Analysis of Single Cell ATAC-seq Data

Yang (Eric) Li

CSHL Sequencing Technologies and Bioinformatics Analysis

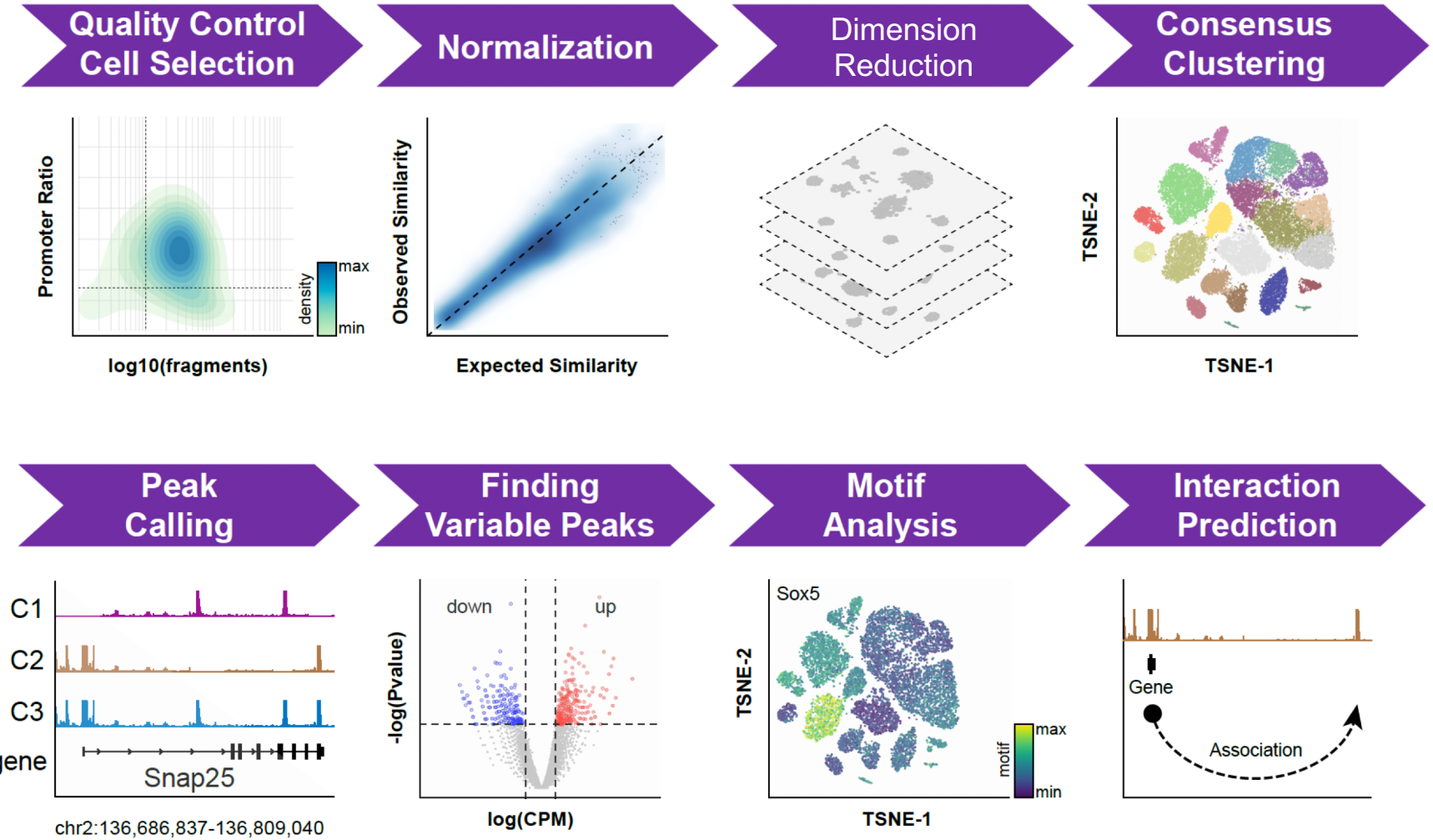
Nov 13, 2023

Yang (Eric) Li Lab
Website: yelilab.wustl.edu
Email: yeli@wustl.edu

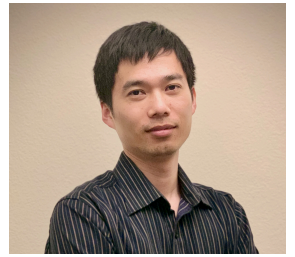
Download

wget http://3.86.234.121/snATACseq_cshl_seqtec_2023.zip

Single Nucleus Analysis Pipeline for ATAC-seq



Rongxin Fang, Ph.D.

