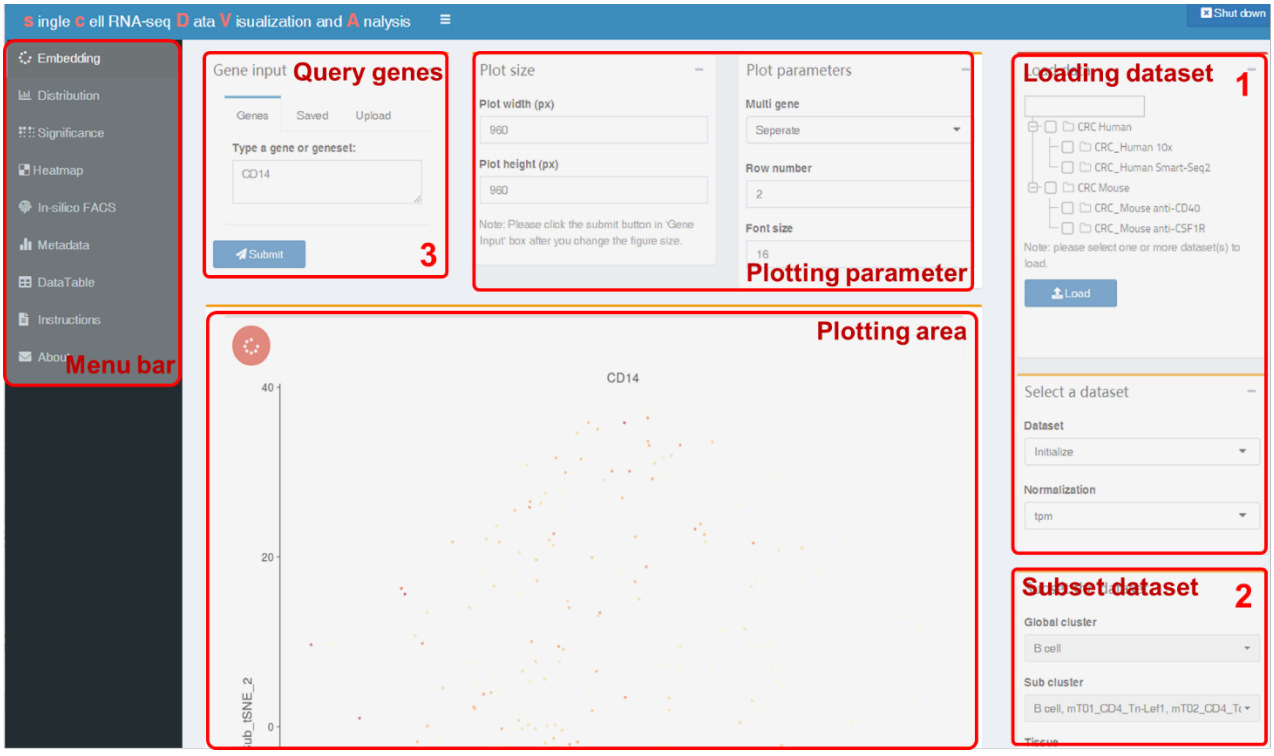
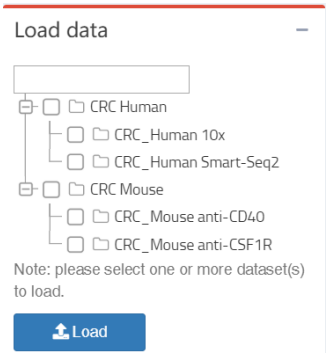
User Manual of scDVA

scDVA (short for single cell RNA-seq data visualization and analyzation) is an interactive web server developed for users to explore and analyze the single cell RNA-seq data. scDVA is developed based on R package *shiny*.

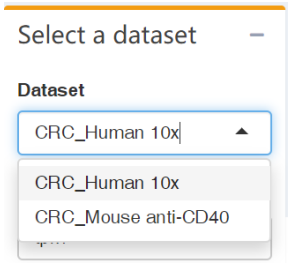
1. Page layout



2. Load datasets



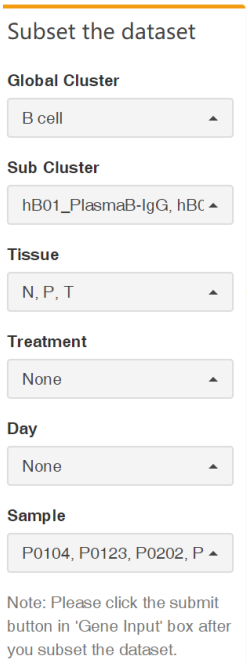
All datasets available are listed in the upper right-hand corner of the page. Users should first select the dataset of interest and then click the "**Load**" button. It will take some time to load the expression matrix with a large number of cells. You can load different datasets for many times, but deselecting a loaded dataset will not free it from the memory.



When loading is done, the loaded datasets are available in "**Select a dataset**" panel below. Users could also choose a normalization method. "*Counts*" refers to $\log_2 counts$ normalized by size factor calculated with *scran* package, while "*tpm*" refers to $\log_2 tpm$ normalized by library size.

