Title
DEEP LEARNING MODELS FOR THE ANALYSIS OF SINGLE CELL GENOMICS

Author
Johansen, Nelson Jamse

Publication Date
2022

Peer reviewed|Thesis/dissertation

DEEP LEARNING MODELS FOR THE ANALYSIS OF SINGLE CELL GENOMICS

By

Nelson Johansen

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Gerald Quon, Chair

Fereydoun Hormozdiari

Ian Korf

Committee in Charge

2022

Abstract

Deep Learning Models for Single Cell Genomics

Nelson Johansen

Doctor of Philosophy

Graduate Department of Computer Science

University of California, Davis

Single cell transcriptomic technologies which capture high dimensional measurements of gene expression in individual cells have been exponentially scaling in the number of cells that can be sequenced and analyzed simultaneously. Capturing a snapshot of the landscape for possible gene expression measurements from a collection of cells enables researchers to observe the space of molecular variation inherent to specific biological systems, termed atlasing. A challenge to building deeply characterized atlases of complex biological systems such as the human brain is in the identification and correction of confounding factors which do not relate to the underlying biology but instead arise from technical confounders. In this dissertation I present deep learning models applied to single cell genomics which remove unwanted technical variation and contamination as well as perform novel analysis not previously possible using standard methods.

The construction of single cell genomics atlases leverages recent advances in single cell RNA sequencing technologies such as 10X and SmartSeq which can capture thousands of cells in single experiment. When the sequencing of individual cells is performed on different technologies this introduces unwanted technical variation (bias) specific to the technology and confounds attempts to merge scRNA-seq experiments into more complete atlases. To address this challenge, we developed scAlign to remove the effects of unwanted technical variation on gene expression specifically, scRNA-seq alignment

based on advances in computer vision. scAlign, an unsupervised deep learning method, performs data alignment that can incorporate partial, overlapping or a complete set of cell labels, and estimate per-cell differences in gene expression across datasets or conditions to characterize specific expression changes due to conditions such as age or disease.

With the recent surge of atlases efforts across complex tissues, conditions and species another challenge is how to integrate the deep characterizations of cell state with lower resolution assays of single cell or bulk genomics. Specifically, spatial and multi-omics assays do not collect RNA from a single cell but instead from a spot containing multiple cells or in the later contamination from the unintended collection of additional cells. We developed scProjection to join deeply sequenced atlases with lower resolution genomic assays to address the unwanted heterogeneity in mixed samples and project such samples in a way that recovers the underlying single-cell measurements. scProjection is demonstrated to accurately estimate the abundance of cell types that compose a mixed RNA sample while simultaneously identifying the gene expression measurements consistent for each cell type in the sample to identify cell type specific changes due spatial location of cells or disease state.

Acknowledgements

I would like to acknowledge the people who have been instrumental in the creation and support of the work presented here.

First and most importantly, I would like to express my gratitude to my advisor Dr. Gerald Quon for taking me on as hist first Ph.D. student at UC Davis. Gerald provided me the opportunity to work on a wide range of open challenges in computation biology. Our frequent discussions, tinkering and brainstorming have shaped and define my research and now career. I am truly grateful to have learned from such a knowledgeable advisor with whom I shared great overlap in research interests. I would also like to thank Gerald for fostering collaborations with researchers and fellow lab members who were interested in my research as well as providing to me the opportunity to present at international conferences and national institutions.

Also, I want to thank the members of the Quon lab, all of whom helped define my research and made my time as a Ph.D. student fun both in and out of the lab. I want to thank all the collaborators who shared their time with me to enable interesting and truly insightful biological applications of the methods presented in this work. Specifically, I would like to thank the Ed Lein and members of the Allen Institute for seeing promise in the methods developed during my Ph.D. research which led to some of the research I am most proud to be a part of.

I'd like to thank Lauren for being my support through the many years of my academic career. Also, I would like to recognize my parents and grandparents for setting a great example and for providing an upbringing that fostered creative thinking and an inquisitive nature that defines who I am today. Finally, I would like to thank all the unnamed friends, family and colleagues who were my support system and at times much needed distraction from academia.

Table of Contents

Chapter 1

Introduction

Advances and commercialization of the molecular biology field has led to the development of sequencing technologies which can characterize the molecular state for millions of individual cells sampled from any tissue in a biological sample. Leveraging these technologies researchers have characterized the gene expression levels across entire mouse and human bodies into deep atlases of cellular state. Specifically, the Allen Institute is leading the charge in mapping the entire mouse and human brain in terms of neuron types as well as functional properties of individual neurons. [CITE, ALLEN] Such detailed atlases of complex organs, for the first time, provides a reference for understanding the deleterious effects of diseases such as Alzheimer's which leads to changes in molecular state of cell types in the brain [CITE, ROSMAP studies, gtex] that can now be quantified to identify candidates for precision therapeutics.

The challenge to mining these cellular atlas lies in understanding and accounting for observed variation in gene expression which is due to technical factors instead of the underlying biology. Technical sources of variation include sample purity, sequencing technology, preparation protocols as well as institution performing the sequencing experiments. Confounding in analysis which does not correctly remove or account for such technical factors can lead to spurious associations between normal and diseased samples while lowering our power to detect changes in gene expression due to underlying biology. Correcting for unwanted technical variation can be difficult as the effect on expression levels may be unknown and highly non-linear.

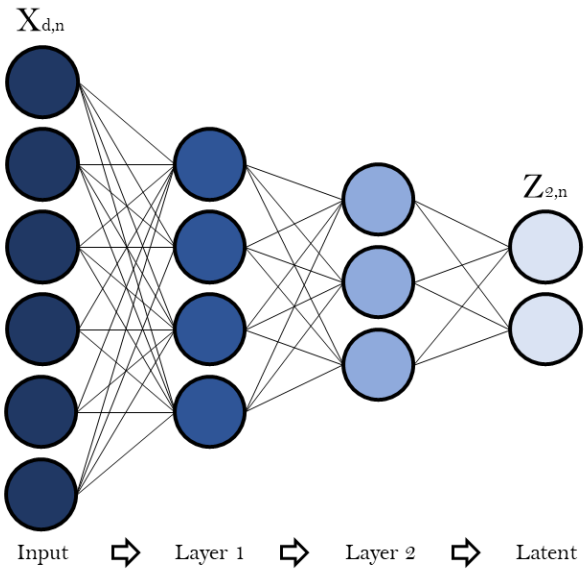
The work in this dissertation is focused on the development of machine learning models, specifically neural networks, which can address technical variation within and between studies of biological samples which otherwise should be comparable. Neural networks provide a framework for handling massive data in terms of feature and sample spaces, while also being flexible enough through modification of the core network building blocks and loss functions to address many critical challenges in the molecular biology field. Neural network models have already been proven in the fields of computer vision and natural language processing to be highly successful in removing technical variation and accurately learning on a variety of unsupervised and supervised learning tasks. In the field of computational biology, we draw inspiration from our colleagues in these standard ML fields and such inspiration is reflected in the models developed within this dissertation to address unwanted technical and biological variation.

The remainder of this chapter provides the background necessary for select topic areas in both machine learning and molecular biology which form the basis for the work presented in the following chapters of this dissertation.

3.1 Artificial neural networks

The field of machine learning has now advanced a set of powerful frameworks which are inherently interested in learning models from data. Models that learn to encode knowledge directly from complex high-dimensional data can perform tasks including prediction of unknown (latent) features, various forms of classification as well as latent (compressed) representation of the original data. Such models formalize the idea of learning from data to build machine knowledge or skill which defines the fields of machine learning and artificial intelligence.

