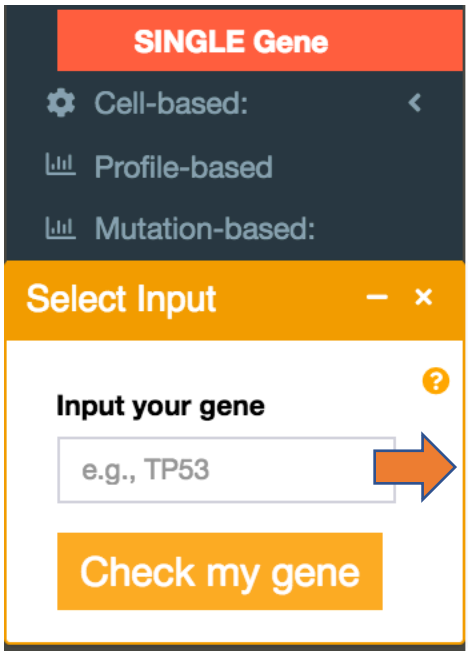
User inputs the numerator gene

User inputs the denominator gene

Users can check and confirm the input genes by clicking on the orange-filled action button **"Check genes"**.

Single Gene

- User inputs a single gene symbol. Here normalized FPKM gene expression values is used for survival analysis. Again, input gene symbol can be checked by clicking on "**Check my gene**" button before running the analysis to make sure it is correct and present in the datasets.



Type your HGNC HUGO gene symbol. Only one gene is accepted.

CIBERSOFT TILs

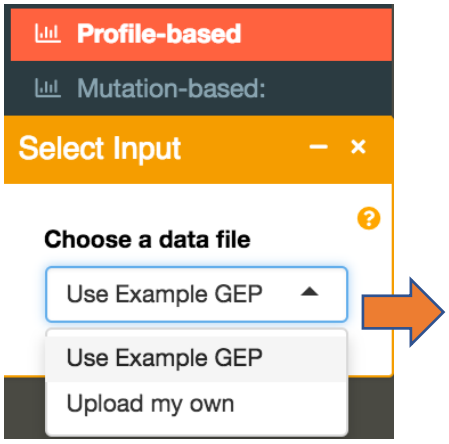
- No input is required for this analysis type.

Digital TILs

- No input is required for this analysis type.

Profile-based

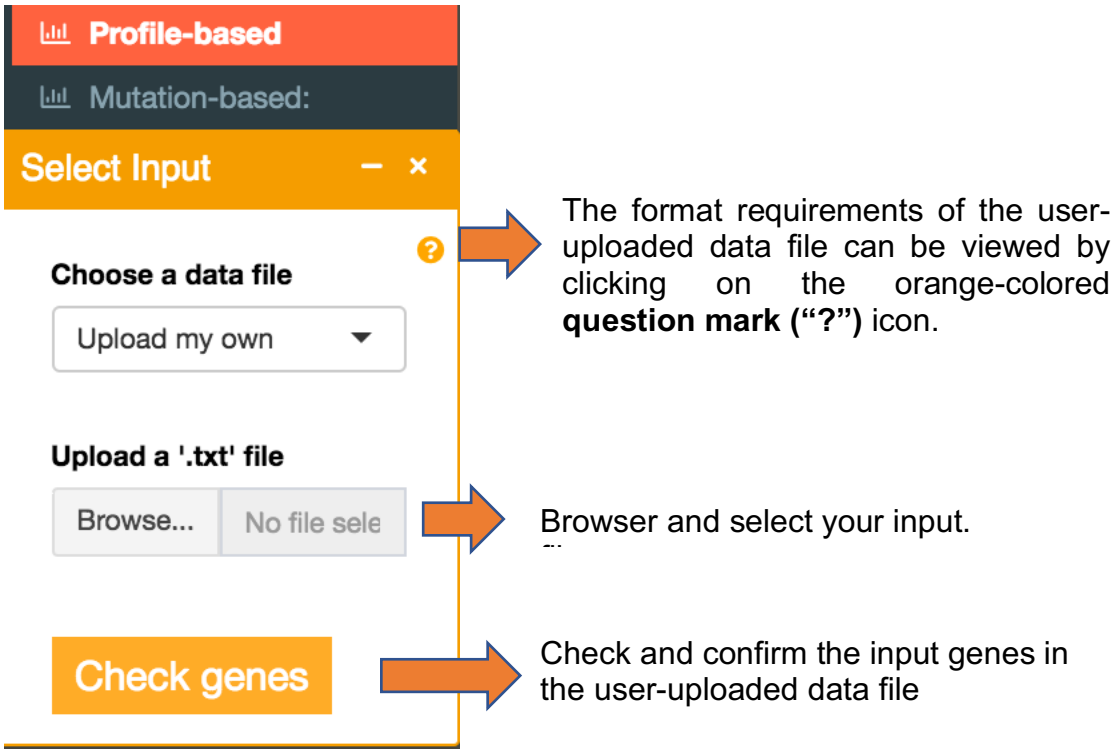
- User inputs the data file consisting of the weightings for each gene in the signature.



