**User Guide for “analyze10xgenomCancer” Function**

**Overview**

The **analyze10xgenomCancer** function is designed to process 10x Genomics single-cell RNA-seq data, specifically aimed at cancer research. This function handles data loading, preprocessing, clustering, and visualization, along with advanced analyses like gene correlation network construction and the calculation of Hausdorff distances for cluster comparisons, and Cancer classification. By comparing the function’s output line plot and norm table with the stacked line plots and norm values from the training datasets, cancer can be predicted.

**Usage**

The function can be used in a default mode with minimal required inputs but also allows for extensive customization through various parameters.

**Input and Parameters**

* **Dataset**: 10x Genomics scRNAseq dataset of cancer (HDF5 format).
* **file\_path**: Path to the file containing the 10x Genomics scRNAseq dataset of cancer (HDF5 format).

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter | Values | Parameter type | Remark |
| min.cells | 3 | Default | Seurat object formation |
| min.features | 100 | Default | Seurat object formation |
| Min features threshold | \* | User defined | Filtration parameter |
| Max features threshold | \* | User defined | Filtration |
| Mitochondrial genes % threshold | \* | User defined | Filtration |
| Scale factor | 10000 | Default | Data Normalization |
| nfeatures | 2000 | Default | High variable features identification |
| dimension | \* | User defined | Number of PCs |
| Resolution Parameter | \* | User defined | Clustering |
| alpha | 0.05 | Default | Correlation network parameter |
| Gene correlation threshold | 0.50-0.68 | Default | One at a time with the increment of 0.03 |
| Max possible RMSE value | 7 | Default | Outcome normalization parameter |
| R function: analyze10xgenomCancer() takes the input (\*) of the said parameters and all other parameters set with the default values | | | |

**Example Usage**

Here are two example commands to run the **analyze10xgenomCancer** function based on different user needs:

For a filtered dataset within the sample size range 2500-5000 obtained from the 10x Genomics Database, set the min\_features=0, max\_features= 25000, mt\_threshold= 100 in the following R-function after running the function to use the whole dataset information. The ‘dims’ and ‘resolution’ are user defined parameters. In this example we chose a breast cancer dataset which is of size 3,813; so we set the parameters as mentioned above. We set dims=1:10 and resolution=0.2 in the function as default value to illustrate this particular example. In this case the following command will be fine:

analyze10xgenomCancer(“file\_path \\V1\_Breast\_Cancer\_Block\_A\_Section\_1\_filtered\_feature\_bc\_matrix.h5”)

Or

analyze10xgenomCancer(“file\_path \\V1\_Breast\_Cancer\_Block\_A\_Section\_1\_filtered\_feature\_bc\_matrix.h5”, 25000, 100, 1:10, 0.2)

For a filtered dataset with the sample size larger than 5000 obtained from the 10x Genomics Database, the min\_features, max\_features, mt\_threshold, dims and resolution all are user defined parameters. In this case the following command will be ok with setting the parameters by the user (See Table 2, S1 from the paper “An integrated computational framework utilizing single-cell genomics for precise classification and prediction of multiple cancer types” by Sudarshan Gogoi, Soumen Bera and Amit Chakraborty for details).

analyze10xgenomCancer(“file\_path \\V1\_Breast\_Cancer\_Block\_A\_Section\_1\_filtered\_feature\_bc\_matrix.h5”, min\_features, max\_features, mt\_threshold, dims, resolution)

**Output**

The above function will return a UMAP plot, a line plot and finally a Frobenius norms table at different gene correlation thresholds. If something wrong happens, an error information will be displayed.

* **UMAPPlot**: A plot object showing the UMAP visualization of the data.
* **Frobenius Norm Table:** A table showing **c**hanges of Frobenius Norm in different threshold.
* **LinePlot**: A plot object showing changes of Frobenius Norm in different threshold.
* By comparing the function’s output **LinePlot** and **Frobenius Norm Table** with the stacked line plots and norm values from the training datasets, cancer can be predicted.

**Error Handling**

If there is an issue with the data input or during any of the processing steps, the function will return an error message detailing what went wrong, ensuring that users can make necessary adjustments.