Benchmarking methods for alignment of scRNA-sea and scATAC-seq data

Contents

1 Introduction

Single-cell Assays for Transposase-Accessible Chromatin by sequencing (scATAC-seq) enable profiling of the chromatin accessibility landscape of thousands of single-cells, and identification of *cis*- and *trans*- acting regulatory elements in heterogeneous cell populations (Cusanovich et al. 2015, 2018). However, common analysis tasks for single-cell RNA-seq (scRNA-seq) datasets, such as clustering and annotation of cell types, are challenging for scATAC-seq data, because of it's extreme sparsity and limited prior knowledge of cell-type specific accessibility patterns. Integration of accessibility datasets with comprehensively annotated scRNA-seq datasets can allow denoising of biological signal, guide cell type annotation and disentangle causal relationships between biological layers of information and how these co-determine complex phenotypes (Buenrostro et al. 2018; Granja et al. 2019).

In most cases, molecular profiles will be measured in parallel from cells sampled from the same tissue/cellular population. Consequently, aligning different datasets requires an at least partial correspondence between profiled features across omics. This is true when integrating different scRNA-seq datasets or scRNA-seq data with spatial transcriptomics. For omic types that measure molecular features that are different than genes, data is usually preprocessed to generate a matrix of gene level features e.g. measuring gene activity from ATAC accessibility peaks [https://www.ncbi.nlm.nih.gov/pubmed/30078726]. Different integration methods use different inference algorithms for the latent space projection (e.g. canonical correlation analysis, non-negative matrix factorization, variational autoencoders), but all allow the mapping of a gene expression (or other molecular feature) vector x on a latent space vector z, via a vector of feature loadings w. Inspection of loadings can distinguish features that allow alignment across datasets and potentially indicate cell and omic specific contributions to the overall data variation.

- Finding which method works best
- but also comparing different preprocessing strategies

Here we implemented an analysis workflow to optimize the performance of methods to transfer labels inferred from a single-cell gene expression dataset to a single-cell accessibility dataset. We compared three published methods based on robustness and performance on a range of metrics. We then applied our workflow to optimize integration of scRNA-seq and scATAC-seq datasets generated from developing human thymus.

2 Results

2.1 Designing a benchmark for label transfer methods: PBMC dataset

We started by comparing the performance of three publicly available methods for label transfer: CCA (Stuart et al. 2019), Conos (Barkas et al. 2019) and Liger (Welch et al. 2019). We used a publicly available dataset of Peripheral Blood Mononuclear Cells (PBMC) from 10XGenomics. After filtering, this comprises of transcriptomic profiles for 5607 cells and accessibility profiles for 8960 cells.

Visualizing the predicted labels on the embedding of genome-wide scATAC-seq profiles, we find that for all the methods reconstruct similar clusters, which agree with the global structure of the accessibility data (Fig.1A). Cell type proportions are in line with those found in the cell population measured with scRNA-seq (Fig.1B). We next evaluated run time and robustess of methods running label prediction including a growing fraction of scATAC-seq cells. Conos is the fastest method, with speed minimally affected by the number of query cells (Fig.1). Both Conos and CCA were equally robust in their label predictions for subsets of cells, with Adjusted Rand Index compared to labels inferred from the full dataset between 0.6 and 0.8 (Fig.1C). Conversely, Liger showed much higher variability in label prediction.

All methods associate labels with a prediction score, that allows to remove labels with high uncertainty (Fig.1D). We found that different methods score with high confidence similar clusters (Fig.1E). We clustered scRNA-seq cells with increasing resolution (i.e. from bigger few clusters to many smaller clusters) and compared the distribution of prediction scores. We found that Conos predicts labels with higher confidence with larger (low-resolution) clusters (Fig.1F)

We wanted to quantify if cells that get annotated with the same label also have a similar genome-wide bin accessibility profile. This is crucial to allow reconstruction of consensus chromatin accessibility profiles of clusters/cell types. We calculate the fraction of nearest neighbors per cell that share the predicted label. For each prediction, we calculate a null distribution of this fraction by randomly shuffling the assigned labels. We define a KNN purity statistic as the deviation between the null and the true distribution (KS test, see Methods 5.2.1). We found that Conos and Liger performed slightly better on this dataset than CCA (Fig.1G).