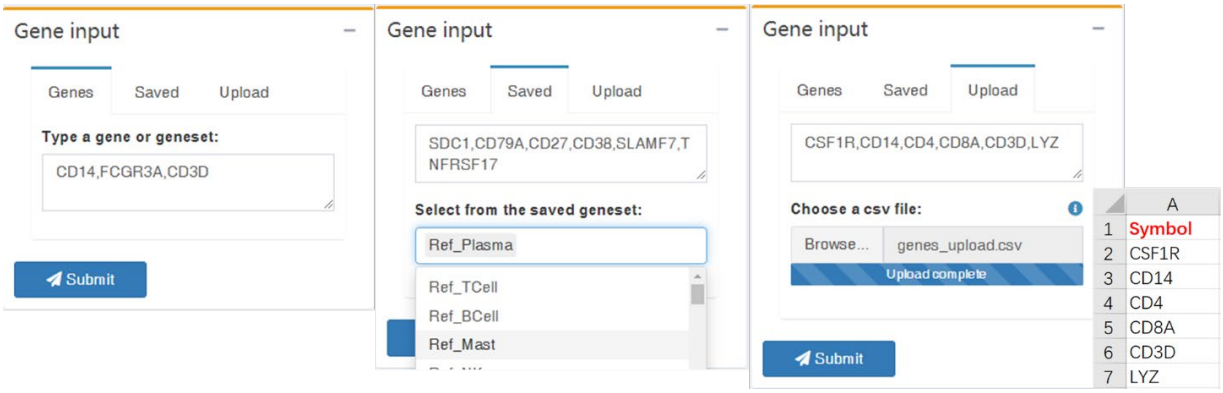
3. Filter the data and input the query gene(s)

Before plotting, it is recommended to filter the dataset in order to pay more attention to the cells you interested in. Users can filter the dataset according to *Global cluster* (B cell, CD4 T cell, CD8 T cell, …), *Sub cluster*, *Tissue* (Normal, Peripheral blood and Tumor), *Treatment*, *Day* and *Sample*. The final subset of the selected dataset is produced by the intersection of all filter conditions.



Then, users should type in the query gene symbol(s) to explore the data. Here we provide three different modes of gene input: keyboard input, pre-existing gene lists or uploading a .csv file. Either way, scDVA only accepts case-insensitive gene symbol or comma-separated gene symbols list as input. When you use a pre-existing gene list or upload a .csv file as input, all the genes in the list will appear in the text input filed and can be edited manually. It should also be noted that the gene symbols in .csv file must be in column named as "Symbol".

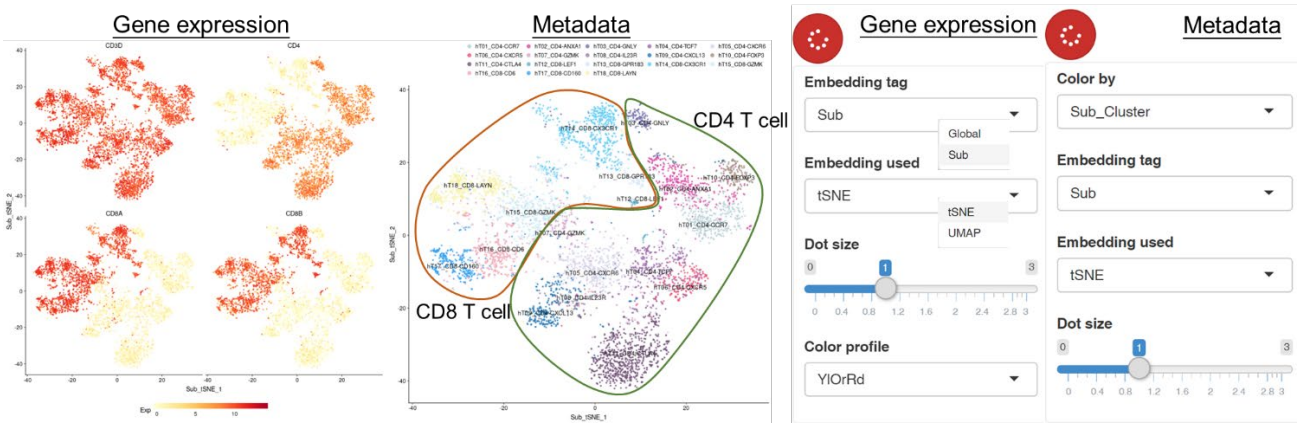
All changes in "**Subset the dataset**" or "**Gene input**" panel will be received by the server after clicking the "**Submit**" button in the "**Gene input**" panel. Never forget it!



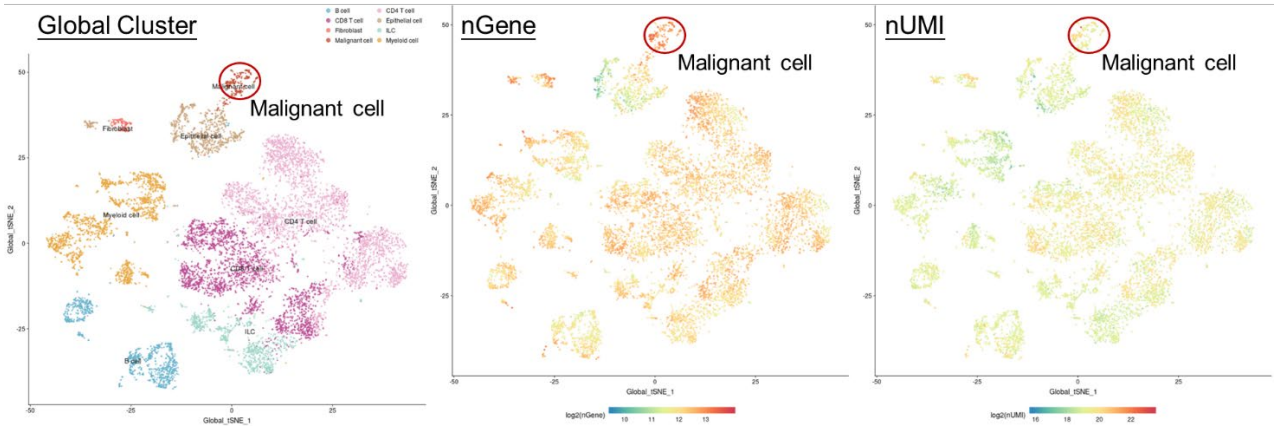
4. Embedding plot and plotting parameters

