Stat 115/215 Lab

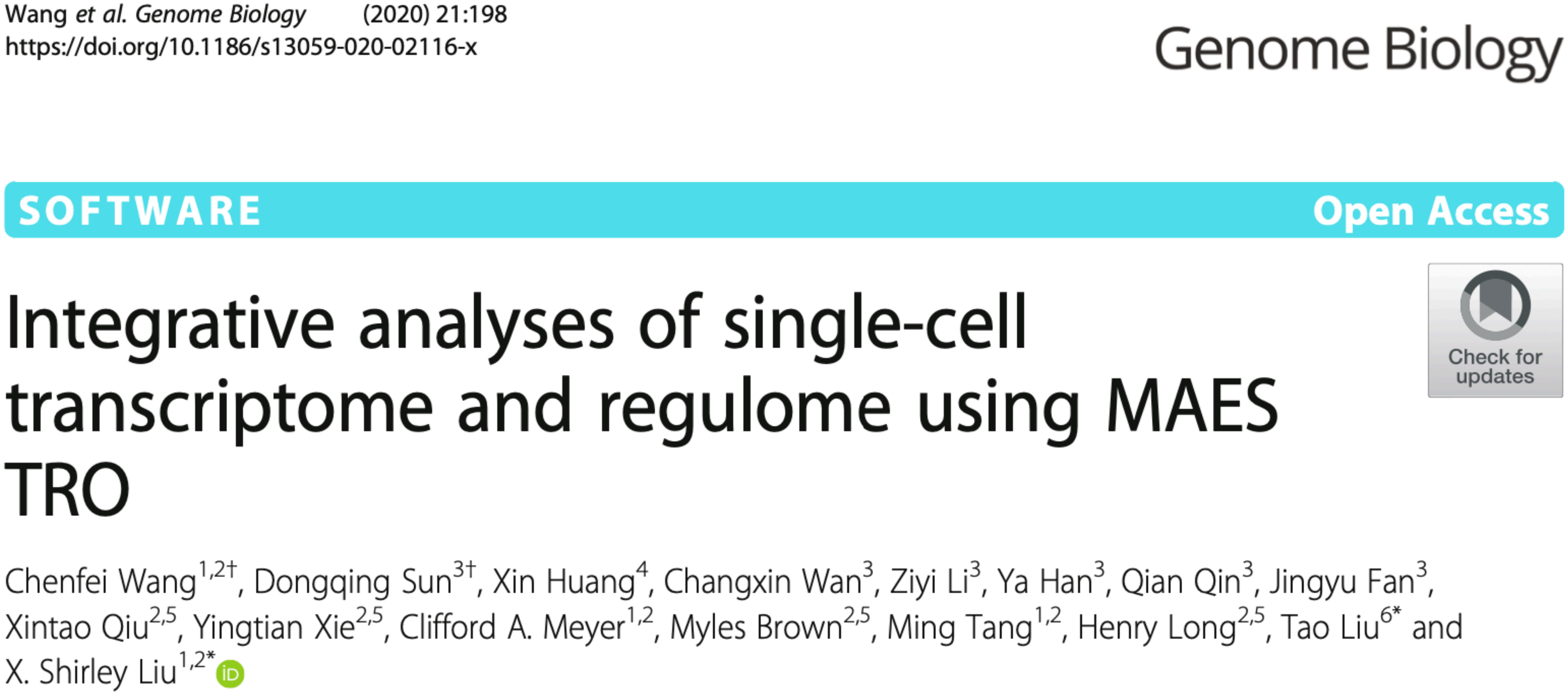
Phillip Nicol

March 30, 2021

Outline

- Introduction to MAESTRO (Part I of HW).
- Integrating scRNA-seq and scATAC-seq data (Part IV of HW).

