Multi-modal alignment of single-cell transcriptomic and epigenomic data

Yuchao Jiang
UNC Chapel Hill

ISMB 2019

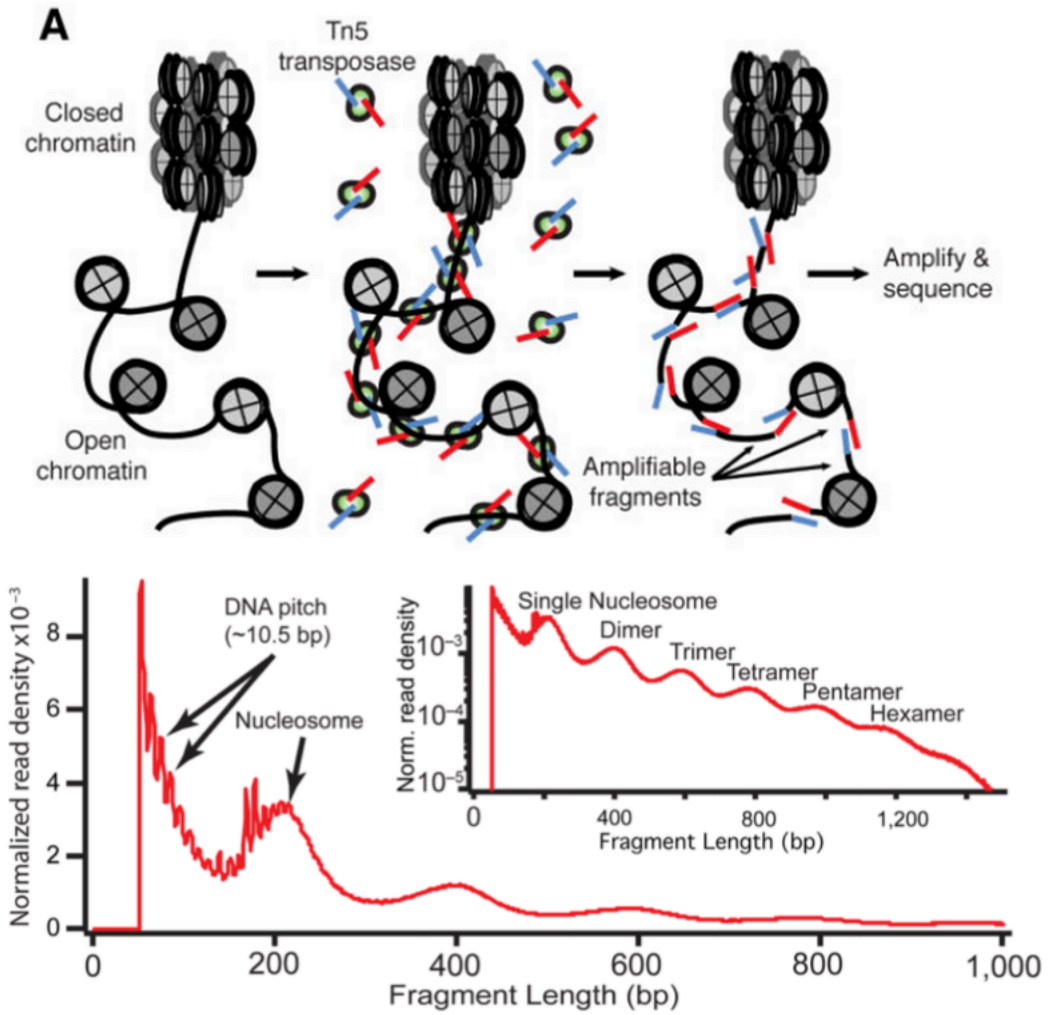
July 21st, 2019

Single-cell chromatin accessibility by scATAC-seq

Chromatin accessibility by ATAC-seq

Tn5 transposase inserts sequencing adapters into open chromatin regions

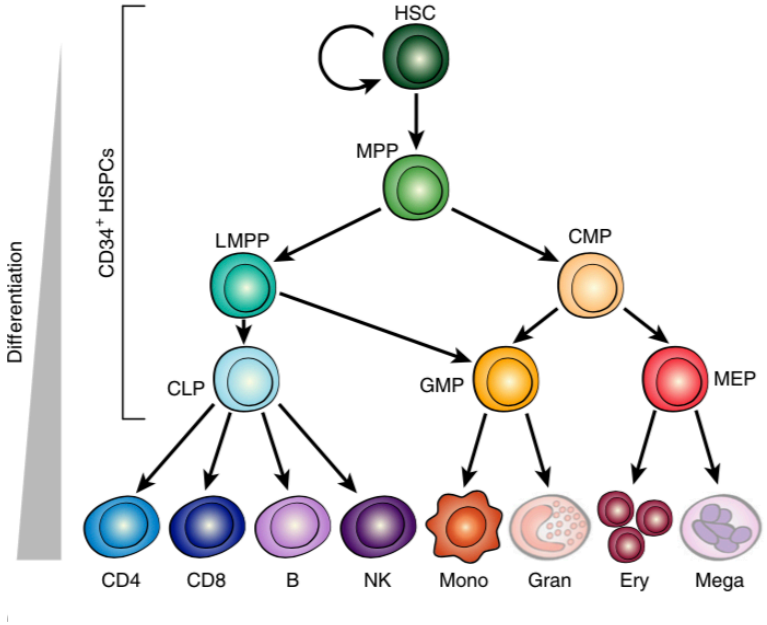
Assay for active transcription regions as well as TF sites and other regulatory sites



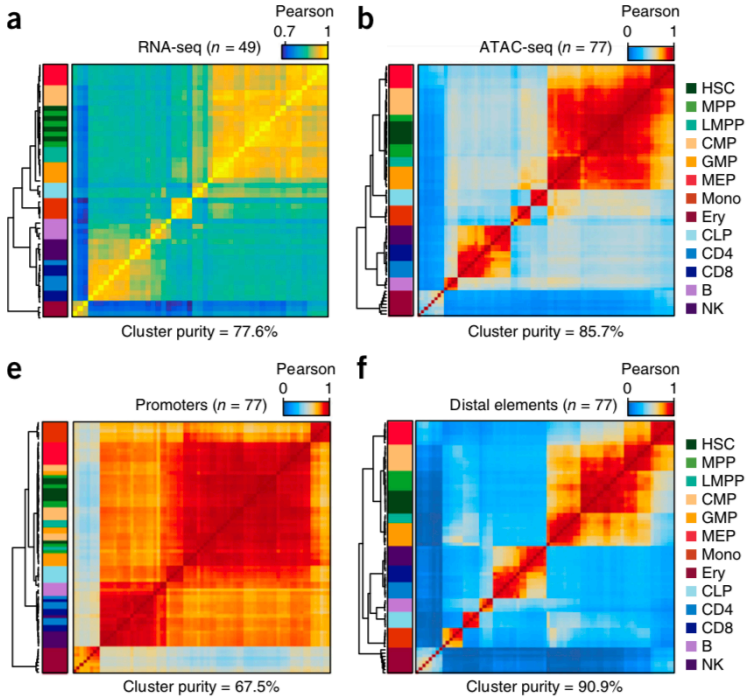
Buenrostro et al. (Nat Methods 2013)

Why single cells?

Human hematopoietic hierarchy



Bulk ATAC-seq and RNA-seq after laborious / expensive FACS



- Interrogate epigenomics at single-cell resolution
- Define cell type and state
- Investigate regulatory mechanisms

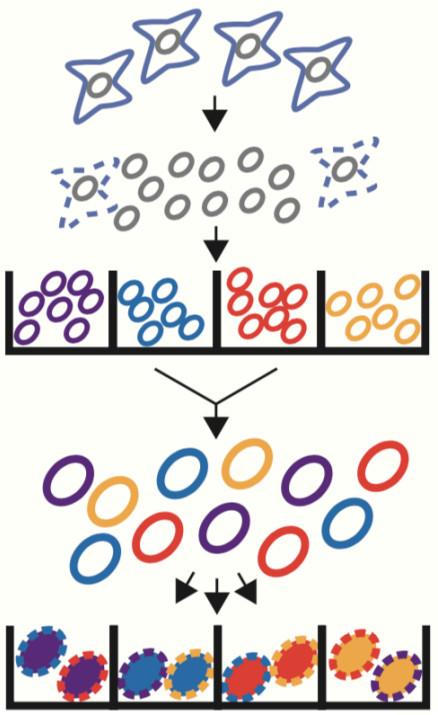
Single-cell ATAC-seq protocols

Cusanovich et al. (Science 2015)
Preissl et al. (Nature Neuroscience 2018)