Users can explore the gene expression level or the metadata of each cell in a 2-D space with tSNE (t-distributed Stochastic Neighbor Embedding) or UMAP (Uniform Manifold Approximation and Projection) coordinates. The expression plot and the metadata plot are arranged vertically in the plotting area.

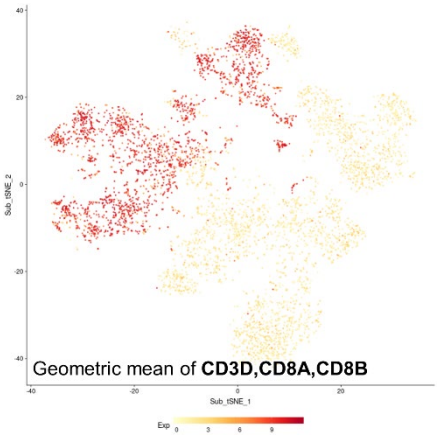
The embedding parameters you can adjust are listed in the red round button in the upper left edge of the plotting area. Two embedding tags, "*Global*" and "*Sub*", are provided. The "*Global*" tag is used when you try to plot all global clusters' cells at the same time, while the "*Sub*" tag is used when plotting only one global cluster's cells (CD4 T cells and CD8 T cells can be plotted together). You can also change the embedding coordinates (tSNE or UMAP), the dot size and the color profile.

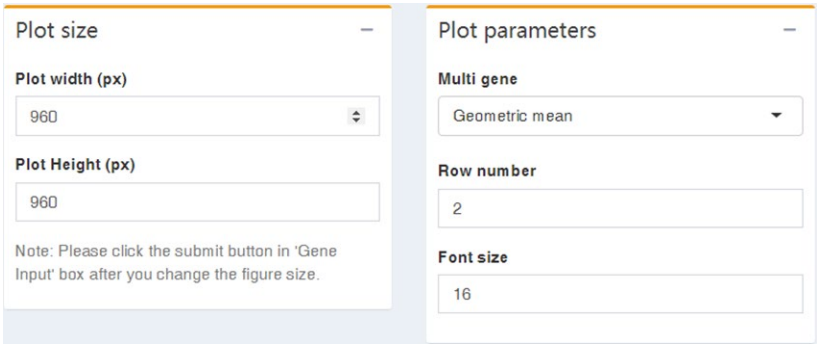


Besides the columns in the metadata (*Sub_Cluster*, *Global_Cluster*, …), you can also color the cells according to the number of genes or UMIs (library size of cells sequenced by Smart-seq2).



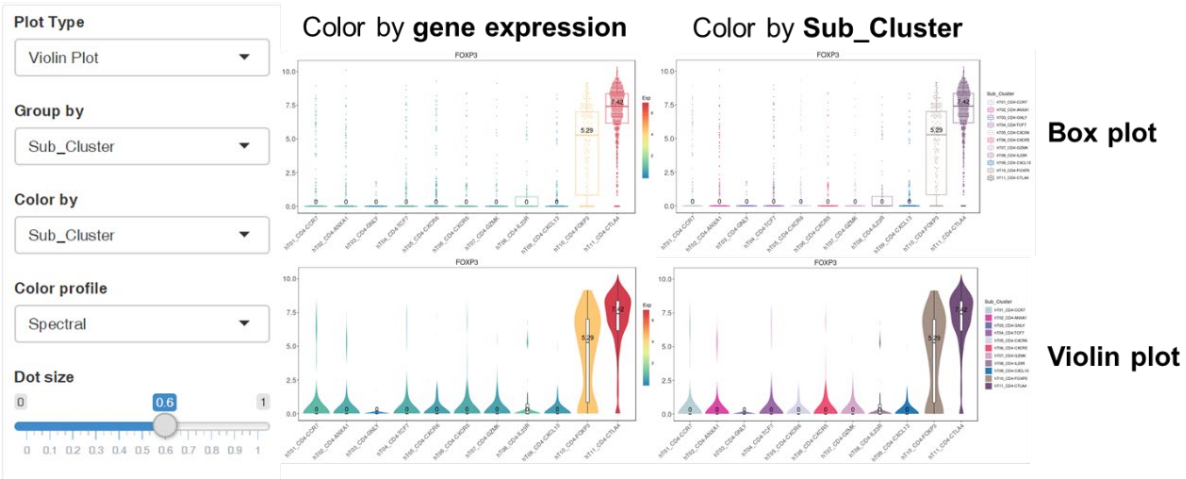
Users can also adjust the plotting parameters at the top of the screen. Just as you do when you subset the data, the changes in "**Plot size**" will only work after clicking the "**Submit**" button. When you type in multiple genes as input, you can change "**Multi gene**" item to "*Geometric mean*", and the geometric mean of all input genes' expression levels will be used as a signature and labeled in the embedding plot. You can also adjust the number of genes plotted in a column with the "**Row number**" item.

