Analysis for Single-Cell Assay for Transposase Accessible Chromatin

Jason Buenrostro
Greenleaf & Chang Lab Stanford University
August 2015

Analysis from: Single-cell chromatin accessibility reveals principles of regulatory variation. Jason D. Buenrostro, Beijing Wu, Ulrike M. Litzenburger, Michael L. Gonzales, Dave Ruff, Michael P. Snyder, Howard Y. Chang & William J. Greenleaf. Nature. Doi: 10.1038/nature14590

Single-cell **A**ssay for **T**ransposase **A**ccessible **C**hromatin (scATAC-seq) uses the prokaryotic Tn5 transposase to tag regulatory regions by inserting sequencing adapters into accessible regions of the genome. With scATAC-seq, individual cells are captured and assayed using a programmable microfluidics platform ($C_1^{™}$ single-cell Auto Prep System, Fluidigm). The following protocol has been developed to analyize scATAC-seq data, however, principles of this anlayis infrastructure may be applied to ensemble data sets as well. The following code has been optimized and implemented for human (hg19) and mouse (mm9) data sets, other genomes are not currently supported.

The following protocol is largely a collection of scripts written by me, which is not necesarily designed to be efficient or robust. This particular protocol is organized in the order I find most intuitive, which is not necessarily the most efficient, please use caution when implementing this code and feel free to modify for your specific use.

I. Requirements

Common tools available online:
Bowtie2
Samtools
Picard tools
Homer
MACS2
Bedtools
MEME/FIMO
MATLAB

Additional data sets available online:
Genome fasta file (hg19 and mm9)

Scripts provided here:
singles_Bowtie_PE_aln.v3.sh
shrunFIMO.sh

pyadapter_trim.py
pyMakeVplot.py
pySinglesGetAlnStats.py
pyGetTopSummits.py
pyScorePeaks.py
pyBamBedSinglesCount.py

Additional data files provided here:
BlacklistFiles
TSSfiles
motifFiles

MATLAB code provided here:
addBiasBins.m
calculateNormVar.m
calculateTFVariance.m
readtext.m
rldecode.m

II. Directory organization

The example folder contains the required code and associated files not currently available online. A description of each directory is provided below:

"00_bin" contains bash and python scripts for single-cell tasks. The code "available online" shown above must be in your current path and python packages must be installed. Be sure that all dependancies are met prior to running this code. Specifically you'll need to update a few PATHS to run on your server, listed here:

The "picardPATH" in "singles_Bowtie_PE_aln.v3.sh"
The "memePATH" in "shrunFIMO.sh"

If further assistance is needed, please contact your system adminstrator and not the author of this protocol.

"01_additionalData" contains required files for processing scATAC-seq data. This includes: i) a custom blacklist set of regions to exclude, ii) motif files for mapping known motifs and iii) TSS files for calculating enrichment at TSSs.

"02_fastq_source" contains fastqs from an example data set (resting GM12878 cells). For convienience, the trimmed fastqs for all 96 wells have been precomputed.

"03_alignment" contains two files, i) README.txt which describes the steps used to do the alignments and ii) 140905_GMdata.xls which contains basic information of the cells captured. When using this code base, be sure that your annotation file is saved as tab delimited and contains unix compatible new line characters.

"04_motifProcessing" contains a README.txt used for mapping motifs in peaks.

"05_MATLAB" contains: i) runAnalysis.m, code for calculating variability and making useful plots and ii) 00_JDBcode a directory of useful tools required for calculating variability.

III. Align scATAC-seq data

1. First modify the "picardPATH" in README.txt. Then simply follow the directions as seen here.
2. Call peaks with MACS2 as shown in README.txt, you'll need at least 50,000 peaks. If you don't get that from MACS2 you'll need to acquire bulk data or you'll need to merge with additonal single-cell data.
3. If you would like to run the entire length of the script, simply type "bash README.txt".

IV. Annotate peaks and process motifs

1. First modify "genomeFasta" in the README.txt directory.
2. If you would like to run the entire length of the script, simply type "bash README.txt".
3. I note here, that we used Tn5 bias for the manuscript, however, we find that GC bias is very correlated to Tn5 bias and is functionally equivalent to Tn5 bias. For simplicity I have provided code to normalize to GC bias instead of Tn5 bias.

V. Run MATLAB to calculate variability

1. To run the MATLAB code download the directory to your personal computer. To save time you only have to download "04_motifProcessing" and "05_MATLAB".
2. Open "runAnalysis.m" in MATLAB.
3. Run the code written in "runAnalysis.m". Here I demonstrate a few simple use cases of the variability analysis which includes QC and data figures.

VI. Additional discussion

See associated supplementary material provided in the manuscript. I have included a version containing the analysis methods in this directory as well. If you haven't already be sure to read the supplementary discussion section (section #7) before running this code.