MAESTRO



Wang *et al. Genome Biology* (2020) 21:198
https://doi.org/10.1186/s13059-020-02116-x

Genome Biology

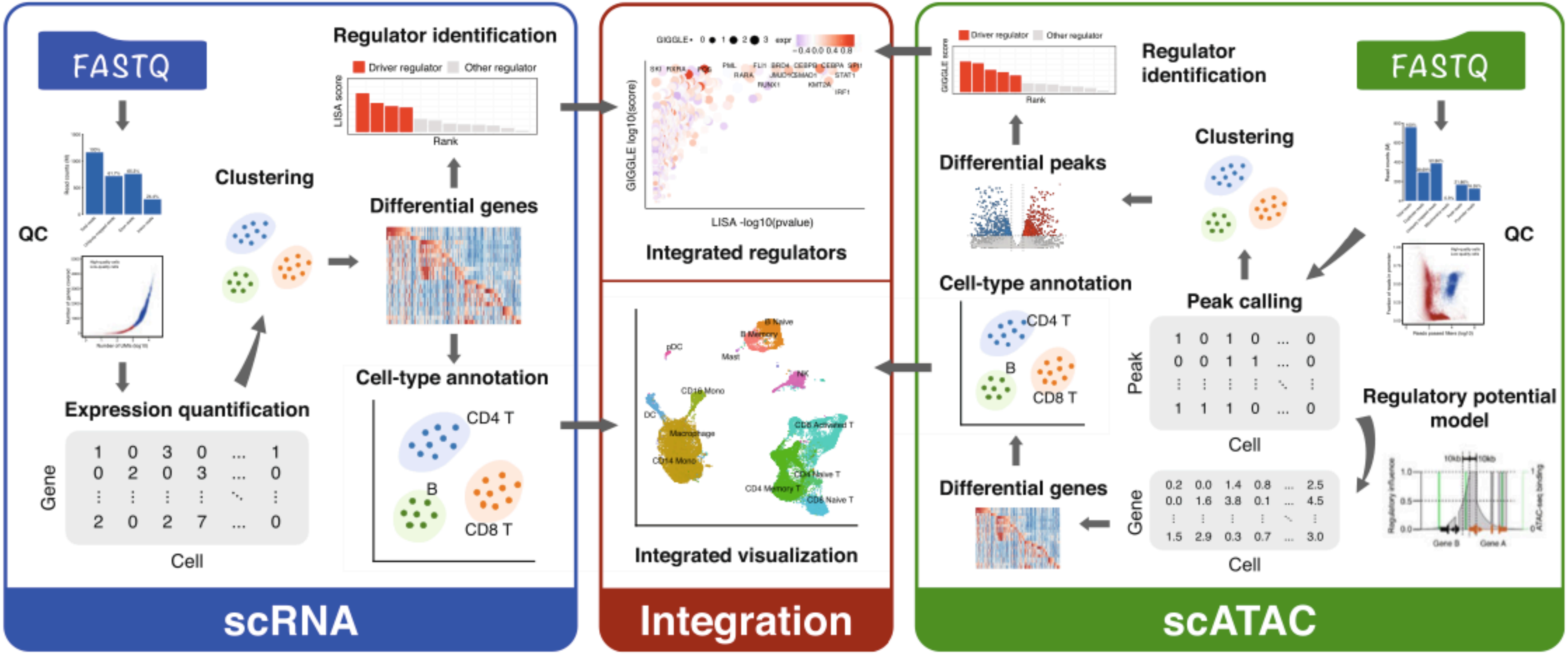
SOFTWARE **Open Access**

Integrative analyses of single-cell transcriptome and regulome using MAES TRO

Chenfei Wang[1,2†], Dongqing Sun[3†], Xin Huang[4], Changxin Wan[3], Ziyi Li[3], Ya Han[3], Qian Qin[3], Jingyu Fan[3], Xintao Qiu[2,5], Yingtian Xie[2,5], Clifford A. Meyer[1,2], Myles Brown[2,5], Ming Tang[1,2], Henry Long[2,5], Tao Liu[6*] and X. Shirley Liu[1,2*]

What is MAESTRO?

- MAESTRO (Model-based Analyses of Single-cell Transcriptome and RegulOme) is a pipeline for the analysis and integration of scRNA-seq and scATAC-seq data.
- MAESTRO can start from a FASTQ file and return a processed Seurat object (thereby performing alignment, demultiplexing of barcodes, cell-type annotation, etc).