The artificial neural network also called (feedforward) neural network or multilayer perceptrons (MLPs) has become the quintessential model and building block in machine learning. The goal of neural network models is to approximate a function f which can be arbitrary complex. For example, compression of high-dimensional feature vectors into latent representations $z - f(x;\theta)$ maps a high-dimension vector $x_{d,n}$ to a low-dimensional latent space $z_{2,n}$. Neural network models define the function f by learning values for θ which are represented by the neural network architecture. A neural networks architecture is defined by a composition of functions on a directed acyclic graph defining a chain structure:



Each layer has a set of parameters θ_l defined by a matrix of weights W_l and a vector of biases b_l which are learnable through minimization of a convex and non-convex loss functions using back propagation. Loss functions encode the task of a neural network model, for example an autoencoding neural network aims to minimizing the reconstruction error of the original input x after mapping to a compressed low dimensional space z.

3.2 Generative neural networks

Generative (probabilistic) models incorporate a level of uncertainty over the unknown (latent) variables θ defining a neural network. Along with the nature of such uncertainties through conditional relationships between variables $p(\theta|x)$ (conditional probability distributions). Specifically in the Bayesian context we define prior distributions over each latent variable θ which is then updated to a posterior distribution given the data. Estimation of these posterior distributions is computed through variational inference which has recently been extended to neural networks in the form of variational autoencoders (VAE).

Variational autoencoders are the generative analog of the autoencoder that assume a high-dimensional variable $x_{d,n}$ is randomly sampled from some underlying generative process whose true probability distribution $p_\theta^*(x|z)$ is not known, where z is the latent representation of $x_{d,n}$. We attempt to learn the underlying generative process with a flexible distribution $x \sim p_\theta(x|z)$, typically binomial, to adapt to the data through estimation of θ. The goal of optimizing θ is to learn the values for which $p_\theta(x_{d,n}) \approx p_\theta^*(x_{d,n})$ so the predictive or classification loss function can be minimized. By incorporating an estimate of uncertainty in the neural network we can learn for each input sample $x_{d,n}$ a variance term that the model can use to adapt focus for specific samples with limited likelihood under the models' generative process, such samples could include outliers or sparse input.

3.3 Molecular biology

The work in this dissertation focuses on novel approaches for the analysis of the molecular state for individual cells drawn from various tissues, organisms and conditions. The molecular state of

a cell is defined by its genome composed of deoxyribonucleic acid (DNA) which encodes individual genes and becomes regulated through epigenetics to control the translation of specific genes into ribonucleic acid (RNA) strands. Capturing the transcriptional profile of a cell provides mode for exploring the molecular state of cells which is the primary focus of the methods presented in this dissertation.

Single cell RNA sequencing (scRNA-seq) technologies have been rapidly advancing and now enable the capture of high-resolution snapshots of gene expression activity in tens of thousands of cells enabling efforts to map whole tissues exposing the underlying cellular state atlases. Briefly, scRNA-seq technologies isolate individual cells through micro-fluidics and captures individual RNA fragments (short reads) which are built into a library and sequenced. These reads are then mapped back to a reference genome to identify the precise location on the genome the transcription originated from along with detailed annotations such as the associated gene symbol. RNA-seq technologies are fallible and do not always capture the entire transcriptome for each cell leading to the random dropout of individual genes transcripts.

The collection of scRNA-seq data has accelerated rapidly leading to the development of tools for the integrative analysis of multiple scRNA-seq datasets. scRNA-seq data integration aims to characterize and eliminate the effect of experimental factors driving gene expression variation between multiple scRNA-seq datasets, so that downstream analyses such as clustering, and trajectory inference performed on all datasets jointly are driven by the underlying biology and not on which technology a cell was sequenced. The models we develop in Chapters 2 and 3 aim to address limitations of sequencing technologies and efforts to perform joint analysis of tens of thousands of cells across technical confounders and data modalities.

Chapter 2

Alignment and rare cell identification in scRNA-seq

2.1 Introduction

Single cell RNA sequencing (scRNA-seq) technologies such as 10X[1] and SmartSeq[2] enable the capture of high-resolution snapshots of gene expression activity in individual cells. As the generation of scRNA-seq data accelerates, integrative analysis of multiple scRNA-seq datasets[3–10] is becoming increasingly important. However, the technologies used to sequence individual cells transcriptomes introduce non-biological variation (bias) specific to each technology that confounds attempts to integrate scRNA-seq experiments into larger atlases[5,11,12]. As the size and availability of single cell RNA-seq experiments keeps increasing the characterization and removal of unwanted effects of technical factors on measured gene expression across studies is critically important to modern genomics analysis. This chapter focuses on methods developed for the purpose of aligning single cell genomics data into common feature spaces in which batch effects have been corrected.

In this chapter we present our work on developing one of the first neural network alignment approaches in single cell genomics, scAlign[13]. We developed scAlign based on the observation that the scRNA-seq alignment problem is closely related to the problem addressed by domain adaptation in the field of computer vision [14,15]. scAlign is an unsupervised deep learning method for data alignment that can incorporate partial, overlapping or a complete set of cell labels, and

estimate per-cell differences in gene expression across datasets or conditions. Before delving into the scAlign model, I will first introduce batch effects, the basic principles of alignment and review the current computational approaches for mitigating unwanted technical variation.

2.1.1 Batch effects and linear correction

Batch effects or non-biological variation in RNA sequencing experiments is commonly observed between experiments where batches of cells were sequenced days or months apart or under different conditions such as environmental factors or sequencing platform[16–18]. This non-biological variation leads to samples separating first by technical factors, batch effects, and not the underlying biology (**Fig. 2.1**). If left unaccounted for these batch effects can confound the underlying biological relationships between cells leading to reduced power and spurious associations in downstream analysis such as differential expression[19–21] or trajectory inference[22,23].

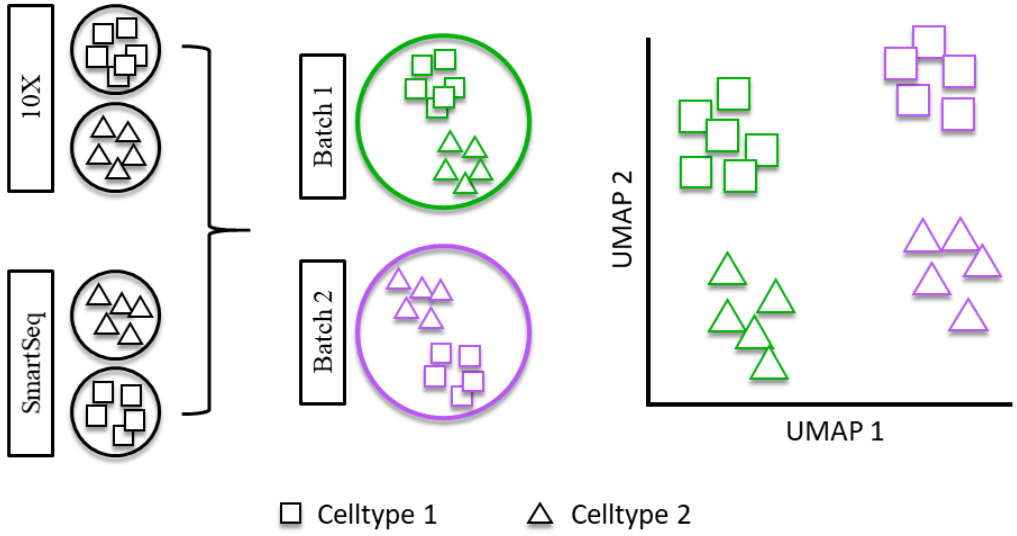


Figure 2.1: Batch effects in single cell experiments. Summary of the effect technical variation due to sequencing platform can have on gene expression measurements of two cell types.