Users can select an example dataset or upload their own gene weightings.

- The “Use Example GEP” option are the weightings of each 18-gene signature from (claim # 21c) as published in the Patent filed under “WO201609437” https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2016094377&tab=PCTCLAIMS used to gene expression profile (GEP) scores.

- When users select “**Upload my own**” option under Choose a dataset they must “Upload” an input data file from their local computer.



File format requirements:

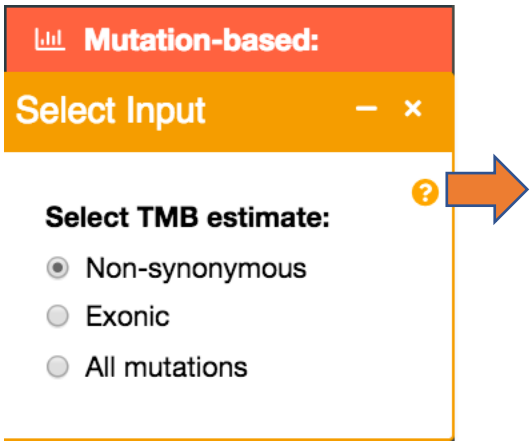
- Maximum file size limit of up to 500 MB
- ASCII formatted tab-delimited .txt file
- Here is a screenshot of a user-uploaded ‘File Input’ for profile-based analysis

gene	scoring_weight
CCL5	0.008346
CD27	0.072293
CD274	0.042853
CD276	-0.0239
CD8A	0.031021
CMKLR1	0.151253
CXCL9	0.074135
CXCR6	0.004313
HLA.DQA1	0.020091
HLA.DRB1	0.058806
HLA.E	0.07175
IDO1	0.060679
LAG3	0.123895
NKG7	0.075524
PDCD1LG2	0.003734
PSMB10	0.032999
STAT1	0.250229
TIGIT	0.084767

- First column must be Human HGNC gene symbols, labeled as "**gene"**
- Second column must be corresponding gene weightage or score, column label does not matter

Mutation-based

- User selects a Tumor Mutation Burden (TMB) estimate from the three choices:
  - Non-synonymous somatic mutations
  - Exonic somatic mutations
  - All somatic mutations

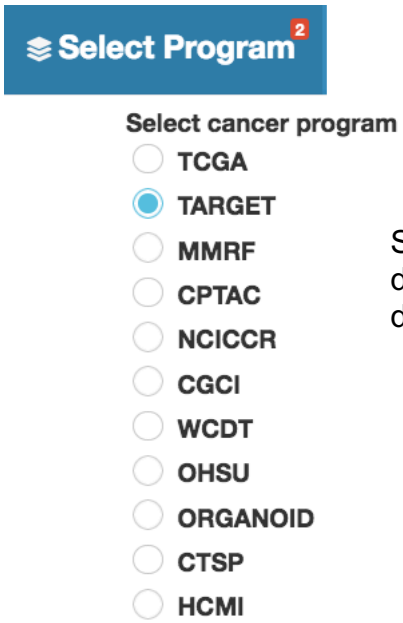


TMB estimate definition can be viewed by clicking on the orange-colored **question mark ("?")** icon.

PANEL 2

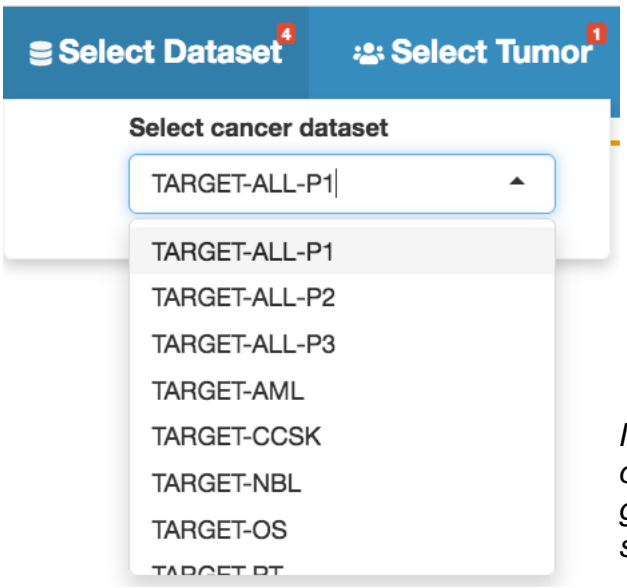
This panel includes Step 3 to Step 9. User provides inputs in terms of cancer dataset, tumor type, partition method etc for the selected analysis type from Panel 1. A help page can be viewed by clicking on the question mark orange icon under each drop-down-menu button of the panel 2 to understand the options.

STEP 3: Click on the "**Select Programs**"



Select a cancer program under which your cancer dataset of interest falls. There are a total of 11 different cancer programs to choose from[5].

