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/

CIMAC-CIDC
Immuno-Oncology
Biomarkers Network

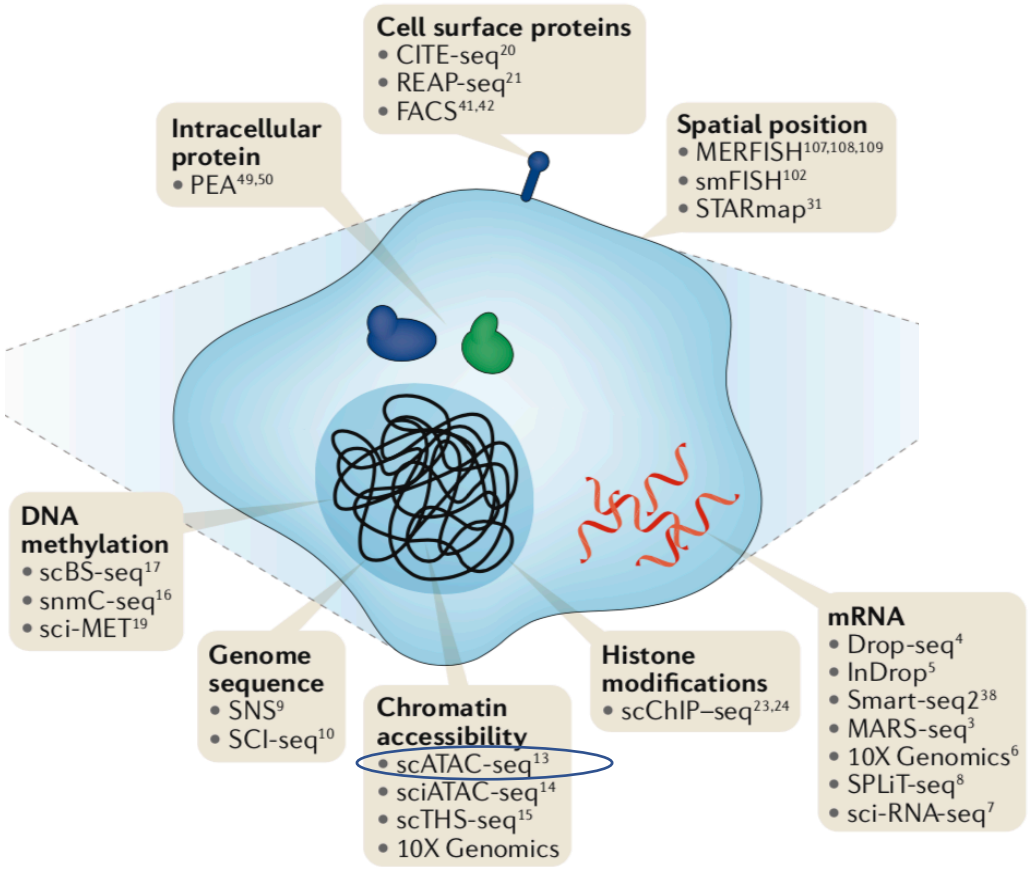
NIH NATIONAL CANCER INSTITUTE

Dana-Farber
Cancer Institute

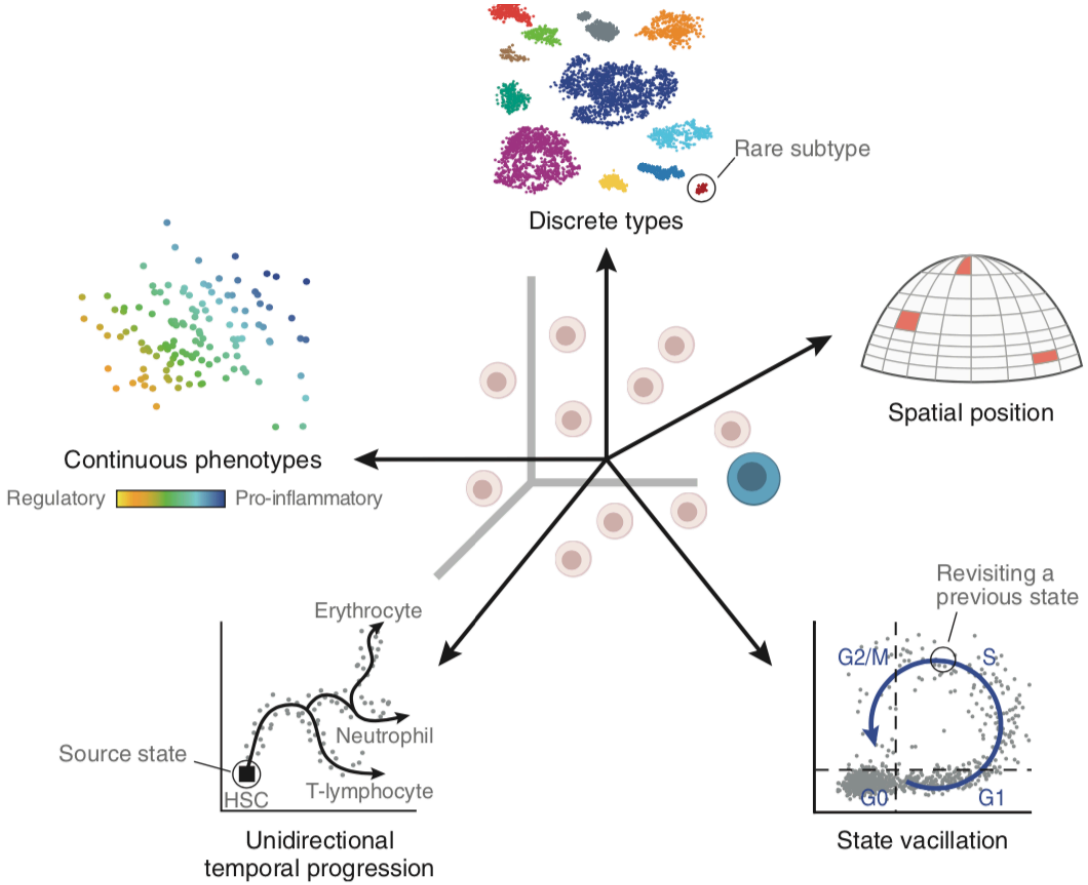
https://cimac-network.org/
Cancer Immunological Data Commons (CIDC)