2.2 Optimizing label transfer on Thymus dataset

We moved on to optimizing label transfer and integration on a dataset capturing cell populations with less distinct molecular signatures. We analyzed a dataset generated from developing thymus at 10 post-conceptional weeks. Libraries for scRNA-seq and scATAC-seq were generated on different samples from the same donor. Clustering and annotation of the transcriptomes identified 15 distinct cell populations (Fig|2). The main cluster represents T cells undergoing maturation, from the Double Negative (DN) state, to Double Positive (DP) to CD4+ or CD8+ Single Positive (SP) T cells. Other smaller clusters were identified as other components of the thymic microenvironment, such as thymic epithelial cells (TEC) or other immune cell populations (e.g. Dendritic Cells, Macrophages, Innate like cells). Clustering of genome-wide scATAC-seq profiles also partitioned the data with a similar structure, with smaller clusters and a main population divided into subclusters (Fig. ?). Repeating the label transfer workflow as for the PBMC dataset we noticed that while all methods consistently assigned the smaller clusters, only with CCA transfered annotations resemble

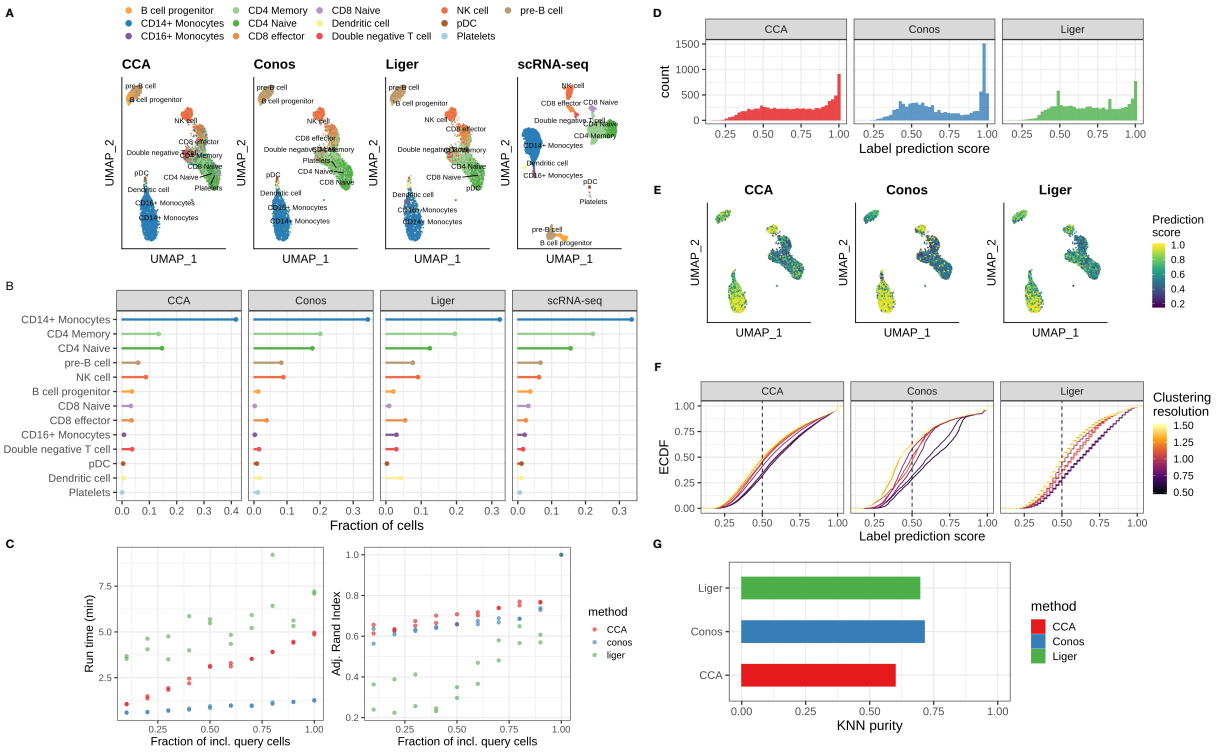

Figure 1: Label comp

chromatin accessibility clusters, and this is reflected in the KNN purity score (Fig.2B). We reasoned that the aggregation to gene-level features might be increasing noise in the T cell cluster. To obtain a cell-by-gene matrix that best represents the signal in the genome-wide cell-by-bin matrix, we binarized accessibility by gene so that accessibility is equal to 1 if at least one accessible genomic bin overlaps with the gene body or promoter. We found that binarizing the gene x cell matrix maintains the clustering structure found in the genome-wide accessibility, while this is lost when taking bin counts as accessibility measures of genes (Fig.2B). Consequently, using the binary gene matrix for label transfer greatly increased the KNN purity for all methods (Fig.2D).