However, due to current limitations in experimental protocols these batch effects are inevitable as researchers are now incrementally sequencing millions of single cells as well as pooling scRNA-seq data from multiple labs, sequencing technologies, or conditions to produce high resolution atlases of the gene expression for individual tissues or diseases[12,24–27].

The change in gene expression associated with batch effects can be linear or nonlinear leading to a need for specific tools which can operate alongside standard processing pipelines. Approaches that aim to normalize expression across cells such as transcripts per million (TPM) alone cannot correct for batch effects due to changes in the magnitude of gene expression associated to technical factors. The methods specifically tailored to address linear batch effect in single cell analysis are still in active development yet commonly utilized linear correction methods include limma[28] and ComBat[29] which were both tools developed initially for microarray data but are now being utilized for single cell RNA-seq batch correction. The aim of these approaches is to fit a linear model to the expression of each gene $x_g \in X$ which incorporates a design matrix of technical factors T as covariates.

$$x_g = \alpha_g + \beta_g T + \epsilon$$

where α_g specifies the overall gene expression, β_g is a vector of coefficients for the covariate matrix T and an error term ϵ assumed to follow a Normal distribution. Batch effects are then corrected by computing the residuals of linear model as corrected gene expression measurements for downstream analysis.

$$x_g^* = x_g - (\alpha_g + \beta_g T)$$

Some methods make additional assumptions about the distribution of the gene expression values such as the negative binomial in ComBat or Normal distribution as in Limma. As you may notice, by fitting a separate linear model per gene these approaches do not account for batch effects which affect modules of correlated genes in a similar manner. Leading both Limma and ComBat to employ a hierarchical empirical Bayes approach that shares parameters across genes to shrink batch effect parameters towards a common batch effect estimate leading to better estimating in the presence of small batches and outliers.

2.1.2 Alignment

The goal of scRNA-seq data alignment, similar to batch effect removal, is to characterize and eliminate the effect of experimental factors driving non-biological and expression variation between multiple scRNA-seq datasets. The goal being to ensure that downstream analyses such as clustering[30,31] and trajectory inference[31–33] performed on all datasets jointly are not driven by these factors. Such experimental factors include both technical nuisance factors such as batch or sequencing protocol[11,34–38], as well as biological factors of interest such as in case-control studies[39–42] or speciation[25].

Dataset alignment can be viewed as mapping one dataset onto another by warping the data in a manner that aims to preserve biological association and remove technical variation. For example, in case-control studies for which a pair of scRNA-seq datasets are generated from biological replicate populations before and after stimulus, functionally matched cell types across datasets must be identified and aligned to estimate cell type-specific response to stimulus. The more differential the response of the individual cell types, the more complex a mapping is required. Therefore, integrative tools must be able to freely scale up or down the complexity of their

mapping functions to successfully perform alignment depending on the heterogeneity of cell type-specific response to stimulus. In the extreme case where some cell types are present in only a subset of conditions being integrated, this poses additional mapping challenges since there may not be a 1-1 correspondence between types across conditions.

2.1.3 Current alignment methods and limitations

Current alignment tools can be separated into two exclusive sets: those that require all cells from all datasets to have known cell type labels (supervised), and those that do not make use of any cell type labels (unsupervised).

Method	Main parameters	Input	Dim. reduction
scAlign	Neural network architecture, hyperparameters, kernel	Expr., HVG, CCA, PCA, etc.	Yes
scVI	Neural network hyperparameters	Counts	Yes
Seruat	Number of components, iterations	HVG	Yes
ZINB-Wave	Number of factors	HVG	No
scMerge	Number of clusters, factors	Expr.	No
Scanorama	Number of neighbors, HVGs	HVG	No
MNN	Number of neighbors, kernel	HVG	No

Table 2.1: Comparison of alignment methods. Overview of alignment methods main parameters, accepted forms of input and where the method produces an aligned dimensionality reduction of the integrated data.

More common are the unsupervised approaches (**Table 2.1**) which include: (1) mutual nearest neighbors (MNN)[4] which tries to find the most similar cells (or mutual neighbors) across data batches with the assumption that the matched cells are of the same type. (2) Seurat[11] which utilizes canonical correlation analysis (CCA) to find a linear adjustment of the data that maximizes

correlation across batches then finds mutual neighbors (anchors) between datasets to quantify the strength of the batch effect and align the datasets. (3) scVI[9] a deep generative neural network approach based on a hierarchical Bayesian model where the transcriptome of each cell is compressed into a latent representation and decoded through nonlinear transformations that accounts for batch effects to compute posterior estimates for a ZINBwave[43] distribution per gene for each cell.

While effective for specific tasks the current approaches for alignment either make explicit assumptions about the distribution of single cell gene expression data or cannot flexibly scale in computation complexity and efficiently to handle a wide range of alignment problems. Additionally, these approaches are either unsupervised or fully supervised which consequently cannot handle when only a subset of cells can be labeled with high accuracy, or if only one dataset is labeled (as is the case when reference annotated cell atlases are available[44–49]). We identified the critical importance of the development of a method which is scalable to millions of cells, expressive enough to handle complex nonlinear warping across datasets due to batch effects and generalizable to a wide range of application areas are of critical importance to the field of single cell genomics.

2.2 scAlign

Here we present scAlign, a deep learning-based method for scRNA-seq alignment. scAlign performs single cell alignment of scRNA-seq data by learning a bidirectional mapping between cells sequenced within individual datasets, and a low-dimensional alignment space in which cells group by function and type, regardless of the dataset in which it was sequenced. This bidirectional map enables users to generate a representation of what the same cell looks like under each

individual dataset, and therefore simulate a matched experiment in which the exact same cell is sequenced simultaneously under different conditions.

Compared to previous approaches, scAlign can scale in alignment power due to its neural network design, and it can optionally use partial, overlapping, or a complete set of cell type labels in one or more of the input datasets. We demonstrate that scAlign outperforms existing alignment methods including Seurat[5,50], scVI[9], MNN[51], scanorama[10], scmap[52], MINT[3] and scMerge[6], particularly when individual cell types exhibit strong dataset-specific signatures such as heterogeneous responses to stimulus. While misalignment of cell types unique to one dataset is an inherent challenge for any alignment technique, we show that scAlign produces minimal false positive matchings. Furthermore, we show that our bidirectional map enables identification of changes in rare cell types that cannot be identified from alignment and data analysis steps performed in isolation. We also demonstrate the utility of scAlign in identifying changes in expression associated with sexual commitment in malaria parasites and posit that scAlign may be used to perform alignment in domains other than single cell genomics as well.

2.2.1 Overview of alignment with scAlign

The overall framework of scAlign is illustrated in **Figure 2.2**. While this paper is written in the context of aligning multiple datasets representing cell populations exposed to different stimuli or control conditions, scAlign can be readily used for any data alignment context discussed in the introduction. The premise of alignment methods is that when similar cell populations are sequenced under different conditions, some (possibly large) separation can be observed between cells of the same functional type but sequenced in different conditions (**Fig. 2.2a**). The first component of scAlign is the construction of an alignment space using scRNA-seq data from all

conditions, in which cells of the same functional type are indistinguishable, regardless of which condition they were sequenced in (**Fig. 2.2b**).