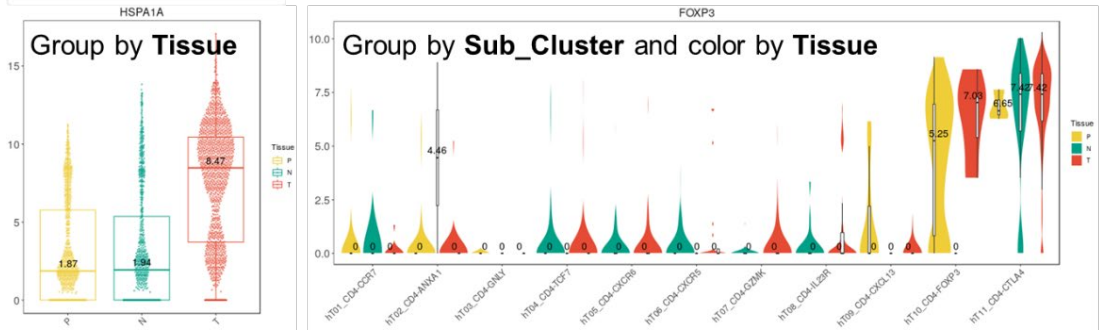




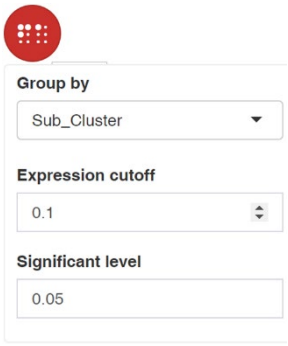
5. Distribution plot

Users can check the gene expression level and distribution pattern under the "**Distribution**" menu. You can switch between box plot and violin plot through the **"Plot type**" item and group all cells by the metadata using "**Group by**" item. When using the box plot, the median expression level will be labeled in the plot. Besides coloring the plot with the information in metadata, users are also allowed to color each group with the mean expression level of cells in it with the "*Exp*" option in "**Color by**" item. If the data are not colored according to their group, then cells in each group will be further divided into different groups following the color option.





6. Significance plot



We allow users to analyze the differences among group mean expression of one single gene (with multiple genes input, the geometric mean signature score will be used) using ANOVA (analysis of variance) model. The groups, which can be selected from the "**Group by**" item, are arranged in the table and renamed from Grp01. The percentage of cells with the gene expressed (defined as the expression level higher than "**Expression cutoff**" item), the mean value and the standard deviation of expression level in each group are also calculated. Tukey's HSD (honestly significant difference) test is used to compare all possible pairs of means and calculate the p-value.

Show 10 entries Search:

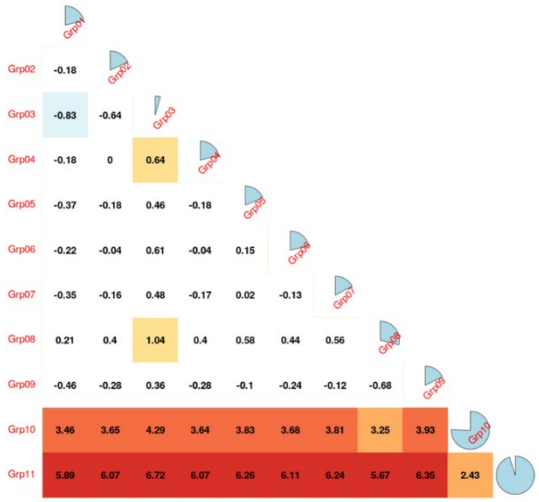
	Grp01	Grp02	Grp03	Grp04	Grp05	Grp06	Grp07	Grp08	Grp09	Grp10
Cluster	hT01_CD4-CCR7	hT02_CD4-ANXA1	hT03_CD4-GNLY	hT04_CD4-TCF7	hT05_CD4-CXCR6	hT06_CD4-CXCR5	hT07_CD4-GZMK	hT08_CD4-IL23R	hT09_CD4-CXCL13	hT10_CD4-FOXP3
Cell Percentage	0.213	0.182	0.044	0.211	0.187	0.214	0.172	0.289	0.177	0.755
Expression Mean	0.88979883	0.70589361	0.06351108	0.70836002	0.52403105	0.66912130	0.54259228	1.10465299	0.42524121	4.35162735
Expression Sd	2.0896841	1.8375538	0.3077708	1.8172533	1.5263244	1.7068519	1.5755338	2.1847660	1.3511737	3.0720139

Showing 1 to 4 of 4 entries Previous 1 Next

Significance

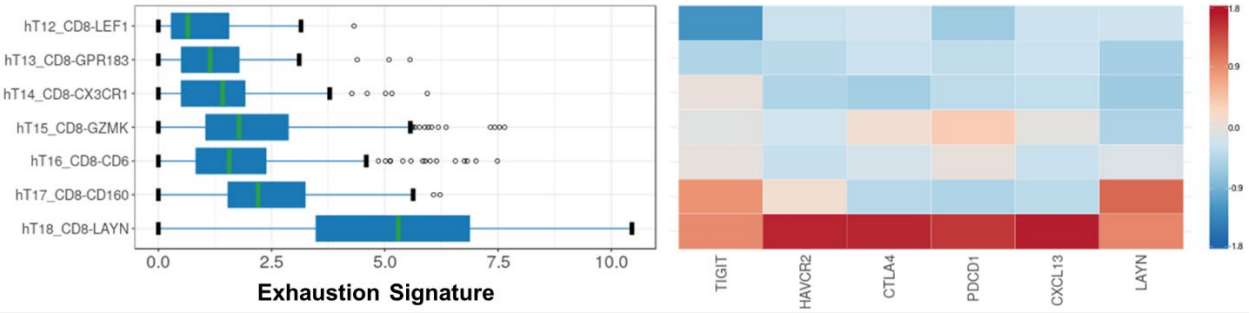
```
F-value is 548.156095202756
p-value is 0
```

Each box in the figure below shows the result of a paired test of group on the given row versus the group on the given column. The number labeled in the box denotes log fold change, while the filled color denotes significance level. The deeper the color, the more significant on expression change is, and only the significant comparisons are marked (default by 0.05, which can be adjusted in the "**Significant level**" item). Red or blue color depends on the sign of logFC only. The fan chart refers to the percentage of cells with the gene expressed in each group.