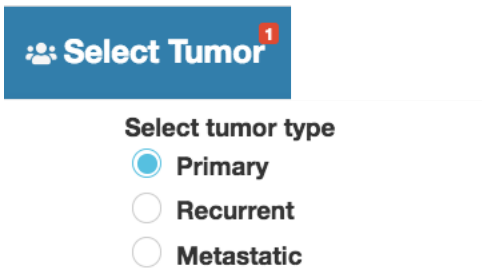
STEP 4: Click on the "**Select Dataset**"



Select the cancer dataset of interest within the selected cancer program on which you would like to perform the survival analysis on. For example, here all datasets under selected TARGET cancer program are listed to choose

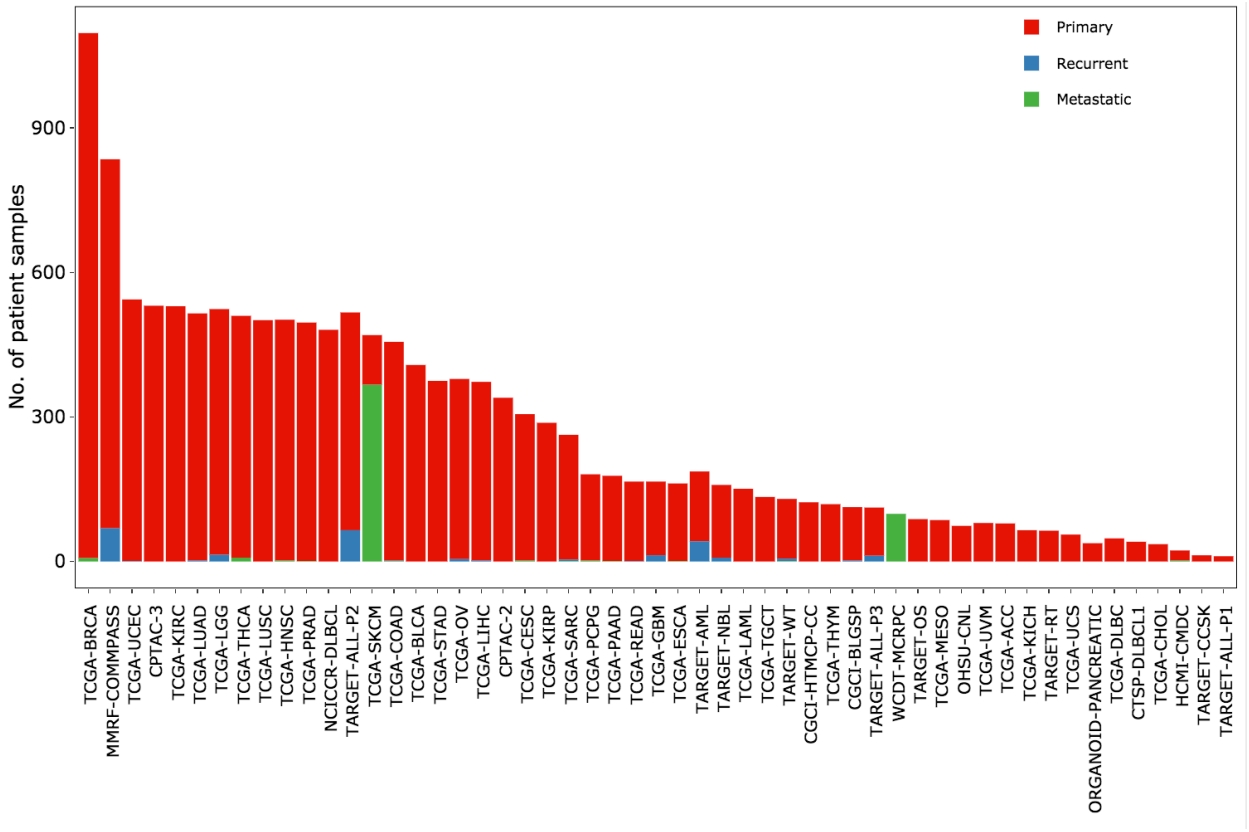
Note: At present, users can only select one dataset at a time, except for single-gene analysis type where users can select up to 5 datasets

STEP 5: Click on the "**Select Tumor**"



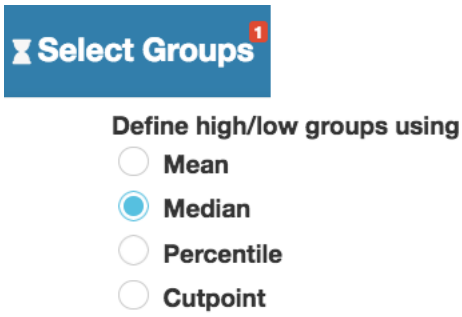
Select drop-down-menu and select the tumor type for the selected dataset from Step (4).

