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

Contents

1 Introduction

Following technological advances and discoveries with single-cell transcriptomics, new single-cell sequencing strategies are emerging to profile other molecular layers in thousands of cells, from chromatin accessibility, to expression of surface proteins, to methylation. These methods have already been employed to discover patterns of epigenetic heterogeneity in a variety of tissues [ref]. However, at present these methods produce noisier data and at lower throughput. Integration of such datasets with more detailed and comprehensively annotated scRNA-seq datasets can allow denoising of biological signals, guide cell type annotation and disentangle causal relationships between biological layers of information and how these co-determine complex phenotypes.

Methods for simoultaneous profiling of biological layers are starting to emerge [refs], but these are mostly low-throughput and labor intensive. 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 bechmark for label transfer methods: PBMC dataset

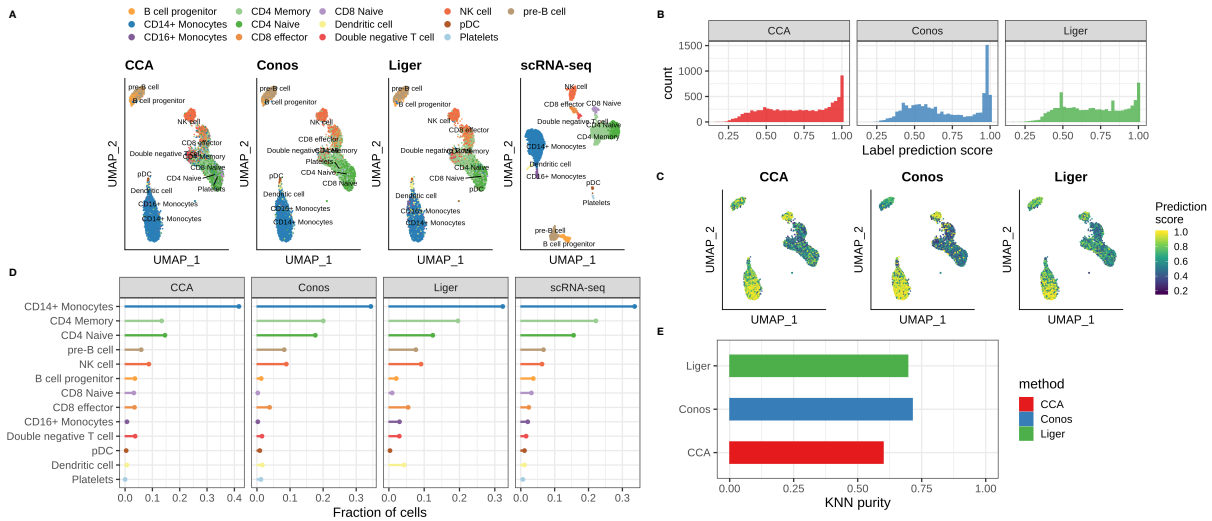
We run our initial benchmark on a publicly available dataset of Peripheral Blood Mononuclear Cells (PBMC) downloaded from the 10XGenomics website.

We first evaluate the ability of different methods to transfer cell type annotations derived from scRNA-seq data to scATAC-seq data of the same tissue (See Methods 5.5.1 for details about label transfer procedure for each method).

Visualizing the predicted label assignments on the embedding of ATAC cells based on the bin x cells matrix, we find that for all the methods the label assignment have some coincidence with the clusters from ATAC data alone.

All integration methods measure the uncertainty of their assignment (Fig.**??**A). Setting a threshold on the label prediction score allows to remove from downstream analysis cells with a low confidence prediction label, and mark them as unasssigned. We found that at the cutoff of 0.5, suggested by (**???**), Liger excludes the least amount of cells, while Conos scores the most cells with higher confidence (Fig.**??**B).

Taking a closer look at predicted label composition and confidence score for each cell type, we found that the different methods called similar fractions of cells for the same labels. On average Liger score with lower confidence smaller cell populations, such as pDCs, platelets, CD8+ naive cells.