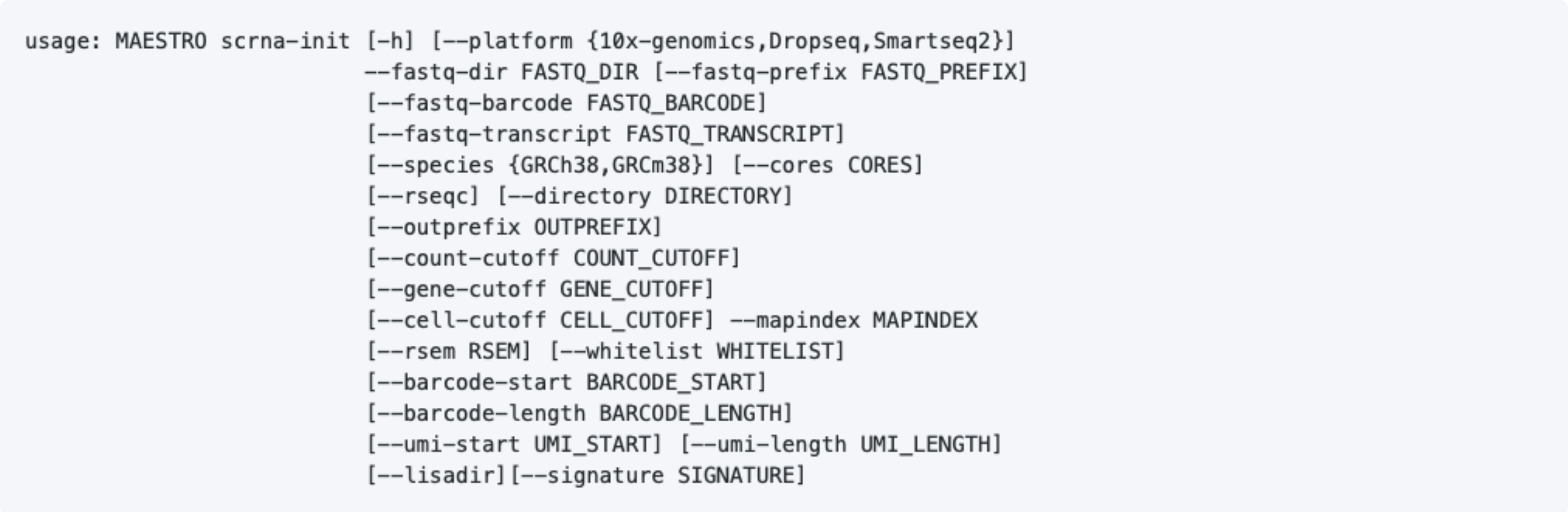


Wang et al (2020).

Running MAESTRO

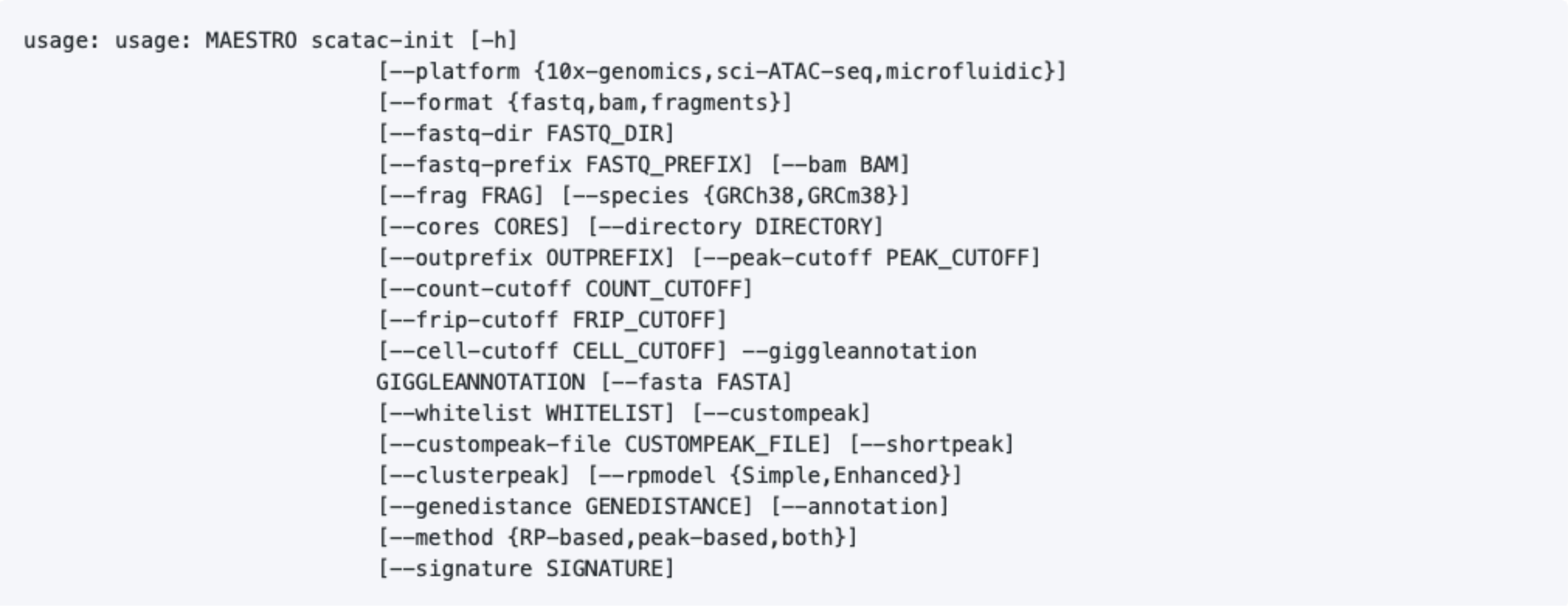
- Maestro is built using Snakemake (Mölder et al., 2021). Snakemake is a tool for workflow management.
- Four steps for running on Cannon:
  - Source the conda environment.
  - Configure the pipeline using `MAESTRO scrna-init` or `MAESTRO scatac-init`. You do not need to run this on a compute node.
  - Run the analysis as a batch job. I specified 60 GB of RAM, 5 hours of run-time, and 16 CPUs.
  - Analyze the results (copy and paste text files and transfer images to your local computer).