Chenfei Wang et al. Genome Biology 2020

Analyzing single-cell omics data give insights to biological functions

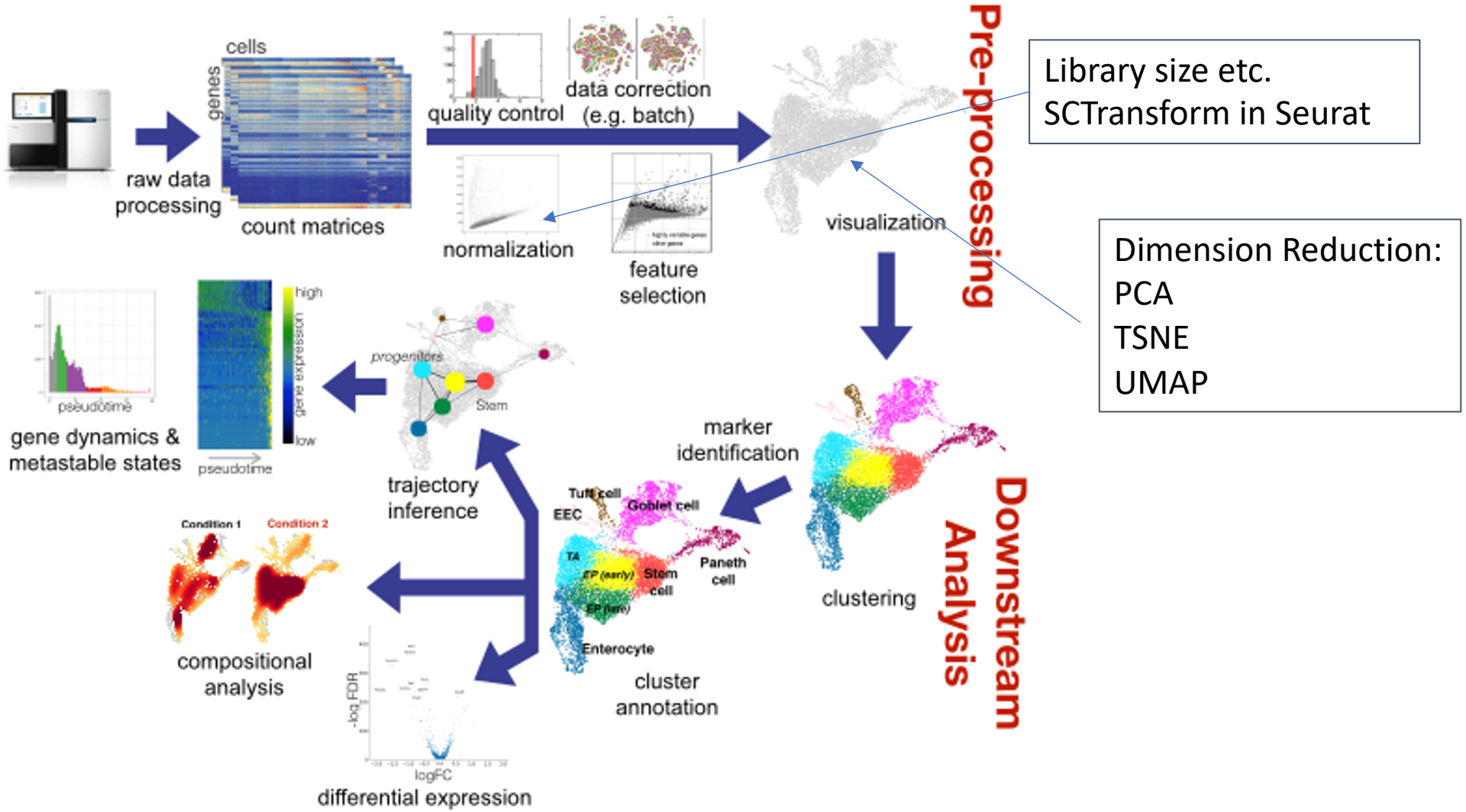


Tim Stuart & Rahul Satjia,
Nat Rev Genet, 2019



Wager et al,
Nat Biotech, 2016

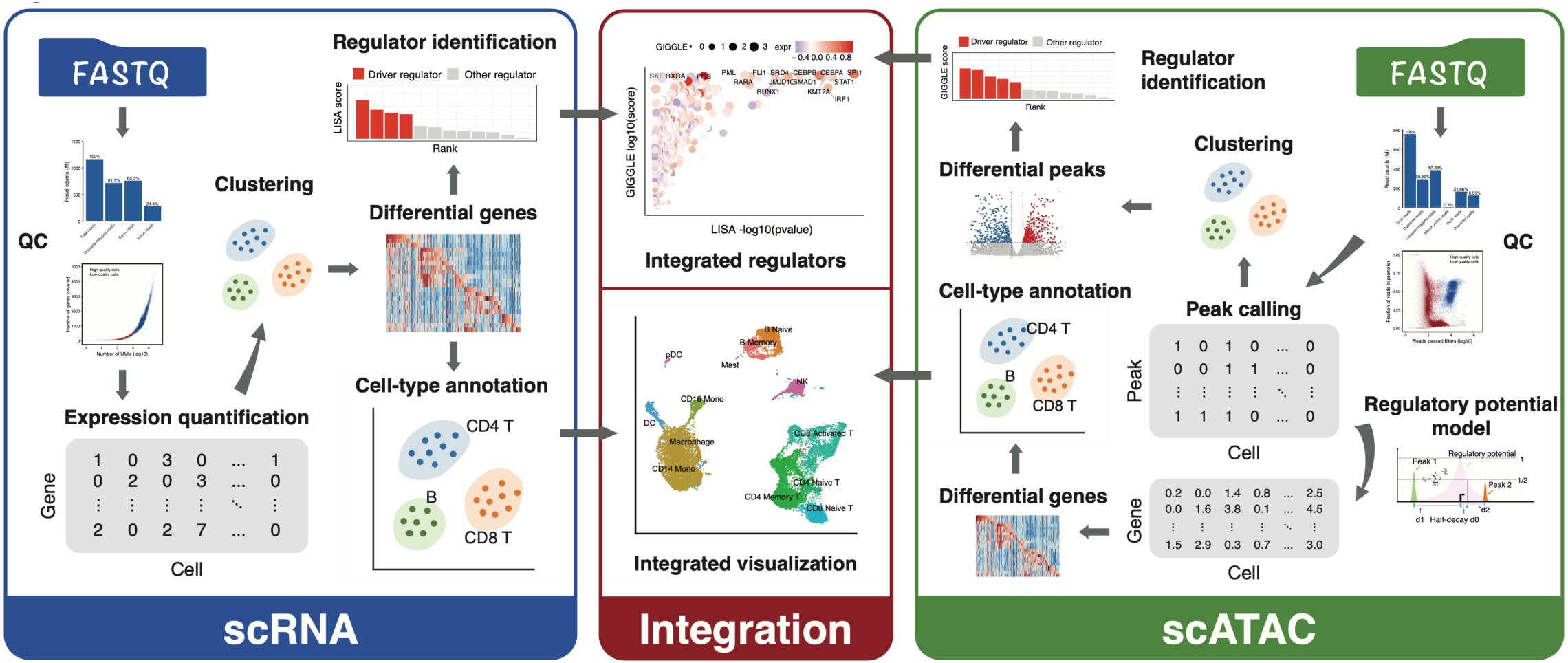
Workflow of a typical* scRNA-seq analysis



Credit to Peter Hickey

Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, (2019).