- Majority of the cancer datasets are comprised of "Primary" tumor samples, with the exception of few such as skin melanoma (TCGA-SKCM), which is dominated by "Metastatic" tumor samples. Below bar plot shows the distribution of tumor types across all datasets. The default is set to "Primary".

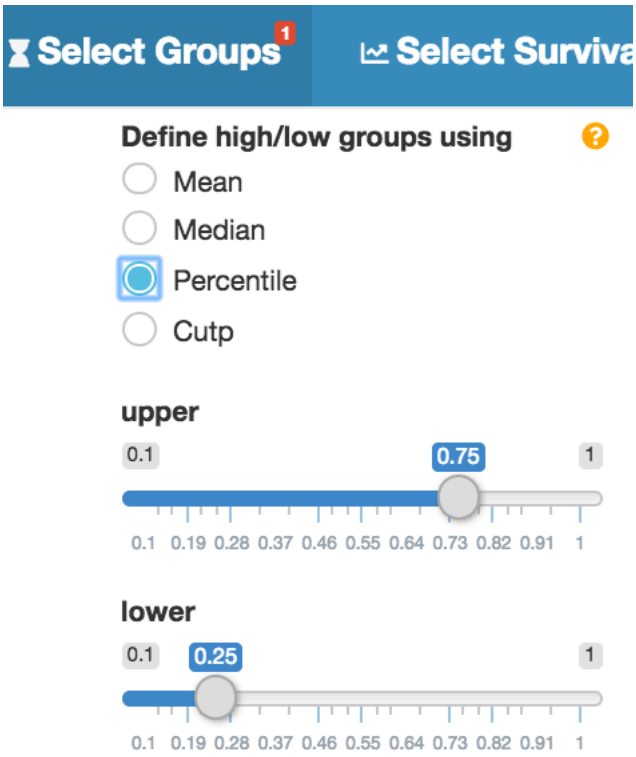


STEP 6: Click on the ***"Select Groups"***

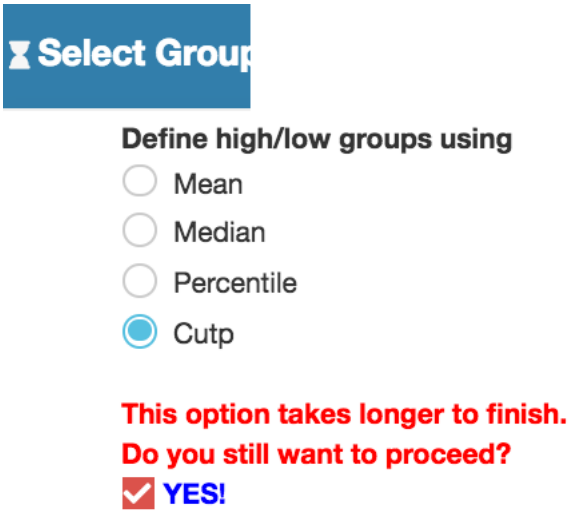
There are four different partitioning methods to divide the patient tumors into high and low groups to study their effect on the survival. The default is set to "median".



With *Percentile* option, users are able to select upper and lower threshold of their choice (e.g., quartiles or 10^{th} vs. 90^{th} percentile) to categorize low and high groups, respectively.

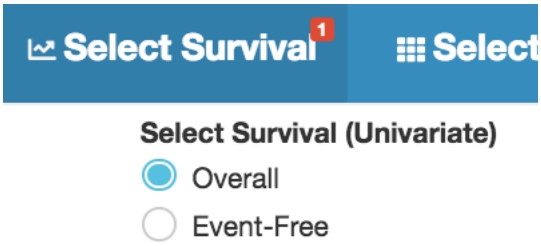


The last option, *Cutp* is estimated based on martingale residuals[6] using the 'survMisc' package[7] to separate the patients into high and low groups. *Note that this option takes longer to finish, so click the box next to "YES!" if you wish to proceed and please be patient.*



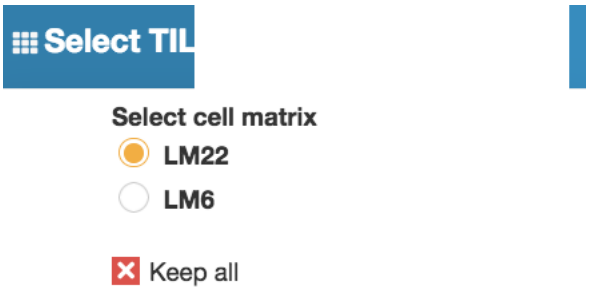
STEP 7*:* Click on the ***"*Select Survival"**

User can select either “Overall” or “Event-free” survival analysis. *Note that the Event-free survival analysis is only applicable to TARGET datasets.* The default is set to “Overall”.



STEP 8: Click on the ***“Select TILs”***

Here select for the tumor-infiltrating immune cell types matrix (LM6 or LM22). The relative fractions of cell types are estimated for validated LM6 and LM22 immune cell gene signature from bulk tumors FPKM gene expression data using the CIBERSOFT deconvolution method[2]. This step is to correlate the deconvoluted cell proportions of tumors to molecular profile estimated for the selected analysis type.