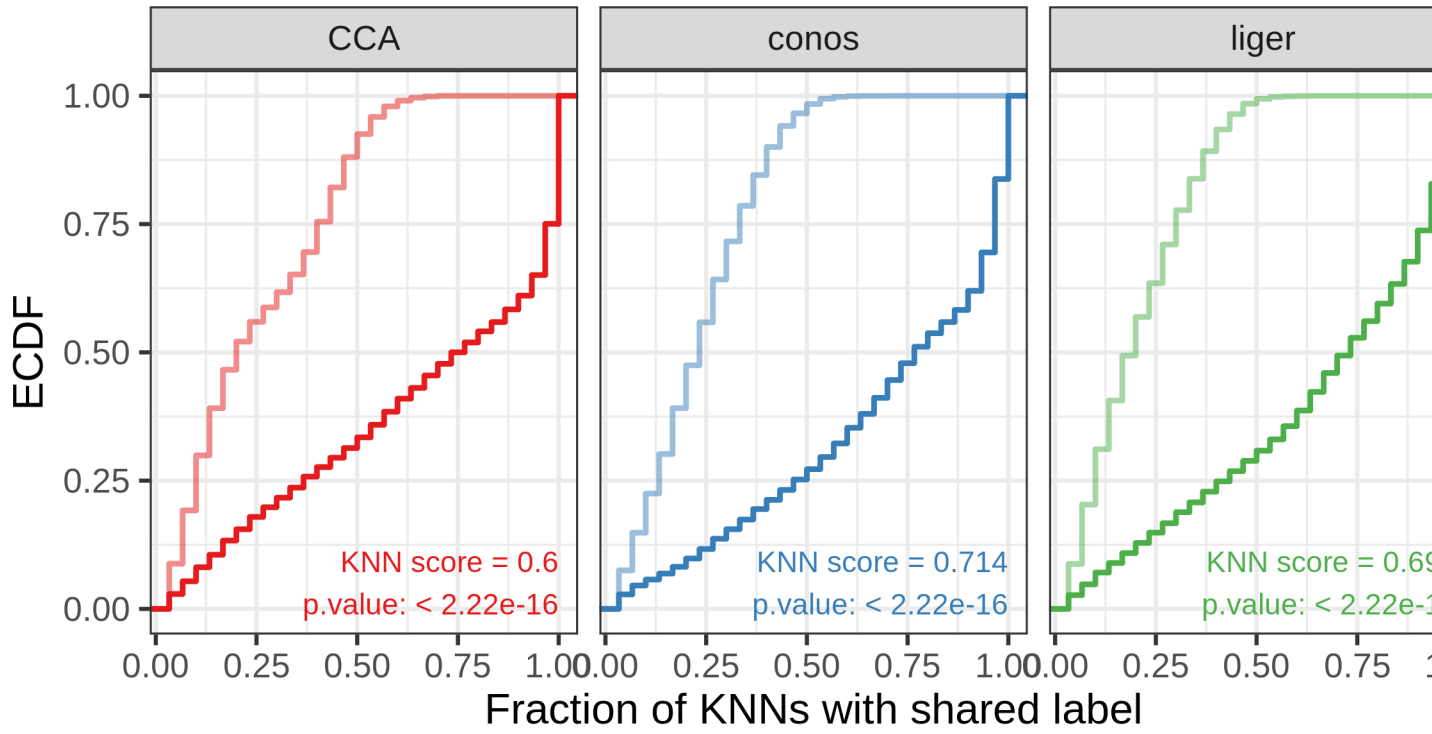


Figure 1: Cumulative distributions of KNN scores

We wanted to evaluate if the predicted annotations maintain the structure found when clustering cells based on bin accessibility alone, i.e. if cells that have a similar bin accessibility profile also get annotated with the same cell type. For each cell we measure the fraction of the K nearest neighbors that share the same predicted label (KNN score). Looking at the distributions of KNN scores, we find that Liger maintains the original connectivity structure better, followed by Conos and finally CCA, even if the differences are not that substantial. Also the KNN score is dependent on the assigned cell type.

We checked accessibility of PBMC marker genes in the called cell types. We found that cell populations called with CCA show accessibility of expected marker genes from transcriptomics, for NK cells, monocytes, CD8+ cells, and the platelets even.