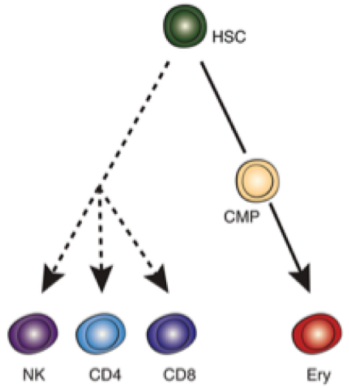


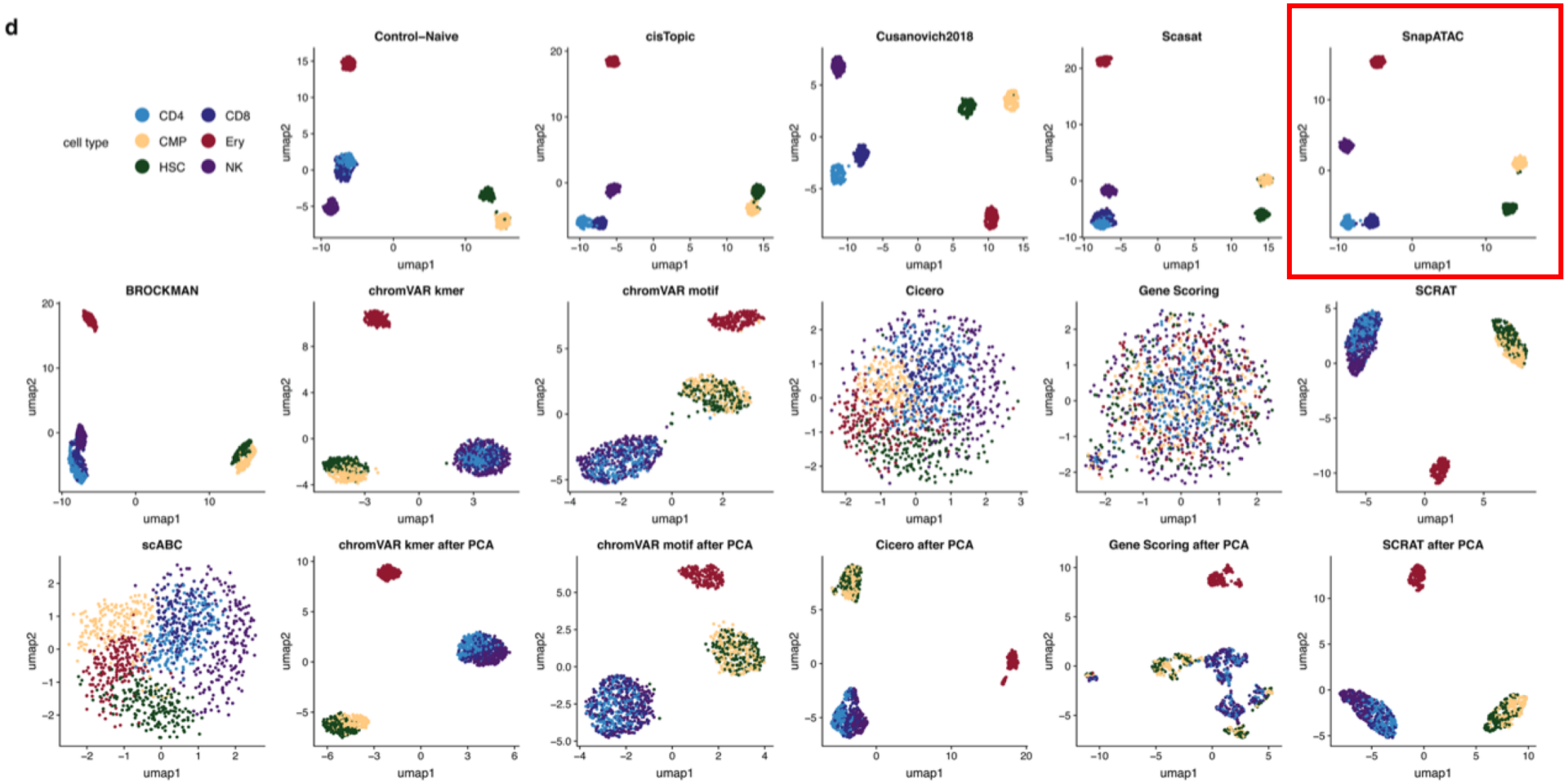
Kai Zhang, Ph.D.

SCAN ME

Fang et al, Nature Communication 2021.

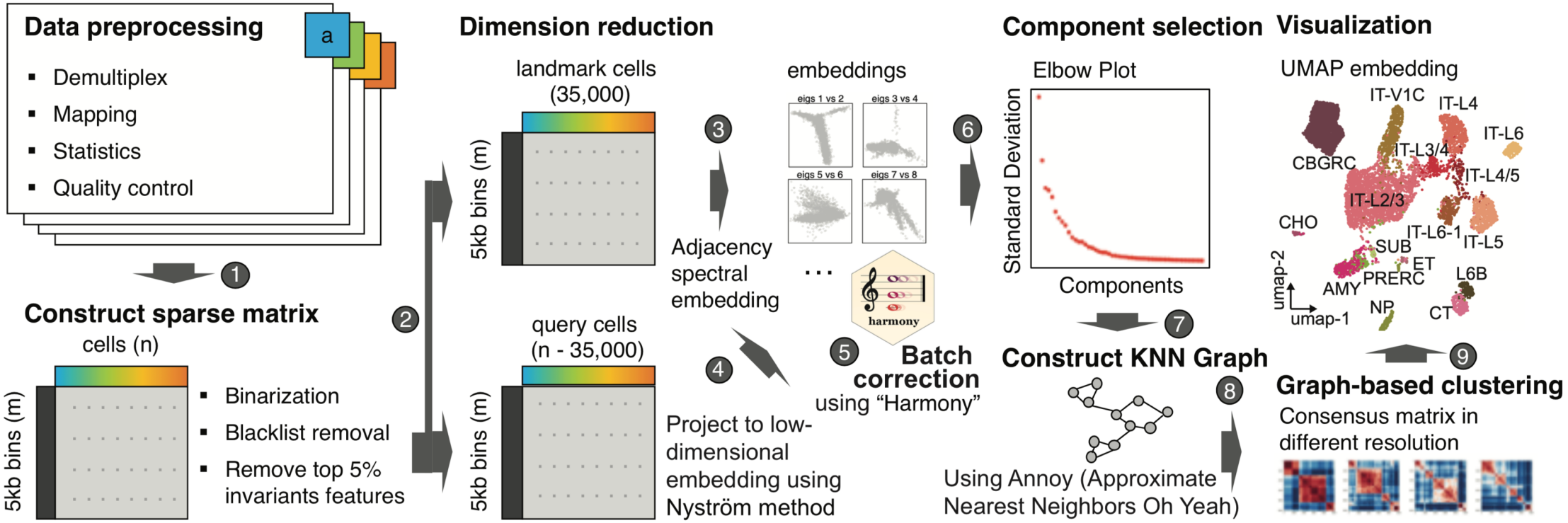
Benchmarking results in simulated bone marrow datasets