2.3 Integration of accessibility and gene expression profiles of T cells

Our comparison of label transfer outcomes indicated that integration with a binary cell-by-gene accessibility matrix and using CCA is the best integration strategy, by our metrics. Using these indications, we performed integration af scATAC-seq and scRNA-seq cells focusing on the putative T cell populations in the thymic datasets (Fig.3A-B). We modelled the differentiation trajectory of T cells using pseudotime analysis (Fig.3C). This ordered the cell populations in both datasets mostly according to the expected order for conventional T cell maturation (Fig.3D). We were then able to explore changes in accessibility patterns along the inferred differentiation trajectory. We observed that genome-wide accessibility tends to decrease during T cell maturation (Fig.3E), as it is expected for cells moving towards a "primed" state (ref? Maybe **???** ?). We then performed motif accessibility analysis with ChromVAR (Schep et al. 2017), which measures enrichment or depletion of accessibility at transcription factor (TF) motifs in scATAC-seq profiles. Among motifs with most variable accessibility in our dataset, we identify motifs for many TFs that have been shown to be involved in T cell maturation (Fig.3), such as TCF, RUNX, REL and TBX transcription factors (Hosoya et al. 2018, Park et al. 2019, unpublished). These show distinct changes in accessibility along pseudotime, especially at the very early and final stages of maturation. Motif accessibility showed less variability between

