snATACClustering Pipeline Manual

Dependencies:

Java Dependencies:

Java (version 8 update 231 or higher)

snATACClusteringTools: https://github.com/UcarLab/snATACClusteringTools

R Dependencies:

Seurat: https://cran.r-project.org/web/packages/Seurat/index.html

EdgeR: https://bioconductor.org/packages/release/bioc/html/edgeR.html

Matrix: https://cran.r-project.org/web/packages/Matrix/index.html

matrixStats: https://cran.r-project.org/web/packages/matrixStats/index.html

Python Dependencies:

MACS2: https://github.com/taoliu/MACS

Installation:

1. Download snATACClustering.sh, snATACClustering.R, and snATACClustering_CONFIG_Default.txt and add them to the same directory.

2. Download or compile the snATACClusteringTools jar from https://github.com/UcarLab/snATACClusteringTools and add this jar (snATACClusteringTools.jar) file to the same directory as the shell and R script.

Configuration File:

The configuration file controls various variables used throughout the pipeline. Please use *snATACClustering_CONFIG_Default.txt* as a template to edit these variables. This file is passed as an argument to the pipeline shell script.

Note: *Ensure that configuration files (and other input files) include a trailing newline.*

Variables and descriptions:

Variable	Description
START	The starting step to run. (Between 1 and 11)
END	The final step to run. (Between 1 and 11)
BINSIZE	The size of the bins (in base pairs (bp)) for computing read counts in pass 1 clustering.
TOPBINS	The top number of bins (those with the most read counts) used for pass 1 clustering.
CLUSTERRES	The initial cluster resolution used by Seurat.
P1MINCLUSTERSIZE	The minimum cluster size for pass 1 clustering.
P1NUMPCA	The number of components for SVD clustering in pass 1 clustering.
PEAKEXT	The peak extension used for defining peaks in pass 2 clustering. If 250 is used, peaks are defined by extending 250bp for both upstream and downstream, defining peaks of 500bp in length
TOPPEAKS	The top number of peaks used for pass 2 clustering.
P2NUMPCA	The number of components for SVD clustering in pass 2 clustering.
TOPVARIABLEPEAKS	The top variable peaks selected for pass 2 clustering.
MAXJOBS	The maximum number of child processes created by the pipeline.
JOBCHECKRATE	The rate (in seconds) to check to see if a job is available.

Running the Pipeline:

To run the pipeline, execute the shell script with the following:

```
snATACClustering.sh Arg1 Arg2 Arg3 Arg4 Arg5
```

Arg1: The full path of a tab delimited file containing the full paths of all the bam files and cell ids. (i.e., the possorted.bam and singlecell.csv file from cellRanger). The format should be as follows (delimited by tabs):

Sample ID1 Path to bam file1 Path to cell id file1
Sample ID2 Path to bam file2 Path to cell id file2

Remember to include the trailing newline after the last sample. The sample id can be any label that can be used for a directory name.

Arg2: The full path of a tab delimited file for the size of each chromosome:

Chromosome Size

See hg38ChromSizes.txt as an example. Note: you can use this file to restrict chromosomes used in the pipeline (such as excluding chromosomes X and Y).

Arg3: The full file path of the configuration file.
Arg4: The full file path of the output directory.
Arg5: The full file path of the directory containing the shell script.

To run in the background, use nohup:

Example:

```
nohup snATACClustering.sh A1 A2 A3 A4 A5 > nohup.out 2>&1 & disown
```

Pipeline Step Information:

Step 1: Generate bin counts from the bam files for each. Read counts are counted within a specified window size across the genome and the number of reads are reported for each bin.

Step 2: Merges bin counts from all samples.

Step 3: Generates a sparse matrix of the top N bins.

Step 4: Performs the first pass clustering based on bin counts. Cells are clustered using Seurat and plotted using UMAP.

Step 5: Splits the bam files into separate bam files based on pass 1 clustering.

Step 6: Calls peaks using MACS2 for each cluster.

Step 7: Peak processing steps. Peak summits are used to define uniform peaks and merged. For overlapping peaks, peaks with the highest quality score are kept, discarding the remaining peaks.

Step 8: Similar to the bin count step, the number of reads within each cell at the selected peak positions are obtained.

Step 9: Merges peak counts across all samples.

Step 10: Generates a sparse matrix of the top N peaks.

Step 11: Performs pass 2 clustering.