Combinatorial barcoding

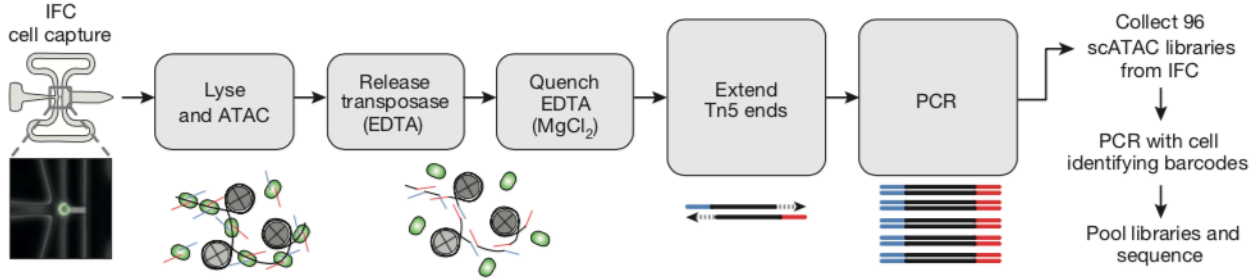
Nuclei isolated tagged and with barcoded Tn5 transposases in wells

Then pooled and redistributed to another set of wells, where second barcode introduced during PCR



Buenrostro el al. (Nature 2015)

Optimized microfluidic platform (Fluidigm)

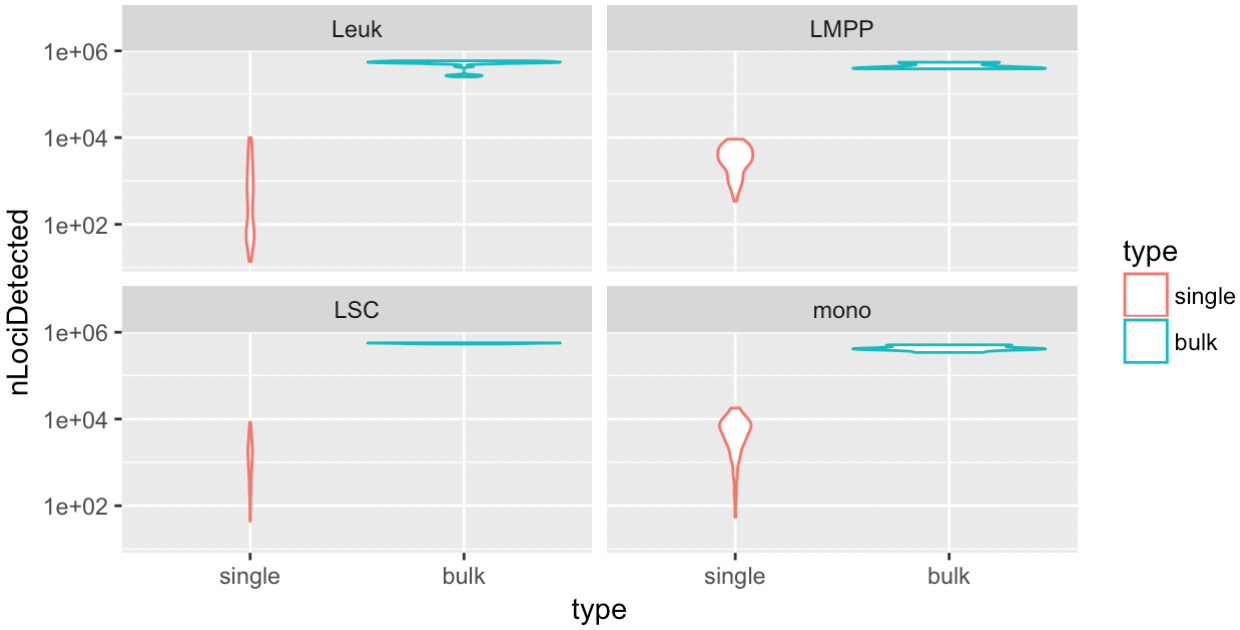


scATAC-seq has higher throughput than scChIP-seq, scDNase-seq, scHi-C

Bulk versus single cells

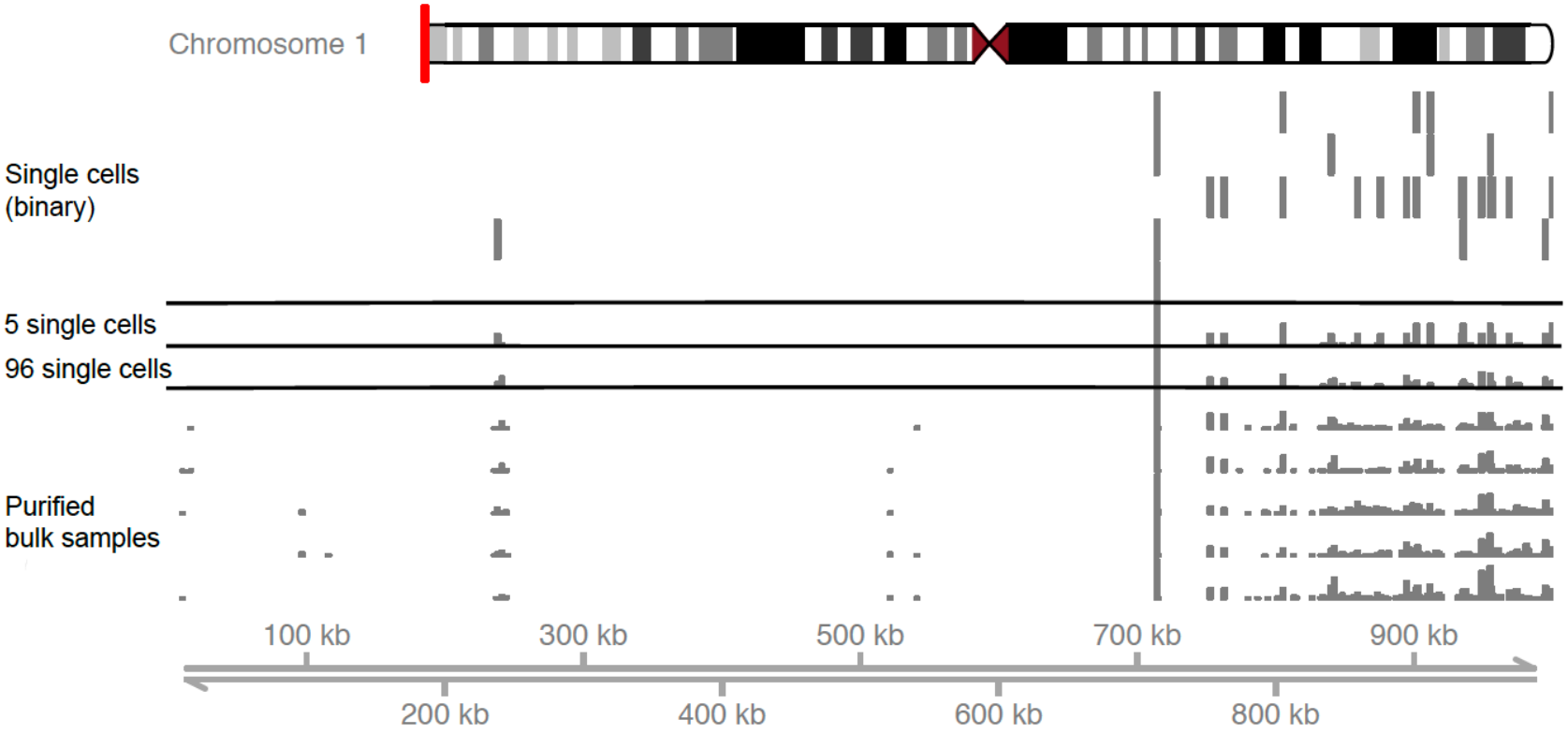
	Bulk ATAC-seq	scATAC-seq	scRNA-seq
Coverage	50M	Combinatorial: 3K Fluidigm C1: 70K	10X Genomics: 2-3M Fluidigm C1: 10-30M
Range	potentially many copies of RNA transcripts	0, 1, or 2 accessible DNA regions (diploid)	potentially many copies of RNA transcripts
Dropout (compared to bulk)		Yes	Yes

Loci detected Single vs Bulk
Data from Corces et al. (Nature Genetics 2016)



Bulk versus single cells, cont'd

Supplementary Fig. 1: Snapshot of single-cell and bulk-tissue ATAC profiles. scATAC-seq data from 5 human monocyte cells, aggregate of 5 and 96 single cells, as well as purified bulk samples, are shown. scATAC-seq data are binary, while bulk ATAC-seq data are on an integer scale (max 60 to 100). Sum of scATAC-seq data recapitulates the purified bulk ATAC profile of the same cell type.



(Urrutia et al., Bioinformatics, 2019)

scATAC-seq datasets

Supplementary Table 1: Summary statistics across different single-cell/nucleus ATAC-seq platforms. Mean number of loci detected per cell, mean number of cells detected per locus, total number of cells, as well as percentages of 0, 1, and ≥ 2 read counts from the read count matrix were shown.