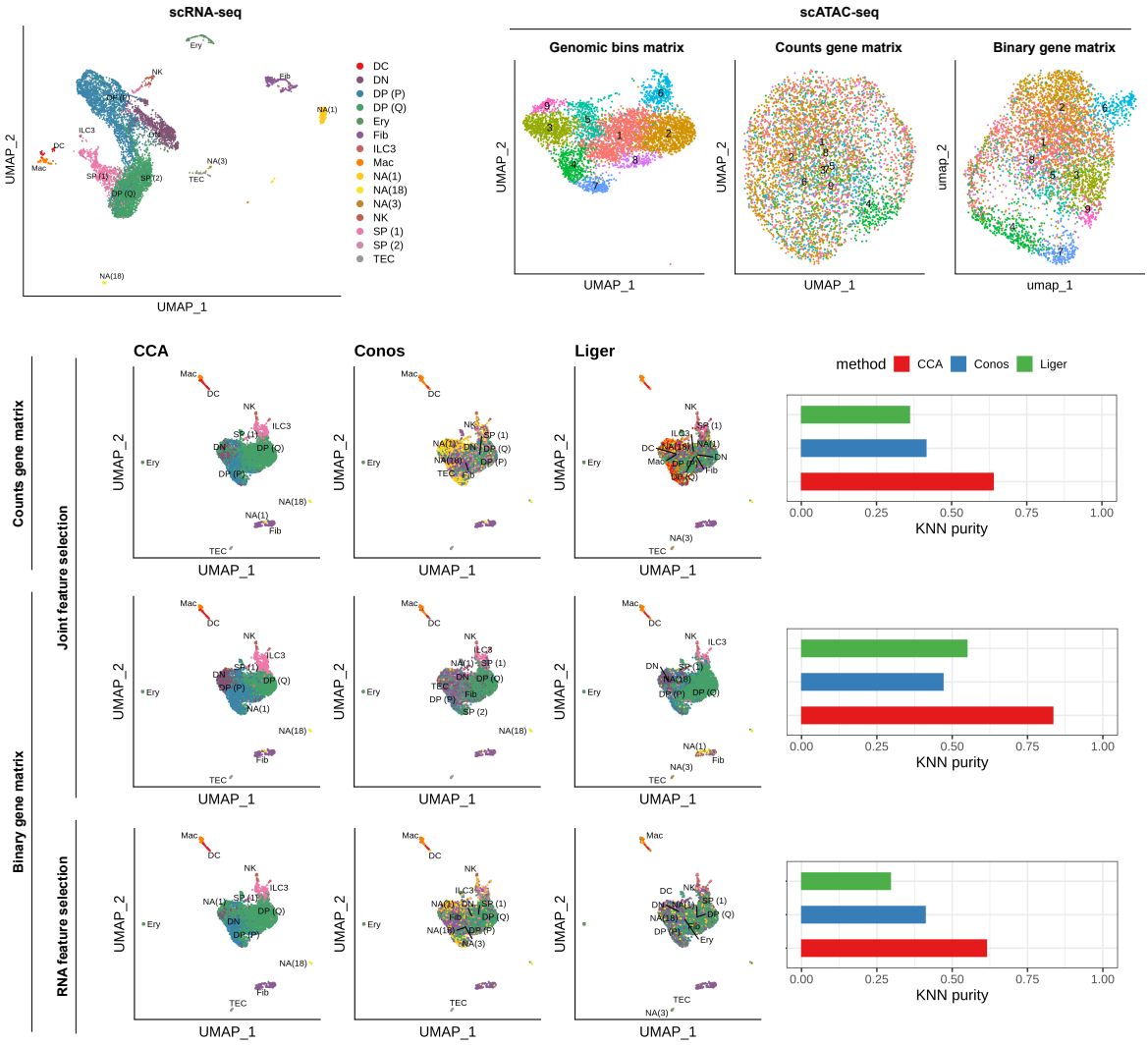

Figure 2: Thymus dataset

DP cell states. Data integration enabled us to compare patterns of TF motif accessibility with patterns of TF expression along pseudotime. We identified instances in which TF expression correlates with motif accessibility, as well as cases in which motif accessibility decreases as TF expression increases (3G).

3 Discussion

- Methods perform similarly on a dataset with defined clusters
- Does it make a difference how do you reduce ATAC to gene - features? Scoring that takes into account enhancer activity (e.g. Cicero) might be more informative to align differentiating cells.

- Testing performance on datasets from joint profiling protocols

4 Conclusion

- CCA seems to outperform the other label transfer methods
- Binarization of the gene level accessibility data allows assignment of subclusters
- Selecting features based on both datasets is better

5 Methods

5.0.1 Datasets

PMBC (10X genomics): this dataset is composed of Peripheral Blood Mononuclear Cells (PBMCs) from one donor. Raw scRNA-seq counts were downloaded with the Seurat R package [(Satija et al. 2015)] (download link: https://www.dropbox.com/s/3f3p5nxrn5b3y4y/pbmc_10k_v3.rds?dl=1). Raw scATAC-seq fragments were downloaded with the SnapATAC R package (Fang et al. 2019) (download link: http://renlab.sdsc.edu/r3fang//share/github/PBMC_ATAC_RNA/atac_pbmc_10k_nextgem.snap). A total of 5607 scRNA-seq cells and 8690 scATAC-seq cells were used for analysis after QC.

Thymus dataset: this dataset consists of cells from one fetal thymus at 10 post-conception weeks (Park et al. unpublished, Dominguez-Conde et al. unpublished). A total of 8321 scRNA-seq cells and 5793 scATAC-seq cells were used for analysis after QC. Cell type annotations for the scRNA-seq cells was provided by the authors and based on expression of marker genes.

5.1 Data preprocessing and dimensionality reduction

We preprocessed and normalized scRNA-seq using the standard pipeline from the R package Seurat (v3.1.1). ScATAC-seq reads were aligned and preprocessed using CellRanger (10X genomics). We used the SnapATAC pipeline for quality control and preprocessing (**???**). This generates genome-wide single-cell accessibility profiles by binning the genome into fixed-size windows (selected bin size: 5 kb) and constructing a cell-by-bin binary matrix, estimating accessibility for each bin.

For dimensionality reduction, we used Principal Component Analysis for scRNA-seq datasets and Latent Semantic Indexing for dimensionality reduction of accessibility matrices, as proposed by Cusanovich et al. (2015). For data visualization, we used Uniform Manifold Approximation and Projection (UMAP) algorithm (McInnes, Healy, and Melville 2018) on low-dimensional data representation.