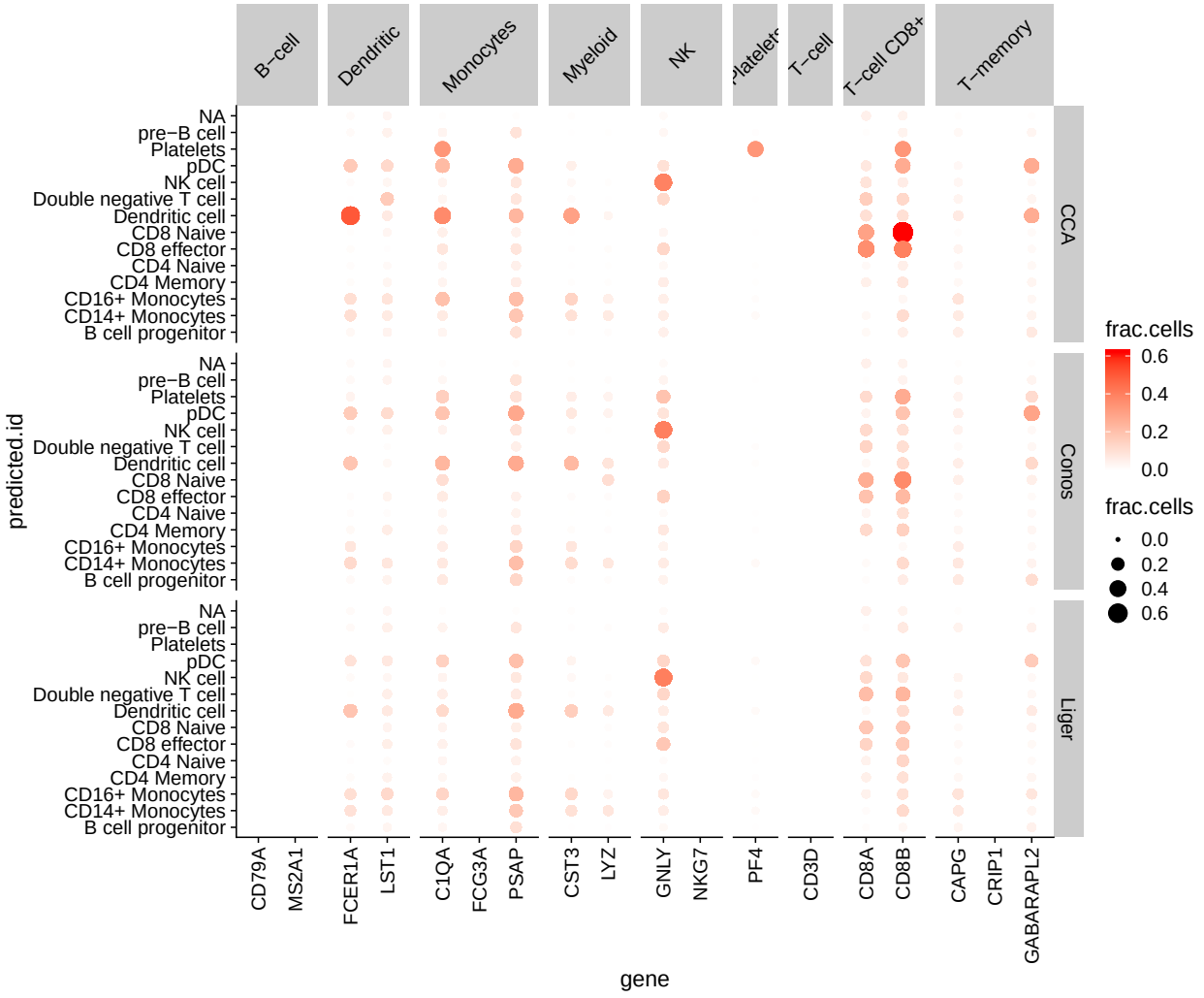


Figure 2: Accessibility of markers in predicted cell clusters