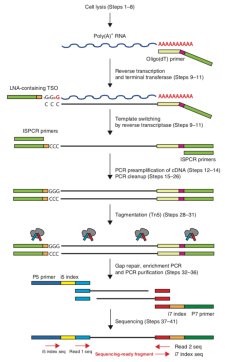
MAESTRO, an integrative analysis workflow based on Snakemake for scRNA-seq and scATAC-seq



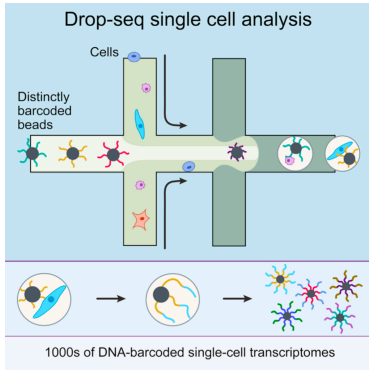
https://github.com/liulab-dfci/MAESTRO

MAESTRO supports data from multiple scRNA-seq and scATAC-seq protocols

scRNA-seq



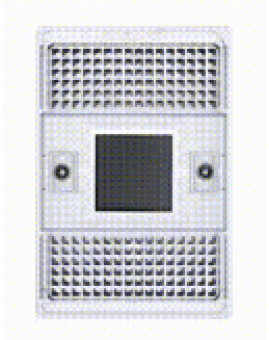
Smart-seq2
Picelli et al., 2014



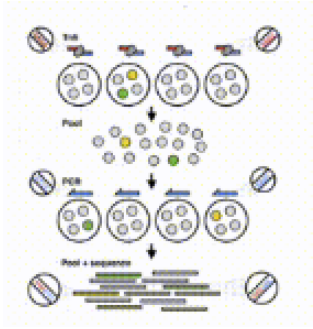
Drop-seq/indrop
Macosko et al., 2015

10x genomics
2016

scATAC-seq



Fluidigm C1
Buenrostro et al., 2015



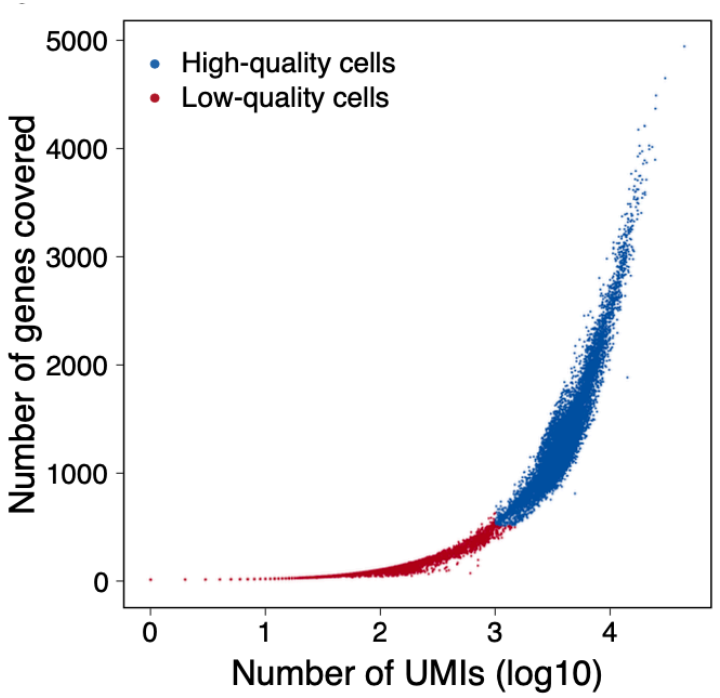
sci-ATAC-seq/dsci-ATAC-seq
Buenrostro et al., 2015, 2019

10x genomics
2018

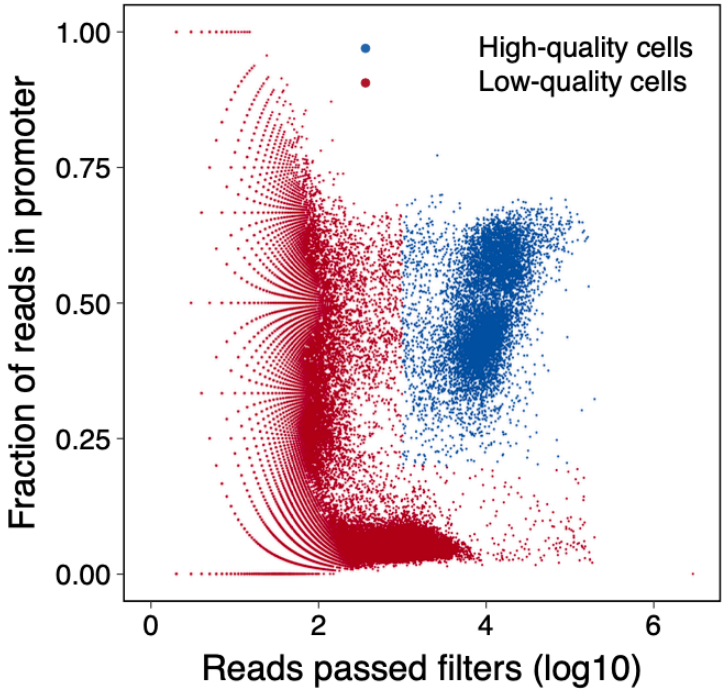
MAESTRO performs quality control at both bulk and single cell level

- 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.

scRNA single-cell QC



scATAC single-cell QC



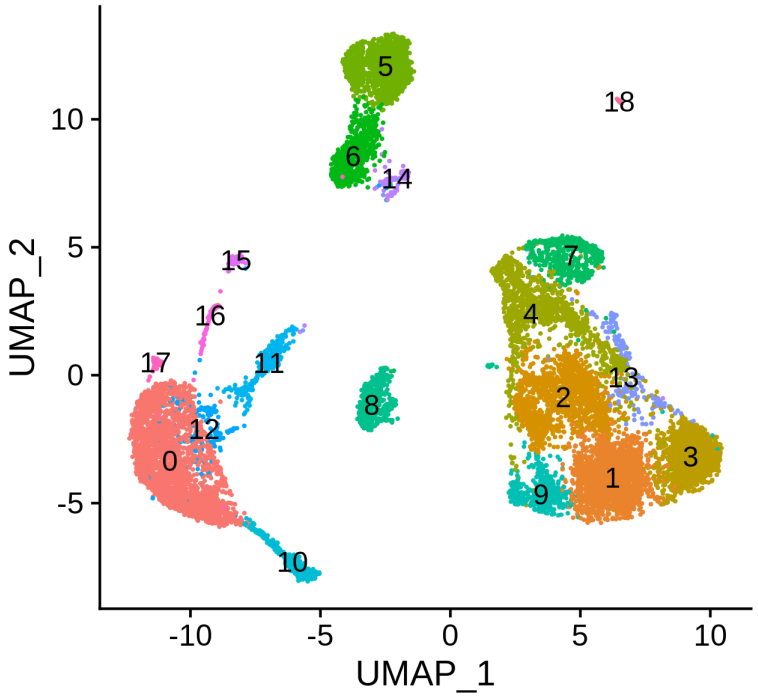
Normalization, expression index and peak calling in MAESTRO

- 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.

