Optimizing integration of scRNA-seq and scATAC-seq datasets

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). Different methods for integration of multi-modal single-cell datasets have been recently proposed (Barkas et al. 2019; Lopez et al. 2019; Stuart et al. 2019; Welch et al. 2019). These require the scATAC-seq data to be reduced to represent accessibility of genes, usually by simply counting the number of accessible regions overlapping with the promoter and gene body (Stuart et al. 2019; Welch et al. 2019). Then, assuming positive correlation between gene accessibility and expression, a common latent space projection is inferred, to find similarities between cells in different data modalities.

Currently there is no consensus on best practices for integration of scATAC-seq and scRNA-seq, regarding methods or preprocessing steps such as the featurization to genes.

Here we implemented an analysis workflow to optimize label transfer from scRNA-seq datasets to scATAC-seq datasets. We defined metrics to compare performance of three published integration methods (Seurat CCA anchoring, hereafter CCA (Stuart et al. 2019), Conos (Barkas et al. 2019) and Liger (Welch et al. 2019)). We then applied our workflow to optimize integration of scRNA-seq and scATAC-seq datasets generated from developing human thymus.

2 Results

We started by comparing the performance methods for label transfer on a publicly available dataset of Peripheral Blood Mononuclear Cells (PBMC) from 10XGenomics.

We first run label prediction including a growing number of scATAC-seq cells to evaluate method run time and robustness. 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 (Fig.1C). Conversely, Liger showed much higher variability.
A crucial step to reconstruct consensus chromatin profiles of cell types is to assess whether cells of the same predicted cell type are also similar in genome-wide accessibility. Visualizing the predicted labels on the embedding of genome-wide scATAC-seq profiles suggests that all methods reconstruct the expected clustering structure (Fig.1A). To quantify this agreement we formulate a KNN purity statistic that measures the fraction of neighbors that share the same label against a permutation based null (see Methods 4.3). We found that Conos and Liger reconstructed slightly purer clusters in this dataset (Fig.1G).

While all methods predict similar labels in steady-state cell populations, we reasoned that datasets capturing differentiating cell states, where the chromatin landscape undergoes extensive remodeling, might more challenging to integrate. We analyzed scRNA-seq and scATAC-seq datasets generated from one fetal thymus at 10 weeks post-conception. We identified 15 cell populations by clustering the transcriptomes, including components of the thymic microenvironment and a main cluster representing T cells undergoing maturation, from the Double Negative (DN) state, to Double Positive (DP) to CD4+ or CD8+ Single Positive (SP) T

cells (Fig.2A). Also when visualizing the scATAC-seq cells we find a large central cluster and many smaller cell populations (Fig.2B). Repeating the label transfer workflow as for the PBMC dataset we noticed that, while all methods consistently assigned the smaller clusters, only with CCA transferred annotations for the T cells resembled chromatin accessibility clusters, as also quantified with the KNN purity score (Fig.2C). We observed that the transformation from genome-wide to gene accessibility significantly increases the noise in the T cell cluster (Fig.2D). We reasoned that summing up sparse accessibility profiles over gene bodies might lead to artifactual inflation of differences between similar cells. We found that binarizing the gene accessibility profiles allows to maintain the clustering structure found in genome-wide accessibility (Fig.2F) and that using the binary gene matrix for label transfer greatly increased the KNN purity for all methods (Fig.2E).

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 of scATAC-seq and scRNA-seq cells focusing on the putative T cell populations in the thymic datasets (Fig.3A-B). We modeled the differentiation trajectory of T cells using pseudotime analysis (Fig.3C). This ordered cell populations in both datasets 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 expected for cells moving towards a "primed" state (Gaspar-Maia et al. 2011). 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 the top variable 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. 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

We have developed a framework to quantitatively compare performance of different methods for label transfer. We believe that this work could lay the ground for a more systematic comparison of integration methods across datasets and modalities. Ideally this could assess integration on recently published multi-omic datasets generated from joint profiling of scRNA-seq and scATAC-seq in the same cells (Chen, Lake, and Zhang 2019), as a ground-truth for correlations between modalities.

We show that the strategy used for aggregation of accessibility profile at the gene level can significantly impact the performance of integration methods. Counting accessible regions that overlap the gene body might be a good enough solution for datasets with strongly distinct accessibility patterns between cell populations. More refined models have been developed to estimate accessibility of sites linked to specific genes, also accounting for action of distal regulatory elements (Pliner et al. 2018). Our framework to compare integration performance using similar models for featurization to genes.

We have demonstrated how optimized integration can be used to align multi-omic datasets along a common differentiation trajectory, such as for T cell maturation. This allows to relate patterns of accessibility to patterns of gene expression along maturation. Interestingly, we observed instances of TFs that showed coordinated, but anti-correlated, changes between expression and motif accessibility. A recent study showed that this can indicate a repressive function of the TF on regulatory elements (Berest et al. 2019).