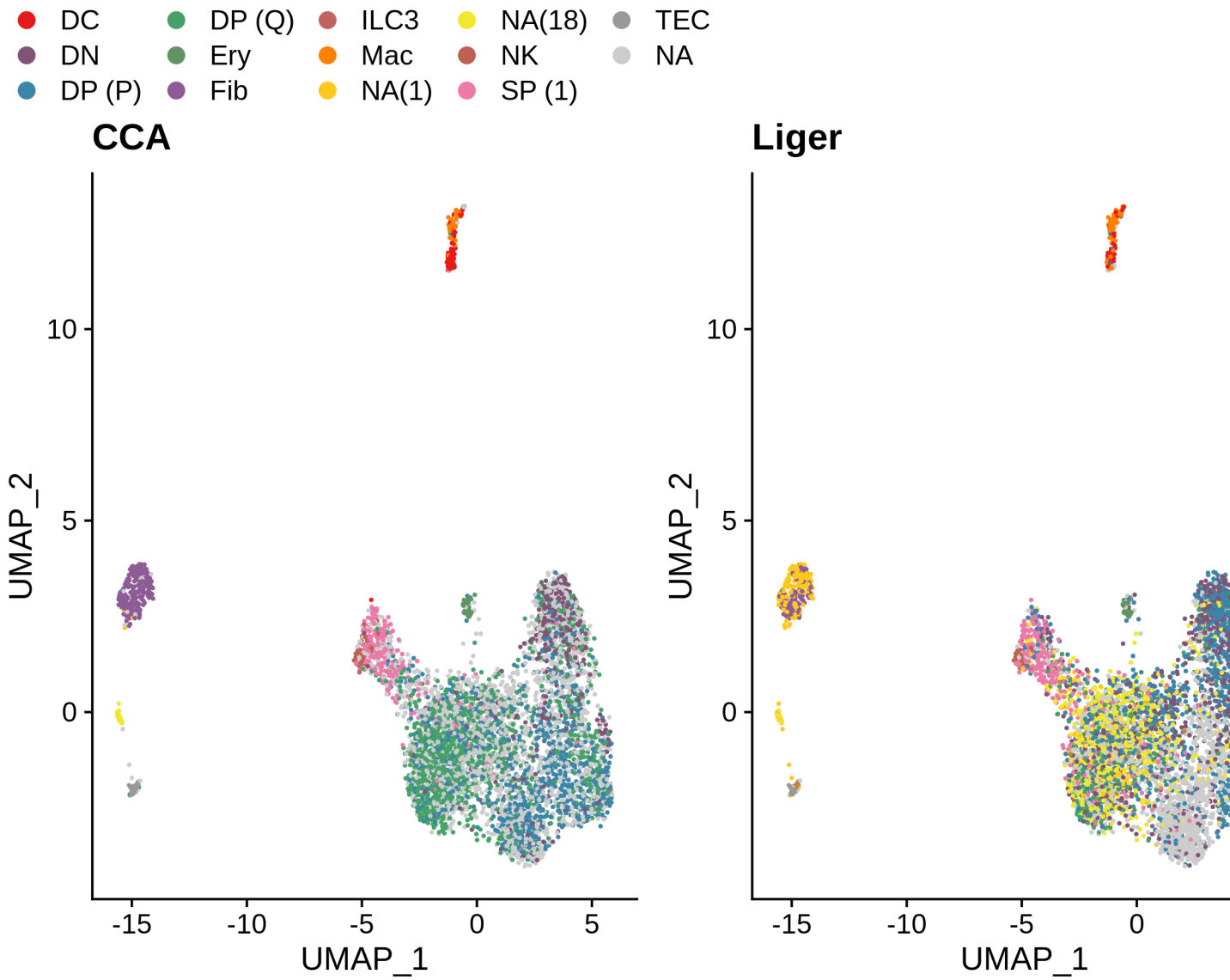


Figure 3: UMAP of predicted labels for thymus dataset