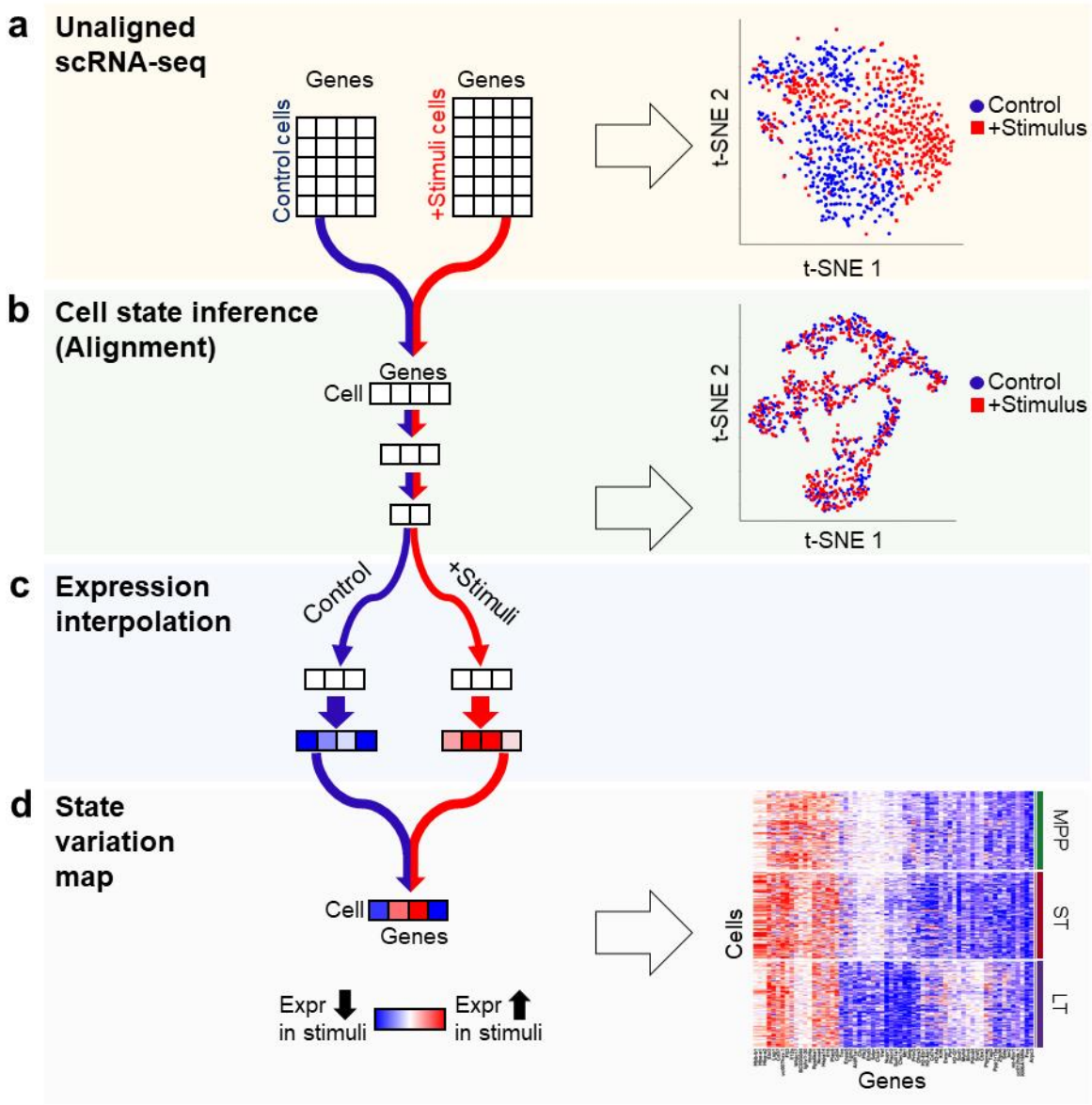


Figure 2.2: Schematic of unsupervised alignment and state variation mapping with scAlign. (**a**) The input to scAlign consists of cells sequenced across multiple scRNA-seq conditions. Expression can be represented as either gene-level expression, or embedding coordinates from dimensionality reduction techniques such as PCA or CCA. (**b**) A deep encoding network learns a low-dimensional alignment space that simultaneously aligns cells from all conditions. (**c**) Paired decoders project cells from the alignment space back into the gene expression space of each condition, and can be used to interpolate the expression profile of cells sequenced from any

condition into any other condition. (**d**) For a single cell sequenced under any condition, we can calculate its interpolated expression profile in all conditions, then measure the predicted variance across all input conditions to calculate a state variation map for the same cell state under different conditions to identify cells whose expression profiles vary significantly across condition.

This alignment space represents an unsupervised dimensionality reduction of scRNA-seq data from genome-wide expression measurements to a low dimensional manifold, using a shared deep encoder neural network trained across all conditions. Unlike autoencoders, which share a similar architecture to scAlign but use a different objective function, our low dimensional manifold is learned by training the neural network to simultaneously encourage overlap of cells in the state space from across conditions (thus performing alignment), yet also preserving the pairwise cell-cell similarity within each condition (and therefore minimizing distortion of gene expression). Optionally, scAlign can take as input a partial or full set of cell annotations in one or more conditions, which will encourage the alignment to cluster cells of the same type in alignment space.

2.2.2 Paired alignment with scAlign

We define the alignment task as identifying a low dimensional embedding space (termed the alignment space) in which functionally similar cells map to the same coordinates. Viewed from the lens of perturbation studies, if sequencing a cell immediately before and after stimulus were possible, alignment would bring cells post-stimulus into the same region of alignment space as the cell before stimulus, therefore removing the effect of the stimulus.

scAlign encodes the alignment space by extending the recent approach of learning by association for neural networks[14,53] into a unified framework for both unsupervised and supervised applications. For notational simplicity, we will assume we are aligning scRNA-seq data from a pair of conditions, though the framework extends to multiple conditions (see below). Let $\vec{x}_i^s$ and $\vec{x}_j^t$ be vectors of length G that represent the gene expression profiles of cells i and j in conditions

s and t, respectively. Similarly, let $\vec{e}_i^s$ and $\vec{e}_j^t$ be vectors of length K that represent that alignment space embedding of cells i and j in conditions s and t, respectively, where the embeddings represent the linear activations of the final output layer of an encoder neural network.

scAlign trains an encoder neural network (parameterized by weights $\boldsymbol{W}$) that defines the alignment space by optimizing the network weights used to calculate $\vec{e}_i^s$ and $\vec{e}_j^t$ to minimize the following objective function:

$$f = \left[\frac{1}{|S|}\sum_i \text{cross-entropy}(\vec{P}_{i,\cdot}^s, \vec{Q}_{i,\cdot}^s)\right] + \left[\frac{1}{|T|}\sum_j \text{cross-entropy}(\vec{P}_{j,\cdot}^t, \vec{Q}_{j,\cdot}^t)\right] + \lambda\|\boldsymbol{W}\|_F^2$$

where

$$\boldsymbol{P}^s = \boldsymbol{P}^{s\to t}\boldsymbol{P}^{t\to s}$$

$$\boldsymbol{P}^t = \boldsymbol{P}^{t\to s}\boldsymbol{P}^{s\to t}$$

$$Q_{i,k}^s = \frac{\exp(-0.5\|\vec{x}_i^s - \vec{x}_k^s\|^2/\sigma_i^2)}{\sum_{k\prime\neq i}\exp\left(-0.5\left\|\vec{x}_i^s - \vec{x}_{k\prime}^s\right\|^2/\sigma_i^2\right)}$$