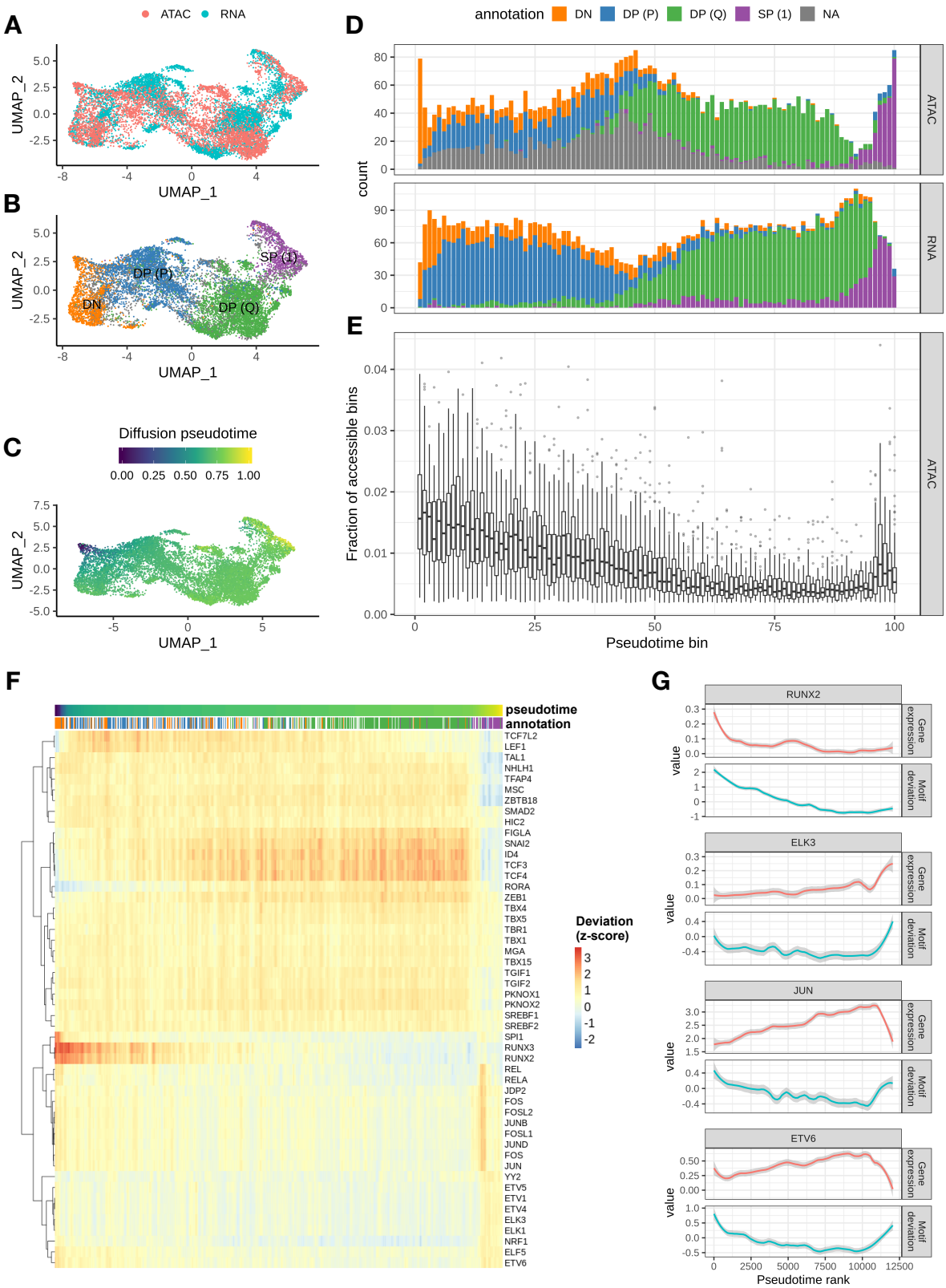

Figure 3: **Integrative analysis of T cell maturation in developint thymus:** Joint visualization of scRNA-seq data with scATAC-seq cells for the T cell clusters, colored by (A) technology, (B) inferred cell type labels and (C) diffusion pseudotime; (D) Distribution of cells along 100 equal-sized pseudotime rank bins, colored by inferred cell type; (E) Distribution of the fraction of accessible genomic bins per cell for each pseudotime rank bin; (F) Heatmap of TF motif deviation for the top 50 most variable TF motifs inferred by chromVAR. Cells are ordered by pseudotime. Values are smoothed with a running average function (step = 30); (G) Smoothed values for log-normalized gene expression counts (top, pink) and TF motif deviation (bottom, cyan) along pseudotime ranks. Values are smoothed with the LOESS function (span = 0.2).