Reference	Platform	No. loci per cell	No. cells per locus	Total no. cells	0 count	1 count	≥ 2 counts
Buenrostro *et al.* (2015)	Fluidigm C1	3,521	37	1,056	96.5%	2.8%	0.7%
Cusanovich *et al.* (2015)	Combinatorial indexing	1,273	4	497	99.2%	0.2%	0.6%
Preissl *et al.* (2018)	Single-nucleus ATAC-seq	1,276	28	2,088	99.1%	0.9%	0.0%

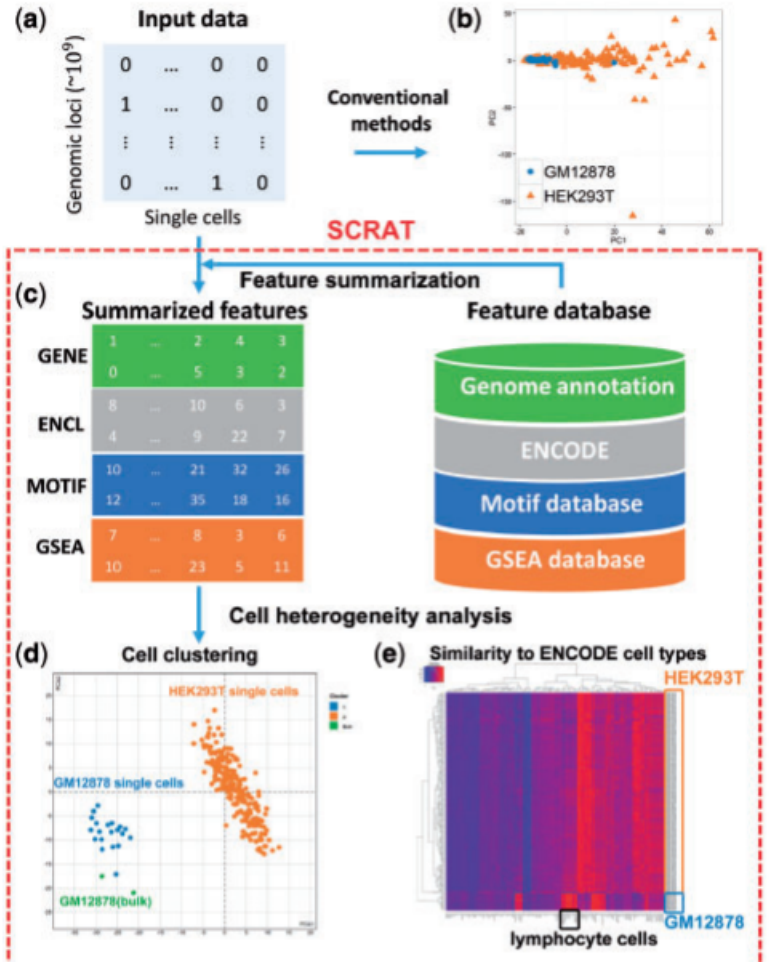
Supplementary Table 2: Single-cell and bulk-tissue ATAC-seq data sets adopted. Data sets across different platforms, species, and cell types were collected. Single-cell and bulk-tissue (after downsampling) data were used for benchmark against other existing methods.

Reference	Platform	Species	Cells	GEO	No. cell types	No. cells	No. peaks
Buenrostro *et al.* (2015)	Fluidigm C1	Human	H1, K562, GM12878, TF-1, HL-60, BJ	GSE65360	6	1,056	184,270
Buenrostro *et al.* (2015)	Fluidigm C1	Mouse	ES, EML cells	GSE65360	2	192	146,080
Cusanovich *et al.* (2015)	Combinatorial indexing	Combined	GM12878, Patski	GSE67446	2	497	157,770
Cusanovich *et al.* (2015)	Combinatorial indexing	Human	GM12878, HEK293T	GSE67446	2	714	104,260
Cusanovich *et al.* (2015)	Combinatorial indexing	Human	GM12878, HL-60	GSE67446	2	656	105,233
Corces *et al.* (2016)	Fluidgim C1	Human	Leukemic cells	GSE74310	4	576	130,448
Corces *et al.* (2016)	FACS + bulk ATAC-seq	Human	Purified hematopoietic cells	GSE74912	13	-	590,650
Preissl *et al.* (2018)	Single-nucleus ATAC-seq	Mouse	Adult forebrain	GSE100033	8	2,088	132,506

(Urrutia et al., Bioinformatics, 2019)

SCRAT – scATAC-seq method

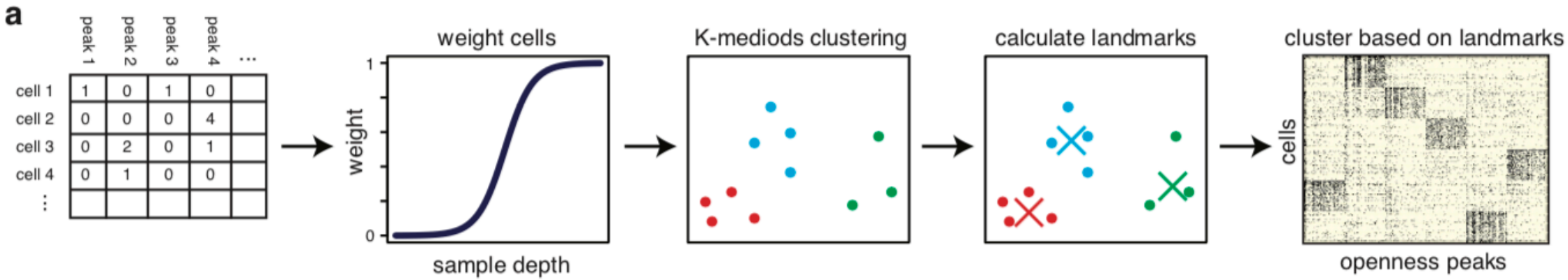
- *SCRAT* aggregates scATAC-seq read counts across biological features such as transcription factor binding motifs, DNase I hypersensitivity sites, genes or gene sets of interest.
- This is followed by a further dimension reduction step and clustering.



Ji et al. (Bioinformatics, 2017)

scABC – scATAC-seq method

- *scABC* constructs a matrix of read counts over peaks, then weights cells by sample depth and applies a weighted K-medoids clustering.
- The clustering defines a set of K landmarks, which are then used to reassign cells to clusters.



Zamanighomi et al. (Nature Communications, 2018)

Destin – scATAC-seq method

- For cell-type clustering, instead of directly aggregating peaks based on existing genomic annotations, *Destin* adopts weighted principal component analysis (PCA), with peak-specific weights calculated based on the distances to transcription start sites (TSSs) as well as the relative frequency of chromatin accessibility peaks based on reference regulomic data from the ENCODE Project (Consortium et al., 2012).

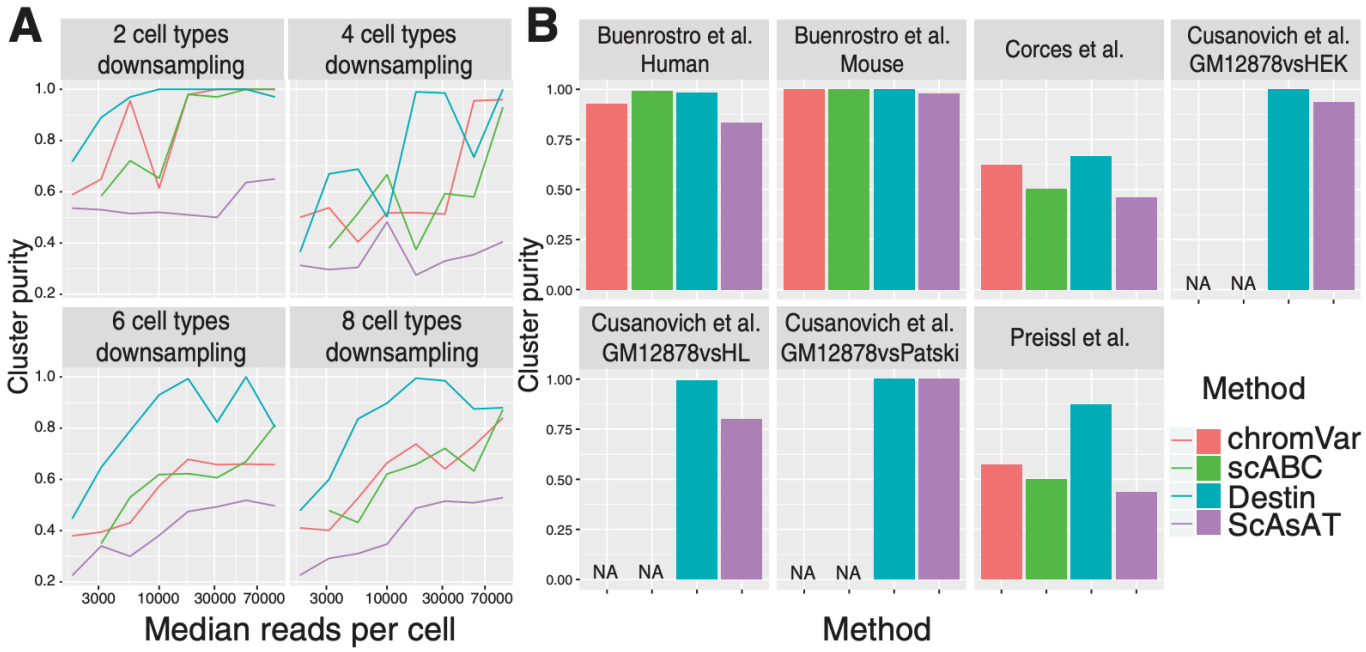
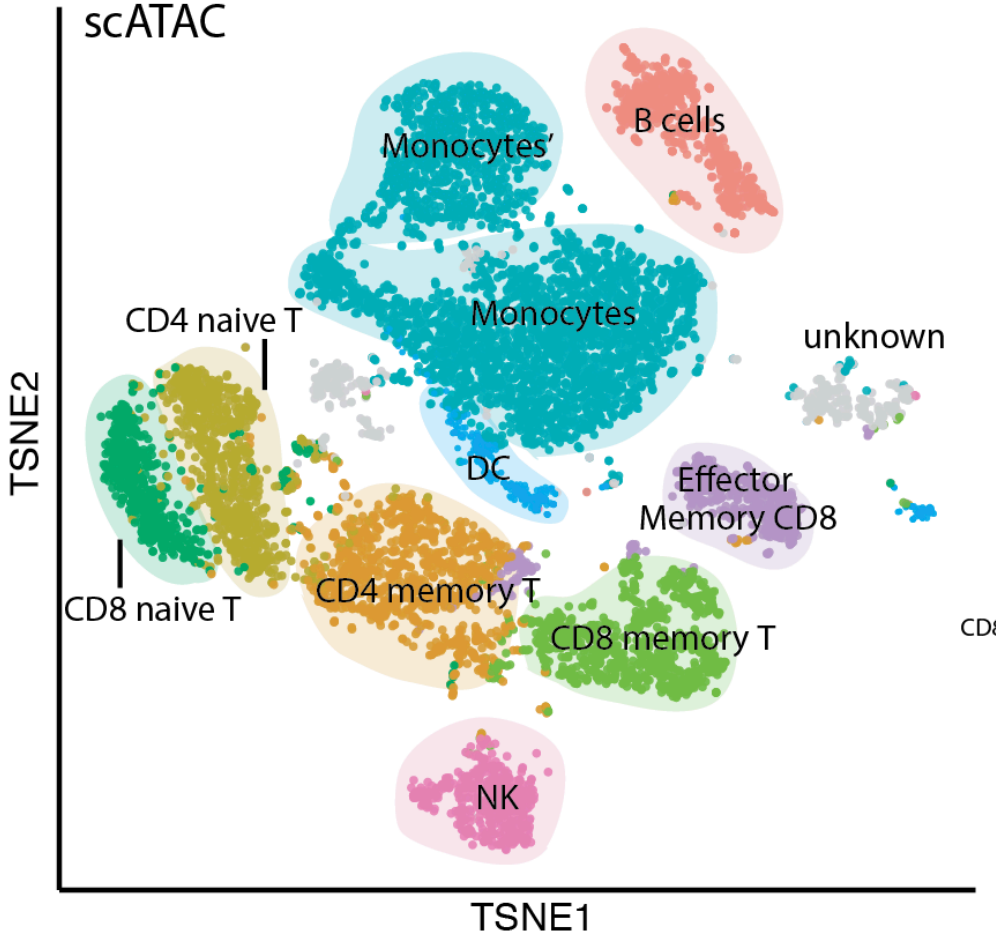