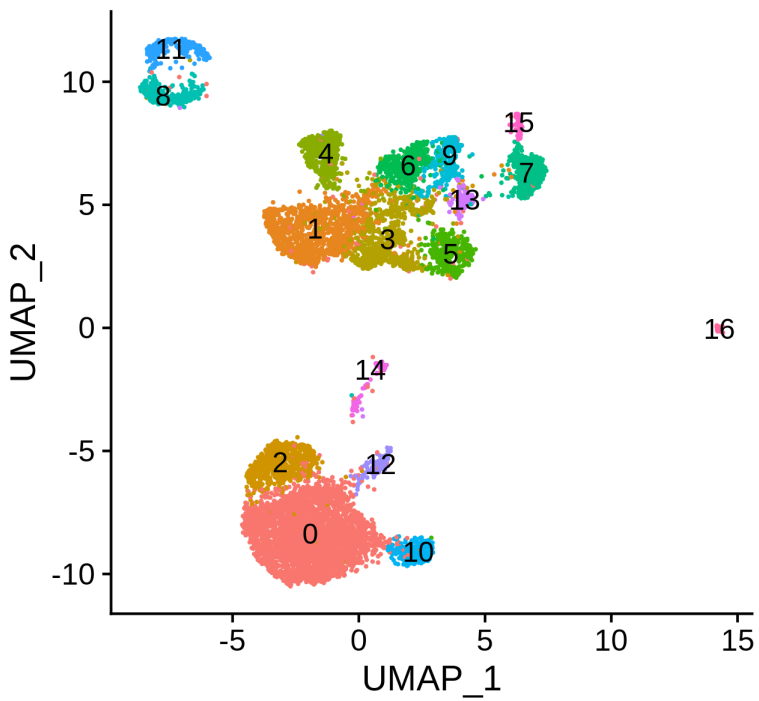
MAESTRO uses the graph-based clustering for scRNA-seq and scATAC-seq

- Dimension reduction
  - ScRNA: PCA
  - ScATAC: Latent semantic index (LSI)
- Build KNN graphs
- Louvain algorithm to detect communities and identify clusters
- Umap visualization

ScRNA
Human pbmc 12k from 10x
res = 0.6



ScATAC
Human pbmc 10k from 10x
res = 0.6



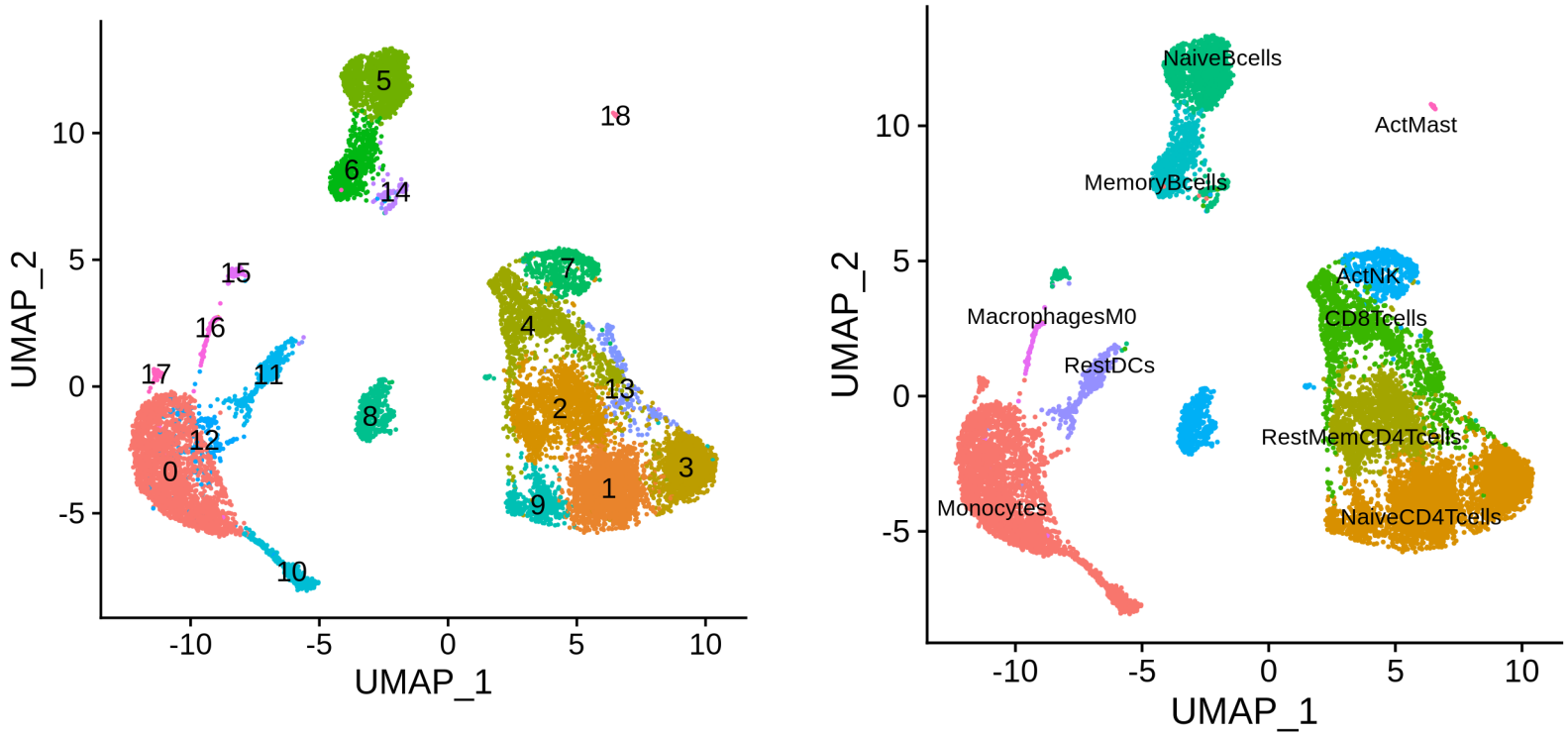
MAESTRO carries out differential expression analysis and supports automatic cell type annotation based on gene signatures