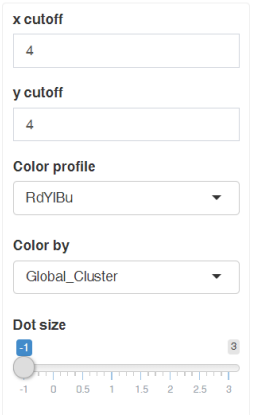
7. Heatmap plot

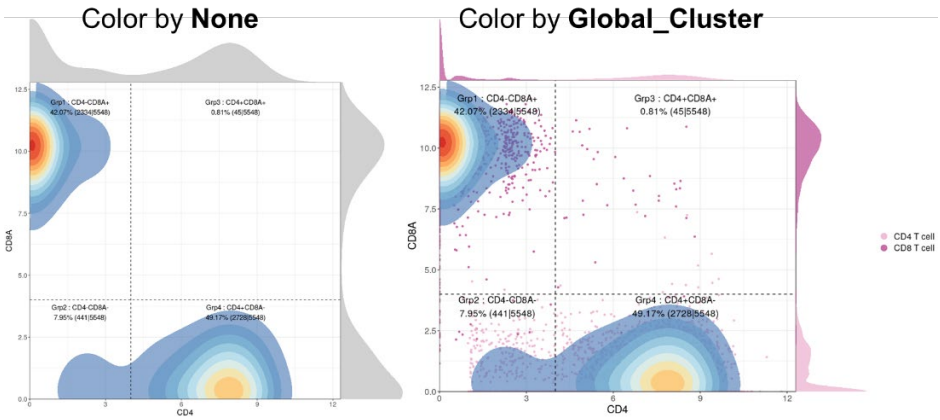
Heatmap plot is recommended when a list of genes (as a signature) are used as input. The expression levels for the signature, per cell, are calculated with the geometric mean of all genes' expression levels and summarized as a boxplot to display the variation of cells in each group (left panel). And the cluster median of each gene is taken per group, and the cluster medians are z-scored across groups (right panel) (Azizi, E., et al., 2018, Cell).



8. In-silico FACS plot

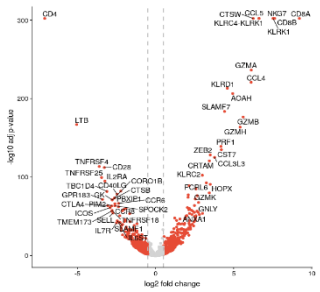
In-silico FACS plot only works when users type in two genes. The cells will be separated into four groups according to the expression levels of the two query genes (adjusted in the "**x cutoff**" and "**y cutoff**" item). The marginal density plot and the points can be colored by the metadata.





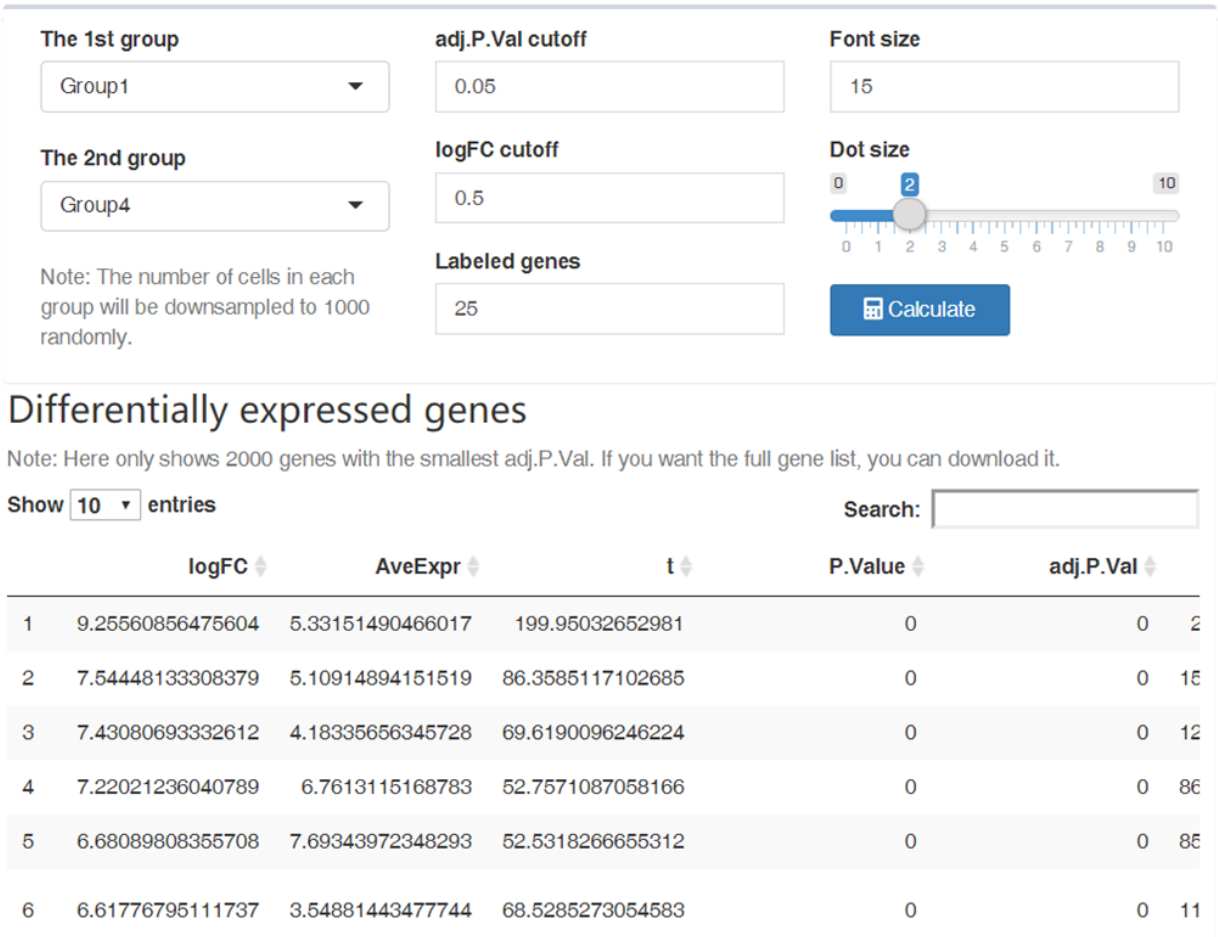
In-silico FACS will separate all cells into Grp1-Grp4:

