Single-cell ATAC-seq Signal Extraction and Enhancement with SCATE

Zhicheng Ji

Johns Hopkins University,
Baltimore, Maryland, USA
zji4@jhu.edu

Weiqiang Zhou

Johns Hopkins University,
Baltimore, Maryland, USA
wzhou14@jhu.edu

Hongkai Ji

Johns Hopkins University,
Baltimore, Maryland, USA
hji@jhsph.edu

December 25, 2018

Contents

1 Introductions

Single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq) is a new technology for measuring genome-wide regulatory element activities in single cells. With the ability to analyze cells' distinct behaviors in a heterogeneous cell population, this technology is rapidly transforming biomedical research. Data produced by scATAC-seq are highly sparse and discrete. Existing computational methods typically use these data to analyze regulatory pathway activities in single cells. They cannot accurately measure activities of individual cis-regulatory elements (CREs) due to data sparsity. SCATE is a new

statistical framework for analyzing scATAC-seq data. SCATE adaptively integrates information from co-activated CREs, similar cells, and publicly available regulome data to substantially increase the accuracy for estimating activities of individual CREs. We show that one can use SCATE to identify cell subpopulations and then accurately reconstruct CRE activities of each subpopulation. The reconstructed signals are accurate even for cell subpopulations consisting of only a few cells, and they significantly improve prediction of transcription factor binding sites. The accurate CRE-level signal reconstruction makes SCATE an unique tool for analyzing regulatory landscape of a heterogeneous cell population using scATAC-seq data.
The main functions of SCATE is demonstrated using the following example of 10 GM12878 and 10 K562 scATAC-seq samples.

2 Read in and Preprocessing Data

The following chunk of data read in the bam files into R as GRanges object.

```
suppressMessages(library(SCATE))

## Warning:  no function found corresponding to methods exports from
'XVector' for:  'concatenateObjects'

set.seed(12345)
bamlist <- list.files(paste0(system.file(package="SCATE"),"/extdata/example"),full.names = T
satac <- sapply(sapply(bamlist,readGAlignmentPairs),GRanges)
```

The function satacprocess preprocesses the scATAC-seq reads required for other functions of SCATE. It transforms the reads into the midpoint of the reads and filter out samples with small library size.

```
satac <- satacprocess(satac)
```

3 Cell Clustering (Optional)

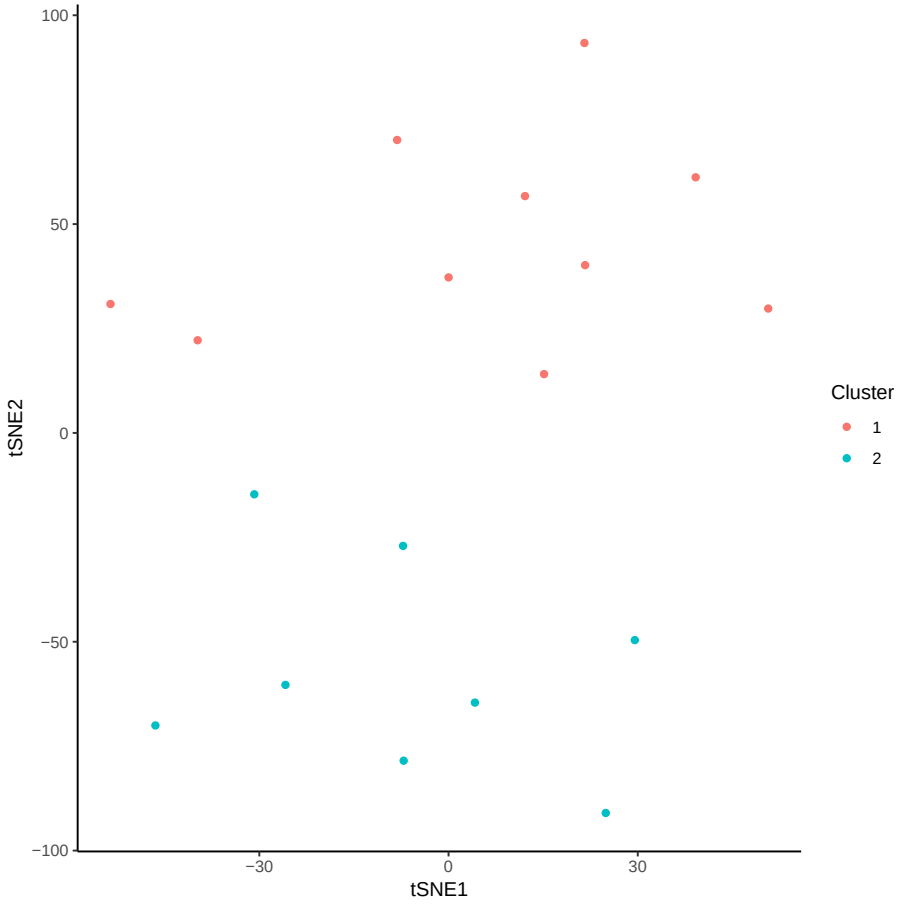
The function cellclu will cluster cells based on averaged signal of CRE clusters.

```
cellclu <- cellcluster(satac,genome="hg19",perplexity=5)

## Warning in mclustBIC(data = structure(c(-8.14851344810963, 21.5448818684518,
:  The presence of BIC values equal to NA is likely due to one or more
of the mixture proportions being estimated as zero, so that the model
estimated reduces to one with a smaller number of components.
```

Check the results of clustering:

```
library(ggplot2)
ggplot(data.frame(tSNE1=cellclu$tsne[,1],tSNE2=cellclu$tsne[,2],Cluster=as.factor(cellclu$cl
```



4 SCATE

The function SCATE will perform SCATE on one or multiple scATAC-seq samples.

```
SCATEres <- SCATE(satac[cellclu$cluster==1],genome="hg19")
summary(SCATEres)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.5255  1.0816  1.0319  1.4562  8.9668
```

5 Peak Calling

The function peakcall will perform peak calling on the SCATE results.

```
peakres <- peakcall(SCATEres,genome="hg19")
peakres[[1]]

## GRanges object with 12348 ranges and 0 metadata columns:
##           seqnames            ranges strand
##              <Rle>         <IRanges>  <Rle>
##       [1]     chr1     713600-714599      *
##       [2]     chr1     762400-763199      *
##       [3]     chr1     858800-859399      *
##       [4]     chr1     935200-936599      *
##       [5]     chr1     948400-949199      *
##       ...      ...               ...    ...
##   [12344]    chr21 47648200-47649399      *
##   [12345]    chr21 47705600-47706999      *
##   [12346]    chr21 47743400-47744999      *
##   [12347]    chr21 47878200-47879199      *
##   [12348]    chr21 48054600-48056199      *
##   -------
##   seqinfo: 23 sequences from an unspecified genome; no seqlengths
```