- 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

ScRNA

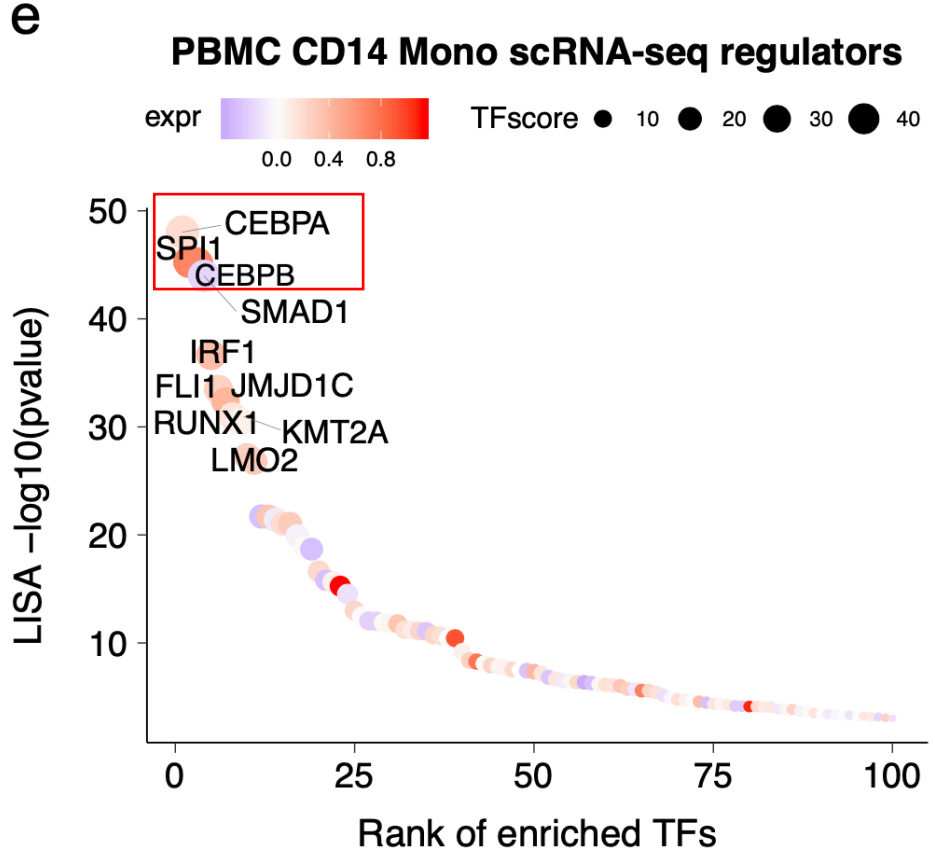
Human pbmc 12k from 10x



Annotated using CIBERSORT signatures

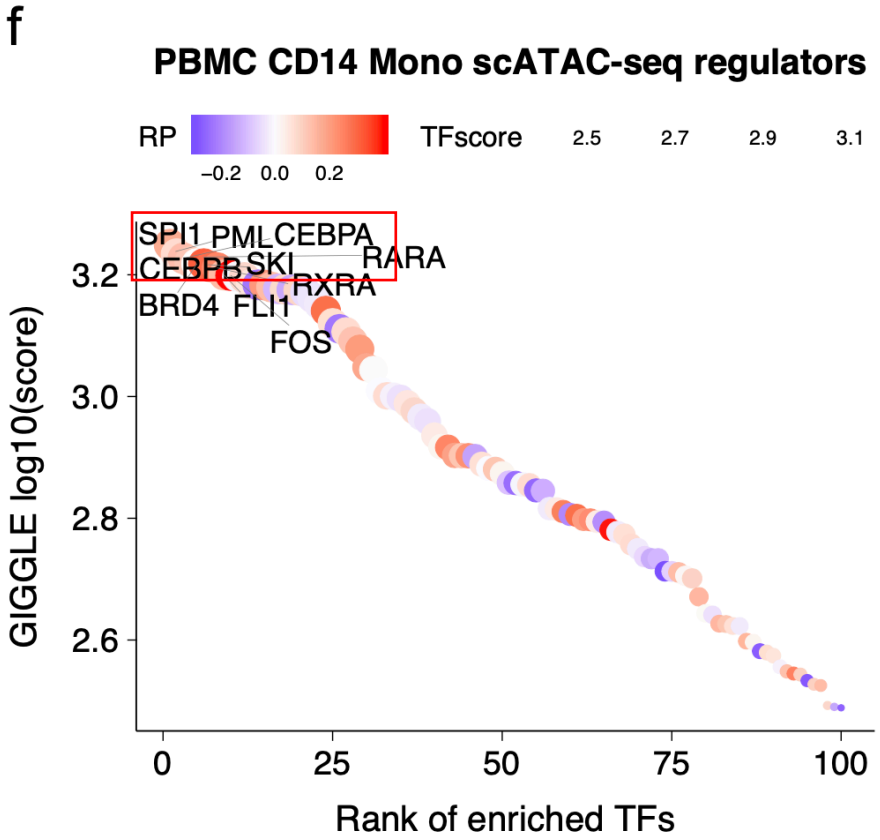
MAESTRO can identify important transcription regulators for both scRNA-seq and scATAC-seq