Initialization for scRNA-seq



```
usage: MAESTRO scrna-init [-h] [--platform {10x-genomics,Dropseq,Smartseq2}]
                          --fastq-dir FASTQ_DIR [--fastq-prefix FASTQ_PREFIX]
                          [--fastq-barcode FASTQ_BARCODE]
                          [--fastq-transcript FASTQ_TRANSCRIPT]
                          [--species {GRCh38,GRCm38}] [--cores CORES]
                          [--rseqc] [--directory DIRECTORY]
                          [--outprefix OUTPREFIX]
                          [--count-cutoff COUNT_CUTOFF]
                          [--gene-cutoff GENE_CUTOFF]
                          [--cell-cutoff CELL_CUTOFF] --mapindex MAPINDEX
                          [--rsem RSEM] [--whitelist WHITELIST]
                          [--barcode-start BARCODE_START]
                          [--barcode-length BARCODE_LENGTH]
                          [--umi-start UMI_START] [--umi-length UMI_LENGTH]
                          [--lisadir][--signature SIGNATURE]
```

Initialization for scATAC-seq

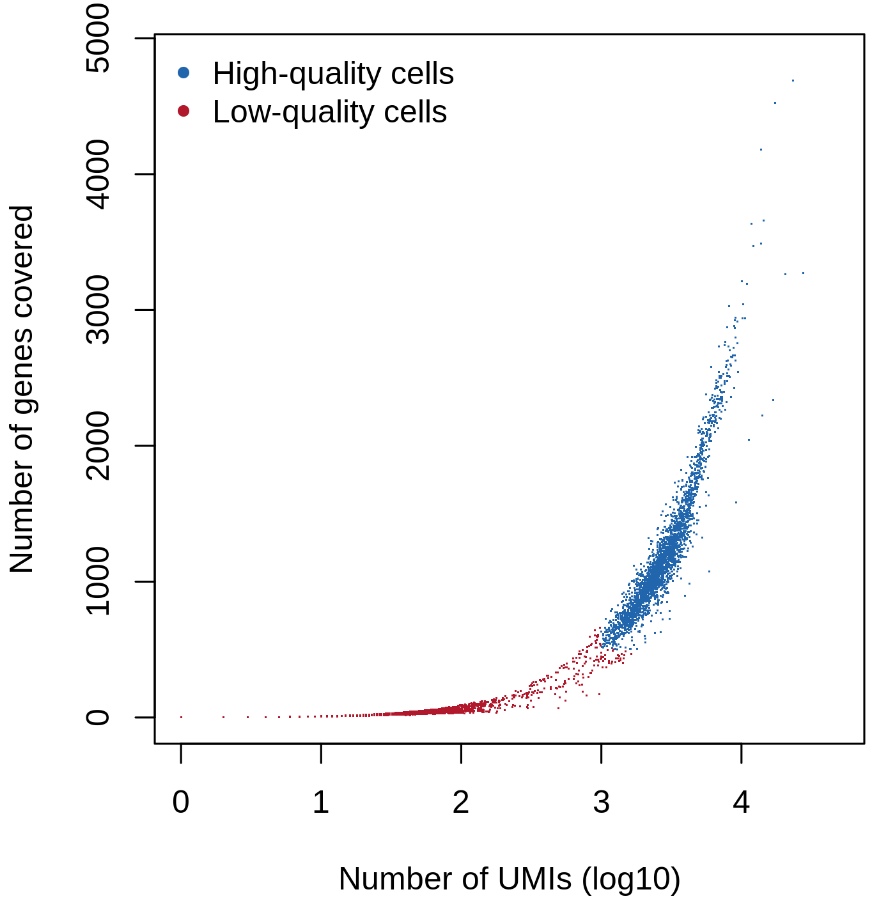


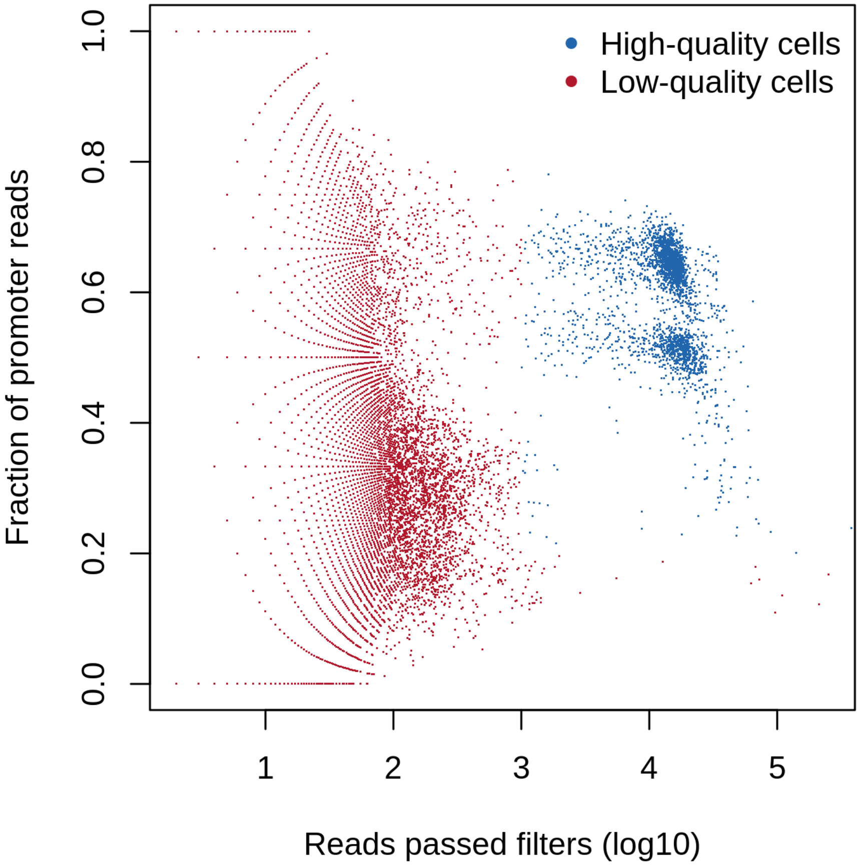
```
usage: usage: MAESTRO scatac-init [-h]
                               [--platform {10x-genomics,sci-ATAC-seq,microfluidic}]
                               [--format {fastq,bam,fragments}]
                               [--fastq-dir FASTQ_DIR]
                               [--fastq-prefix FASTQ_PREFIX] [--bam BAM]
                               [--frag FRAG] [--species {GRCh38,GRCm38}]
                               [--cores CORES] [--directory DIRECTORY]
                               [--outprefix OUTPREFIX] [--peak-cutoff PEAK_CUTOFF]
                               [--count-cutoff COUNT_CUTOFF]
                               [--frip-cutoff FRIP_CUTOFF]
                               [--cell-cutoff CELL_CUTOFF] --giggleannotation
                               GIGGLEANNOTATION [--fasta FASTA]
                               [--whitelist WHITELIST] [--custompeak]
                               [--custompeak-file CUSTOMPEAK_FILE] [--shortpeak]
                               [--clusterpeak] [--rpmodel {Simple,Enhanced}]
                               [--genedistance GENEDISTANCE] [--annotation]
                               [--method {RP-based,peak-based,both}]
                               [--signature SIGNATURE]
```