Here for a selected cell matrix, users has an option to filter tumor samples by significant global deconvolution p-value threshold (<0.05) or not to filter and keep all samples. The default is set to “Keep all”.

STEP 9: ***The LAST STEP!*** Click the “**RUN”** button

Run the analysis on the selected analysis type, dataset, and parameters.

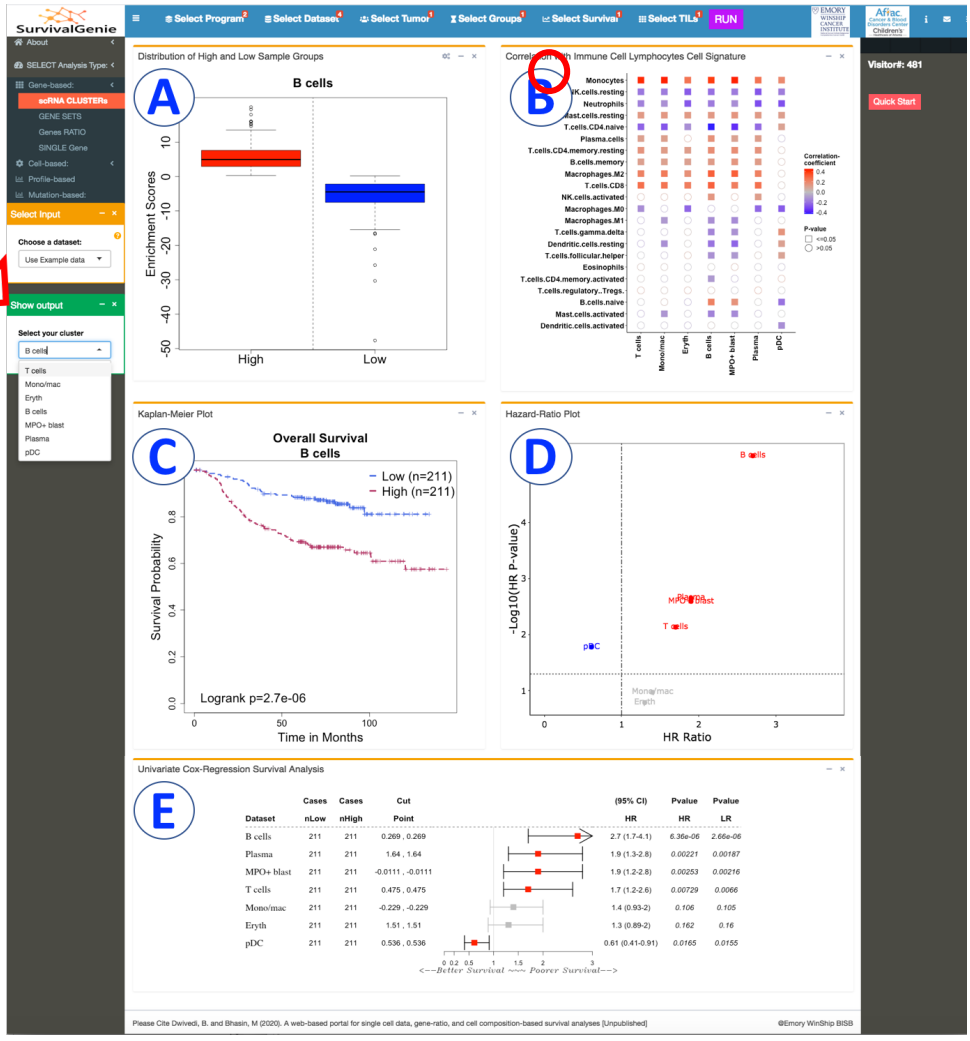
Note: Each time you change an option in Panel 1 or Panel 2, you must hit the RUN button again to update the results.

OUTPUT:

For each analysis type, the results are output in five separate tabs

A. Distribution of High and Low sample groups
B. Correlation with Immune Cell Lymphocytes Cell Signature
C. Kaplan-Meier Plot
D. Hazard-Ratio Plot
E. Forest Plot

