MAESTRO: Model-based AnalysEs of Single-cell Transcriptome and RegulOme

Ming (Tommy) Tang
Twitter: @tangming2005
X Shirley Liu group
Senior scientist at Dana-Farber Cancer Institute
https://divingintogeneticsandgenomics.rbind.io/

https://cimac-network.org/
Cancer Immunological Data Commons (CIDC)
Analyzing single-cell omics data give insights to biological functions

Tim Stuart & Rahul Satjia, Nat Rev Genet, 2019

Wager et al, Nat Biotech, 2016
Workflow of a typical* scRNA-seq analysis

Library size etc.
SCTransform in Seurat

Dimension Reduction:
PCA
TSNE
UMAP

Credit to Peter Hickey

MAESTRO, an integrative analysis workflow based on Snakemake for scRNA-seq and scATAC-seq

https://github.com/liulab-dfci/MAESTRO
MAESTRO supports data from multiple scRNA-seq and scATAC-seq protocols

scRNA-seq
- Smart-seq2
  Picelli et al., 2014
- Drop-seq/indrop
  Macosko et al., 2015

scATAC-seq
- Fluidigm C1
  Buenrostro et al., 2015
- sci-ATAC-seq/dsci-ATAC-seq
  Buenrostro et al., 2015, 2019

10x genomics
- 2016
- 2018
MAESTRO performs quality control at both bulk and single cell level

- Bulk level
  - Mapping summary
  - Duplicated ratio
  - Mitochondria ratio
  - Reads distribution
  - Fragment size distribution
  - Fraction of reads in peaks, promoters

- Single-cell level
  - ScRNA: Number of UMIs and genes covered
  - ScATAC: total number of reads per cell and fraction of reads in promoters.
Normalization, expression index and peak calling in MAESTRO

• scRNA
  • STARsolo to calculate UMI count. (much faster than Cellranger: hours vs days)
  • Gene count by cell matrix as output.

• scATAC
  • Add cell-barcode to fastq read name, align with minimap2. (much faster than cellranger: hours vs days)
  • Aggregate single-cell samples, perform peak calling using MACS2.
  • Support user defined peak regions.
  • Support peak calling from short fragments (less than 150bp).
  • peak by cell matrix as output.
MAESTRO uses the graph-based clustering for scRNA-seq and scATAC-seq

- **Dimension reduction**
  - ScRNA: PCA
  - ScATAC: Latent semantic index (LSI)
- **Build KNN graphs**
- **Louvain algorithm to detect communities and identify clusters**
- **Umap visualization**
MAESTRO carries out differential expression analysis and supports automatic cell type annotation based on gene signatures

- Differential gene analysis
  - Wilcoxon rank sum test
  - DESeq2
  - MAST
  - Presto

- Differential Peak analysis
  - Presto
  - https://github.com/immunogenomics/presto

- Celltype annotation
  - Gene signature based celltype annotation
  - Logfc based celltype scoring
  - Support user defined gene signatures

Annotated using CIBERSORT signatures
MAESTRO can identify important transcription regulators for both scRNA-seq and scATAC-seq

Based on up-regulated genes in each cluster

LISA@ http://lisa.cistrome.org/

Cistrome Data Browser http://cistrome.org/db/#/
http://dbtoolkit.cistrome.org/
MAESTRO provides integrated clustering of scRNA-seq and scATAC-seq

ScRNA and scATAC integrated Human pbmc from 10x
MAESTRO provides a simple regulatory potential (RP) model to estimate gene activity for scATAC-seq

- Gene activity
  - Single-cell regulatory potential (ScRP)
  - Decay distance $d_0 = 10\text{kb}$

$$S_g = \sum_{i=1}^{k} 2 \frac{d_i}{d_0}$$
MAESTRO provides an additional enhanced regulatory potential (RP) model to estimate gene activity.
Enhanced RP-model better model the gene activity compared with other methods

Chenfei Wang et.al Genome Biology 2020
Summary

• MAESTRO is an integrative scRNA-seq and scATAC-seq analysis workflow supporting multiple experimental protocols.
• MAESTRO provides utilities from the basic alignment, QC to high level functional analysis
• MAESTRO follows the best practice for single cell clustering.
• MAESTRO enables transcription regulation analysis for both scRNA-seq and scATAC-seq data based on CistromeDB.
• ScATAC-seq regulatory potential (RP) score outperforms other existing methods in predicting gene expression level and integration with scRNA-seq data.
The future of MAESTRO

• keep adding new features and fixing bugs.
• faster processing scATACseq data.
• multi-sample scRNAseq and scATACseq processing.

https://github.com/liulab-dfci/MAESTRO
Full solution of MAESTRO can be installed using Conda
Acknowledgements