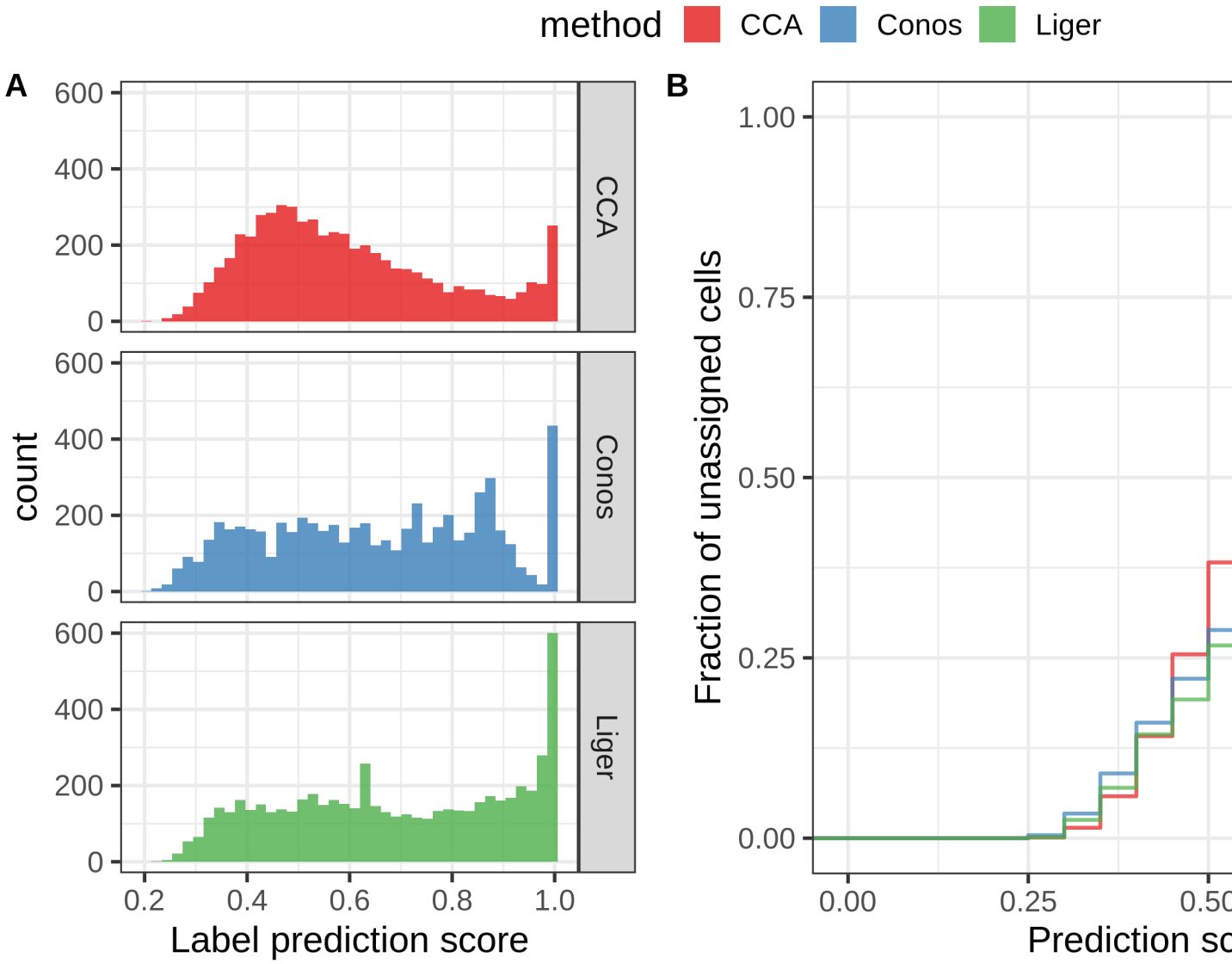
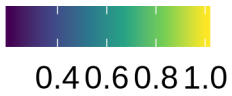
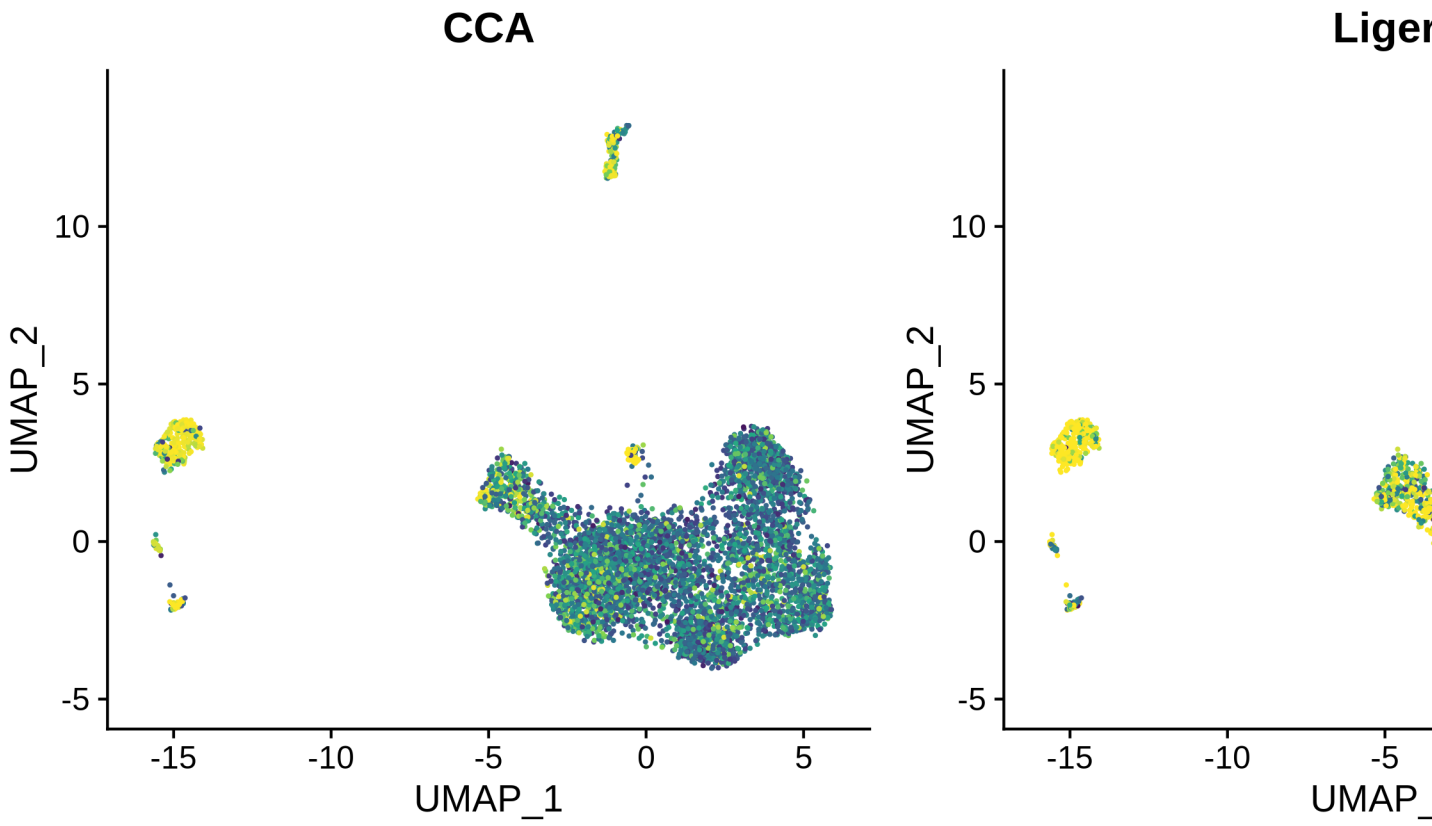