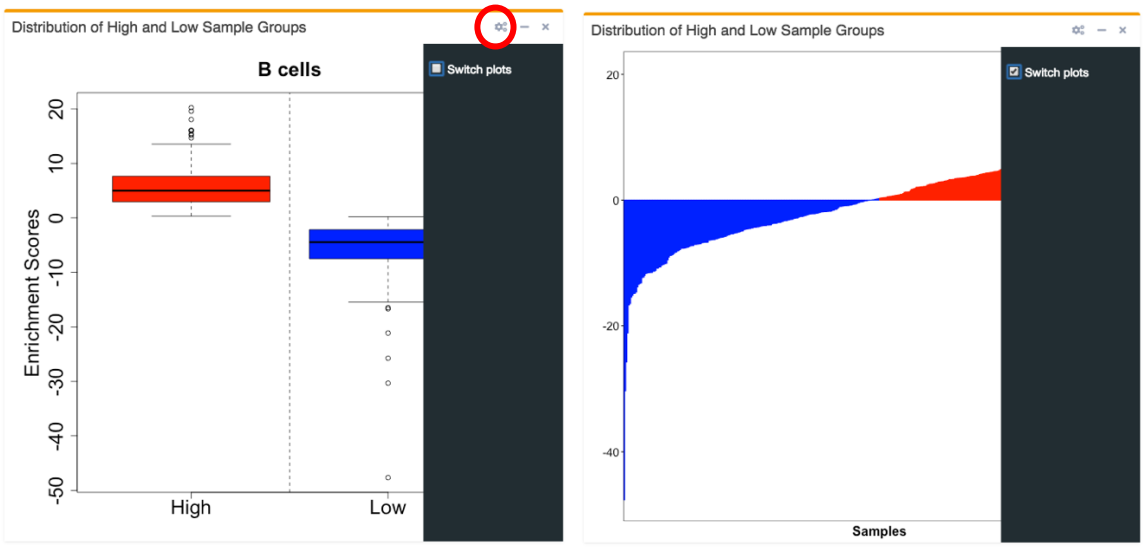
The "**Show Output**" green box allows users to select and view the plots (A) and (C) for the selection (e.g., cluster gene set) within the analysis.



A screenshot of results output from single-cell RNA-seq cluster analysis of example data file on TARGET-ALL-P2 primary tumors with median cutoff.

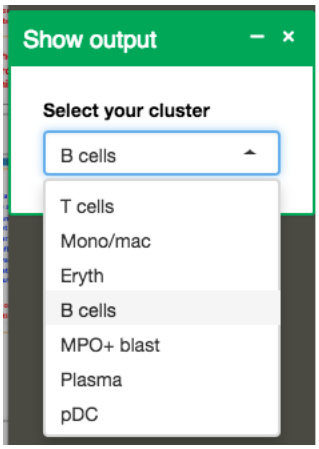
A. Distribution of High and Low sample groups

A box plot showing distribution of estimated molecular data profile (e.g., gene expression, expression ratio of two genes, gene set enrichment scores, cell proportion or mutation burden) separated by the low (blue-filled) and high (red-filled) sample groups. The settings icon (indicated with red-circle) on the top-right corner of the plot allows the users to "**Switch plots"** from box-plot to bar-plot view, later showing data points for each tumor sample by the two-groups.



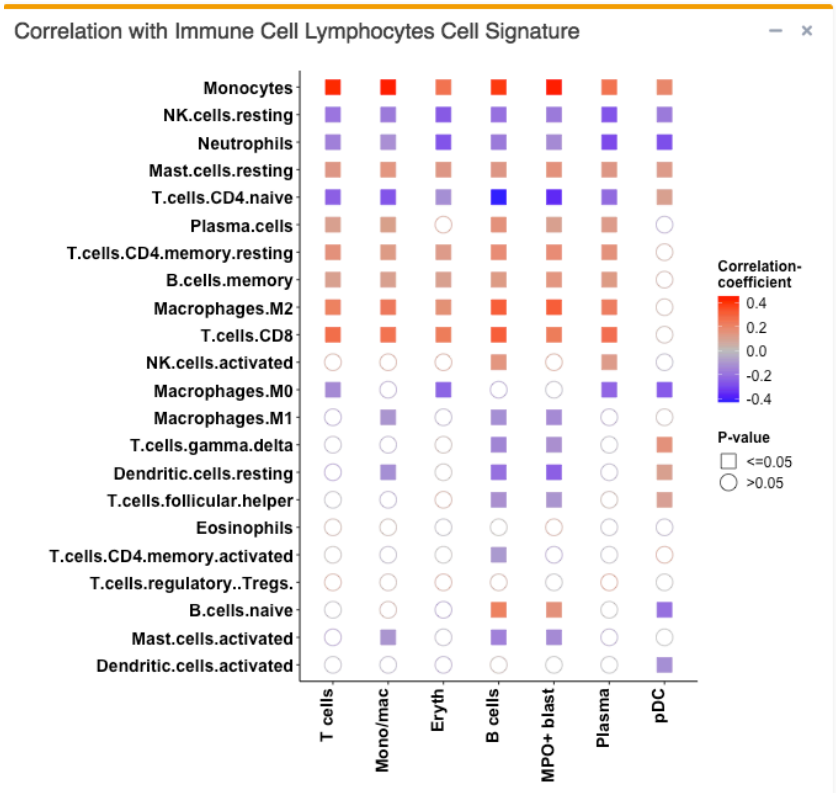
For example, the screenshot above shows the box plot of gene set enrichment scores over all marker genes for "B cells" single-cell cluster across TARGET-ALL-P2 primary tumor samples. Within each cluster, the enrichment scores are categorized into low (in blue) and high (in red) groups based on the selected median cutoff.