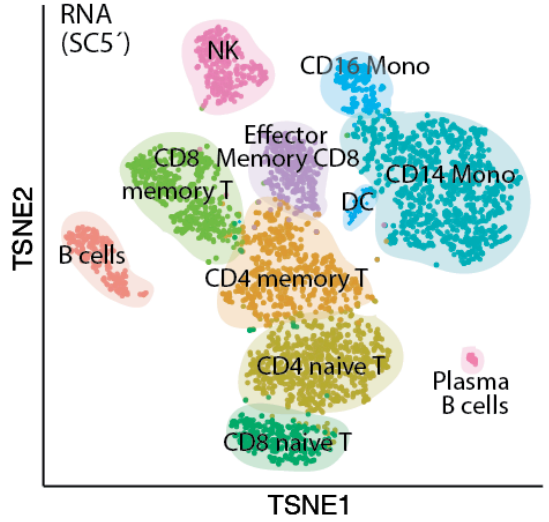
Fig. 1. Benchmark results against existing methods via downsampling and empirical data analysis. (A) Purified bulk ATAC-seq data from Corces *et al.* (2016) were downsampled with different numbers of cell types and different median reads per cell. (B) Cluster results across seven scATAC-seq datasets. chromVAR and scABC cannot be applied to the three datasets from Cusanovich *et al.* (2015) due to unavailability of required input as bam files

Urrutia et al. (Bioinformatics, 2019)

Alignment of scRNA-seq and scATAC-seq data

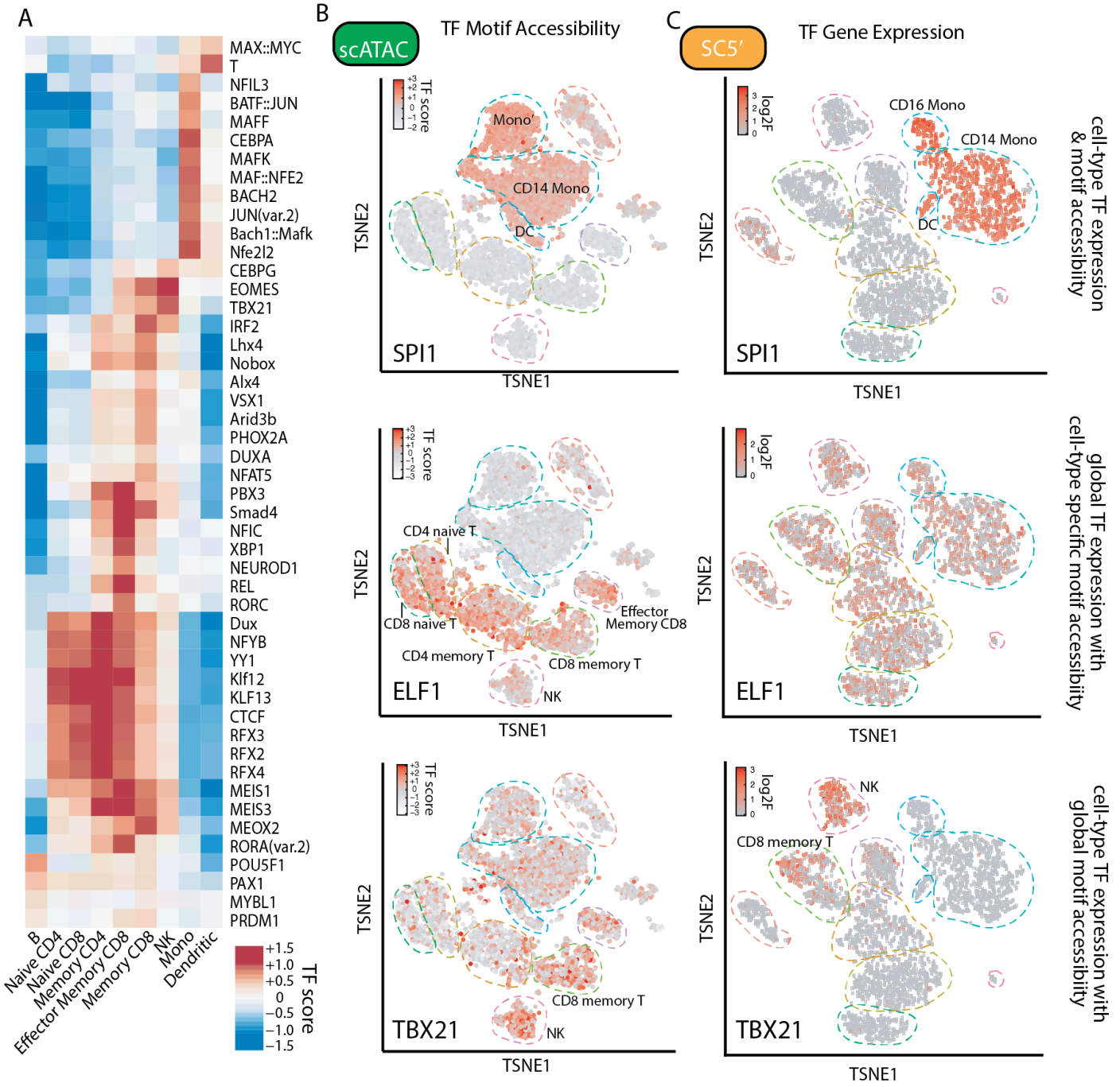
PBMC scATAC-seq and scRNA-seq (5’ and 3’)