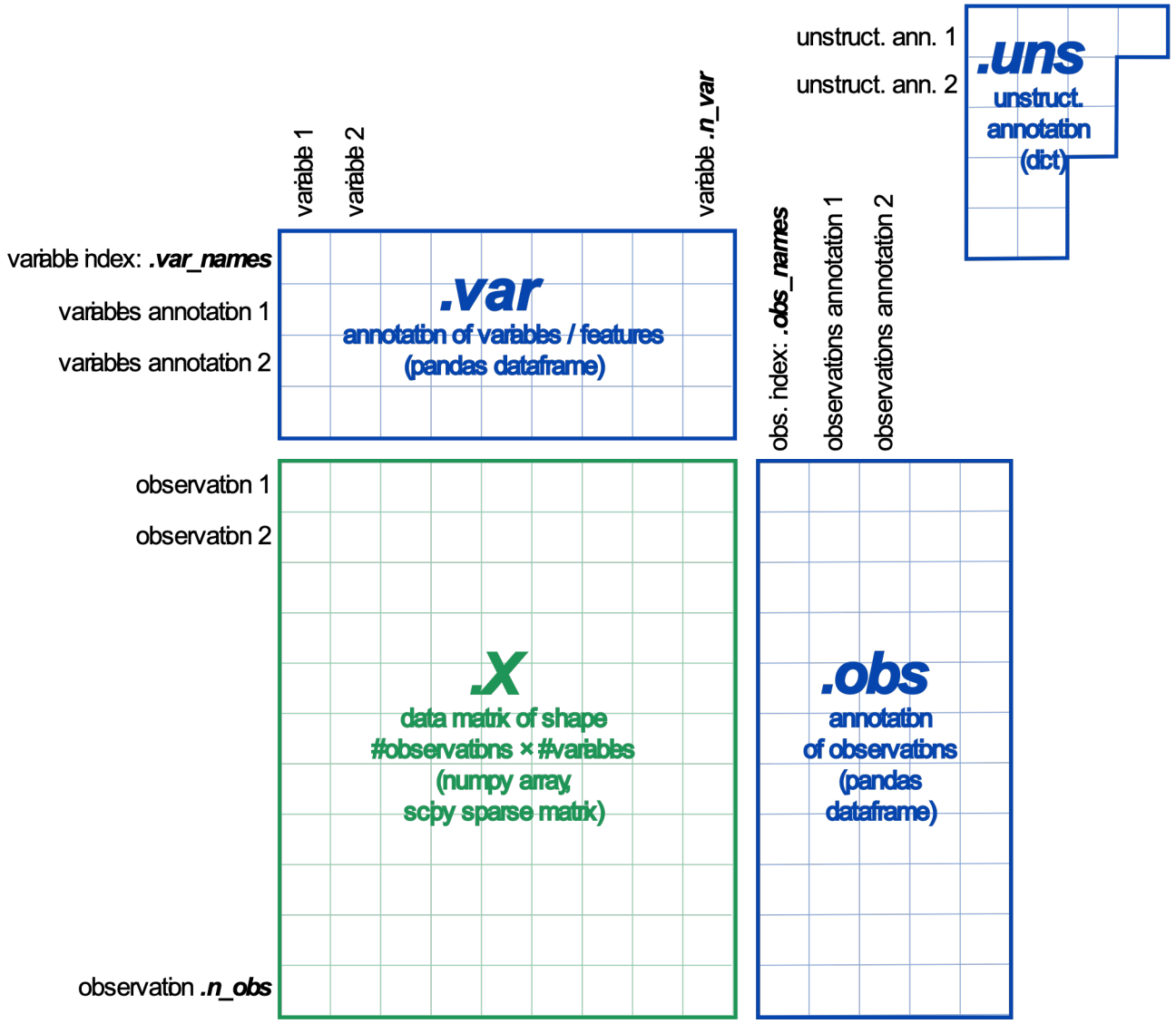


Chen, et al. Genome Biology, 2019

Cell clustering pipeline



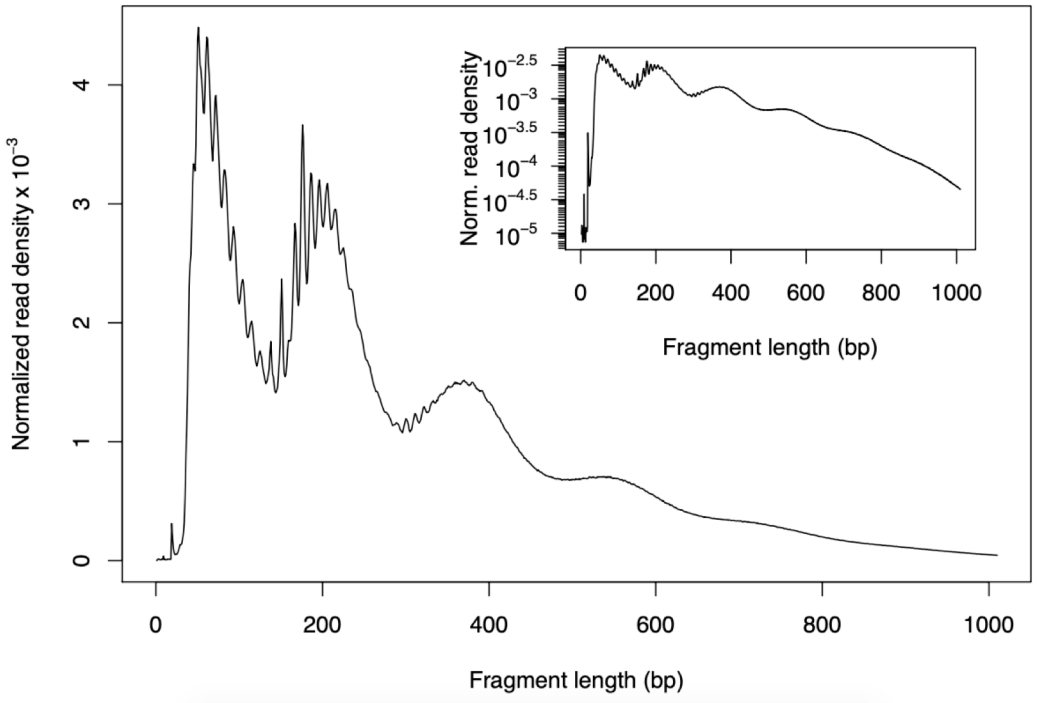
Single cell dataset in h5ad (AnnData) format