The cluster-to-view is selected from the left panel drop-down-menu labeled as "**Select your cluster**" (indicated with a red-arrow above). Note this option is visible only after the data file is uploaded and analysis is finished running. The cluster are listed based on the clusters identified in the uploaded input data file.



B. Correlation with Immune Cell Lymphocytes Cell Signature

Pearson correlation matrix of deconvoluted immune cells RNA-seq gene expression data and the analysis molecular data profile. Correlation coefficients are indicated with a color-gradient, blue-filled squares represent negative correlation; while red-filled squares represent positive correlation. The shape denotes the significance of the correlation; squares are significant while circle are not.



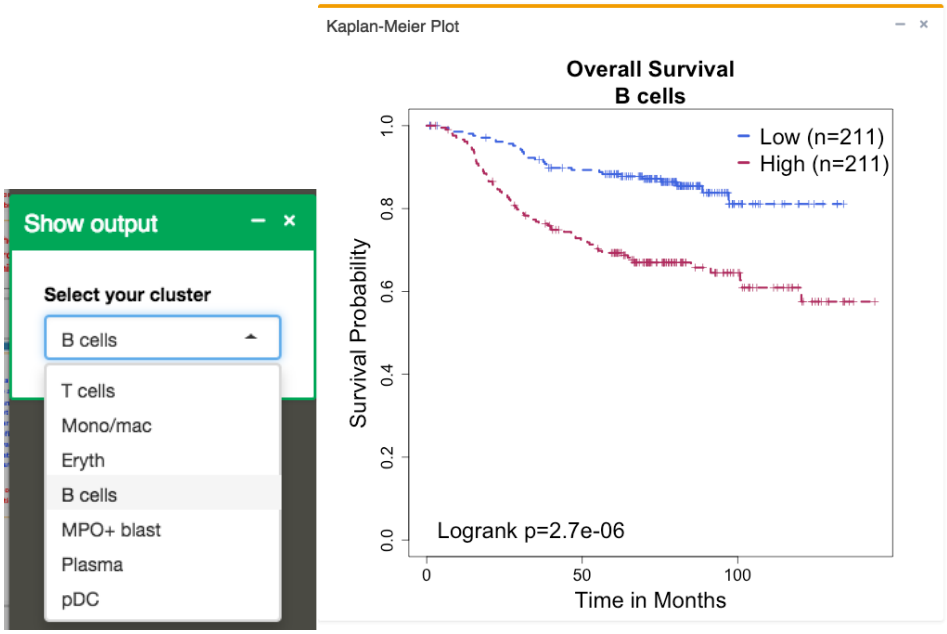
For example, the screenshot above shows the correlation of 22 immune cells proportions and 7 single-cell clusters enrichment scores across primary tumors of

TARGET-ALL-P2 for the scRNA-seq cluster analysis. Here the y-axis represents the selected 22 immune cells signature (LM22; Step (8) of input Panel 2), and x-axis the 7 single-cell clusters from the input data file for scRNAseq cluster analysis. As shown below, Monocytes are significantly positively correlated with all clusters, while resting NK cell are negatively correlated.

C. Kaplan-Meier Plot

Kaplan-Meier (KM) survival curves in the stratified high (red) and low (blue) groups of patients with log-rank test using survival package[7]. A log-rank p-value<0.05 is considered statistically significant.

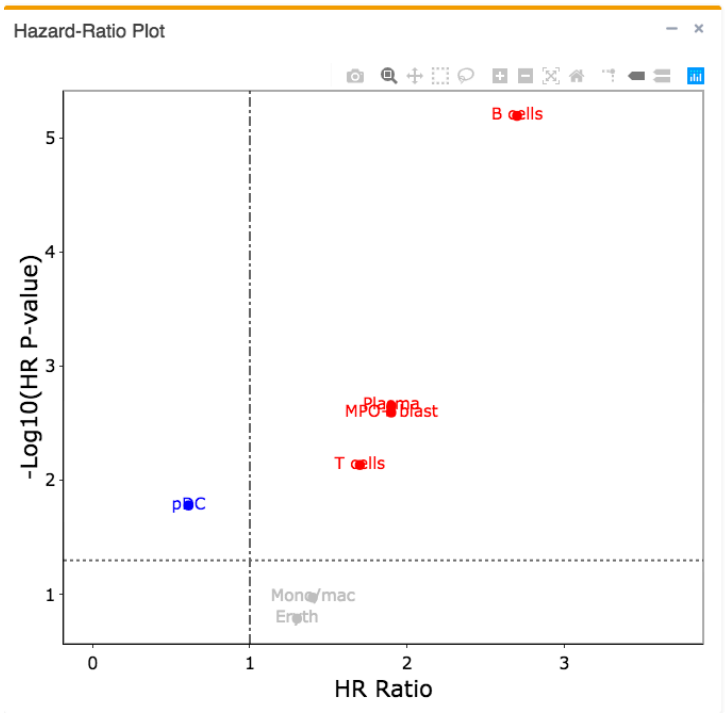
For example, the screenshot below shows the KM curve of selected "B cells" single-cell cluster (out of 7 input clusters) in TARGET-ALL-P2 primary tumors. The patients were divided into two groups based on the enrichment scores. The plot shows that patients with high (red line; n=211 samples) scores had a higher overall survival (log-rank p-value = 2.7e-06) than those with the low scores.