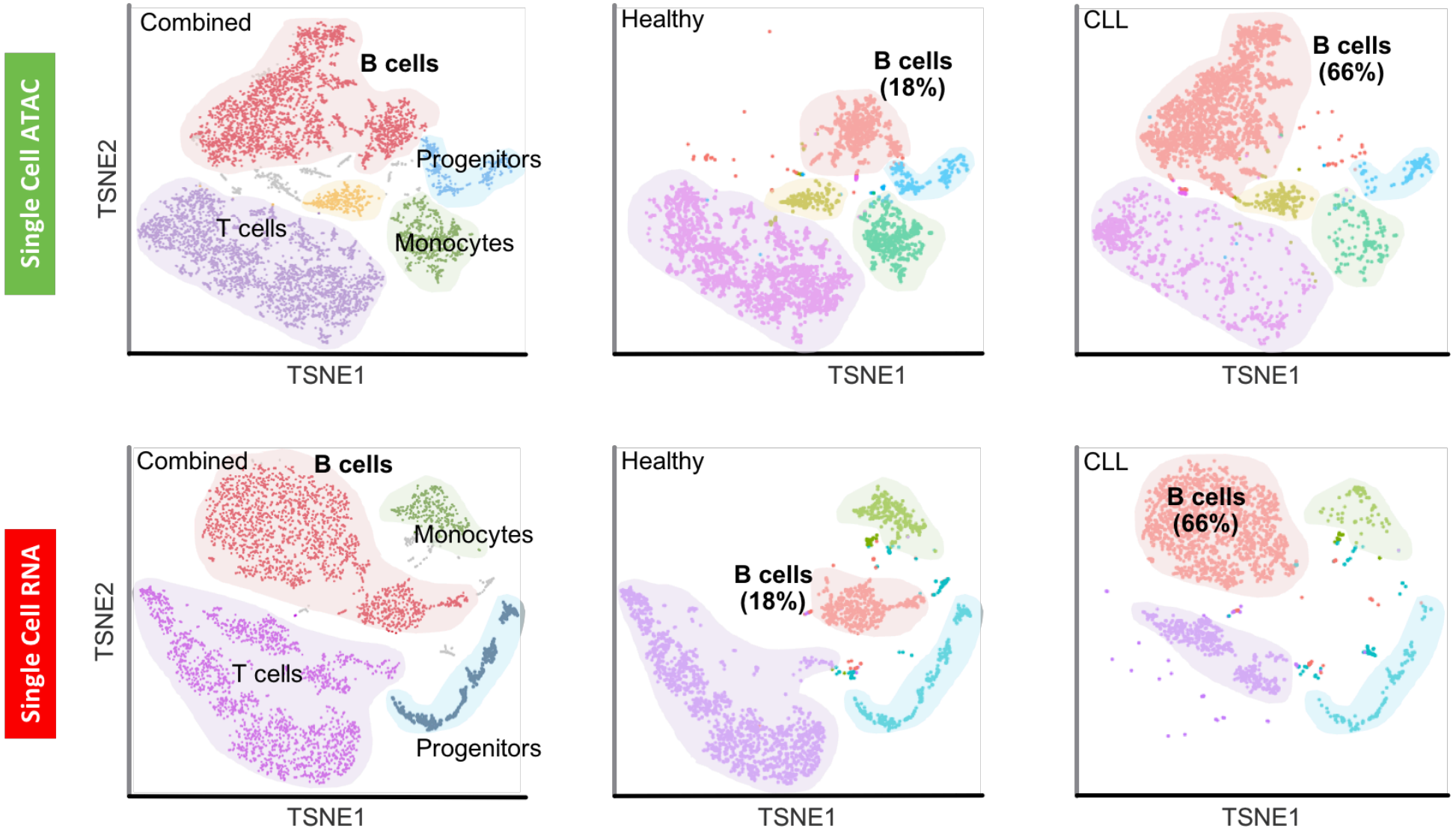


	SC5'	SC3'	scATAC
B cells	247 (5.3%)	158 (6.5%)	687 (6.8%)
CD4 memory T	637 (13.7%)	245 (10.2%)	1,080 (10.7%)
CD4 naive T	794 (17%)	479 (19.9%)	878 (8.7%)
CD8 effector memory T	350 (7.5%)	143 (5.9%)	549 (5.4%)
CD8 memory T	541 (11.6%)	241 (10%)	1,496 (14.8%)
CD8 naive T	320 (6.9%)	–	545 (5.4%)
Monocytes'	181 (3.9%)	70 (2.9%)	918 (9.1%)
DC	73 (1.6%)	63 (2.6%)	323 (3.2%)
Monocytes	1,143 (24.5%)	855 (35.4%)	2,605 (25.8%)
NK	344 (7.4%)	138 (5.7%)	509 (5%)
Plasma B	36 (0.8%)	21 (0.9%)	–
unknown	–	–	498 (4.9%)
Total	4,666	2,413	10,088

Transcription factor



Bone marrow mononuclear cells (BMMCs)



Existing methods: coupled NMF

- "Soft" clustering based on nonnegative matrix factorization.
- To couple two matrix factorizations, we introduce a term $tr(W_2^T A W_1)$, where A is a "coupling matrix." The construction of A is application specific but depends on the assumption that, based on scientific understanding or prior data, it is possible to identify a subset of features in one sample that are linearly predictable from the features measured in the other sample.

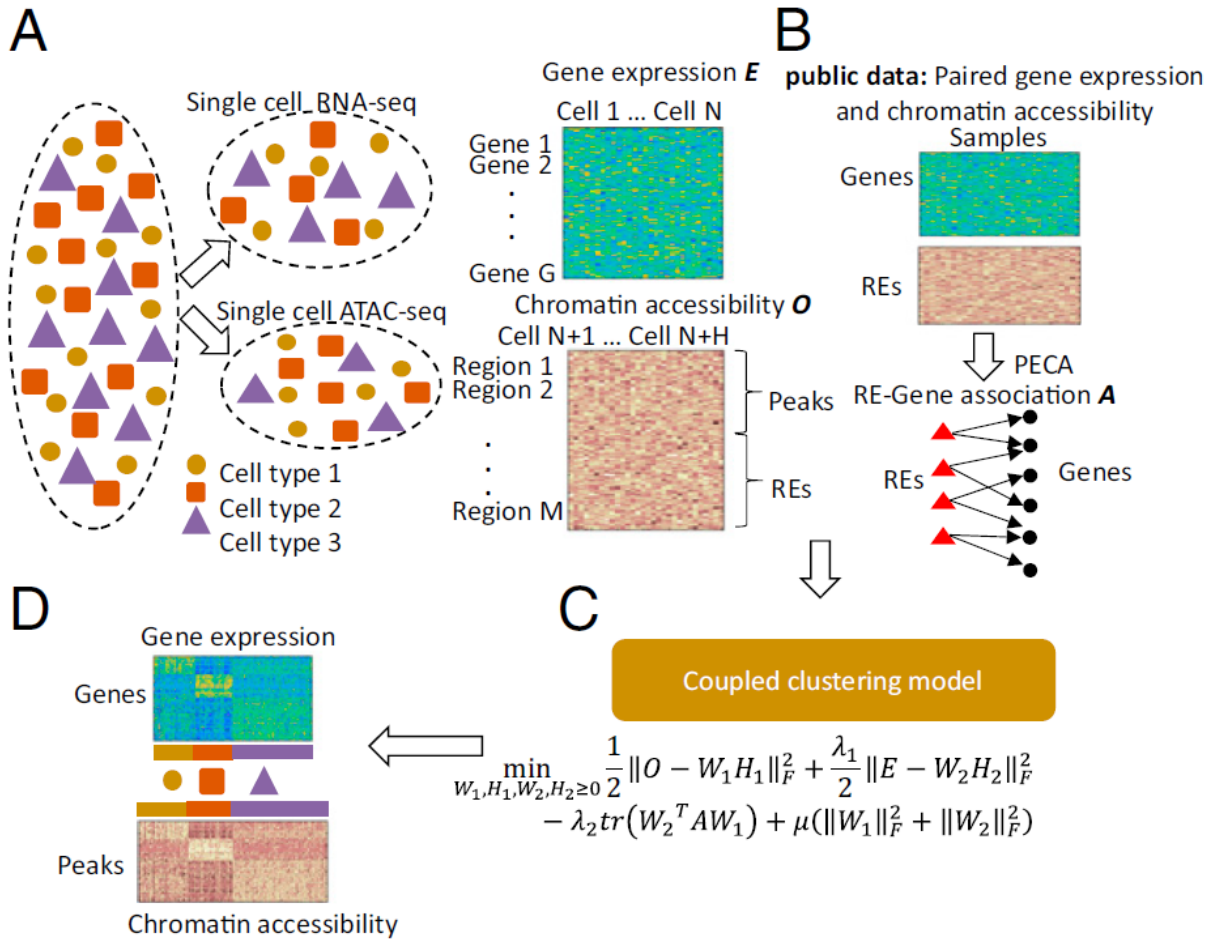


Fig. 1. Overview of the coupled-clustering method. (*A*) Single-cell gene expression and single-cell chromatin accessibility data. (*B*) Learning coupling matrix from public data. (*C*) Coupled clustering model. (*D*) Cluster-specific gene expression and chromatin accessibility.

Duren et al. (PNAS, 2018)

Existing methods: coupled NMF

- In analysis of big data this type of **circularity** is becoming more and more common and problematic: You take another data set and formulate a prior (in this case, ATACseq and RNAseq relationship). Then you use that prior to analyze your own new data (do alignment). Then you test for the same relationships on which you had assumed a strong prior, and draw conclusions. How do you know whether your conclusions are due to your own data or simply transferred from the preexisting data set?