5.2 Label transfer

Accessibility in gene-level features: We generated cell-by-gene count matrices by aggregating bin coverage over the gene bodies and promoters (2kb upstream of transcriptional start site). Count gene matrices were converted to binary gene matrices by substituting counts with 1 if counts > 0.

Feature selection: unless otherwise specified, we select genes for label transfer by taking the union of the most highly variable genes in scRNA-seq (using the Seurat function `FindVariableGenes`, with `selection.method = "mvp"`) and the most frequently covered genes in the scATAC-seq datasets (covered in at least 10% of cells). We found that joint feature selection performs significantly better that selection based on the reference dataset only (data not shown).

Data integration methods: details of published integration methods for single-cell data and the rationale behind exclusion from the benchmark are detailed in Table 1.

Table 1: Details of published data integration methods considered for comparison.

Method	Reference	Model for embedding	Label/feature propagation	Reason for Excluding
Seurat CCA	Stuart et al. (2019)	Canonical Correlation analysis	mNN pairing	/
LIGER	Welch et al. (2019)	Joint Non-Negative Matrix factorization	KNN graph	/
Conos	Barkas et al. (2019)	Joint PCA	Inter/Intra-dataset edges	/
scGen	Lotfollahi, Wolf, and Theis (2019)	Variational Autoencoder	Decoder	Requires cell type annotation in both datasets
totalVI	Gayoso et al. (2019)	Variational inference	Generative model	Requires multi-omic data from the same single-cells
BBKNN	Polański et al. (n.d.)	PCA	Batch balanced graph construction	Bad alignment during testing
gimVI	Lopez et al. (2019)	Variational inference	Generative model	No implementation for right log-likelihood distribution

5.2.1 KNN purity score

We construct the k nearest neighbor (KNN) graph of scATAC-seq cells on the LSI reduction matrix ($k = 30$). For each prediction, we measure the fraction of NNs per cell that have the same predicted label. We do the same after randomly permuting the predicted labels, to estimate a null distribution. The purpose of this step is to avoid giving a high score to a prediction that assigns many cells to just a few clusters. We then compute the Kolmogorov-Smirnov deviation statistic between true and null distribution and define this as KNN purity. KNN purity score is calculated retaining only labels with prediction score > 0.5.

5.2.2 Analysis of T cell maturation

We selected putative T cell clusters based on independent clustering of scATAC-seq and scRNA-seq. We performed co-embedding and label trasfer using the Seurat CCA method (Stuart et al. 2019), with preprocessing as previously described. We imputed gene expression values for scATAC-seq cells and computed diffusion pseudotime (Haghverdi et al. 2016) on all cells as implemented in Scanpy (v1.4.4) (Wolf, Angerer,

and Theis 2018). For analysis of TF motif enrichment, we used the ChromVAR package (Schep et al. 2017), with default settings.

5.2.3 Code availability

All analysis for this report was carried out in R and python. Notebooks and scripts can be accessed at https://github.com/EmmaDann/multiOmic_benchmark.

6 Bibliography

Barkas, Nikolas, Viktor Petukhov, Daria Nikolaeva, Yaroslav Lozinsky, Samuel Demharter, Konstantin Khodosevich, and Peter V. Kharchenko. 2019. "Joint Analysis of Heterogeneous Single-Cell RNA-Seq Dataset Collections." *Nature Methods* 16 (8): 695–98. https://doi.org/10.1038/s41592-019-0466-z.

Buenrostro, Jason D., M. Ryan Corces, Caleb A. Lareau, Beijing Wu, Alicia N. Schep, Martin J. Aryee, Ravindra Majeti, Howard Y. Chang, and William J. Greenleaf. 2018. "Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation." *Cell* 173 (6): 1535–1548.e16. https://doi.org/10.1016/j.cell.2018.03.074.

Cusanovich, Darren A., Riza Daza, Andrew Adey, Hannah A. Pliner, Lena Christiansen, Kevin L. Gunderson, Frank J. Steemers, Cole Trapnell, and Jay Shendure. 2015. "Multiplex Single-Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing." *Science* 348 (6237): 910–14. https://doi.org/10.1126/science.aab1601.