	GeneA-	GeneA+
GeneB+	Grp1	Grp3
GeneB-	Grp2	Grp4



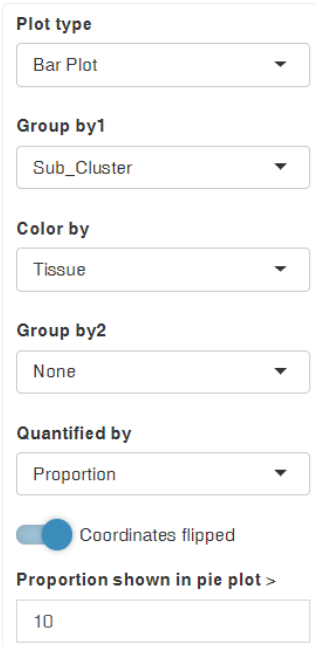
Users can further perform differential expression analysis between two different groups using *limma* package. It should be mentioned that we randomly downsample the number of cells in each group to 1000 to reduce the burden of the server. Users can adjust the cutoff of adjusted p-value and logFC to define the significant genes. And by default, we will label the gene symbol of 25 genes with the highest and lowest (negative) logFC value in the volcano plot, respectively. Users can change this number under the "**Labeled genes**" item. In the data table showing all statistics, we only show 2000 genes with the smallest

adjusted p-value. To get the full gene list, users can click the "**Download**" button and download a .csv file.