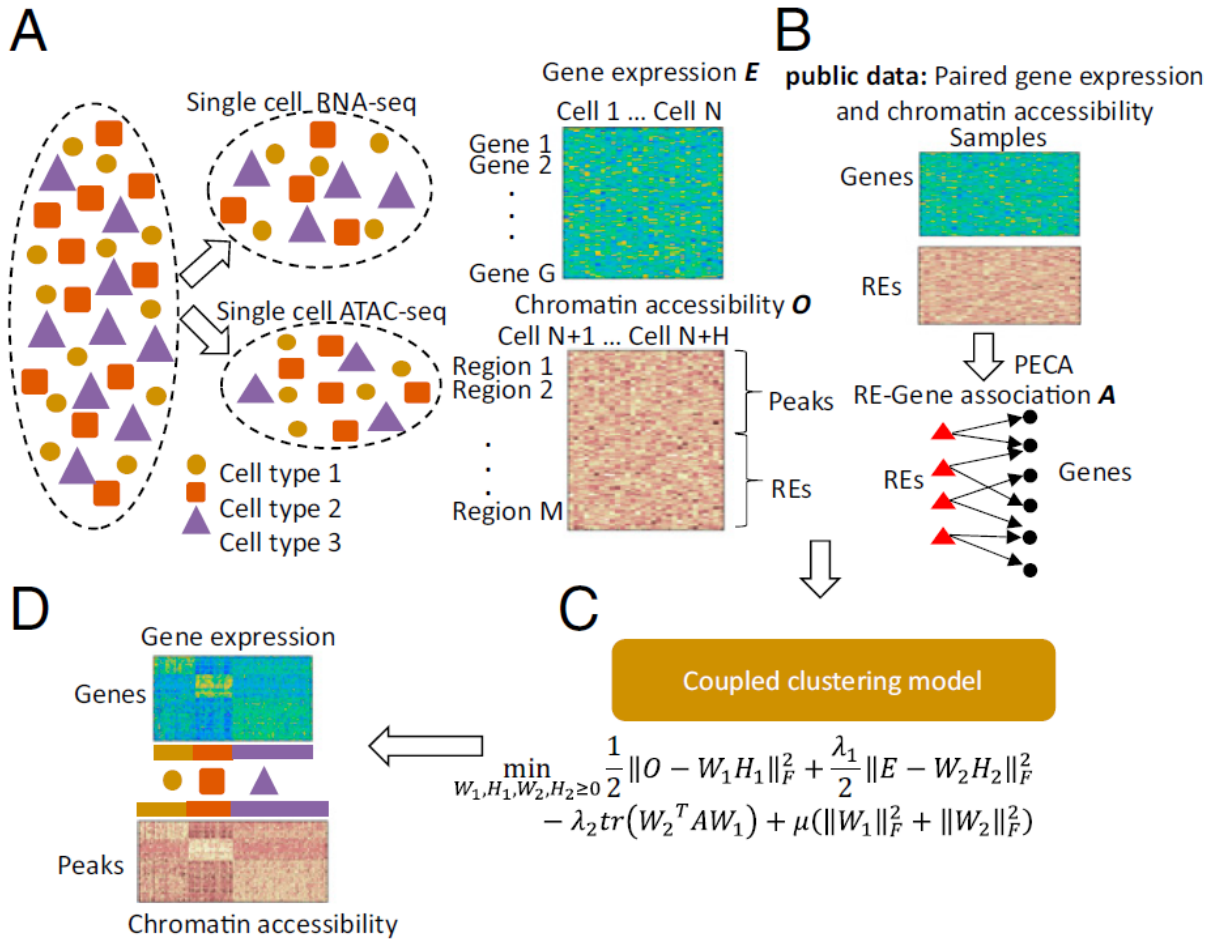


Fig. 1. Overview of the coupled-clustering method. (*A*) Single-cell gene expression and single-cell chromatin accessibility data. (*B*) Learning coupling matrix from public data. (*C*) Coupled clustering model. (*D*) Cluster-specific gene expression and chromatin accessibility.

Duren et al. (PNAS, 2018)

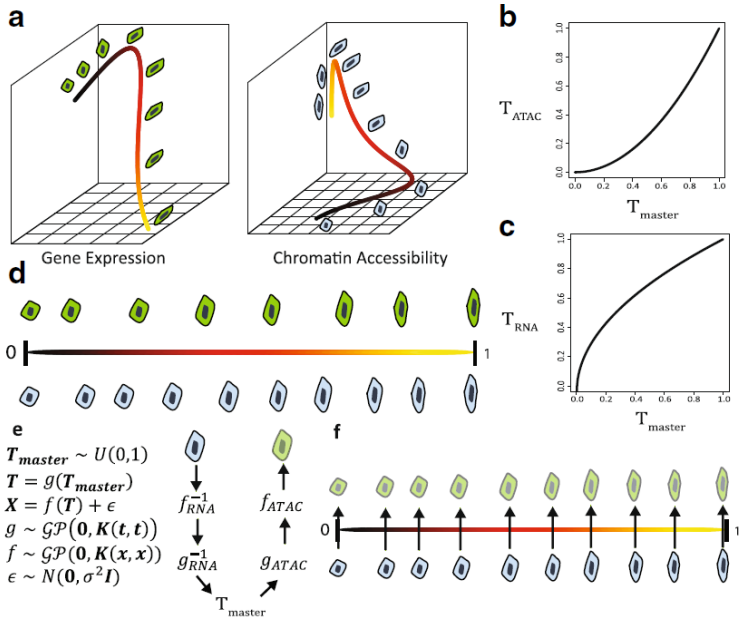
Existing methods: MATCHER

- MATCHER first reconstructs two pseudotime, from two sources of data separately. It then aligns the two pseudotime into **a linear trajectory (master pseudo time between 0 and 1)**. If looking from the trajectory reconstructions standpoint, I think the key issue is that it **doesn't allow branching and implicitly assumes that cells come from a one-dimensional manifold.**

METHOD Open Access

MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics

Joshua D. Welch[1,2], Alexander J. Hartemink[3] and Jan F. Prins[1,2*]



Existing methods: CCA

- Stuart et al., Cell, 2019.
- “Derive a ”gene activity matrix“ from the scATAC-seq profiles, utilizing observed reads at gene promoters and enhancers as a **prediction of gene activity** (by Cicero, which is **correlated with gene expression**), representing a ***synthetic scRNA-seq dataset*** to leverage for integration.”
- Identified anchors (via CCA) between the scRNA-seq and scATAC-seq using the gene activity matrix derived from scATAC-seq profiles.
- Problem: scATAC-seq is too sparse to predict gene expression.