Cusanovich, Darren A., Andrew J. Hill, Delasa Aghamirzaie, Riza M. Daza, Hannah A. Pliner, Joel B. Berletch, Galina N. Filippova, et al. 2018. "A Single-Cell Atlas of in Vivo Mammalian Chromatin Accessibility." *Cell* 174 (5): 1309–1324.e18. https://doi.org/10.1016/j.cell.2018.06.052.

Fang, Rongxin, Sebastian Preissl, Xiaomeng Hou, Jacinta Lucero, Xinxin Wang, Amir Motamedi, Andrew K. Shiau, and Eran A. Mukamel. 2019. "Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-Regulatory Elements in Rare Cell Types," 41.

Gayoso, Adam, Romain Lopez, Zoë Steier, Jeffrey Regier, Aaron Streets, and Nir Yosef. 2019. "A Joint Model of RNA Expression and Surface Protein Abundance in Single Cells." *bioRxiv*, October, 791947. https://doi.org/10.1101/791947.

Granja, Jeffrey M., Sandy Klemm, Lisa M. McGinnis, Arwa S. Kathiria, Anja Mezger, M. Ryan Corces, Benjamin Parks, et al. 2019. "Single-Cell Multiomic Analysis Identifies Regulatory Programs in Mixed-Phenotype Acute Leukemia." *Nature Biotechnology*, December, 1–8. https://doi.org/10.1038/s41587-019-0332-7.

Haghverdi, Laleh, Maren Büttner, F. Alexander Wolf, Florian Buettner, and Fabian J. Theis. 2016. "Diffusion Pseudotime Robustly Reconstructs Lineage Branching." *Nature Methods* 13 (10): 845–48. https://doi.org/10.1038/nmeth.3971.

Hosoya, Tomonori, Ricardo D'Oliveira Albanus, John Hensley, Greggory Myers, Yasuhiro Kyono, Jacob Kitzman, Stephen C. J. Parker, and James Douglas Engel. 2018. "Global Dynamics of Stage-Specific Transcription Factor Binding During Thymocyte Development." *Scientific Reports* 8 (1): 5605. https://doi.org/10.1038/s41598-018-23774-9.

Lopez, Romain, Achille Nazaret, Maxime Langevin, Jules Samaran, Jeffrey Regier, Michael I. Jordan, and Nir Yosef. 2019. "A Joint Model of Unpaired Data from scRNA-Seq and Spatial Transcriptomics for Imputing Missing Gene Expression Measurements." *arXiv:1905.02269 [Cs, Q-Bio, Stat]*, May. http://arxiv.org/abs/1905.02269.

Lotfollahi, Mohammad, F. Alexander Wolf, and Fabian J. Theis. 2019. "scGen Predicts Single-Cell Perturbation Responses." *Nature Methods* 16 (8): 715. https://doi.org/10.1038/s41592-019-0494-8.

McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv:1802.03426 [Cs, Stat]*, December. http://arxiv.org/abs/1802.03426.

Polański, Krzysztof, Matthew D. Young, Zhichao Miao, Kerstin B. Meyer, Sarah A. Teichmann, and Jong-Eun Park. n.d. "BBKNN: Fast Batch Alignment of Single Cell Transcriptomes." *Bioinformatics*. Accessed October 3, 2019. https://doi.org/10.1093/bioinformatics/btz625.

Satija, Rahul, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. 2015. "Spatial Reconstruction of Single-Cell Gene Expression Data." *Nature Biotechnology* 33 (5): 495–502. https://doi.org/10.1038/nbt.3192.

Schep, Alicia N., Beijing Wu, Jason D. Buenrostro, and William J. Greenleaf. 2017. "chromVAR: Inferring Transcription-Factor-Associated Accessibility from Single-Cell Epigenomic Data." *Nature Methods* 14 (10): 975–78. https://doi.org/10.1038/nmeth.4401.

Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. "Comprehensive Integration of Single-Cell Data." *Cell* 177 (7): 1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.

Welch, Joshua, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z. Macosko. 2019. "Single-Cell Multi-Omic Integration Compares and Contrasts Features of Brain Cell Identity." *Cell*.

Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis. 2018. "SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis." *Genome Biology* 19 (1): 15. https://doi.org/10.1186/s13059-017-1382-0.