Figure 4: Prediction scores for thymus dataset

2.2 Optimizing label transfer on Thymus dataset

2.2.1 Label transfer on count gmat





Cell type composition

CCA is also better at maintaining the nearest neighboorhood structure from the ATAC alone in its label assignments

Trying to reproduce accessibility profiles of ordered T-cell populations (see Fig. 2H of JP's manuscript).

- Good: TOX2, ST18, SATB1 on the entry cells, cyclin D genes,

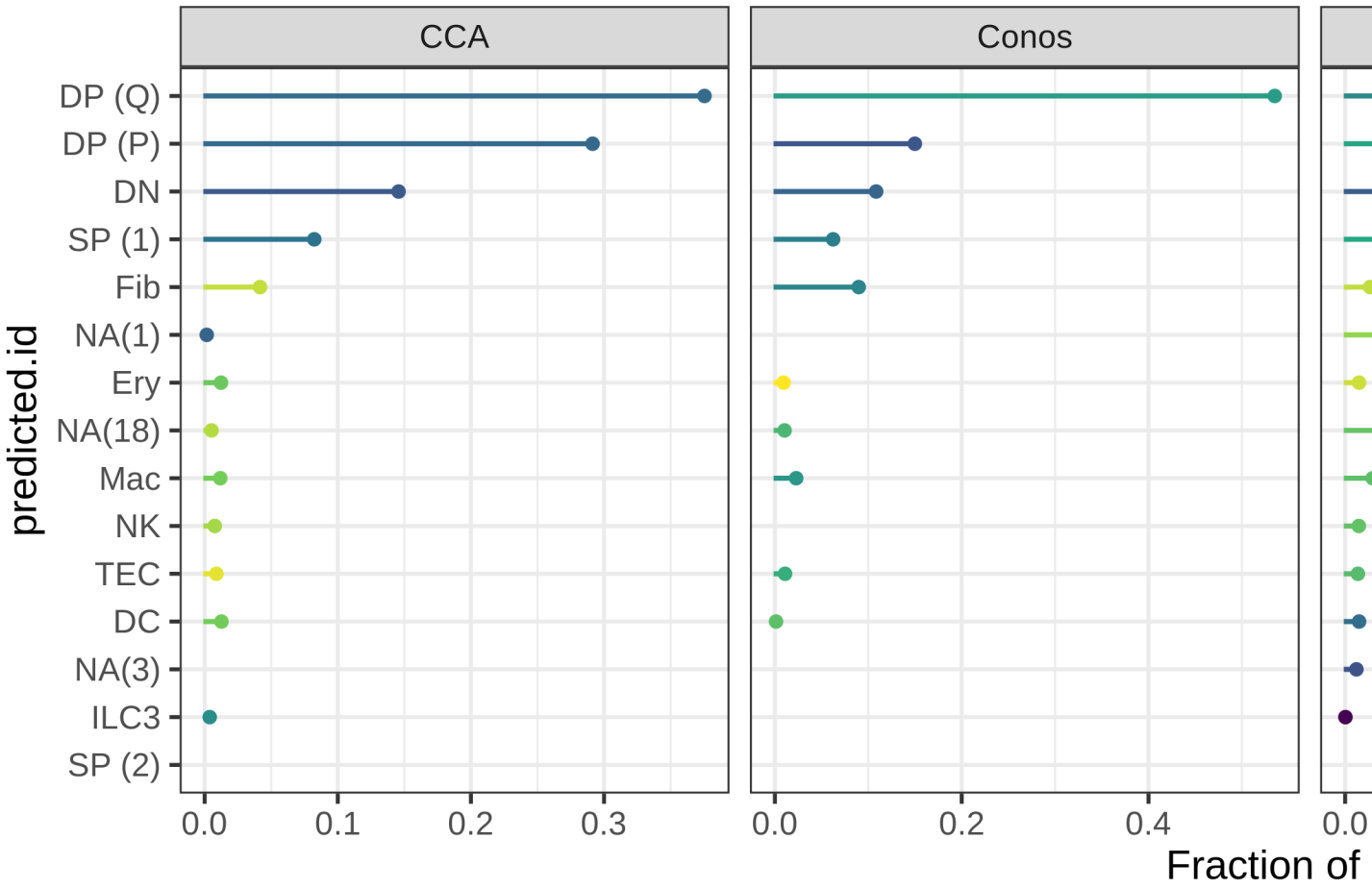


Figure 5: Predicted cell type compositions for thymus dataset

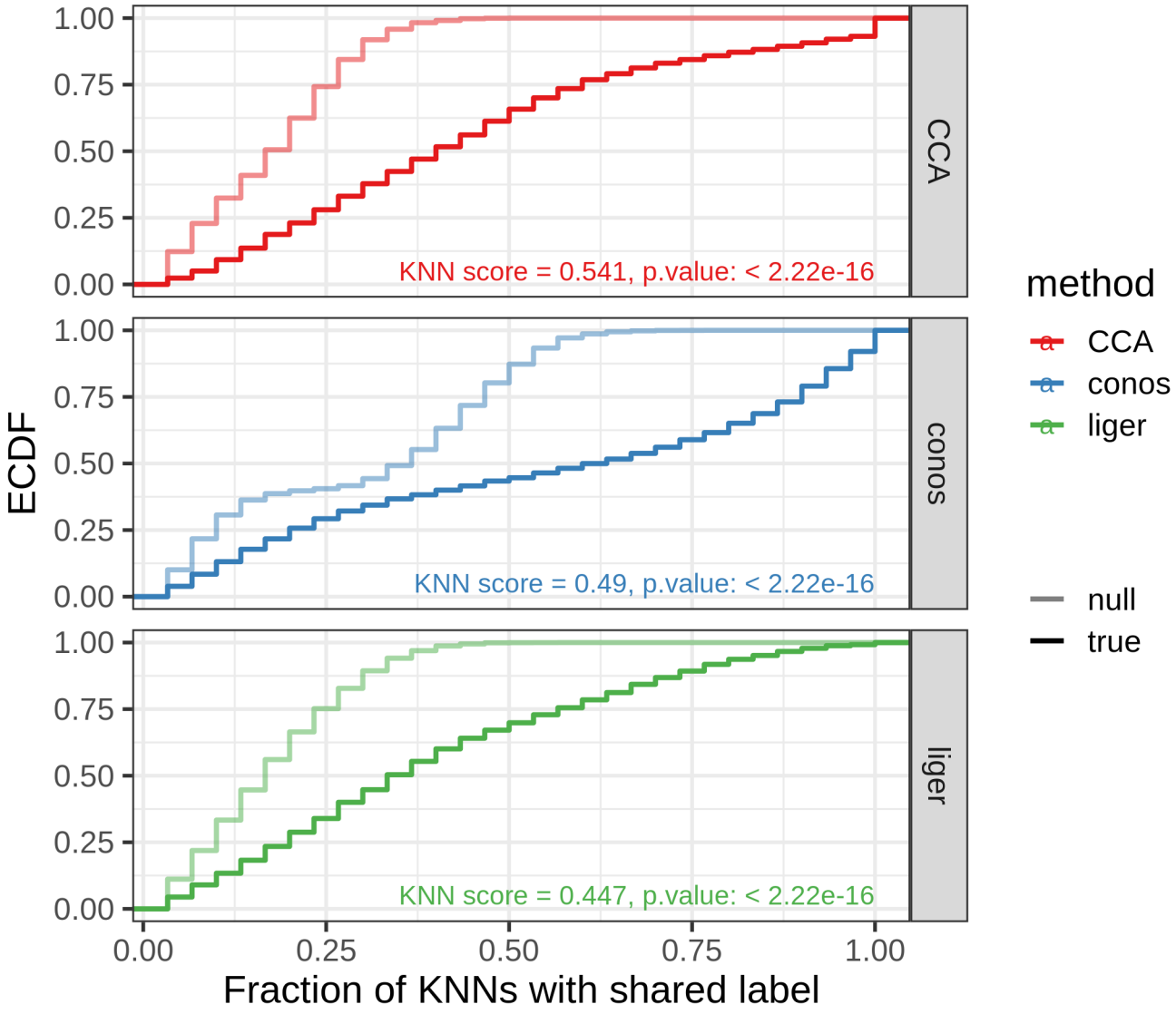


Figure 6: KNN agreement score for thymus dataset

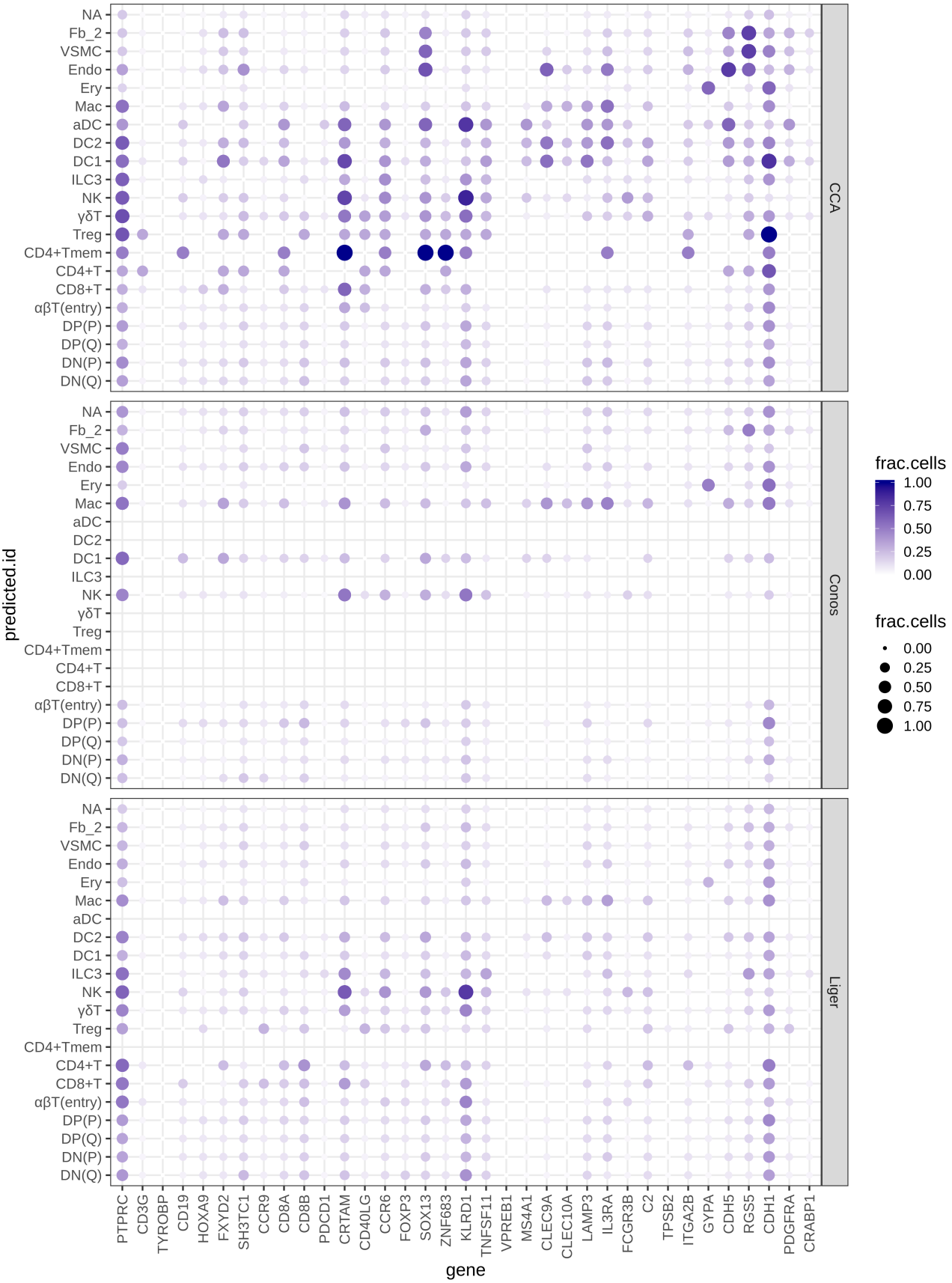


Figure 7: Accessibility of known thymic markers in predicted cell types

- Weird: CD8 accessibility for CD4 cells (possibly wrongly assigned double positives?)

2.3 What next?

- Evaluate label transfer for different cluster sizes
- Label transfer w bigger clusters provided by Cecilia
- MATCHER comparison: CCA seems able to reconstruct the main differences between differentiating T-cells. Will MATCHER find the same relationships between cells?
- Integration w imputation instead of label transfer
- Performance comparison on matched joint profilinf of ATAC - RNA on the same cells

3 Discussion

- 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

5.1 Dataset details

Incl. stats about dataset quality (median no. of fragments per cell, perc. of fragments mapping to peaks, median no. of fragments in peaks per cell)

5.1.1 PBMC dataset

I used datasets of Peripheral Blood Mononuclear Cells (PBMCs) provided by 10X genomics. Raw scRNA-seq counts were extracted from the R object provided by the Seurat package [cit] (downloaded here). Raw scATAC-seq fragments were converted to a `snap` file using `snaptools` (downloaded from here).

5.1.2 Thymus dataset

I used a dataset of developing thymus at n weeks post conception. scATAC-seq and scRNA-seq profiles were measured from tissue samples of the same donor.