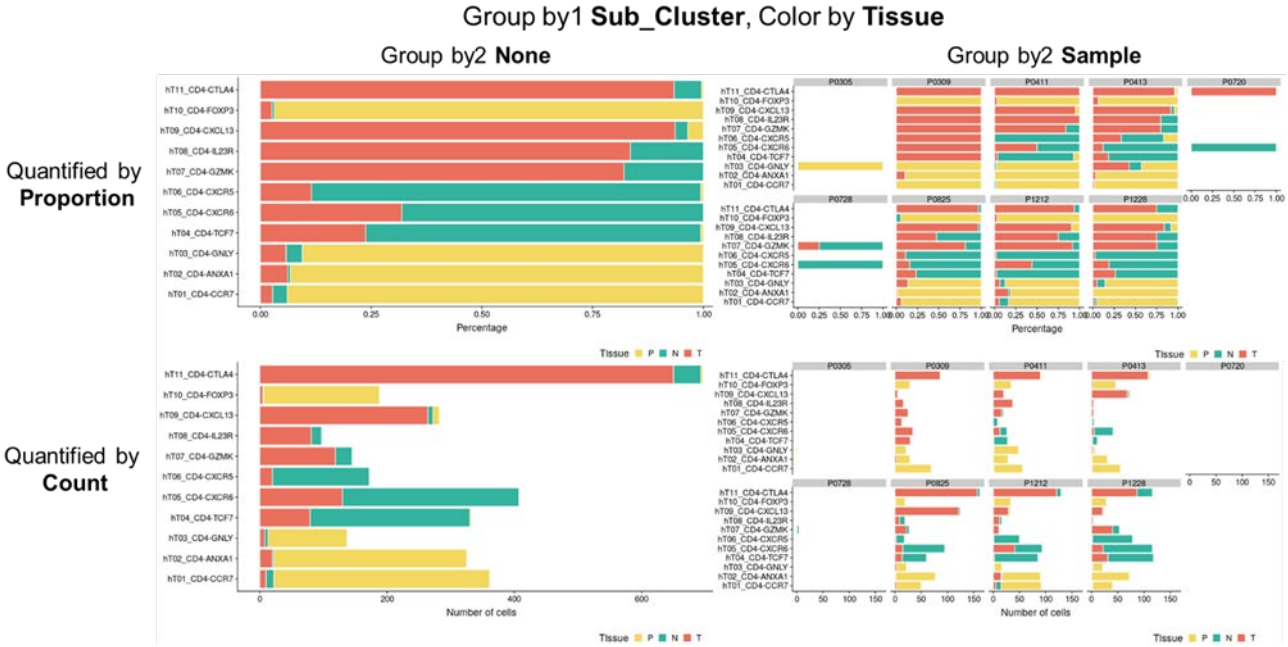


9. Metadata plot

In the metadata plot panel, users can explore the distribution of various metadata combination. For example, users can group the cells by *Sub_Cluster* ("**Group by1**" item) and calculate the tissue distribution in each group ("**Color by**" item). If the "**Group by2**" item is not set to "*None*", then cells will be further subdivided and the plotting area will show a faceted plot. For instance, you can analyze the tissue distribution of cells in each group in different samples as shown in the image below. Sometimes, the proportion can be confusing or misleading when the absolute number of cells is small. That is why we offer the "*Count*" mode in the "**Quantified by**" item, which will show the absolute number of cells.



Users can also explore the distribution of cells using a "*Pie plot*" under "**Plot type**" item. In pie plot mode, the "**Group by2**" and "**Quantified by**" item no longer work.



10.Data table

The input genes and metadata are integrated into a data table which can be searched, rearranged and downloaded. It should be mentioned that the “Exp” column in the data table denotes the geometric mean expression of all the input genes in each cell.

11.Use your own data

It is also easy to explore your own single cell RNA-seq dataset with scDVA. First, you need to download all the R scripts from the GitHub(https://github.com/liziyie/scDVA), including the main code *app.R* and two dependent files *dataprepare_utils.R* and *plot_utils.R*. You also need to maker sure that you have installed all dependent R packages. Users can change the UI, layout or actually displayed contents of the web page through editing *app.R*. This user manual file is saved in the directory **www/**. Some changes of the files in **data/** directory are very essential, including:

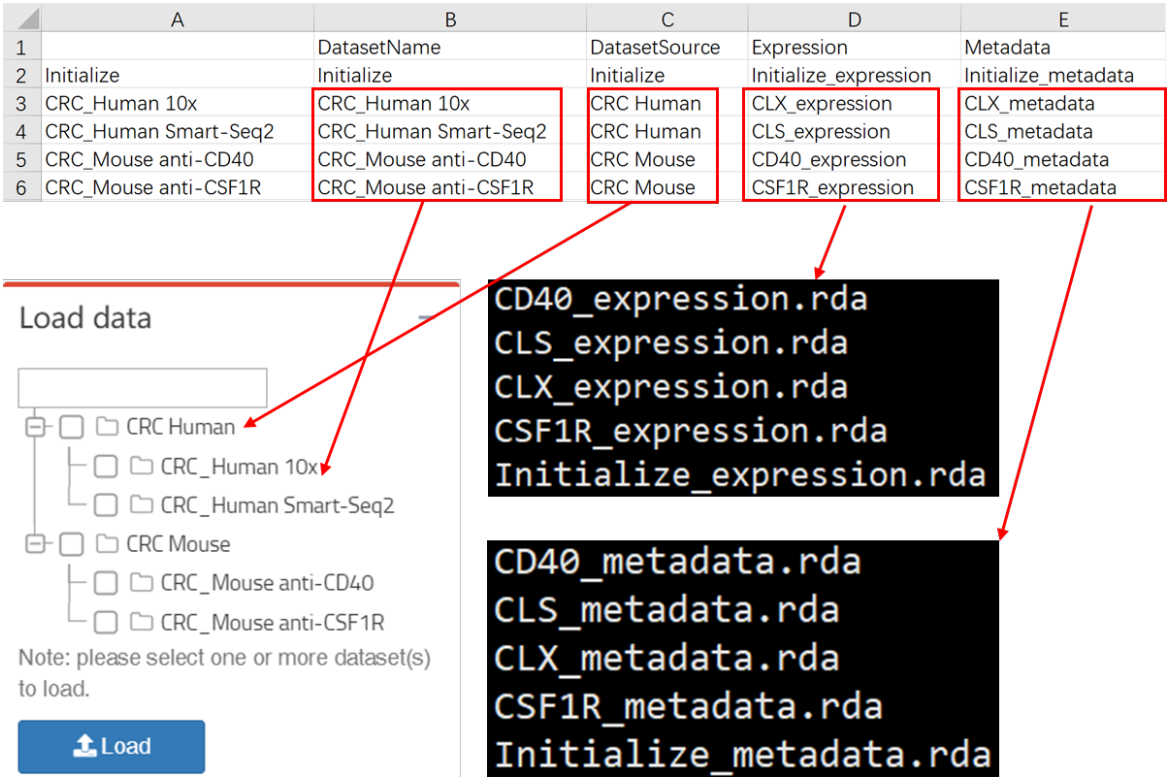
1. *accounts.csv*
   A csv file stores the information of user name and corresponding password. This file must be started with the column name “user” and “password”.

	A	B
1	user	password
2	userA	passwordA
3	userB	passwordB
4	userC	passwordC
5	userD	passwordD

2. *Initialize_expression.rda*, *Initialize_metadata.rda*
   These two R data files are a small dataset used to initialize the website and avoid error reports. The initialization dataset will be hidden from the website after you select a dataset and load it. So please keep these two files in the **data/** directory and do not modify them.

3. *dataset_map.csv*

   This file records the dataset which can be loaded into the scDVA. The "DatasetName" column represents the text rendered in the "**Load data**" panel, and the "DatasetSource" column determines the tree structure. The "Expression" column and "Metadata" column denotes the .rda file name of your own data.