$$Q_{j,k}^t = \frac{\exp\left(-0.5\left\|\vec{x}_j^t - \vec{x}_k^t\right\|^2/\sigma_j^2\right)}{\sum_{k\prime\neq j}\exp\left(-0.5\left\|\vec{x}_j^t - \vec{x}_{k\prime}^t\right\|^2/\sigma_j^2\right)}$$

$$P_{i,j}^{s\to t} = \frac{\exp\left(\vec{e}_i^{s\,T}\vec{e}_j^t\right)}{\sum_{j\prime}\exp\left(\vec{e}_i^{s\,T}\vec{e}_{j\prime}^t\right)}$$

$$P_{j,i}^{t\to s} = \frac{\exp\left(\vec{e}_i^{t\,T}\vec{e}_j^s\right)}{\sum_{i\prime}\exp\left(\vec{e}_{i\prime}^{t\,T}\vec{e}_j^s\right)}$$

$$\vec{e}_i^s = \text{encoder}(\vec{x}_i^s, \boldsymbol{W})$$

$$\vec{e}_j^t = \text{encoder}(\vec{x}_j^t, \boldsymbol{W})$$

The central idea of the alignment procedure of scAlign is that it optimizes the embeddings of cells ($\vec{e}_i^s$ and $\vec{e}_j^t$) such that the scaled, pairwise cell-cell similarity matrix (or formally, a transition matrix) computed between cells within each condition in gene expression space ($\boldsymbol{Q^s}$ and $\boldsymbol{Q^t}$) should be maintained within the alignment space ($\boldsymbol{P}^s$ and $\boldsymbol{P}^t$), respectively. The novel aspect of scAlign compared to other dimensionality reduction methods is in how $\boldsymbol{P}^s$ and $\boldsymbol{P}^t$ are calculated. While $\boldsymbol{P}^s$ would canonically be calculated by transforming the dot product of the embeddings $\vec{e}_i^s$ as is done in the tSNE method[54] for example, scAlign computes roundtrip random walks of length two that traverse the two conditions. $\boldsymbol{P}^{\boldsymbol{s}}_{\boldsymbol{i,k}}$, the transition probability of moving from cell i to cell k within condition s, is calculated as the probability of randomly walking from cell i to cell k in two steps: first from cell i to any cell j in the other condition t in the first step, then from that cell j to cell k (in condition s) in the second step. By forcing the random walk to first visit a cell in the other condition, scAlign encourages the encoder to bring cells from across the two conditions into similar regions of alignment space.

The network weights $\boldsymbol{W}$ are initialized by Xavier[55] and optimized via the Adam algorithm[56] with an initial learning rate of 10^{-4} and a maximum of 15,000 iterations. The neural network activation functions of each hidden layer are ReLU and the embedding layer has a linear activation function. Regularization is enforced through an L2 penalty on the weights along with per-layer batch normalization and dropout at a rate of 30%.The scAlign framework has three tunable parameters: the per-cell variance parameter σ_i^2 that controls the effective size of each cell's neighborhood when defining the similarity matrix in gene expression space, the magnitude of the penalization term λ over $\boldsymbol{W}$ that is fixed at 10^{-4}, and the size of the encoder network architecture.

For the tuning parameter σ_i^2, small values yield more local alignment, whereas larger values yield more global alignment. In our experiments, we train each model with a range of values for σ_i^2. Typically, [5,10,30] provide robust results when training on mini-batches of less than 300 samples. While the per-cell variance parameter σ_i^2 operates on the training mini-batch, we found training is robust to the choice of σ_i^2.

We set the size of the encoder architecture by either automatically constructing a network based on the dimensionality of the input data in conjunction with a complexity parameter, or from a catalog of network architectures which are at most three layers deep. As with other neural networks, the size of the architecture defines the complexity and power of the network. Model complexity is important for alignment because the network must be powerful enough to align cells from conditions that yield heterogeneous responses to stimulus, but not so powerful that any cell in one condition can be mapped to any other cell in another condition, regardless of whether they are functionally related. We have found in our experiments (**Appendix: 3A.S3**) that the combination of cross-entropy loss and shrinkage applied to the network weights yields robustness to generously large network architectures. Namely, by encouraging small weights and minimizing the differences in cell-cell similarity matrices between the expression and embedding spaces, we avoid training the neural network to perform unnecessary complex transformations on the data.

The objective function that scAlign optimizes does not incorporate terms specific to scRNA-seq data such as a negative binomial observation model. We found that computing the principal component and canonical correlates of the normalized scRNA-seq data and using the resulting scores in place of gene expression measurements maintained alignment and interpolation accuracy but sped up training significantly (**Appendix: 3A.S4**). Note that even when the encoder network is given PC or CC dimensions as input instead of gene expression measurements, the

decoder is still trained to transform alignment space coordinates into the original gene expression space.

The training procedure for training a shared autoencoder followed that of scAlign in that the autoencoder was trained on data from all conditions simultaneously. The shared alignment space of the autoencoder was learned by optimizing with respect to the traditional mean squared error of reconstructing the original expression profiles for each condition by simultaneously training condition specific decoder networks.

2.2.3 Multi-way alignment with scAlign

Alignment of three or more conditions simultaneously is implemented in two ways within the scAlign framework. In approach (1) ("all-pairs alignment"), round trip walks are computed between all pairs of conditions and is expected to be the most accurate form of multi-way alignment. In approach (2) ('reference-based alignment'), one condition is defined as a reference, against which all other conditions are aligned.

2.2.3.1 All-pairs alignment

In this strategy, we extend the pairwise alignment approach by performing round trip walks between all pairs of conditions simultaneously, while still sharing a single encoder's neural network parameters across all conditions. Compared to the reference-based alignment approach below, the all-pairs approach will be more robust when there are cell types that are only represented in a subset of the input conditions. The objective function of the pairwise alignment approach is modified to include round trip walks between each condition k and the remaining conditions $l \neq k$:

$$f = \sum_{k} \sum_{l \neq k} \left[\frac{1}{|N|} \sum_{n} \text{cross-entropy}(\vec{P}_{n,\cdot}^{k,l}, \vec{Q}_{n,\cdot}^{k,l}) \right] + \lambda \|\boldsymbol{W}\|_F^2$$

$$\boldsymbol{P}^{k,l} = \boldsymbol{P}^{k \to l} \boldsymbol{P}^{l \to k}$$

$$Q_{i,j}^{k,l} = \frac{\exp\left(-0.5 \left\| \vec{x}_i^k - \vec{x}_j^k \right\|^2 / \sigma_i^2\right)}{\sum_{j' \neq i} \exp\left(-0.5 \left\| \vec{x}_i^k - \vec{x}_{j'}^k \right\|^2 / \sigma_i^2\right)}$$

$$P_{i,j}^{k \to l} = \frac{\exp\left(\vec{e}_i^{k\,T} \vec{e}_j^{\,l}\right)}{\sum_{j'} \exp\left(\vec{e}_i^{k\,T} \vec{e}_{j'}^{\,l}\right)}$$

$$P_{j,i}^{l \to k} = \frac{\exp\left(\vec{e}_j^{l\,T} \vec{e}_i^{\,k}\right)}{\sum_{i'} \exp\left(\vec{e}_j^{l\,T} \vec{e}_{i'}^{\,k}\right)}$$