D. Hazard-Ratio Plot

Hazard-Ratio (HR) plot shows the survival significance by HR ratio and HR p-value. HR is based on a univariate cox proportional hazard regression model with wald-test. An HR ratio above 1 indicates increase in hazard (poor outcome), below

one indicates reduction in the hazard (good outcome), and equal to 1 indicates no effect. A p-value<0.05 is considered statistically significant.



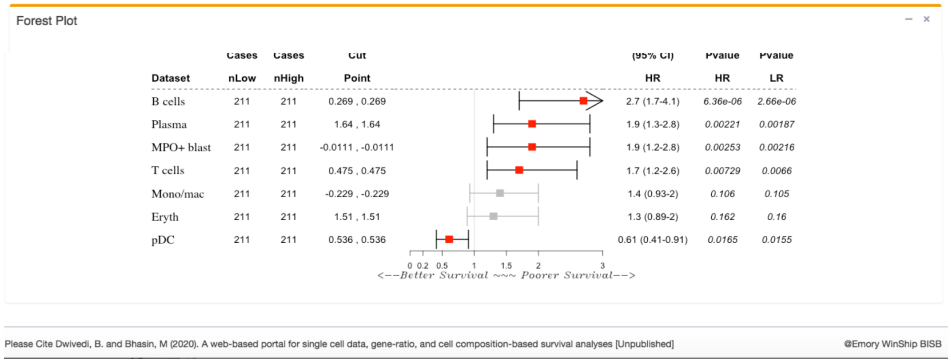
For example, the screenshot above shows the differences in the patient outcome for all 7 clusters in TARGET-ALL-P2 primary tumors. The plot shows a significant association between the patients with high scores and poor outcome for B cells, Plasma, MPO+ Blast, and T cells clusters; while good outcome for pDC cluster. The remaining two clusters Mono/mac, and Eryth are not significant.

E. Forest plot

Forest plot showing the association between the defined two-groups and survival outcome. The survival curves are estimated using Kaplan-Meier and univariate Cox Proportional Hazards regression model based on Survival package[7].

The plot shows the hazard ratio (HR) and 95% confidence intervals associated with two-groups considered in the univariable analyses along with wald-test (HR Pvalue) and log-rank (LR Pvalue) p-values. Also included is the cut-point applied to separate the patients into high and low groups and their corresponding counts. The circles represent the hazard ratio and the horizontal lines extend from the lower limit to the upper limit of the HR 95% confidence interval. Significant associations are highlighted with red-filled circles. An arrow at the end of the

horizontal line indicates higher upper limit of the 95% confidence interval than the maximum shown (i.e., 3).



For example, the screenshot above shows the differences in the patient outcome for all 7 clusters in TARGET-ALL-P2 primary tumors. The plot shows a significant association between the patients with high scores and poor outcome for B cells, Plasma, MPO+ Blast, and T cells clusters, while good outcome for pDC cluster. The remaining two clusters Mono/mac, and Eryth are not significant. Each row illustrates a single cluster (or gene or dataset) result. For example, topmost row shows results for "B cells" cluster, where patients are separated into high (n=211 samples) and low(n=211 sample) groups based on the median partition method (median cut point of 0.269), have a significant association with poor outcome (HR=2.7; 95% CI 1.7-4.1; P=6.36e-06 by wald-test and P=2.66e-06 by log-rank test).

References:

1. Hänzelmann S, Castelo R, Guinney J (2013). "GSVA: gene set variation analysis for microarray and RNA-Seq data." *BMC Bioinformatics*, **14**, 7. doi: 10.1186/1471-2105-14-7, http://www.biomedcentral.com/1471-2105/14/7.
2. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *Methods Mol Biol*. 2018;1711:243-259. doi:10.1007/978-1-4939-7493-1_12
3. Miao YR, Zhang Q, Lei Q, et al. ImmuCellAI: A Unique Method for Comprehensive T-Cell Subsets Abundance Prediction and its Application in Cancer Immunotherapy. *Adv Sci (Weinh)*. 2020;7(7):1902880. Published 2020 Feb 11. doi:10.1002/advs.201902880
4. https://www.cbioportal.org/
5. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics#programs
6. Contal C, O'Quigley J, 1999. An application of changepoint methods in studying the effect of age on survival in breast cancer. *Computational Statistics & Data Analysis* **30**(3):253--70. ScienceDirect (paywall)
7. Chris Dardis (2018). survMisc: Miscellaneous Functions for Survival Data. R package version 0.5.5. https://CRAN.R-project.org/package=survMisc