CellGenIT Docs: https://cellgeni.readthedocs.io/en/0.0.1/visualisations.html

Quality control at bulk level

Fragment size distribution

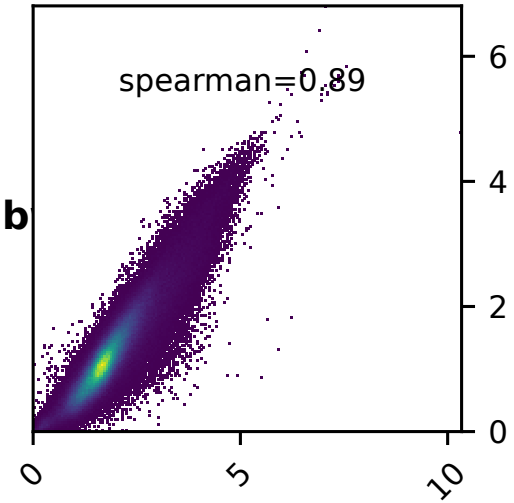


Using ***ATACseqQC*** package

Ou, et al. BMC Genomics. 2018

Correlation between replicates

sample_1.coverage.b



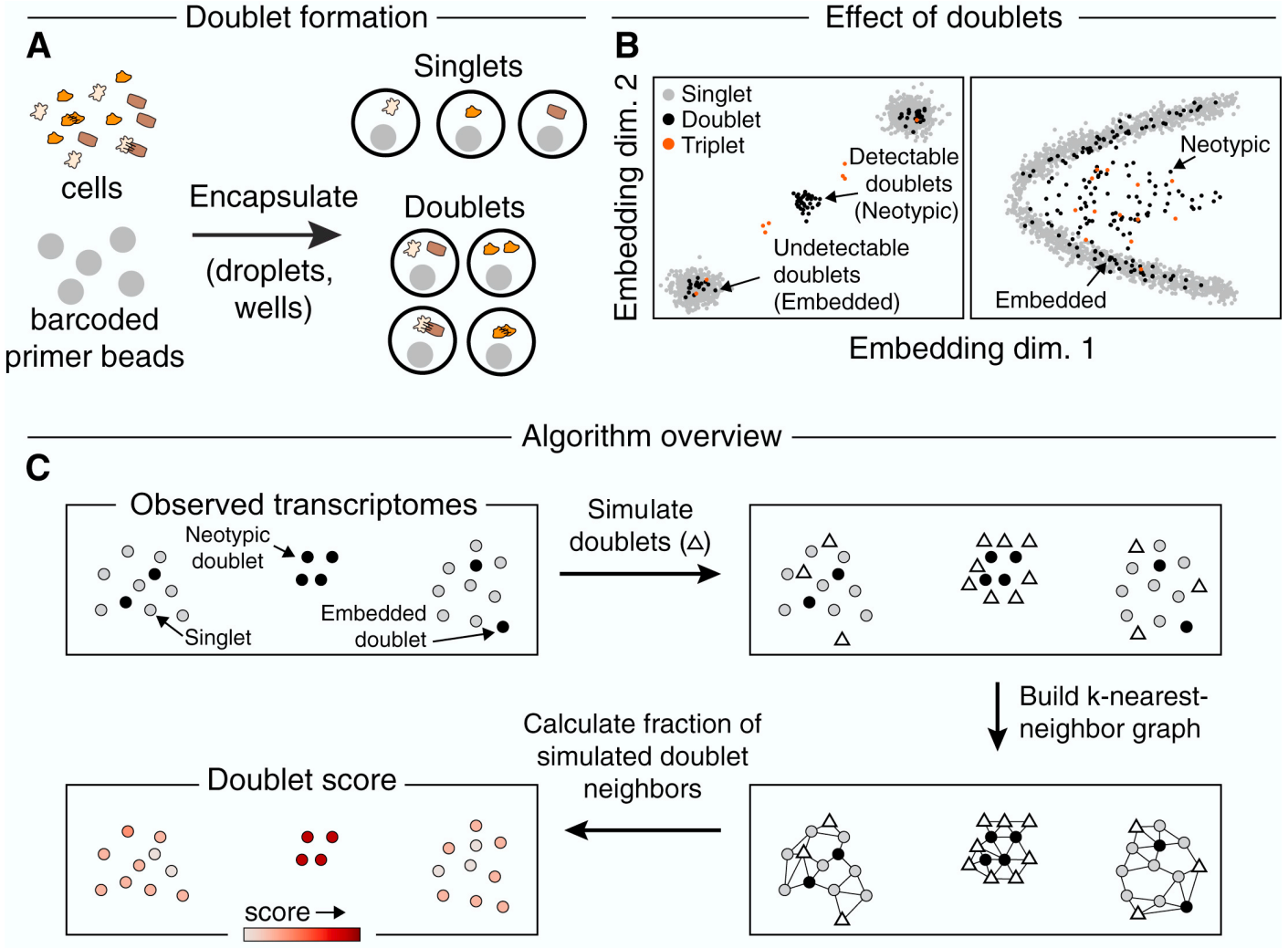
sample_2.coverage.bw

Using ***deepTools*** package

Ramírez, F, Nucleic Acids Research. 2016

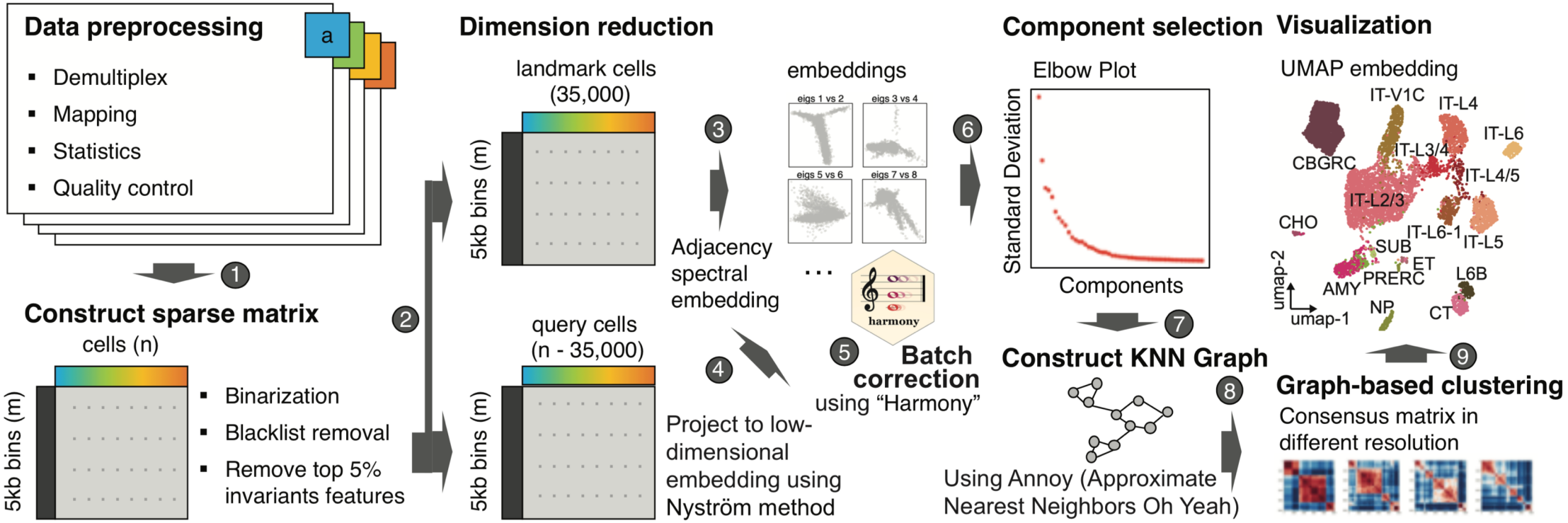