e



Based on up-regulated genes in each cluster

f



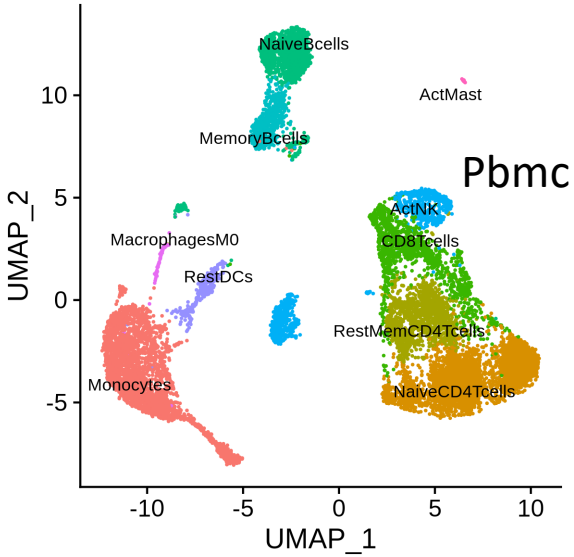
Based on positive peaks in each cluster

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

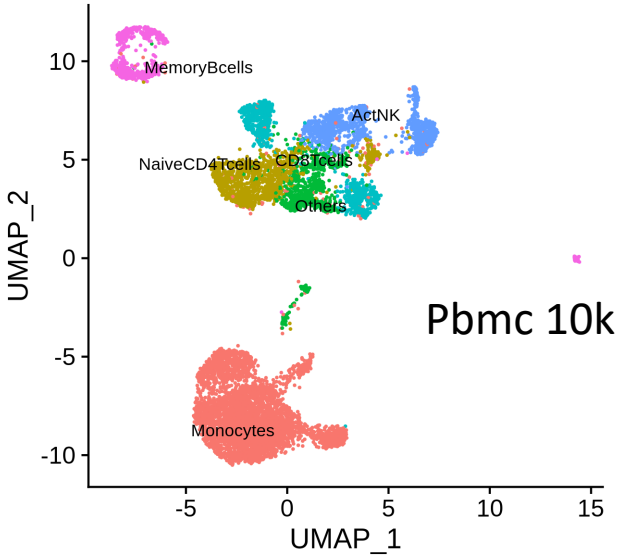
Cistrome Data Browser

http://cistrome.org/db/#/
http://dbtoolkit.cistrome.org/

MAESTRO provides integrated clustering of scRNA-seq and scATAC-seq



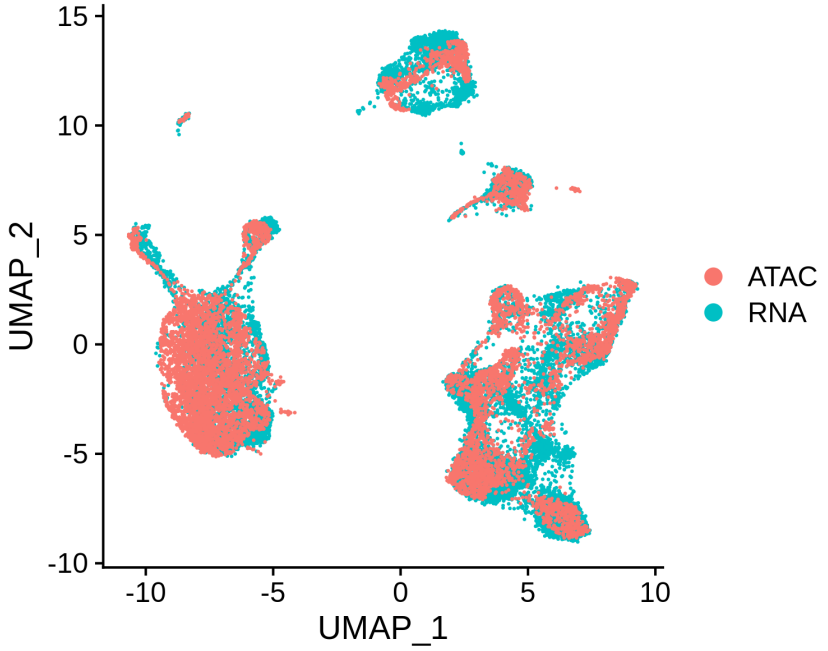
Pbmc 12k scRNA