$$\vec{e}_i^{\,k} = \text{encoder}(\vec{x}_i^k, \boldsymbol{W})$$

$$\vec{e}_j^{\,l} = \text{encoder}(\vec{x}_j^l, \boldsymbol{W})$$

2.2.3.2 Reference-based multi-way alignment with scAlign

In this strategy, multiple conditions are aligned simultaneously by selecting one condition to be a reference (k_{ref}), against which all other conditions ($l \neq k_{\text{ref}}$) are aligned. Compared to the all-pairs approach, reference-based alignment is faster and therefore more scalable, though is expected to perform worse when there are cell types shared amongst non-reference conditions, that are not represented in the reference condition. The objective function for reference-based alignment is as follows:

$$f = \sum_{l \neq k_{\text{ref}}} \left[\frac{1}{|N|} \sum_{n} \text{cross-entropy}(\vec{P}_{n,\cdot}^{k_{\text{ref}}}, \vec{Q}_{n,\cdot}^{k_{\text{ref}}}) \right] + \lambda \|\boldsymbol{W}\|_F^2$$

$$\boldsymbol{P}^{k_{\text{ref}}} = \boldsymbol{P}^{k_{\text{ref}} \to l} \boldsymbol{P}^{l \to k_{\text{ref}}}$$

$$Q_{i,j}^{k_{\text{ref}}} = \frac{\exp\left(-0.5 \left\| \vec{x}_i^{k_{\text{ref}}} - \vec{x}_j^{k_{\text{ref}}} \right\|^2 / \sigma_i^2\right)}{\sum_{j' \neq i} \exp\left(-0.5 \left\| \vec{x}_i^{k_{\text{ref}}} - \vec{x}_{j'}^{k_{\text{ref}}} \right\|^2 / \sigma_i^2\right)}$$

$$P_{i,j}^{k_{\text{ref}} \to l} = \frac{\exp\left(\vec{e}_i^{k_{\text{ref}}{}^T} \vec{e}_j^{l}\right)}{\sum_{j'} \exp\left(\vec{e}_i^{k_{\text{ref}}{}^T} \vec{e}_{j'}^{l}\right)}$$

$$P_{j,i}^{l \to k_{\text{ref}}} = \frac{\exp\left(\vec{e}_j^{l\,T} \vec{e}_i^{k_{\text{ref}}}\right)}{\sum_{i'} \exp\left(\vec{e}_j^{l\,T} \vec{e}_{i'}^{k_{\text{ref}}}\right)}$$

$$\vec{e}_i^{k_{ref}} = \text{encoder}(\vec{x}_i^{k_{ref}}, \boldsymbol{W})$$

$$\vec{e}_j^{l} = \text{encoder}(\vec{x}_j^{l}, \boldsymbol{W})$$

The remaining details for optimizing scAlign's objective function in the multi-way case are identical to the paired alignment task described previously. We note that in our experiments the number of embedding dimensions had to be increased for three or more conditions to accommodate the increased information in the embeddings of the encoder shared across all k condtions.

2.2.4 Using partial or complete cell type labels with scAlign

The objective function optimized by scAlign can naturally incorporate partial, overlapping, or complete cell type labels for the cells, in one or more conditions. Suppose there are C cell type labels available, in a pairwise alignment scenario. Then define matrix $\boldsymbol{A}^s$ such that $A_{i,c}^s = 1$ if cell i in condition s has cell type label c, else $A_{i,c}^s = 0$. Similarly, define matrix $\hat{A}^s$ containing the predicted class labels for all cells in condition s. The scAlign objective function then becomes:

$$f = \left[\frac{1}{|S|}\sum_{i}\left(\alpha\ \text{cross-entropy}\left(\vec{P}^{s}_{i,\cdot}, \vec{Q}^{s}_{i,\cdot}\right) + \beta\sum_{c} A^{s}_{i,c}\ \text{cross-entropy}\left(\vec{A}^{s}_{i,\cdot}, \vec{\hat{A}}^{s}_{i,\cdot}\right)\right)\right] + \left[\frac{1}{|T|}\sum_{j}\left(\alpha\ \text{cross-entropy}\left(\vec{P}^{t}_{j,\cdot}, \vec{Q}^{t}_{j,\cdot}\right) + \beta\sum_{c} A^{t}_{j,c}\text{cross-entropy}\left(\vec{A}^{t}_{j,\cdot}, \vec{\hat{A}}^{t}_{j,\cdot}\right)\right)\right] + \lambda\|\boldsymbol{W}\|^{2}_{F}$$

We incorporate partial, overlapping, or complete label information by introducing an extra set of terms corresponding to classification loss and weighted by the factor β. The classifier loss terms minimize the mean cross-entropy of the predicted and actual cell labels as defined by the second term within each summation of f. The adaptation and classifier components f are balanced by hyperparameter weights α and β respectively. Adjusting α and β allows emphasis to be placed individually on the pairwise cell similarity or known labels; in this work both weights were fixed to 1.0 when label information is provided.

2.2.4 Introduction to interpolation with scAlign

In the second component of scAlign (**Fig. 2.2c**), we train condition-specific deep decoder networks capable of projecting individual cells from the alignment space back to the gene expression space of each input condition, regardless of what condition the cell is originally sequenced in. We use these decoders to measure per-cell and per-gene variation of expression across conditions, which we term the cell state variation map. In the case of aligning two conditions, this cell state variation map estimates a paired difference in expression of the same cell across conditions (**Fig. 2.2d**). scAlign therefore seeks to re-create the ideal experiment in which the exact same cell is sequenced before and after a stimulus in a case-control study, for example.

The interpolation component of scAlign trains a condition-specific decoder to map cells from the alignment space back into each of the individual condition-specific gene expression spaces. The decoder network architecture is chosen to be symmetric with the encoder network trained during the alignment process, with weights randomly initialized and optimized again via the Adam optimizer[56] with learning rate set at 10^{-4} and trained for at most 30,000 iterations.

After interpolating every cell (sequenced in any condition) from the alignment space back to every input condition, for each cell, we obtain multiple condition-specific representations for each cell. Then, per cell, we compute the variance of the interpolated expression patterns for that cell across the input conditions. The result is a matrix, termed the state variance map, which illustrates the variance in each gene-specific expression level for each cell predicted across conditions. In the special case where two conditions are being aligned, this state variance map can be viewed as a (predicted) paired differential expression map, where differences are calculated per cell.

2.3 Benchmarking and validation of scAlign

This section details the benchmarking and validation experimentation published in Johansen and Quon 2019 with the goal of reporting a robust and stable model of alignment which outperforms current state-of-the-art alignment tools.

2.3.1 Capturing cell type specific response to stimulus

We first benchmarked the alignment component of scAlign using data from four publicly available scRNA-seq studies for which the same cell populations were sequenced under different conditions, and for which the cell type labels were obtained experimentally (**Fig. 2.3**, **Appendix: 3A.S1**). Our

first benchmark is CellBench[57], a dataset consisting of three human lung adenocarcinoma cell lines (HCC827, H1975, H2228) that were sequenced using three different protocols (CEL-Seq2, 10x Chromium, Drop-Seq Dolomite) as well as at varying relative concentrations of either RNA content or numbers of cells in a mixture. While the alignment of the homogeneous cell populations sequenced across protocols was trivial and did not require data alignment methods (**Appendix: 3A.S2**), alignment of RNA mixtures across protocols was more challenging and more clearly illustrated the performance advantage of scAlign (**Fig. 2.3a**). We additionally benchmarked alignment methods using data generated by Kowalczyk et al.[58] and Mann et al.[59] on three hematopoietic cell types (LT-HSC, ST-HSC, MPP) collected from the C57BL/6 mouse strain at approximately 2 months ("young") and 2 years ("old") of age. Mann et al. additionally challenged the mice with an LPS or a control stimulus. Similar to our results with CellBench, scAlign outperforms other approaches on both of these benchmarks (**Fig. 2.3b,c**). The results of scAlign in these comparisons were robust to network depth, width and input features (**Appendix: 3A.S3, 3A.S4**) along with choice of hyper parameters.