Doublet removal

Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data

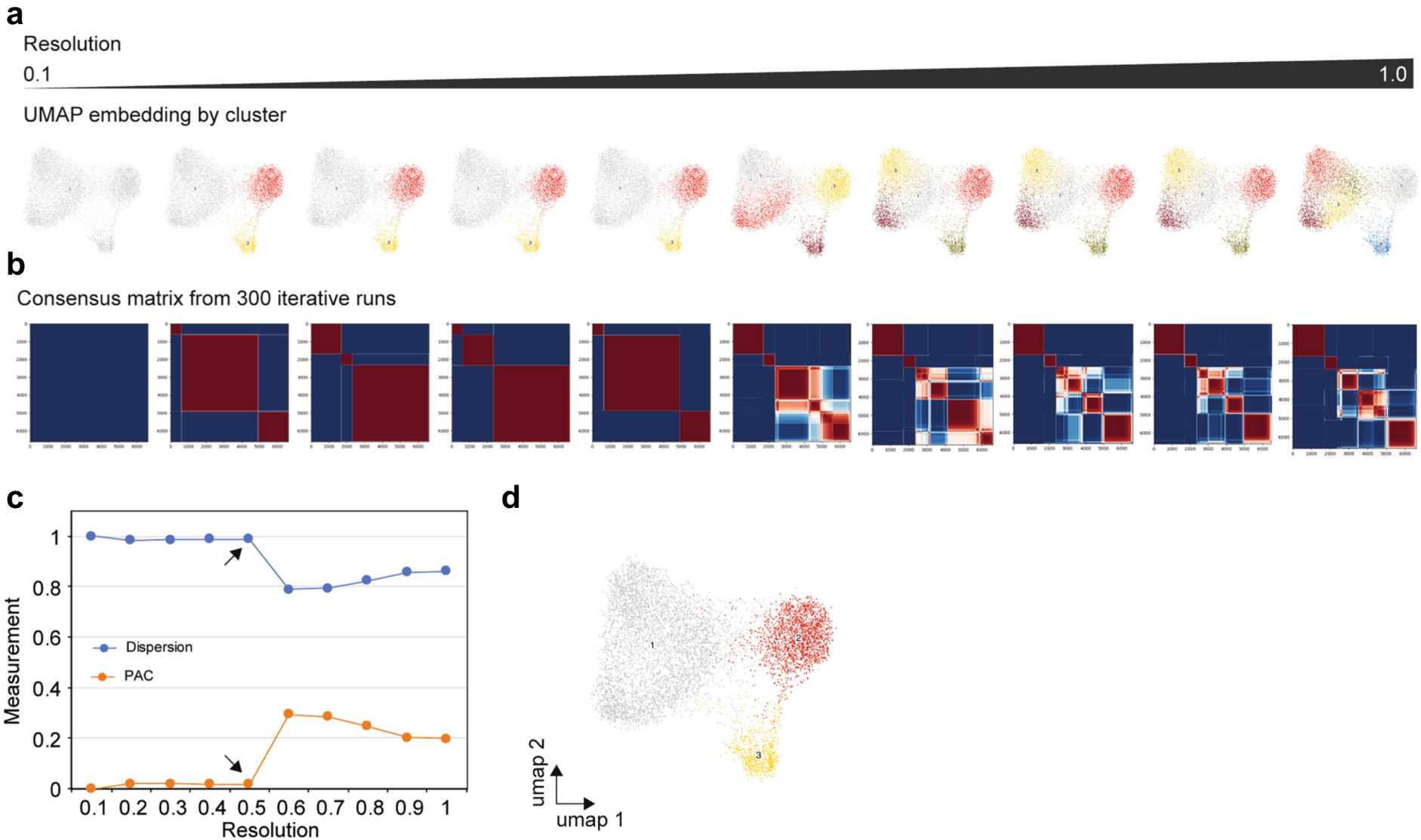


Wolock, S. L., et al. Cell Systems. 2019

Cell clustering pipeline



Cell clustering pipeline



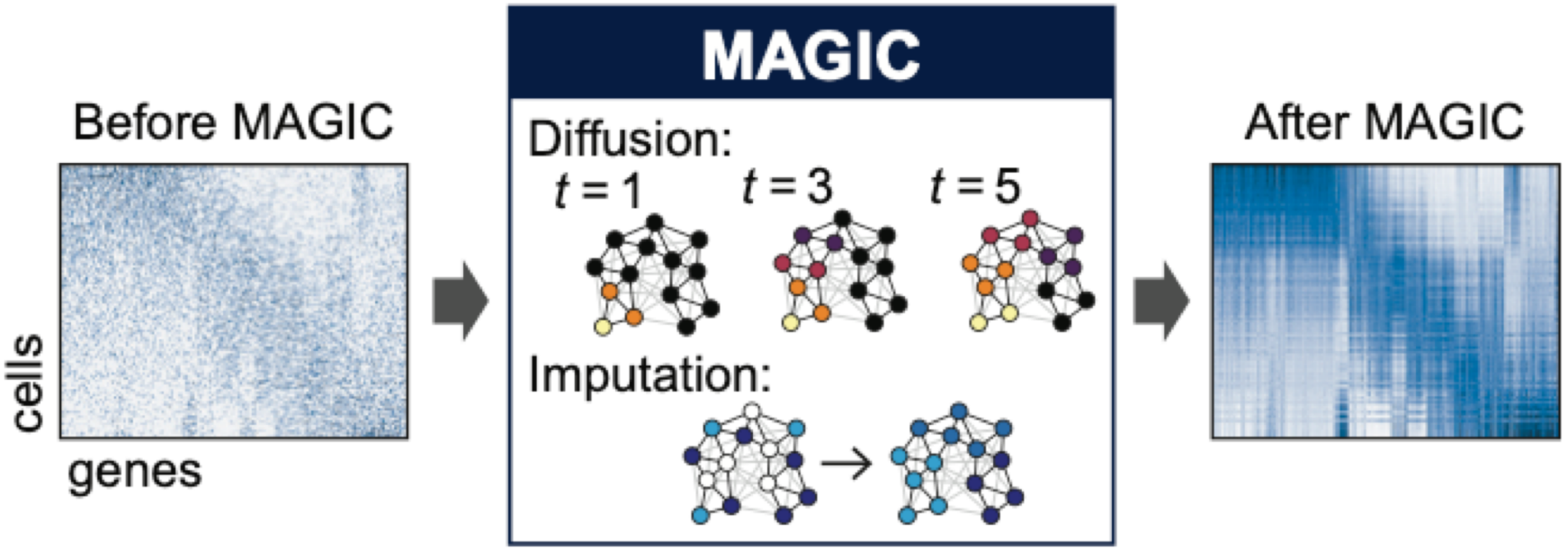