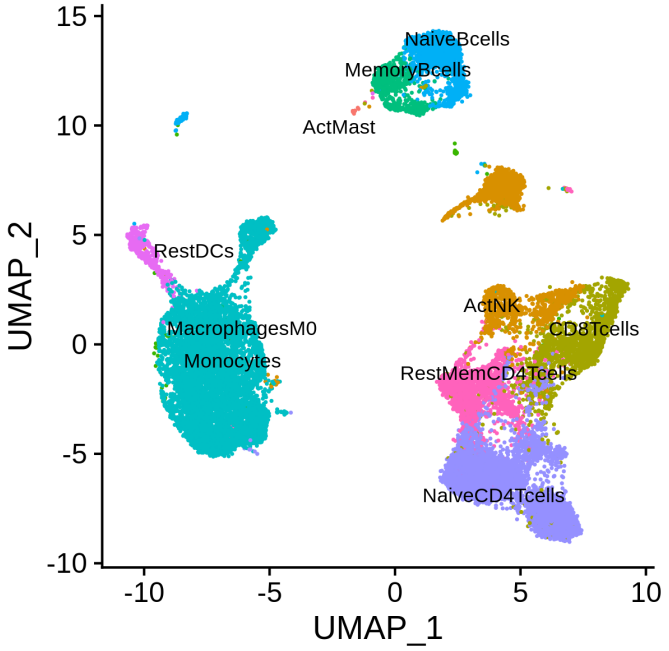


Pbmc 10k scATAC

CCA
MNN

ScRNA and scATAC integrated
Human pbmc from 10x

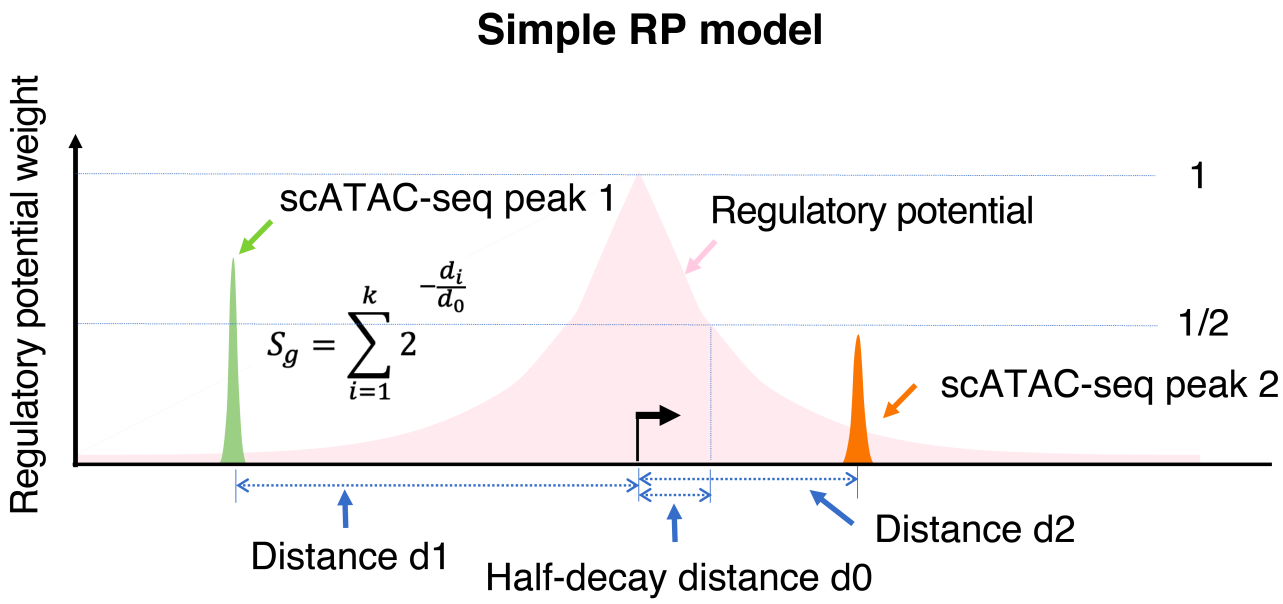




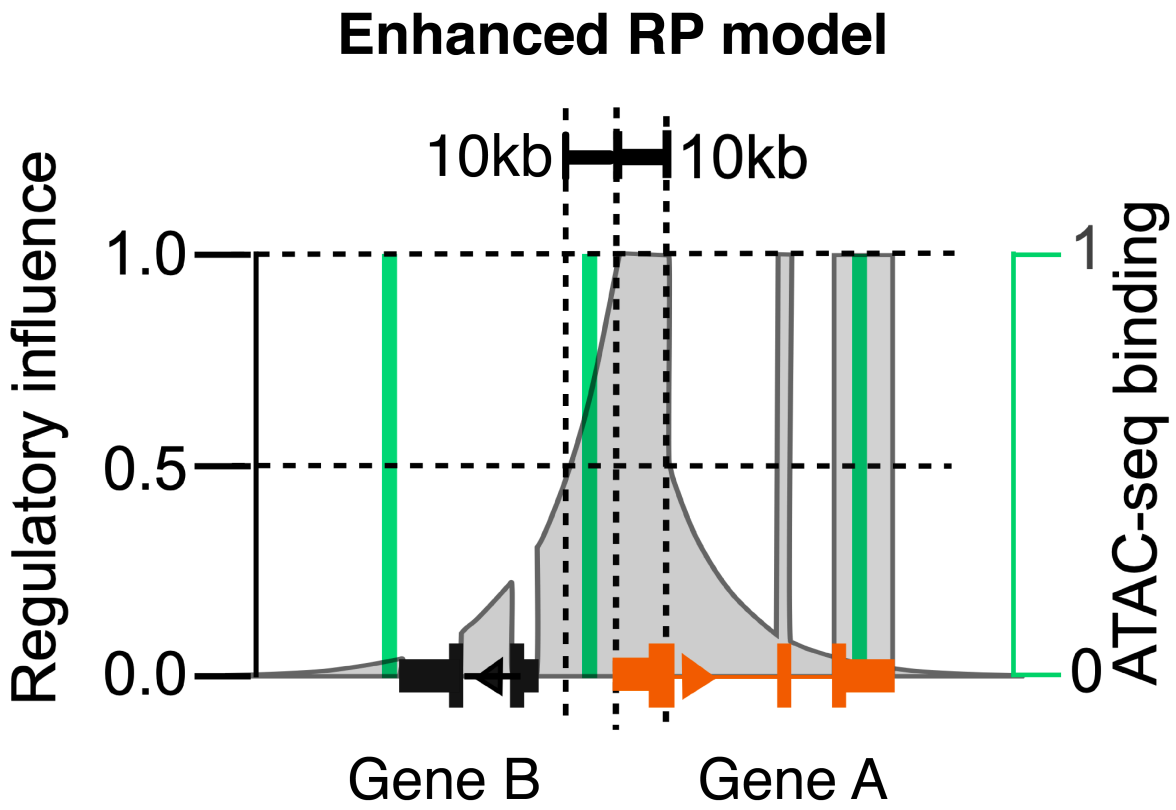
MAESTRO provides a simple regulatory potential (RP) model to estimate gene activity for scATAC-seq

- Gene activity
  - Single-cell regulatory potential (ScRP)
  - Decay distance d0 = 10kb

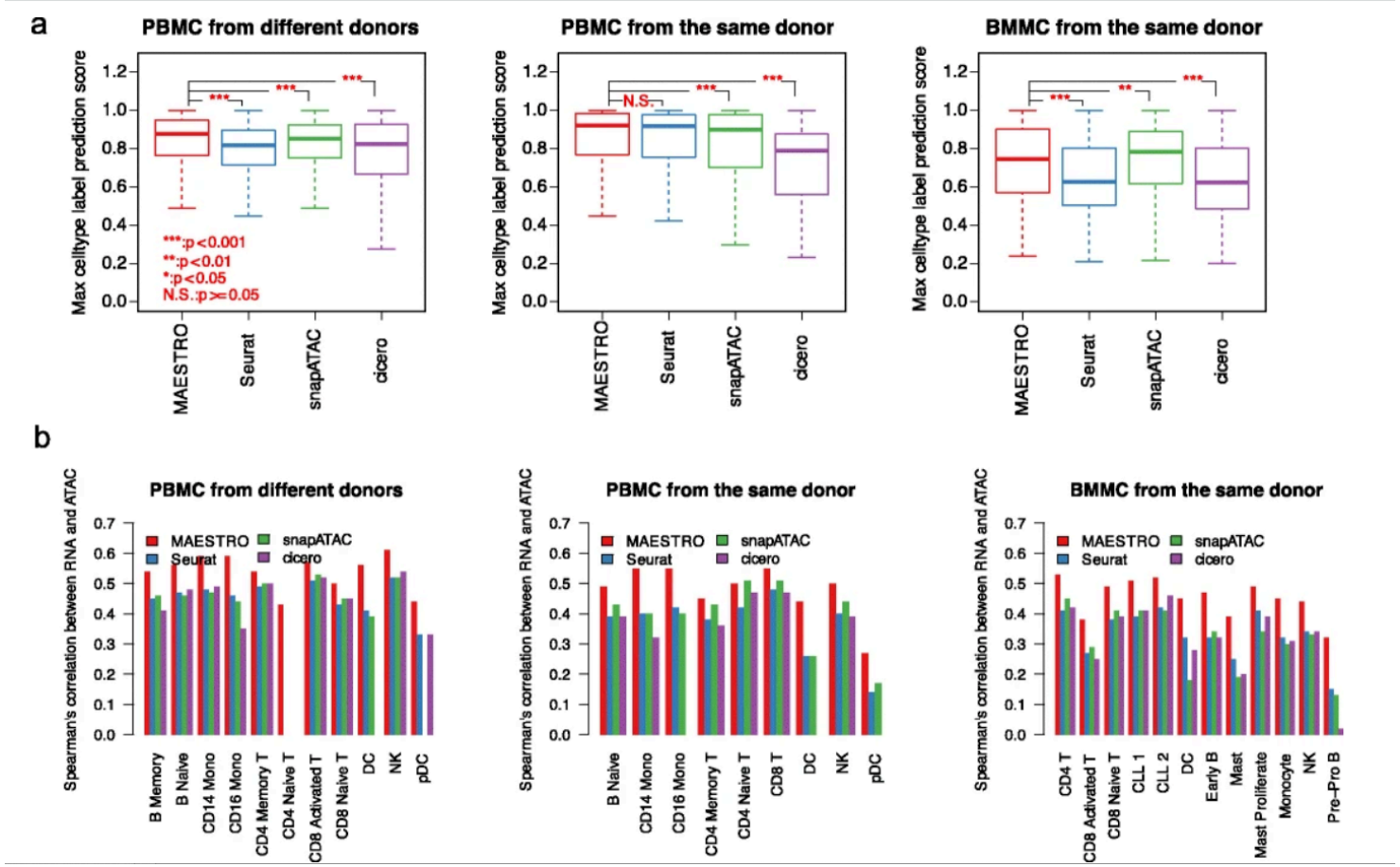
$$S_g = \sum_{i=1}^{k} 2^{-\frac{d_i}{d_0}}$$