CIDC Bioinformatics team:

- Clara Cousins
- Len Taing
- Gali Bai
- Yang Liu

CIDC Software team:

- Ethan Cerami
- James Lindsay
- Pavel Trukhanov
- Roshni Biswas
- Jacob Lurye
- Stephen Van Nostrand
- Joyce Hong

Tao Liu lab:

- Tao Liu

Liu lab:

- X Shirley Liu
- Chenfei Wang
- Dongqing Sun
- Xin Huang
- Changxin Wan
- Ziyi Li
- Li Song
- Allen Lynch
- Cliff Meyer

DFCI CIO:

- Mohamed Uduman
- Jason Weirather

DFCI CFCE:

- Henry Long
MAESTRO is easy to install and generates an html report for various QC metrics

https://github.com/liulab-dfci/MAESTRO

Full solution of MAESTRO can be installed using Conda.
| Dataset 1: PBMC scATAC-seq, 173,477 peaks x 9,361 cells, 8 cores |

<table>
<thead>
<tr>
<th>Time (minutes)</th>
<th>MAESTRO</th>
<th>scABC</th>
<th>cisTopic</th>
<th>chromVAR</th>
<th>Cicero</th>
<th>Seurat</th>
<th>snapATAC</th>
<th>Garnett</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene activity</td>
<td>7.1</td>
<td>180.0</td>
<td>2.2</td>
<td>150.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dimensionality reduction</td>
<td>11.5</td>
<td>59.3</td>
<td>26.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clustering</td>
<td>0.2</td>
<td>186.0</td>
<td>1.2</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Differential peak</td>
<td>0.2</td>
<td></td>
<td></td>
<td>41.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell-type annotation</td>
<td>0.1</td>
<td></td>
<td></td>
<td>5.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regulator identification</td>
<td>4.9</td>
<td>9.2</td>
<td>6.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Memory (GB)</th>
<th>MAESTRO</th>
<th>scABC</th>
<th>cisTopic</th>
<th>chromVAR</th>
<th>Cicero</th>
<th>Seurat</th>
<th>snapATAC</th>
<th>Garnett</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene activity</td>
<td>10.4</td>
<td></td>
<td></td>
<td>103.8</td>
<td>23.6</td>
<td>36.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dimensionality reduction</td>
<td>8.6</td>
<td>5.6</td>
<td>34.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clustering</td>
<td>3.9</td>
<td>95.4</td>
<td>5.7</td>
<td>8.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Differential peak</td>
<td>8.7</td>
<td></td>
<td></td>
<td>40.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell-type annotation</td>
<td>8.4</td>
<td></td>
<td></td>
<td>11.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regulator identification</td>
<td>12.1</td>
<td>5.4</td>
<td>2.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Dataset 2: BCC scATAC-seq, 530,771 peaks x 37,818 cells, 8 cores |

<table>
<thead>
<tr>
<th>Time (minutes)</th>
<th>MAESTRO</th>
<th>scABC</th>
<th>cisTopic</th>
<th>chromVAR</th>
<th>Cicero</th>
<th>Seurat</th>
<th>snapATAC</th>
<th>Garnett</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene activity</td>
<td>44.1</td>
<td></td>
<td></td>
<td>NA</td>
<td>NA</td>
<td>363.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dimensionality reduction</td>
<td>8.9</td>
<td>246.0</td>
<td>54.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clustering</td>
<td>1.2</td>
<td>NA</td>
<td>6.6</td>
<td>0.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Differential peak</td>
<td>1.8</td>
<td></td>
<td></td>
<td>319.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell-type annotation</td>
<td>0.1</td>
<td></td>
<td></td>
<td>9.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regulator identification</td>
<td>8.0</td>
<td>12.9</td>
<td>21.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Memory (GB)</th>
<th>MAESTRO</th>
<th>scABC</th>
<th>cisTopic</th>
<th>chromVAR</th>
<th>Cicero</th>
<th>Seurat</th>
<th>snapATAC</th>
<th>Garnett</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene activity</td>
<td>38.9</td>
<td></td>
<td></td>
<td>NA</td>
<td>NA</td>
<td>63.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dimensionality reduction</td>
<td>39.7</td>
<td>25.3</td>
<td>37.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clustering</td>
<td>16.8</td>
<td>NA</td>
<td>26.7</td>
<td>11.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Differential peak</td>
<td>36.9</td>
<td></td>
<td></td>
<td>41.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell-type annotation</td>
<td>40.3</td>
<td></td>
<td></td>
<td>52.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regulator identification</td>
<td>47.7</td>
<td>16.4</td>
<td>10.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NA: Memory usage larger than 380G and crashed.