Markov Affinity-based Graph Imputation of Cells (MAGIC)

RESOURCE | VOLUME 174, ISSUE 3, P716-729.E27, JULY 26, 2018

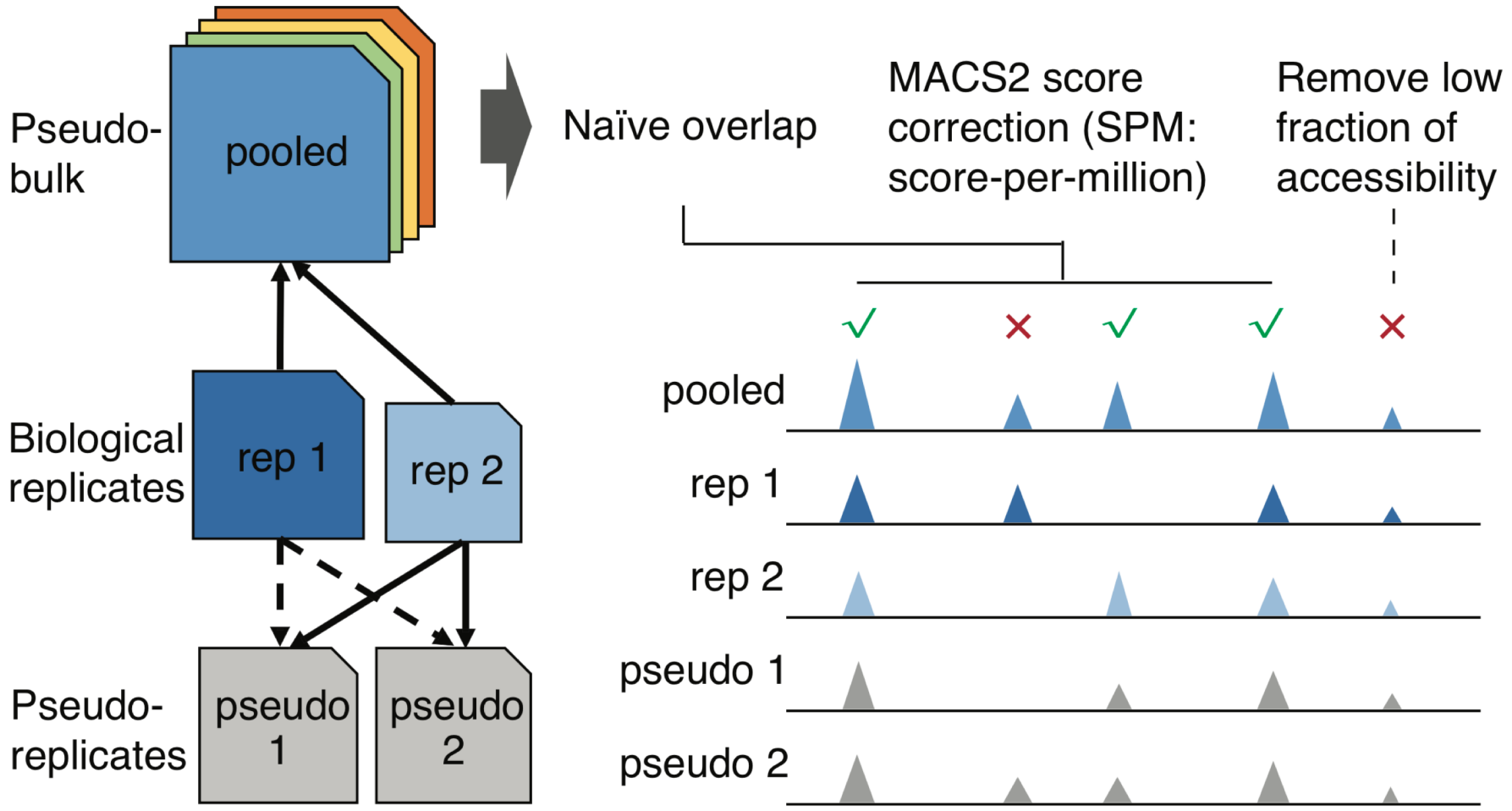
Recovering Gene Interactions from Single-Cell Data Using Data Diffusion