Running the pipeline on Cannon

- Make sure that you have specified sufficient resources.
- Since you only have ~100GB of space in your home directory, you will likely need to delete large files from previous HWs before running MAESTRO.
- When you run `snakemake`, you should be in the `multiome_scrna` folder.

Viewing the QC results

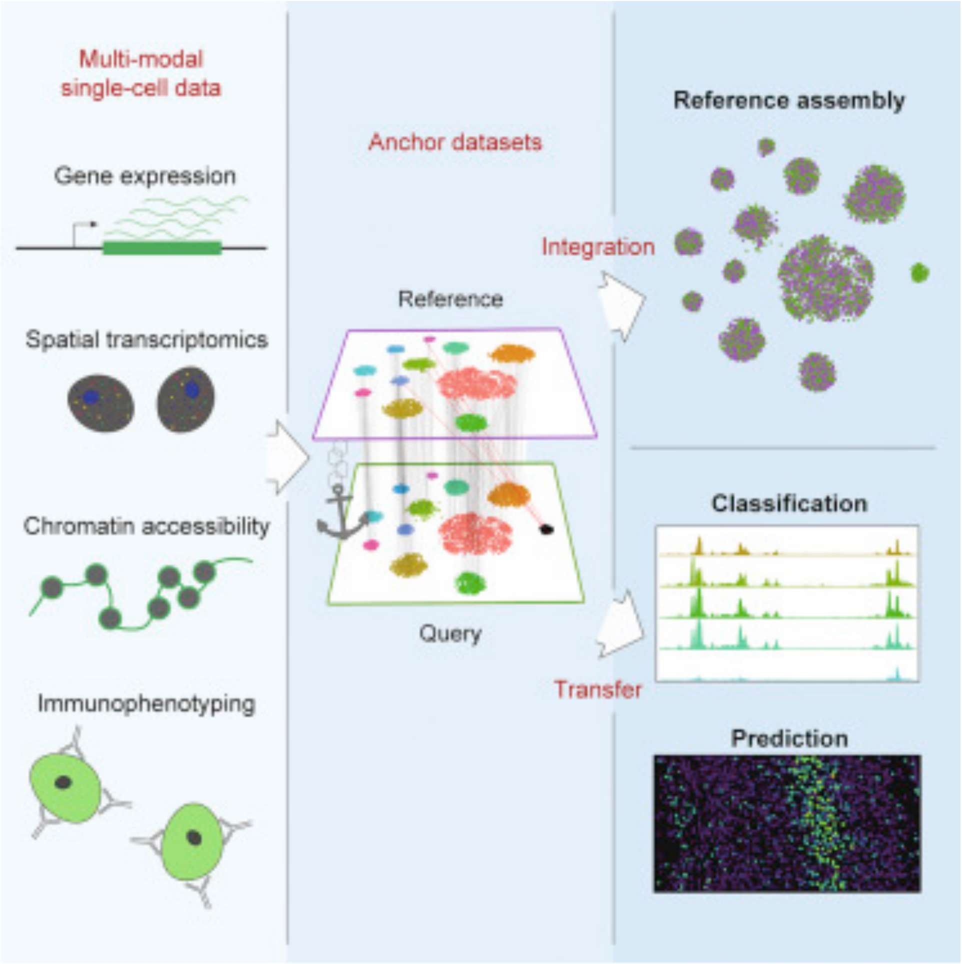




Integrating scRNA-seq and scATAC-seq

- A strength of scRNA-seq is that we have direct knowledge of the gene expression (this is better for things such as cell type annotation).
- In Part IV of the homework, we use "label transferring" to integrate scATAC-seq with scRNA-seq.