To better understand why the relative performance of the other methods was inconsistent across benchmarks (**Fig. 2.3a-c**), we next characterized the difficulty of each benchmark for alignment. For each cell type in each benchmark, we identified cell type marker genes by computing the differentially expressed genes (DEGs) between cell types, individually for each condition. We observed considerable overlap in the cell type marker genes (**Appendix: 3A.S5**), suggesting these benchmarks may be less challenging to align and therefore more difficult to distinguish alignment methods from each other.

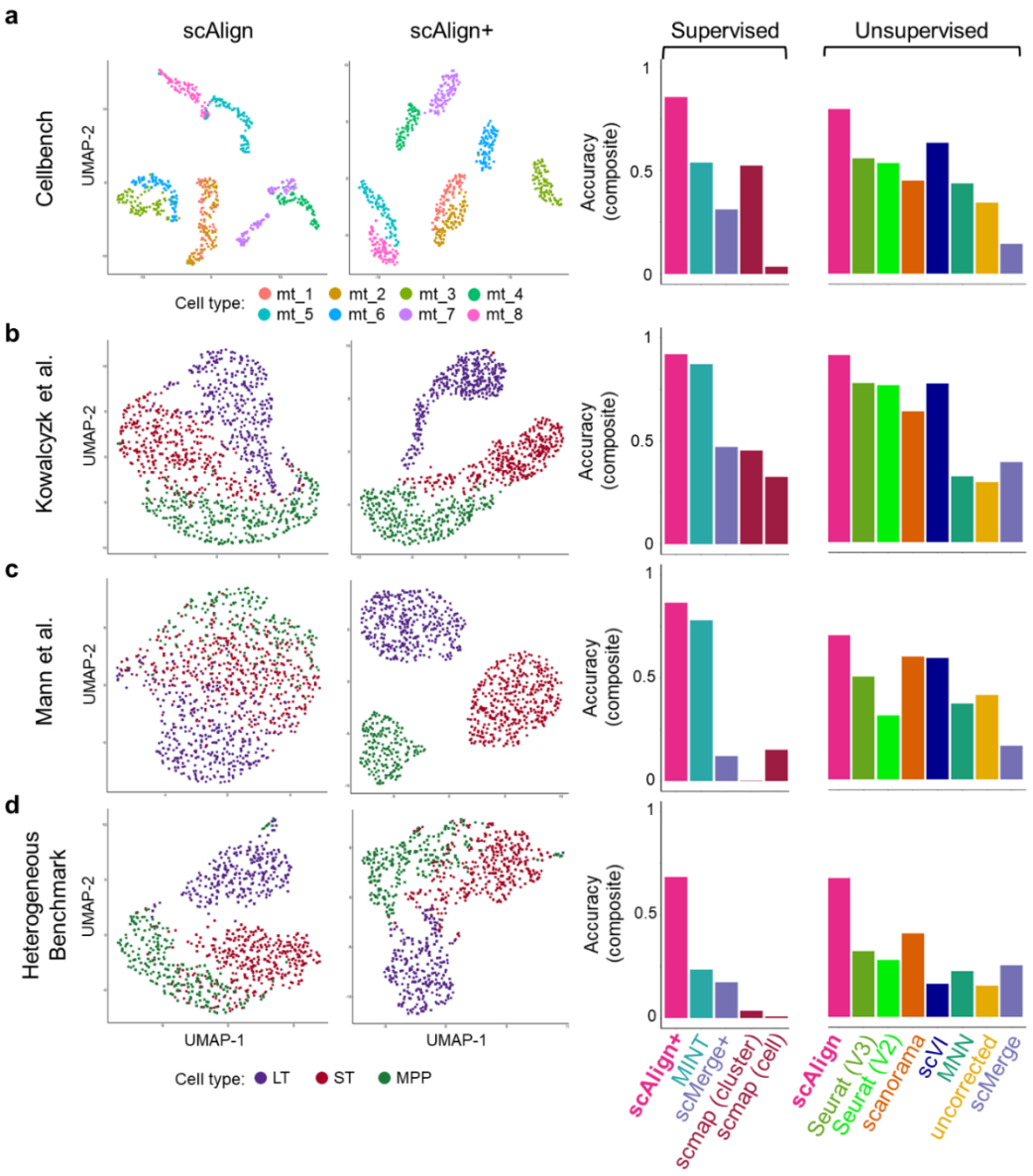


Figure 2.3: scAlign outperforms existing alignment approaches on four benchmarks. (**a**) CellBench, a benchmark consisting of mixtures (mt) of RNA from three cancer cell lines sequenced using multiple protocols. Plots from left to right: (1) UMAP plot of embeddings after alignment with scAlign, where each point represents a cell, and cells are colored according to their mixture type (mt) as reported in Tian et al. (2) UMAP plot of embeddings after alignment with supervised scAlign (scAlign+). (3) Bar plot indicating the accuracy$_{\text{composite}}$ of a classifier, measured as a weighted combination of cross-condition label prediction accuracy and alignment score. (**b**) Same as (a), but with the Kowalczyk et al. benchmark consisting of hematopoietic cells sequenced from young and old mice. Cells are colored according to type (LT, ST, MPP, legend at bottom). (**c**) Same as (a), but with the Mann et al. benchmark consisting of hematopoietic cells sequenced from young and old mice, challenged with LPS. (**d**) Same as (a), but with the HeterogeneousBenchmark dataset consisting of hematopoietic cells responding to different stimuli.

We therefore constructed a novel benchmark termed HeterogeneousBenchmark by combining published scRNA-seq data on hematopoietic cells measured across different studies and stimuli. This benchmark yields smaller overlap in cell type marker genes (**Appendix: 3A.S5**), which makes it more challenging to align. On HeterogeneousBenchmark, we find that scAlign's performance is robustly superior, while Seurat and Scanorama also outperform the remaining methods (**Fig. 2.3d**).

scAlign simultaneously aligns scRNA-seq from multiple conditions and performs a non-linear dimensionality reduction on the transcriptomes. This is advantageous because dimensionality reduction is a first step to a number of downstream tasks, such as clustering into putative cell types[22] and trajectory inference[60–62]. Dimensionality reduction of cell types generally improves when more data is used to compute the embedding dimensions, and so we hypothesized that established cell types will cluster better in scAlign's embedding space in part due to the fact we are defining a single embedding space using data from multiple conditions. We therefore compared the clustering of known cell types in the scAlign embedding space to an autoencoder neural network that uses the same architecture and number of parameters as scAlign, but is trained on each condition separately . In two of the three benchmarks we tested, we found that known cell types cluster more closely and are more distinct in scAlign embedding space compared to that of the corresponding autoencoder (**Fig. 2.4, Appendix: 3A.S6**), suggesting scAlign's embedding space benefits from pooling cells from across all conditions. Furthermore, by pooling cells into a common embedding space scAlign can identify new subpopulations within known cell type clusters (**Appendix: 3A.S7**).