David van Dijk • Roshan Sharma • Juozas Nainys • ... Guy Wolf • Smita Krishnaswamy [9] •
Dana Pe'er [9, 10] • Show all authors • Show footnotes

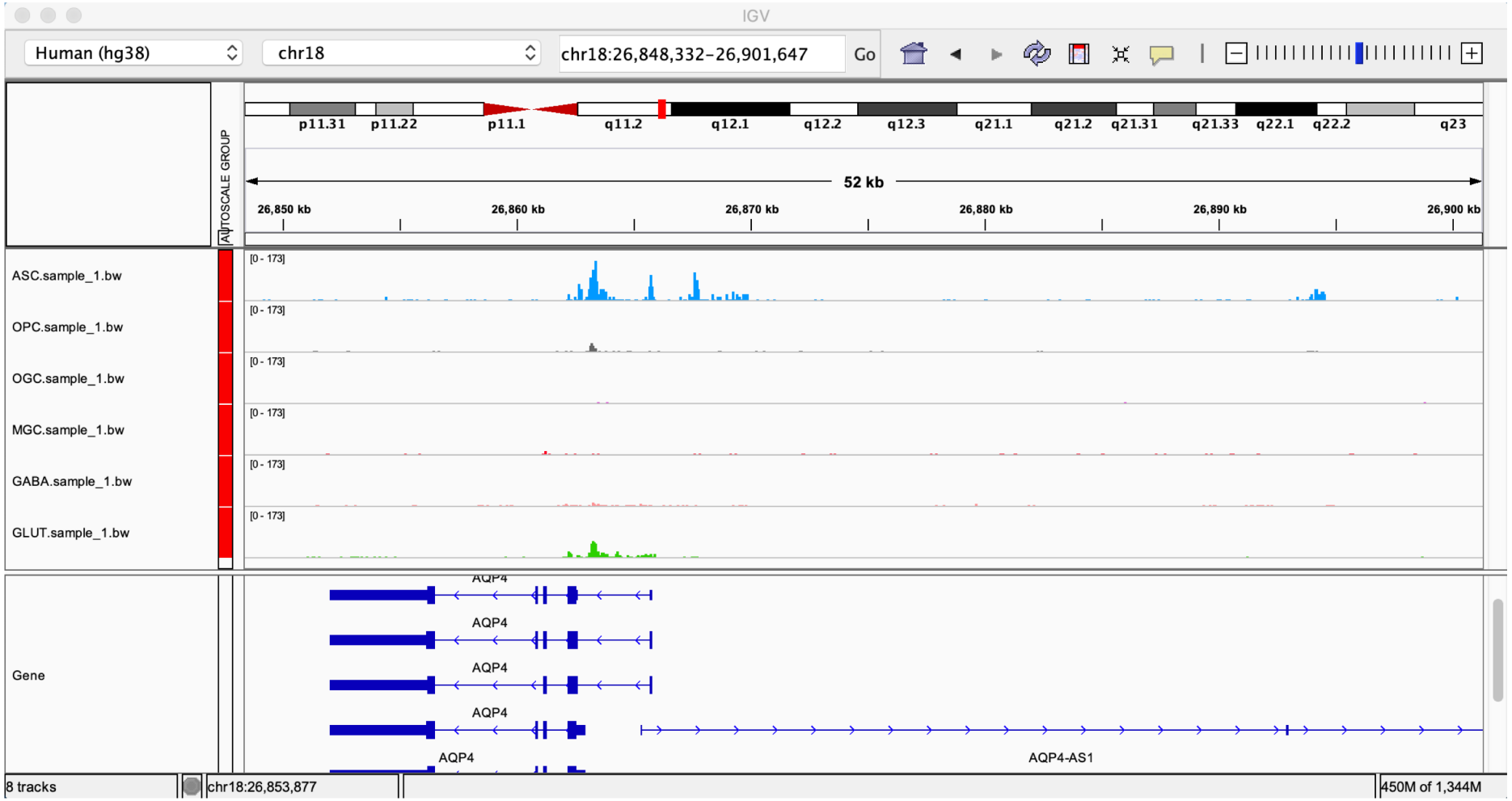
Open Archive • Published: June 28, 2018 • DOI: https://doi.org/10.1016/j.cell.2018.05.061 •



Identification of candidate cis-regulatory elements (cCREs)



Visualization of ATAC-seq signal in Integrative Genomics Viewer (IGV)



Differential analysis in single cell level

Step 1 Identify differential peaks

Full model: $$logit(P_{ij}) = a_j + m_j + r_j + \varepsilon_j$$

Reduced model: $$logit(P_{ij}) = a_j + r_j + \varepsilon_j$$

P: the probability that the i^{th} site is accessible in the j^{th} cell
a: log10(total number of sites observed as accessible for the j^{th} cell)
m: membership of the j^{th} cell in the cluster/region being tested
r: replicate label for j^{th} cell

Then, a likelihood ratio test is used to determine if the full model (including cell cluster membership) provided a significantly better fit of the data than reduced model