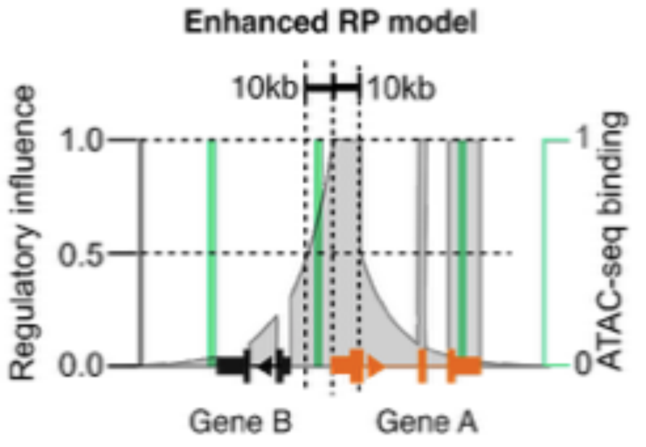
- (Big) Hint: https://satijalab.org/seurat/articles/atacseq_integration_vignette.html contains most of the code needed for this part of the HW.



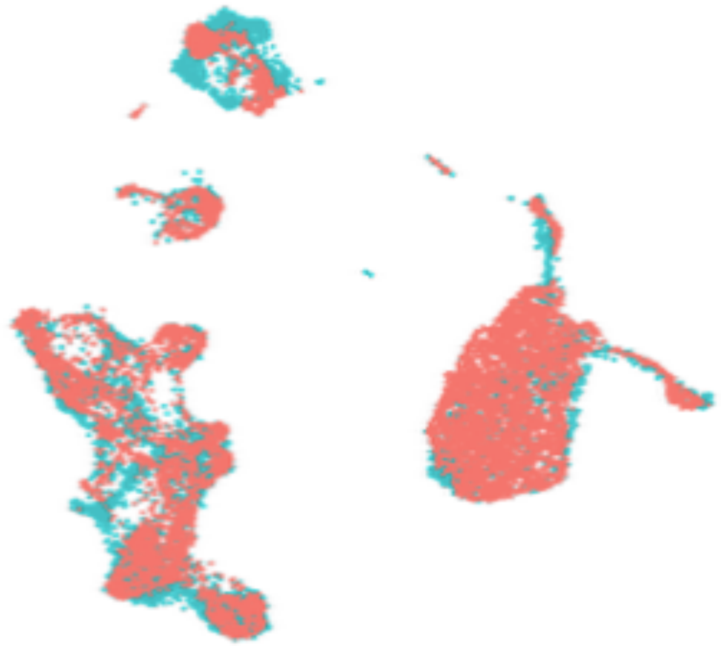
Stuart et al, 2019. *Cell.*

MAESTRO develops a regulatory potential model to predict gene activity in scATAC-seq data. From this, the two datasets can be treated as two different batch and a batch-correction method (CCA) can be applied.