4. *dataset_expression.rda, dataset_metadata.rda*

   Each single cell RNA-seq dataset includes two R data file, dataset_expression.rda and dataset_metadata.rda. To save these .rda files in R, you can use the code like *>save(dataset_expression, file = "dataset_expression.rda", version = NULL)*.

   dataset_expression.rda stores a list named as dataset_expression with two sparse matrix elements, "tpm" and "counts". Some dataset may lack one due to the huge matrix or something else, then just set the missing element as "NA". The sparse matrix can be generated with the function *Matrix(x, sparse = T)* from *Matrix* package. What you should note here is that all the gene symbols as the row names of the matrix should be capitalized, especially in mouse data.

```
> names(CLS_expression)
[1] "tpm"    "counts"
> CLS_expression$tpm[1:4,1:4]
4 x 4 sparse Matrix of class "dgCMatrix"
     N_T_P0104_00001 N_T_P0104_00002 N_T_P0104_00003 N_T_P0104_00004
A1BG               .        1.647398               .               .
NAT2               .        6.701314               .               .
ADA                .        .                      .               .
AKT3               .        .                      .               .
```

```
> CD40_expression$counts
[1] NA
```

   Things are a bit more complicated when generating a dataset_metadata.rda file. There is a data frame named as dataset_metadata in dataset_metadata.rda. All

columns listed in the table below are necessary for scDVA to work.

Column name	Not available	Note
CellName	Essential	The row names of the metadata data frame must be same as the values in the CellName column and the column names of the gene expression matrix
Sample	Fill the column with “None”	The patient ID or library ID
Tissue		The tissue source of the cell
Day		Used in experiments with multiple acquisition time
Treatment		Used in experiments with experimental group and control group, or experiments with different experimental conditions
nUMI, nGene	Essential	The number of UMIs (or total counts of SMART-seq2 data) and genes expressed in each cell
Global_Cluster	Essential	An upper level cluster annotation
Sub_Cluster	Essential	The precised cluster annotation
Global_tSNE_1, Global_tSNE_2	At least one set of coordinates, fill the left one with NA	The tSNE coordinates aiming to show all cells together
Sub_tSNE_1, Sub_tSNE_2		The tSNE coordinates aiming to show all cells in each Global_Cluster respectively
Global_UMAP_1, Global_UMAP_2		The UMAP coordinates aiming to show all cells together
Sub_UMAP_1, Sub_UMAP_2		The UMAP coordinates aiming to show all cells in each Global_Cluster respectively

```
> CLS_metadata[1:4,]
                     CellName Sample Tissue     nUMI nGene  Global_Cluster
N_T_P0104_00001 N_T_P0104_00001  P0104      T 361999.1  1391 Epithelial cell
N_T_P0104_00002 N_T_P0104_00002  P0104      T 528956.2  3498 Epithelial cell
N_T_P0104_00003 N_T_P0104_00003  P0104      T 393885.7  1831 Epithelial cell
N_T_P0104_00004 N_T_P0104_00004  P0104      T 633194.0  1753 Epithelial cell
                       Sub_Cluster Global_tSNE_1 Global_tSNE_2 Sub_tSNE_1
N_T_P0104_00001          hE06_UnIdent     -14.76538      34.42721  8.8460551
N_T_P0104_00002 hE02_Enterocyte-FABP1       2.10684      32.01860 -5.1214092
N_T_P0104_00003 hE02_Enterocyte-FABP1     -11.12333      30.97303  1.4686165
N_T_P0104_00004 hE02_Enterocyte-FABP1     -14.00934      31.04038 -0.6261847
                Sub_tSNE_2 Sub_UMAP_2 Sub_UMAP_1 Global_UMAP_2 Global_UMAP_1
N_T_P0104_00001   12.17789         NA         NA            NA            NA
N_T_P0104_00002   19.14311         NA         NA            NA            NA
N_T_P0104_00003   17.17119         NA         NA            NA            NA
N_T_P0104_00004   16.51511         NA         NA            NA            NA
                 Day Treatment
N_T_P0104_00001 None      None
N_T_P0104_00002 None      None
N_T_P0104_00003 None      None
N_T_P0104_00004 None      None
```

5. *color_panel.R*

   We pre-stored a color panel with 68 different colors named as c68. The users can also use their own color panel by manually setting following vectors’ values: Global_Cluster_color_panel, Sub_Cluster_color_panel, Tissue_color_panel,

Sample_color_panel, Treatment_color_panel and Day_color_panel. All these vectors must be indexed with all the unique elements in the corresponding metadata information. When a figure needs to be colored by the metadata information, the corresponding color panel is preferred. But if any element in the metadata is not found in the index, we'll turn to use c68 to color the figure instead.

6. *Saved_genes_panel.rda*
   A vector named as Saved_genes_panel is saved in this file. Each element in this vector is a character string recording a group of gene symbols separated by commas. And the index of each element is the name of this gene signature. This file is used in the "**Saved**" menu of "**Gene input**" panel.

```
> class(Saved_genes_panel)
[1] "character"
> length(Saved_genes_panel)
[1] 72
> Saved_genes_panel["Ref_TCell"]
                          Ref_TCell
"CD3D,CD3E,CD3G,CD4,IL7R,CD8A,CD8B"
> Saved_genes_panel["Ref_Exhaustion"]
                       Ref_Exhaustion
"TIGIT,HAVCR2,CTLA4,PDCD1,CXCL13,LAYN"
```