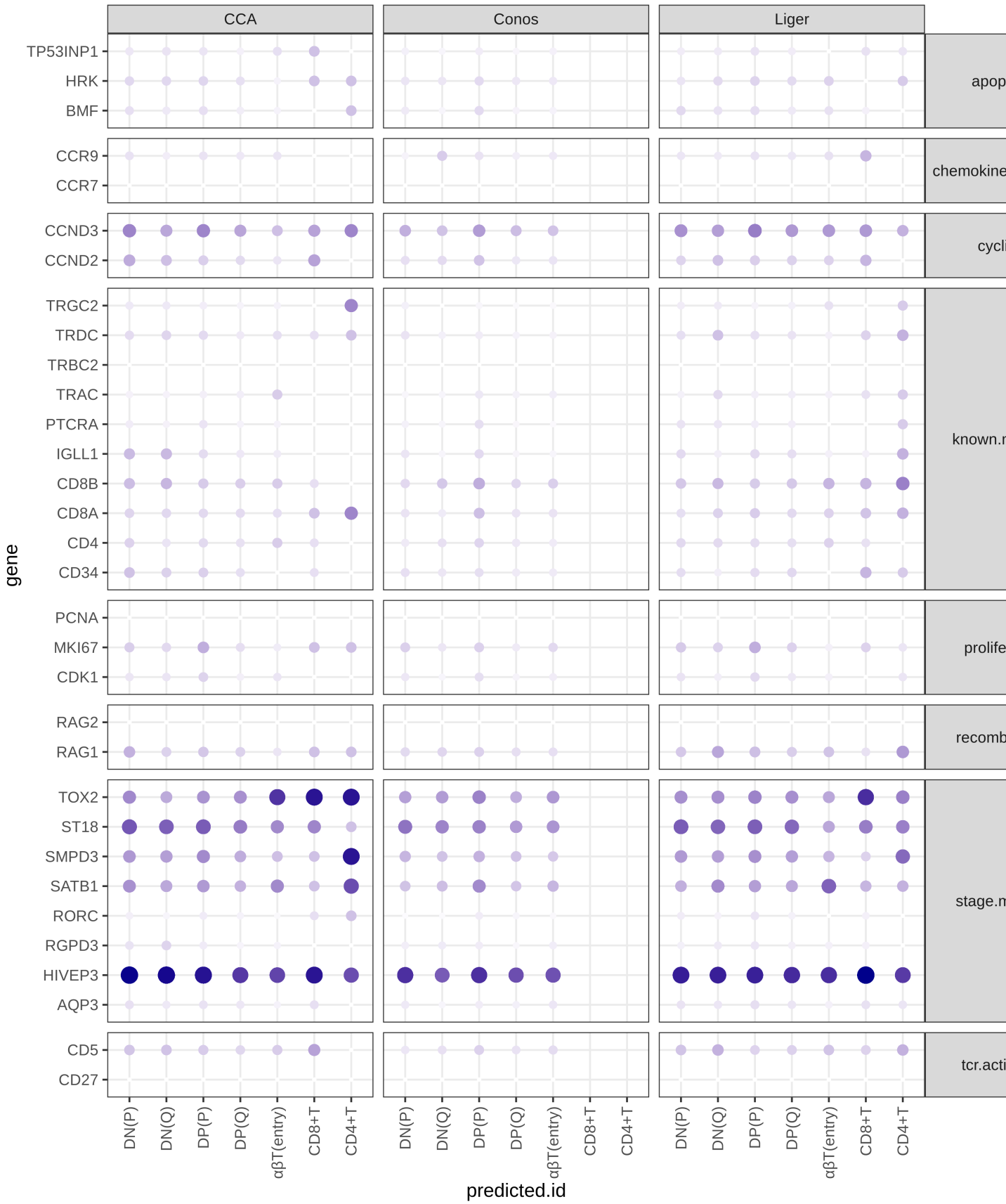


Figure 8: Accessibility of known T-cells markers in predicted cell types

5.2 Data preprocessing

ATAC-seq:

ScATAC-seq reads were aligned and preprocessed using CellRanger (10X genomics). For preprocessing and quality control of scATAC-seq data we used the SnapATAC pipeline (**???**). Briefly, we selected high-quality cell barcodes based on the number of unique fragments and the ratio of fragments in gene promoters. Then, we binned the genome into fixed-size windows (selected bin size: 5 kb) and generated a bin x cell binary matrix estimating coverage for each bin. We filtered out bins that overlap ENCODE-defined blacklist regions, as well as bins with excidingly high or low z-scored coverage.

We generated cell-by-gene count matrices by aggregating bin coverage over the gene bodies and promoters (2kb upstream of transcriptional start site). We then generated a second cell-by-gene matrix binarizing the count matrix.

RNA-seq: cells with high expression of mitochondrial genes were filtered out. Raw counts c were normalized by total cell coverage and converted to $log(c + 1)$ as a variance stabilizing transformation.

Feature selection For label transfer, we select informative genes by taking the union of the 2000 most highly variable genes (HVGs) in scRNA-seq and scATAC-seq datasets (using the Seurat function `FindVariableGenes`). Other studies have used only the HVGs from the reference dataset (in most cases the scRNA-seq). We found that these two feature selection strategies gave similar label transfer results and we opted for the union to avoid selecting only genes that have very low coverage in the ATAC-seq data.

5.3 Dimensionality reduction and clustering

We used principal component analysis as a low dimensional representation of scRNA-seq datasets. Unless otherwise stated, we used Latent Semantic Indexing for dimensionality reduction of accessibility matrices, as proposed by Cusanovich et al. (2015). For data visualization, we used the Uniform Manifold Approximation and Projection (UMAP) algorithm (McInnes, Healy, and Melville 2018) on low-dimensional data representation.

5.4 scATAC-seq data embedding

We performed dimensionality reduction on the bin x cell accessibility matrix using Latent Semantic Indexing as implemented in Seurat. After selection of signification dimensions, we embedded cells for visualization in 2D using the UMAP algorithm.

5.5 Tested integration methods

Method	Reference	Model for embedding	Label/feature propagation	Included in benchmark	Reason for Excluding
Seurat CCA	Stuart et al. (2019)	Canonical Correlation analysis	mNN pairing	Yes	/
LIGER	Welch et al. (2019)	Joint Non-Negative Matrix factorization	KNN graph	Yes	/
Conos	Barkas et al. (2019)	Joint PCA	Inter/Intra-dataset edges	Yes	/
scGen	Lotfollahi, Wolf, and Theis (2019)	Variational Autoencoder	Decoder	No	Requires cell type annotation in both datasets

Method	Reference	Model for embedding	Label/feature propagation	Included in benchmark	Reason for Excluding
totalVI	Gayoso et al. (2019)	Variational inference	Generative model	No	Requires multi-omic data from the same single-cells
BBKNN	Polański et al. (n.d.)	PCA	Batch balanced graph construction	No	Bad alignment during testing
Cusanovich2018	Cusanovich et al. (2018)	PCA	KNN graph	No	Code unavailable
gimVI	Lopez et al. (2019)	Variational inference	Generative model	No	No implementation for right log-likelihood distribution

5.5.1 Label transfer

One of the key tasks for integration methods is to be able to transfer cell type annotations learnt from a reference to a query dataset. This is especially useful if the query is a scATAC-seq dataset, where calling of cell types based on prior knownledge on marker genes is often not possible. Different models are adapted to transfer discrete cell state labels derived from gene expression to cells measured with scATAC-seq.

5.5.1.0.1 Seurat CCA

Identified anchor pairs are weighted based on the query cell local neighboorhood (the k nearest anchors) and the anchor score. The obtained reference cells x query cells weight matrix is then multiplied by the matrix of annotation x reference cells, to generate a query cell x annotation matrix. This returns a prediction score for each class for every cell in the query dataset, ranging from 0 to 1 (**???**).

5.5.1.0.2 LIGER

While the authors do not describe a method for transferring discrete labels, I adapted their strategy for feature imputation. I build a cross-dataset KNN graph in the aligned factor space, then I assign each query cell to the most abundant label between the k nearest neighbors in the reference dataset (k=30). The prediction score for label l is computed as the fraction of nearest neighbors that have the predicted label.

$$score = \frac{count(l)}{k}$$

5.5.1.0.3 Conos

Label transfer is treated as a general problem of information propagation between vertices of the common graph (detailed in Barkas et al. (2019)). The label score is the label probability updating during the diffusion process.

5.5.2 KNN purity score

Knn purity score is calculated retaining over labels with prediction score > 0.5

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.

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.

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.

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*. https://doi.org/10.1093/bioinformatics/btz625.

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*.