4 Methods

4.0.1 Datasets

PMBC: 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.

4.1 Data preprocessing and dimensionality reduction

We preprocessed and normalized scRNA-seq using the standard pipeline from the R package Seurat (v3.1.1). We clustered cells using the leiden algorithm (Traag, Waltman, and Eck 2019) and annotated populations based on expression of marker genes from the literature. 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.

4.2 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 integration using 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	/

Method	Reference	Model for embedding	Label/feature propagation	Reason for Excluding
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

4.3 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. 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.

4.4 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 transfer using the Seurat CCA method (Stuart et al. 2019), with pre-processing 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). The cell of origin for the pseudotime algorithm was selected by expression of markers of early DN cells (IGLL1, CD34). (Wolf, Angerer, and Theis 2018). For analysis of TF motif enrichment, we used the ChromVAR package (Schep et al. 2017), with default settings.

4.5 Code availability

All code for this analysis can be accessed at https://github.com/EmmaDann/multiOmic_benchmark.

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.

Berest, Ivan, Christian Arnold, Armando Reyes-Palomares, Giovanni Palla, Kasper Dindler Rasmussen, Holly Giles, Peter-Martin Bruch, et al. 2019. "Quantification of Differential Transcription Factor Activity and Multiomics-Based Classification into Activators and Repressors: diffTF." *Cell Reports* 29 (10): 3147–3159.e12. https://doi.org/10.1016/j.celrep.2019.10.106.

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.

Chen, Song, Blue B. Lake, and Kun Zhang. 2019. "High-Throughput Sequencing of the Transcriptome and Chromatin Accessibility in the Same Cell." *Nature Biotechnology*, October, 1–6. https://doi.org/10.1038/s41587-019-0290-0.

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.

Gaspar-Maia, Alexandre, Adi Alajem, Eran Meshorer, and Miguel Ramalho-Santos. 2011. "Open Chromatin in Pluripotency and Reprogramming." *Nature Reviews Molecular Cell Biology* 12 (1): 36–47. https://doi.org/10.1038/nrm3036.

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.

Pliner, Hannah A., Jonathan S. Packer, José L. McFaline-Figueroa, Darren A. Cusanovich, Riza M. Daza, Delasa Aghamirzaie, Sanjay Srivatsan, et al. 2018. "Cicero Predicts Cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data." *Molecular Cell* 71 (5): 858–871.e8. https://doi.org/10.1016/j.molcel.2018.06.044.

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.

Traag, V. A., L. Waltman, and N. J. van Eck. 2019. "From Louvain to Leiden: Guaranteeing Well-Connected Communities." *Scientific Reports* 9 (1): 1–12. https://doi.org/10.1038/s41598-019-41695-z.

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.