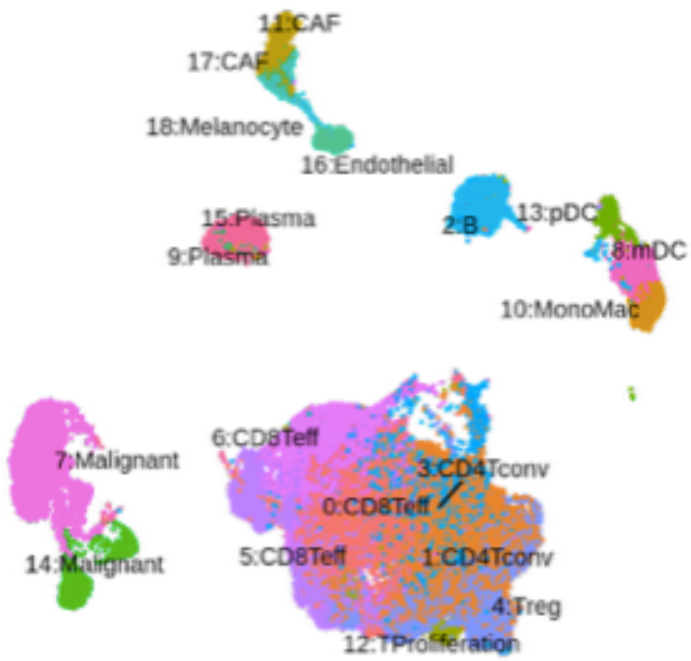
Gene activity modelling from scATAC-seq



Integration of PBMC scRNA-seq and scATAC-seq



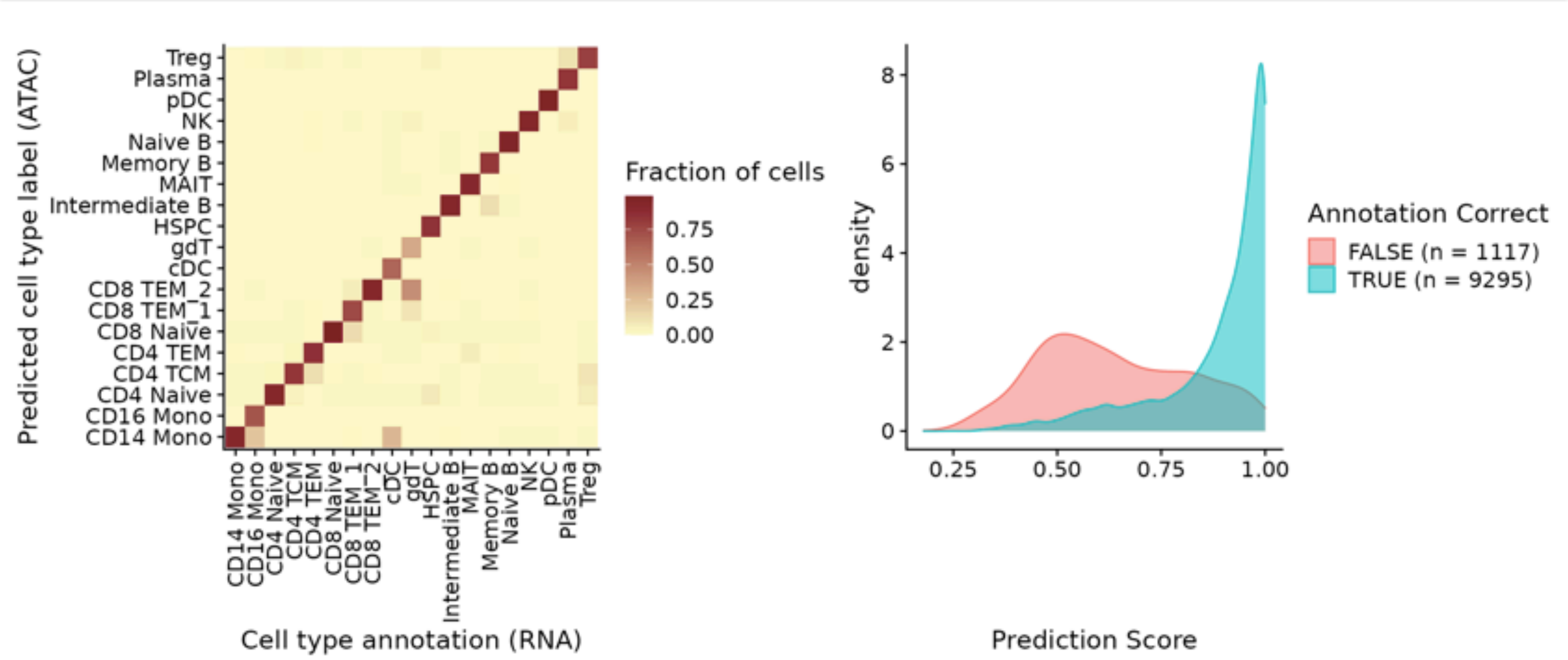
Integration of large-scale scRNA-seq and scATAC-seq



Key steps for Part IV

- Step 1: Match barcodes between scRNA-seq and scATAC-seq (most of the code for this part is already given).
- Step 2: Run `FindTransferAnchors()` to obtain a ”mapping” between the scRNA-seq and the scATAC-seq data.
- Step 3: Run `TransferData()` to annotate the scATAC-seq data using the scRNA-seq data.
- Step 4: Compare the inferred labels to the true labels.

Sample plot for 3.3 and 3.4 (See Vignette for code)



Other hints and suggestions for homework

- Part I, Problem 2: use scATAC_cell_filtering.png instead of scATAC_read_distr.png.
- Part II, Problem 2: Consider the Seurat function `PercentageFeatureSet` and look at the provided example.
- Part II, Problem 5.1: The first answer provided by the link is wrong, make sure you read the follow-up to understand why.