Cell

CellPress

Volume 177, Issue 7, 13 June 2019, Pages 1888-1902.e21

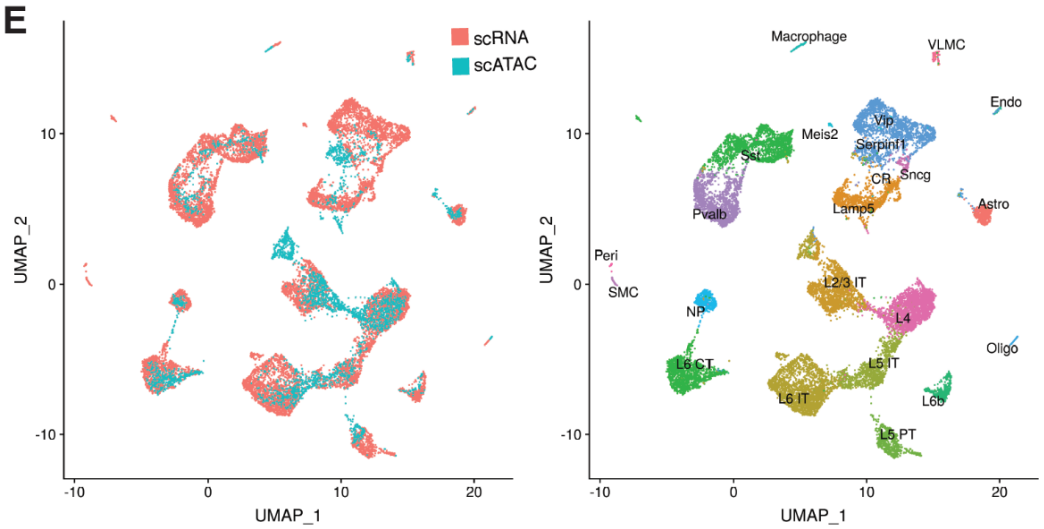
Resource

Comprehensive Integration of Single-Cell Data

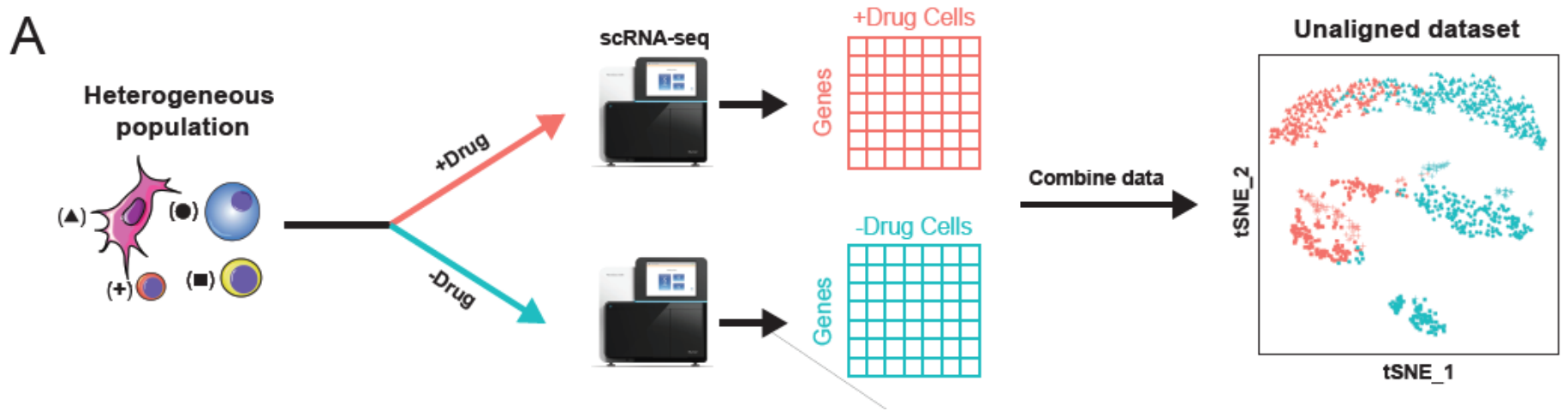
Tim Stuart [1, 4], Andrew Butler [1, 2, 4], Paul Hoffman [1], Christoph Hafemeister [1], Efthymia Papalexi [1, 2], William M. Mauck III [1, 2], Yuhan Hao [1, 2], Marlon Stoeckius [3], Peter Smibert [3], Rahul Satija [1, 2, 5]

Show more

https://doi.org/10.1016/j.cell.2019.05.031

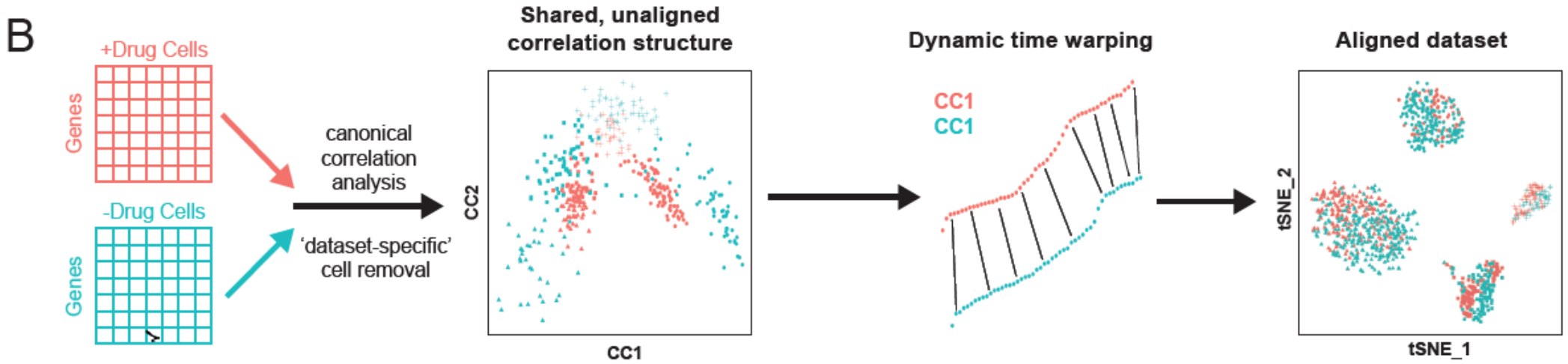


Overview of Seurat alignment of scRNAseq datasets



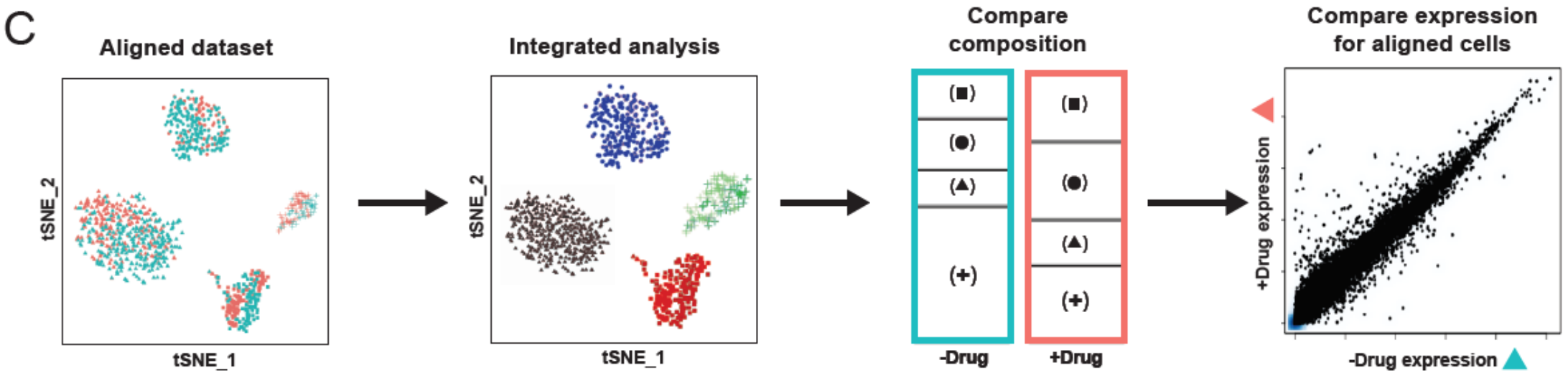
Example: Heterogeneous scRNA-seq datasets are generated in the presence or absence of a drug.

Overview of Seurat alignment of scRNAseq datasets



1. **Shared gene correlation structure** that is conserved between the two datasets, and can serve as a scaffold for the alignment
2. **Identifies and discards** individual cells that cannot be well described by this shared structure, and are therefore 'dataset-specific' (unalignable).
3. It **aligns** the two datasets into a conserved low-dimensional space, using non-linear 'warping' algorithms to normalize for differences in feature scale, in a manner that is robust to shifts in population density.

Overview of Seurat alignment of scRNAseq datasets



4. It proceeds with an **integrated downstream analysis**, for example, identifying discrete subpopulations through clustering, or reconstructing continuous developmental processes
5. It performs **comparative analysis** on aligned subpopulations between the two datasets, to identify changes in population density or gene expression

Identifying shared correlation structures across datasets:

- **Canonical Correlation Analysis (CCA)** is a way of inferring information from cross-covariance matrices

$$X = (x_1, \dots, x_n)^T, Y = (y_1, \dots, y_m)^T$$
$$\Sigma_{XY} = cov(X, Y)$$

Seek vectors a, b s.t. $a'X$ and $b'Y$ maximize:

$$\rho = corr(a'X, b'Y)$$

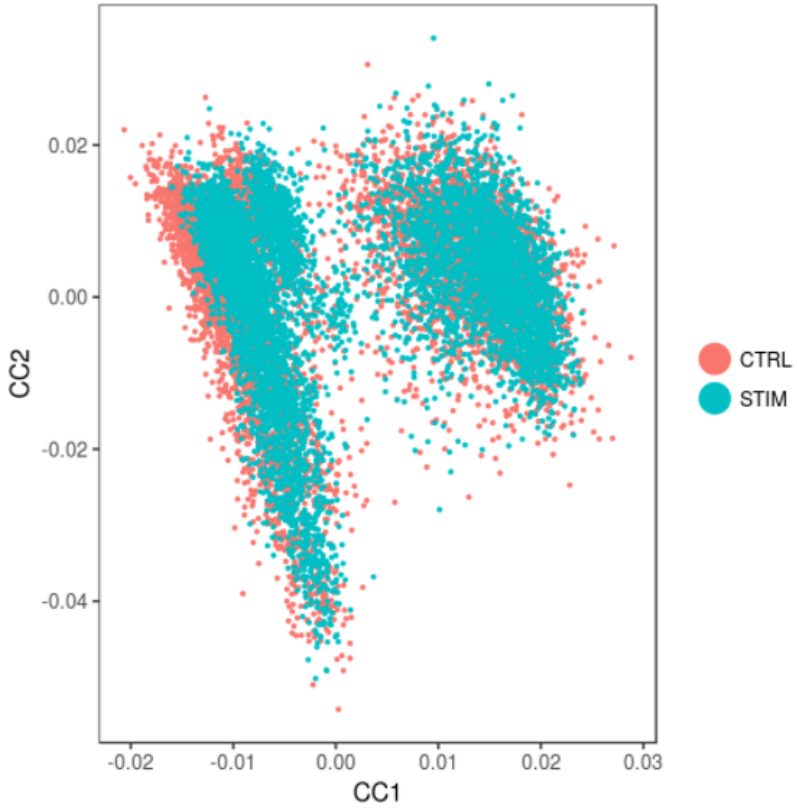
- Find projections of both datasets such that the correlation between the two projections is maximized.