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

Liu lab:

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

CIDC Software team:

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

DFCI CIO:

- Mohamed Uduman
- Jason Weirather

DFCI CFCE:

- Henry Long

Tao Liu lab:

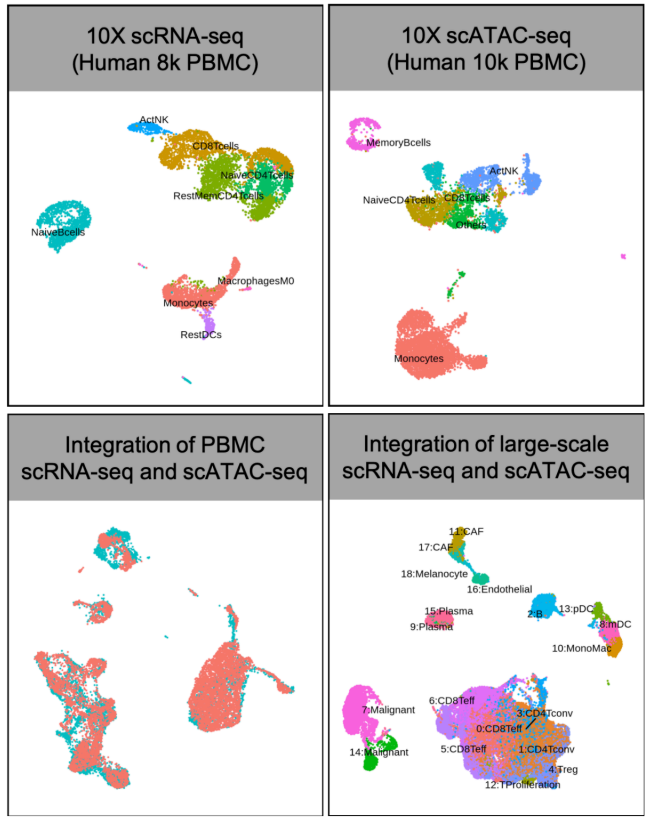
- Tao Liu

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.

Documents @



Html output example @

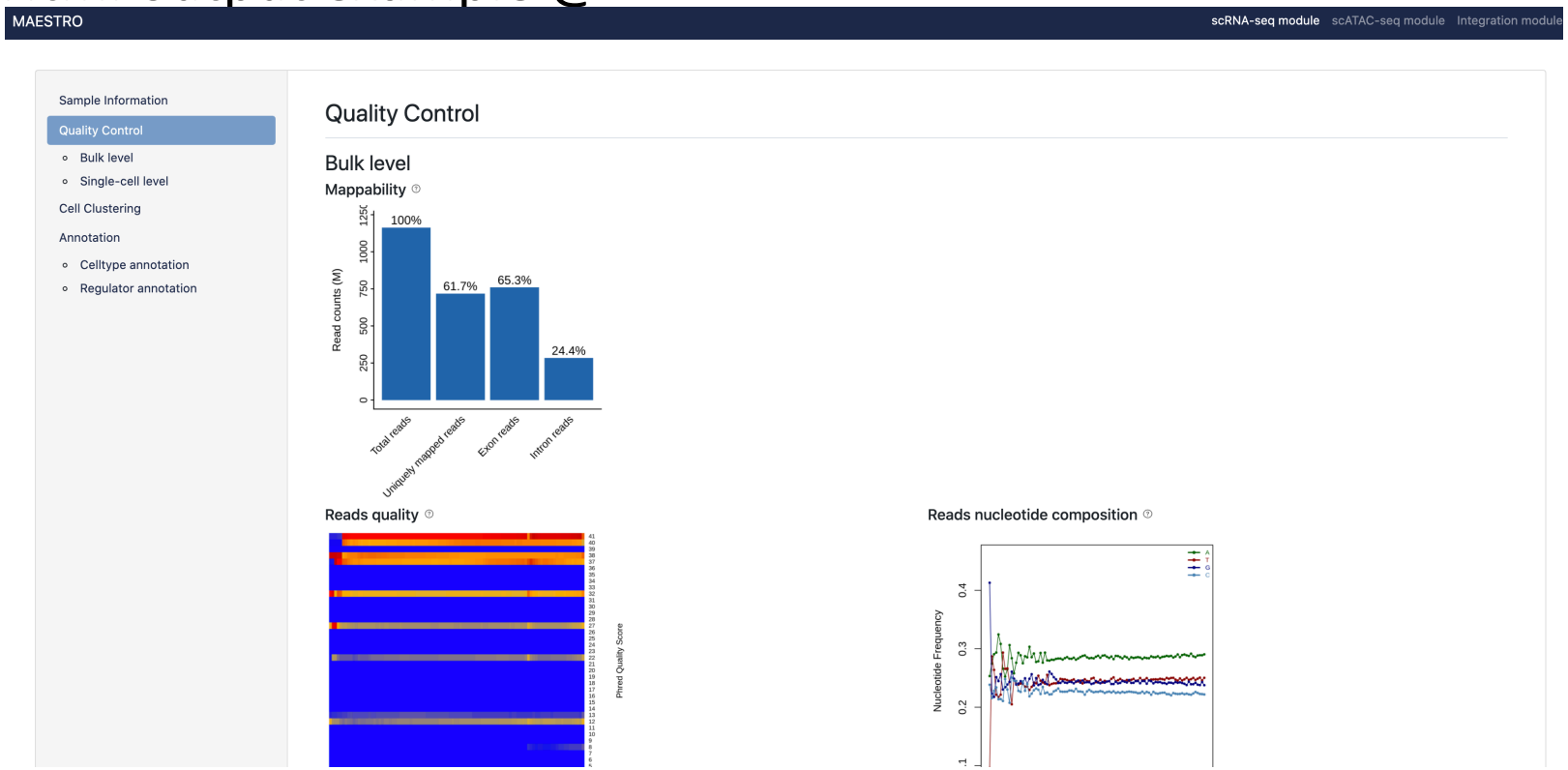


Table S2 Time and memory usage comparison between MAESTRO and other scATAC-seq analysis tools

Dataset 1: PBMC Different Donors scATAC-seq, 173,477 peaks x 9,361 cells, 8 cores

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

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

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

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

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

NA: Memory usage lager than 380G and crashed.