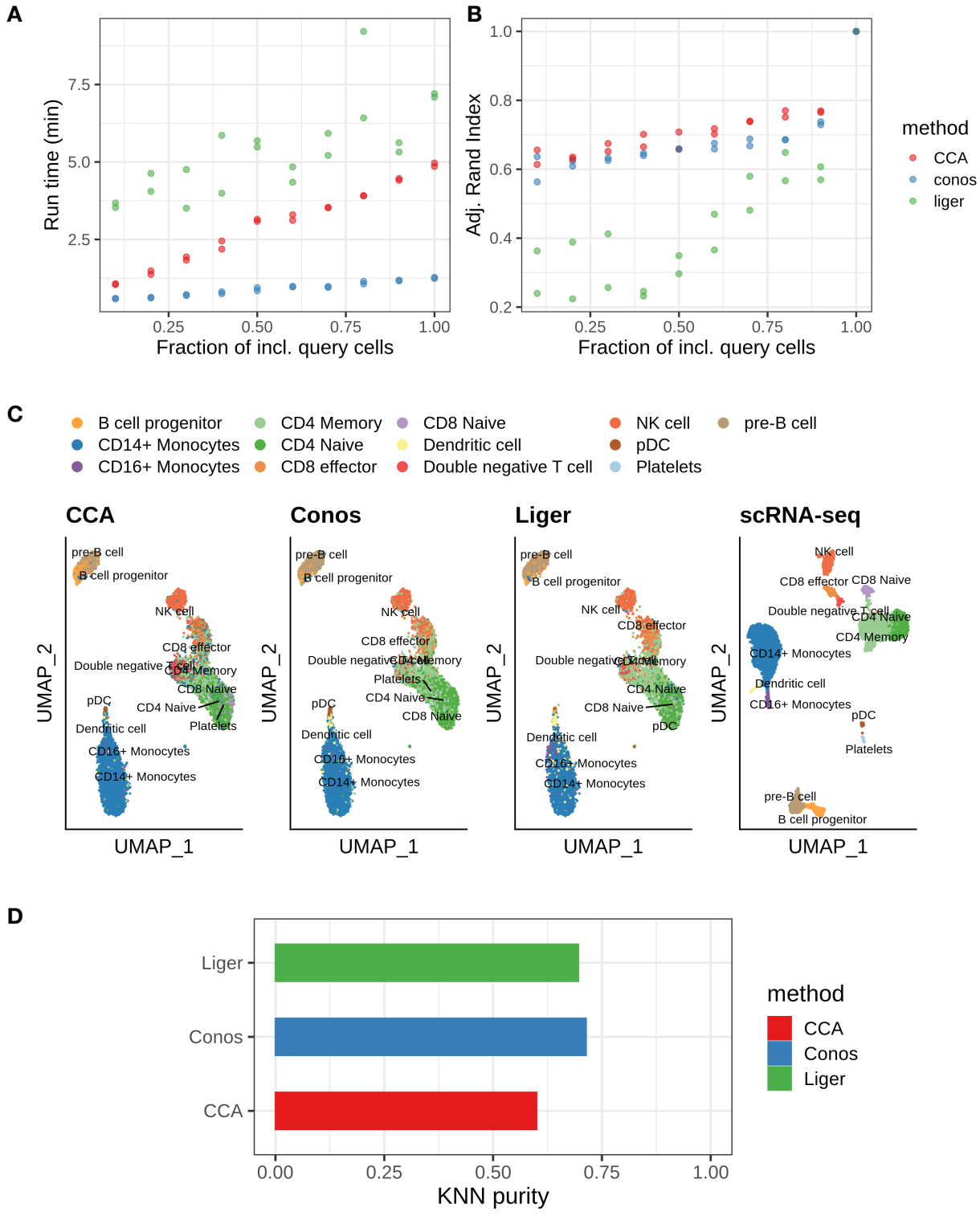

Figure 1: **Comparison of integration methods performance on PBMC dataset:** (A-B) Benchmark for integration with increasing size of scATAC-seq dataset (query cells), each point represents a label transfer run, color indicate the used method (2 runs per method); (A) comparison of run time; (B) similarity of predicted ID, compared to integration with full scATAC-seq dataset, measured by adjusted rand index; (C) UMAP visualization of scATAC-seq cells (as in B), colored by label transfer outcomes integrating on gene accessibility counts; right: quantification of KNN purity

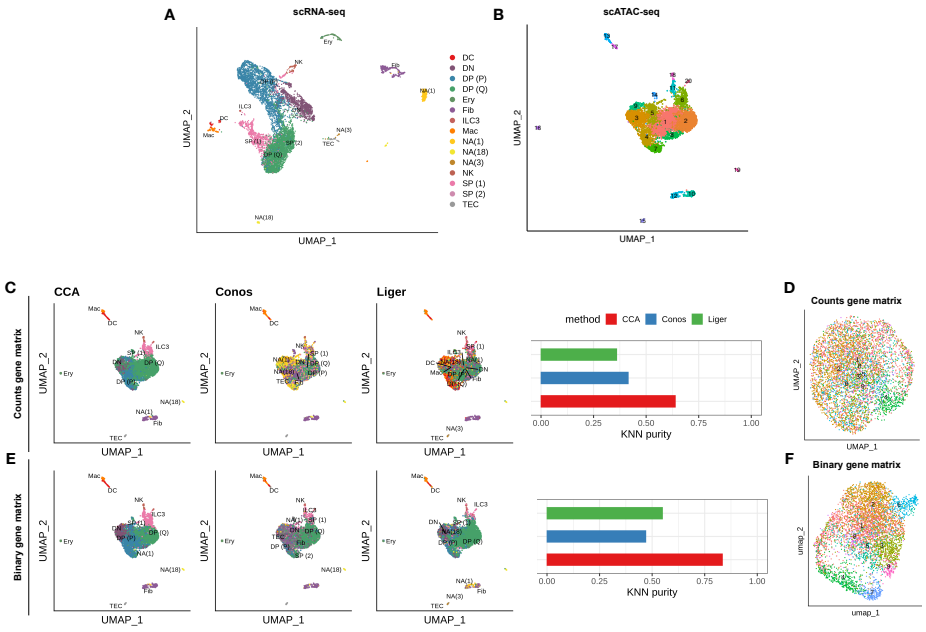

Figure 2: **Optimization of label transfer on developing thymus dataset:** (A) UMAP visualization of scRNA-seq cells in the thymus colored by cell type (DC: dendritic cells, DN: double negative T cells, DP (P): Proliferative double positive T cells, DP (Q): quiescent double positive T cells, Ery: eruthrocytes, Fib: fibroblasts, ILC3: innate lymphoid cell types, Mac: macrophages, NK: natural killer cells, SP: single positive T cells, TEC: thymic epithelial cells); (B) UMAP visualization of scATAC-seq cells in the thymus (genome-wide accessibility), colored by clusters identified using the leiden algorithm; (C) left: UMAP visualization of scATAC-seq cells (as in B), colored by label transfer outcomes; (D) quantification of KNN purity

