An introduction to RecursiveConsensusClustering package

Recursive Consensus Clustering (RCC), is an unsupervised clustering algorithm for novel subtype discovery from both bulk and single-cell datasets. RCC facilitates the generation of new biological insights through intuitive visualization of clustering results.

Input files:
Config file: config.csv file is provided as the initial input file for clustering the data.
Note: *We observed that providing only protein-coding genes for single-cell RNA-seq datasets yielded better results.*

The parameters provided in the config file should be as follows:

Expression Matrix	Input.csv	
SampleInfo file	SampleInfo.csv	
Minimum Slope Threshold	0.175	*This is a slope measure. Decreasing the slope will result in fewer but tighter clusters*
Minimum #samples for clustering	5	
Minimum line length	30	*This measures the line length of for each k in the CDF plot. Increase the length for fewer and tigher clusters*
% of genes/features for clustering	3	
% of genes to be differentially expressed	10	
Type of dataset	singleCell	
Output dir	./	

The gene quantification data provided should be normalized and log transformed for clustering.

Loading Packages

library(RecursiveConsensusClustering)

We have provided a sample single-cell dataset for testing the algorithm. The same data can be downloaded from here.

To run RCC use the following command:

RCC_clus("config.csv")
Note: *This may take some time depending on the size of the dataset. RCC works efficiently for datasets upto 15000 cells/samples. The time to run the clustering may increase based on size of the data.*

Once the clustering is done, RCC will output the following files:
1) OutputRCC.csv: This file contains the cluster assignment information along with the sample/cell annotation provided in the sampleInfo file
2) genesUsed.csv: This file lists all the genes used by RCC clustering at all levels.

To view the cluster assignments use the following command:

clusterAnnotation("OutputRCC.csv", "cell_type2", "Clusters", ".")

The parameters are as follows:
1) RCC output file
2) Name of the annotation column provided in the sampleInfo file
3) Name of the cluster assignment column, (Incase of RCC the assignment information is always provided in the Clusters column)
4) Output directory

To view the tracking plot use the following command:

trackingPlot("Input.csv", "OutputRCC.csv", "genesUsed.csv", "cell_type2", ".")

The parameters are as follows:
1) Expression matrix file
2)RCC output file
3) The genes used for clustering / a file with marker genes of interest
4) Name of the annotation column provided in the sampleInfo file
5) Output directory
This function will output the tracking plot along with the gene markers found at each level.