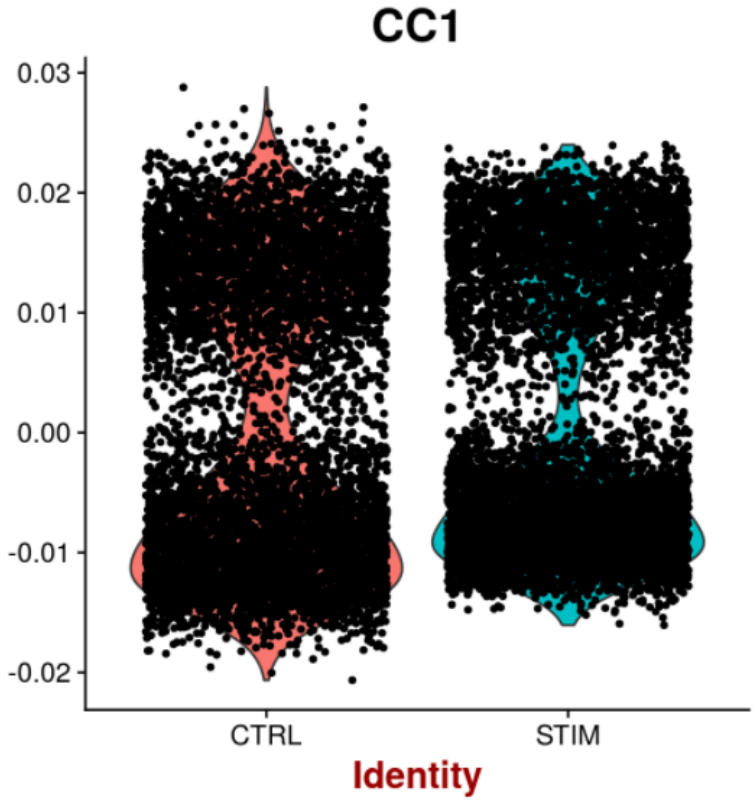
$$\max_{u,v} u^T X^T Y v \text{ subject to } u^T X^T X u \leq 1, v^T Y^T Y v \leq 1$$

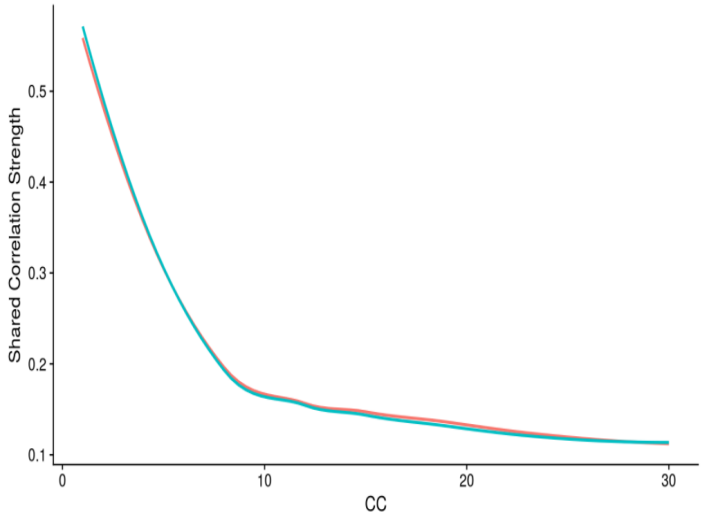
Aligning basis vectors from CCA

- Linearly transform the 'metagenes' to match their 95% reference range, correcting for global differences in feature scale
- Determine a mapping between the metagenes using 'dynamic time warping', which locally compresses or stretches the vectors during alignment to correct for changes in population density

CCA plot before alignment

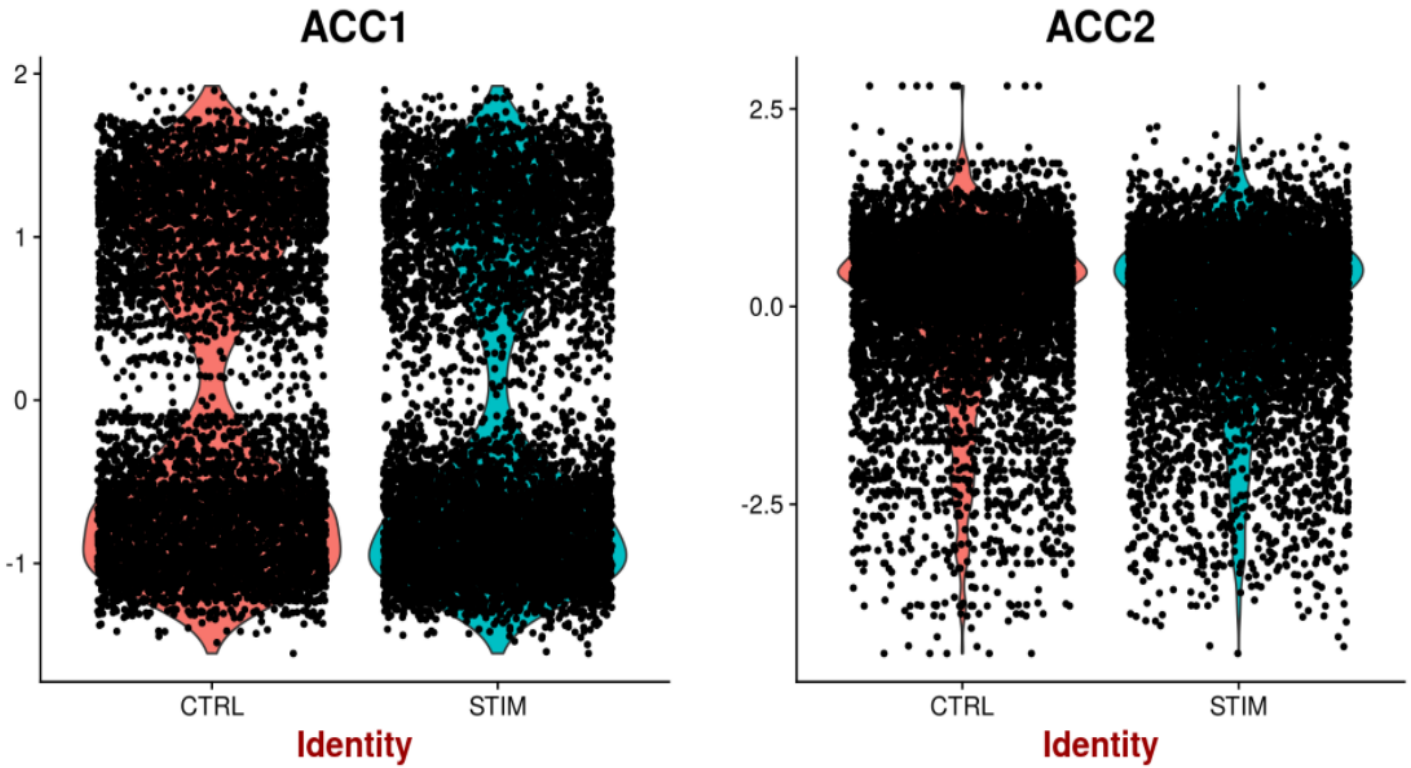






CC1 and CC2 separate myeloid from lymphoid cells in both datasets, but the values remain ‘shifted’ relative to each other. Need to choose CCs for downstream analysis and then ‘align them’.

After alignment



Existing methods: BIRD+MNN

- Predict scATAC-seq profiles by scRNA-seq
- Align scATAC-seq and predicted scATAC-seq data using MNN (i.e., align same data types)

Global Prediction of Chromatin Accessibility Using RNA-seq from Small Number of Cells

Weiqiang Zhou, Zhicheng Ji, Hongkai Ji

doi: https://doi.org/10.1101/035816

References

- Cusanovich, Darren A., et al. "Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing." *Science* 348.6237 (2015): 910-914.
- Buenrostro, Jason D., et al. "Single-cell chromatin accessibility reveals principles of regulatory variation." *Nature* 523.7561 (2015): 486.
- Ji, Zhicheng, Weiqiang Zhou, and Hongkai Ji. "Single-cell regulome data analysis by SCRAT." *Bioinformatics* 33.18 (2017): 2930-2932.
- Zamanighomi, Mahdi, et al. "Unsupervised clustering and epigenetic classification of single cells." *Nature communications* 9.1 (2018): 2410.
- Urrutia, Eugene, et al. "Destin: toolkit for single-cell analysis of chromatin accessibility." *Bioinformatics*, btz141 (2019).
- Duren, Zhana, et al. "Modeling gene regulation from paired expression and chromatin accessibility data." *Proceedings of the National Academy of Sciences* 114.25 (2017): E4914-E4923.

References, cont’d

- Welch, Joshua D., Alexander J. Hartemink, and Jan F. Prins. "MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics." *Genome biology* 18.1 (2017): 138.
- Stuart, Tim, et al. "Comprehensive Integration of Single-Cell Data." *Cell* (2019).
- Butler, Andrew, et al. "Integrating single-cell transcriptomic data across different conditions, technologies, and species." *Nature biotechnology* 36.5 (2018): 411.
- Zhou, Weiqiang, Zhicheng Ji, and Hongkai Ji. "Global prediction of chromatin accessibility using RNA-seq from small number of cells." *bioRxiv* (2016): 035816.
- Haghverdi, Laleh, et al. "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors." *Nature biotechnology* 36.5 (2018): 421.