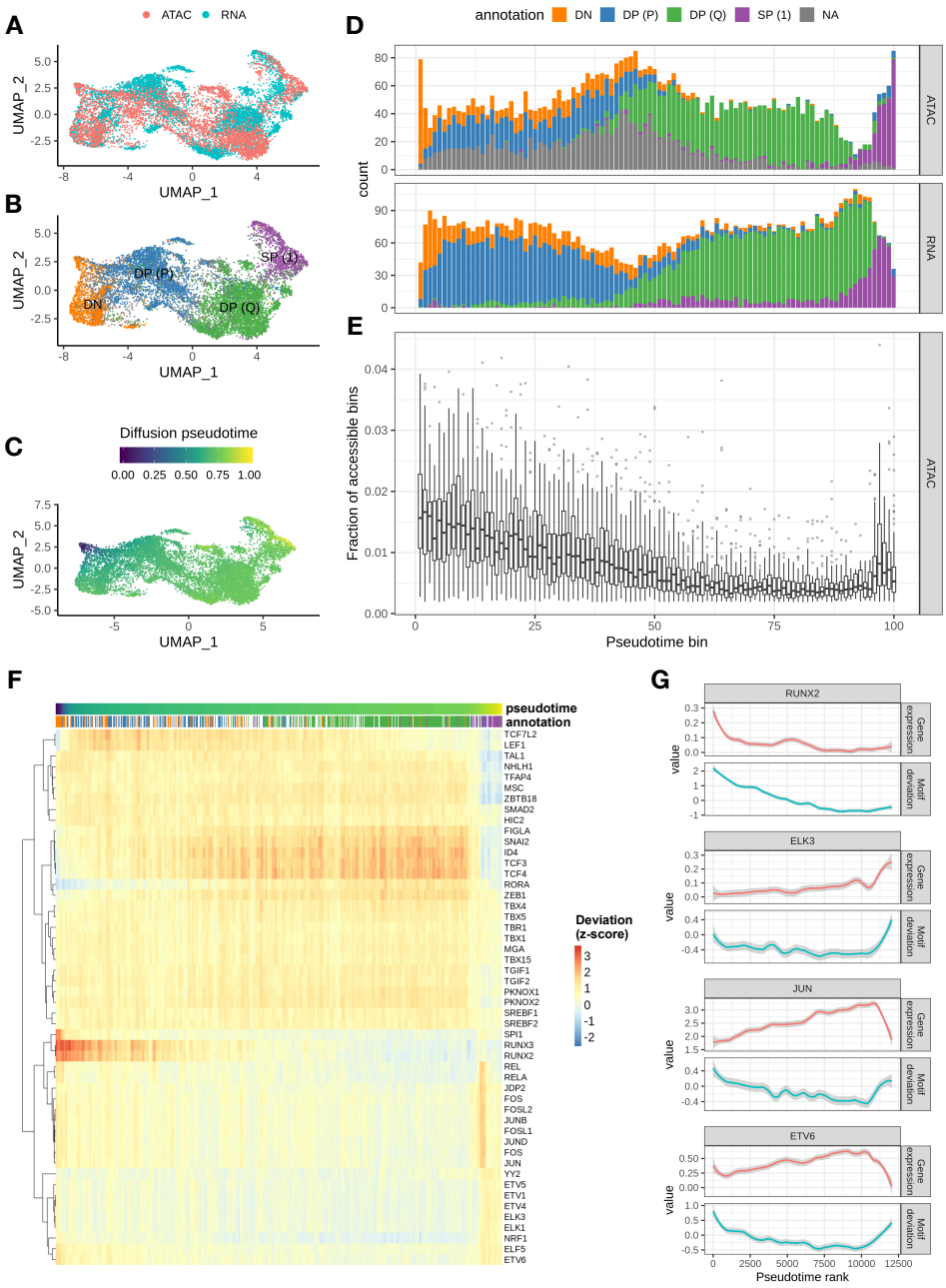

Figure 3: **Integrative analysis of T cell maturation in developing thymus:** Joint visualization of scRNA-seq and scATAC-seq cells for the T cell clusters, colored by (A) technology, (B) inferred cell type labels and (C) diffusion pseudotime; (D) Distribution of cells along 100 equal-sized pseudotime 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).