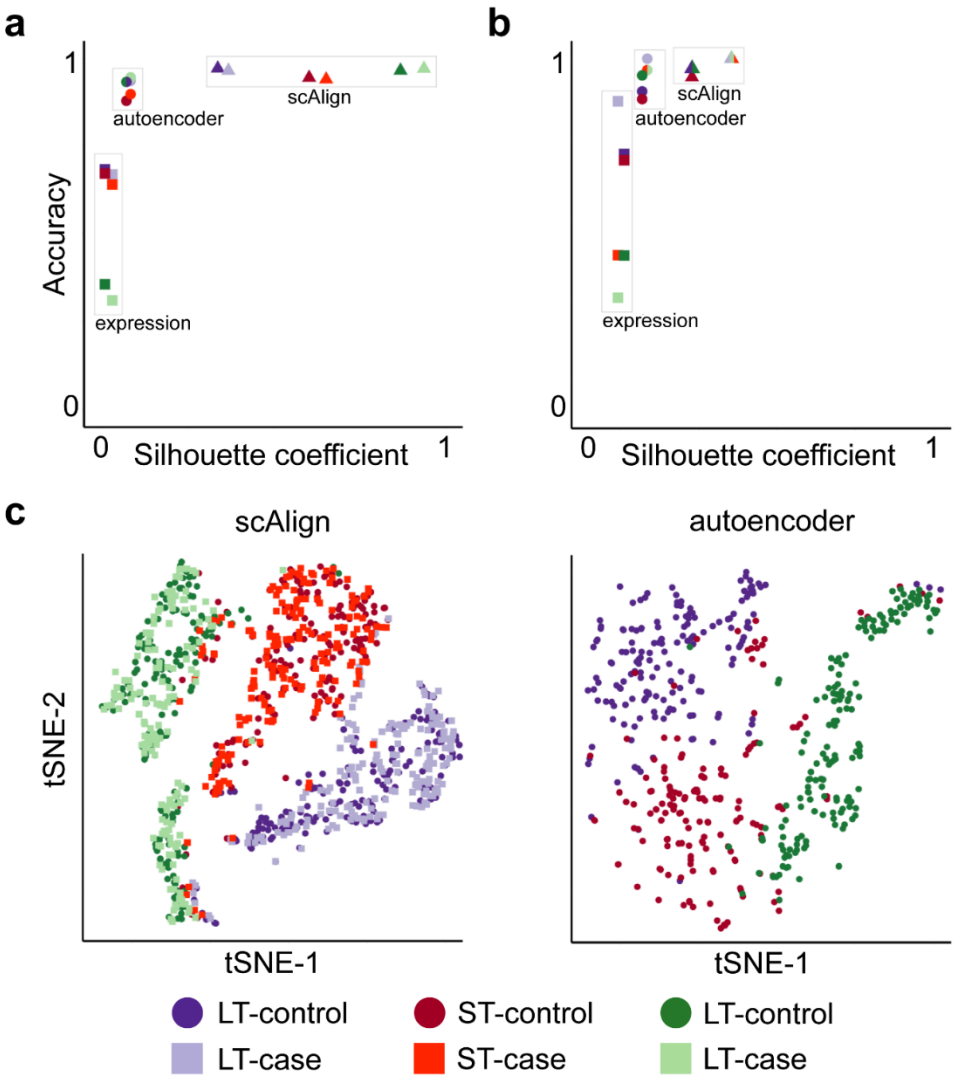


Figure 2.4: Joint analysis of cells from all conditions leads to more accurate clustering of cell types compared to independent analysis of individual conditions. (**a**) Scatterplot illustrating the quality of clustering of cell types within each condition from the Mann et al. benchmark. Each point represents one cell type in one condition, when the embedding is computed using either the original expression data ('expression'), the embedding dimensions of scAlign, or the embedding dimensions of an autoencoder with the same neural network architecture as scAlign. The y-axis represents classification accuracy, while the x-axis represents the silhouette coefficient. (**b**) Same as (a), but for HeterogeneousBenchmark (**c**) tSNE plots visualizing the embedding space of scAlign trained on both conditions and (**d**) an autoencoder trained on a single condition.

A unique feature of scAlign is that it can optionally use cell type labels for a subset of (or all) cells if available, but does not require any labels by default. In other words, scAlign can perform unsupervised, semi-supervised or fully-supervised alignment. One example of a use case would be when a labeled, highly quality cell atlas is available, it can be used to label cells sequenced from a newer, smaller study. **Figures 2.3a-d** illustrate, for each of the four benchmarks, that scAlign performance improves when cell type labels are available at training time, and exceeds the performance of other supervised methods such as MINT[63], scMerge[6] and scmap[52]. Even when only a subset of cells from one condition have labels available for semi-supervised training, scAlign performance improves compared to a strictly unsupervised alignment, though still lower than a fully supervised scAlign+ (**Fig. 2.5, Appendix: 3A.S8**). When provided with labels, the cell-cell similarity matrix of the supervised scAlign method is qualitatively similar to the cell-cell similarity matrix of cells in the original gene expression space as well as the unsupervised scAlign alignment space, suggesting the inferred alignment space is robust to adding labels during alignment (**Appendix: 3A.S9**).

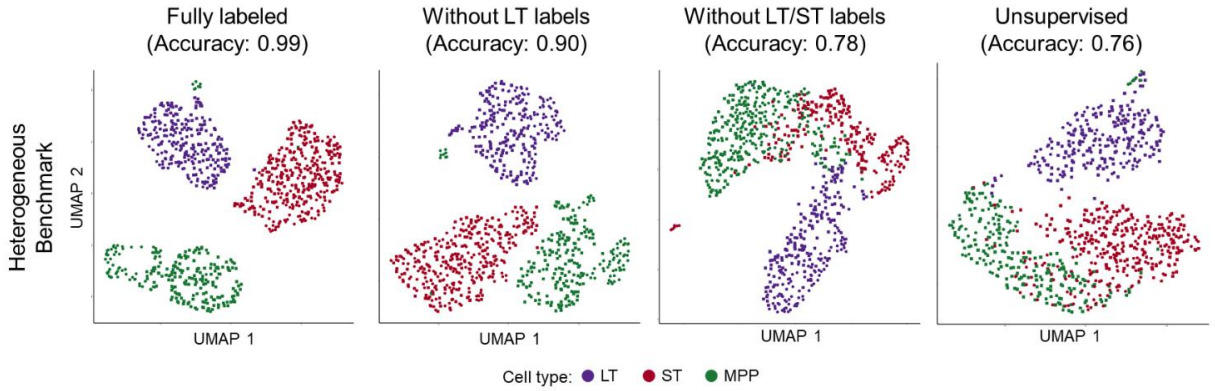


Figure 2.5: Semi-supervised alignment mode of scAlign enables use of partial sets of cell type labels. UMAP visualization of the HeterogenousBenchmark after alignment with scAlign+ trained with (**a**) labels for all cells in both conditions, (**b**) after removal of labels for LT-HSC HSC in the stimulated condition, (**c**) after removal of labels for LT-HSCs and ST-HSCs in the stimulated condition, and (**d**) scAlign trained without cell labels.

2.3.2 Accurate Interpolation of gene expression

One of the more novel features of scAlign is the ability to map each cell from the alignment space back into the gene expression space of each of the original conditions, regardless of which condition the cell was originally sequenced in. The idea of "re-styling" cells under each condition was drawn from the field of style transfer in computer vision where images are mapped onto new feature spaces with unique patterns no different than the patterns of gene expression induced by experimental conditions. This mapping is performed through interpolation: for each condition, we learn a mapping from the alignment space back to gene expression space using cells sequenced in that condition, then apply the map to all cells sequenced in all other conditions. This interpolation procedure enables measurement of variation in gene expression for the same cell state across multiple conditions, and simulates the ideal experiment in which the exact same cell is sequenced before and after a stimulus is applied, and the variation in gene expression is subsequently measured.

To measure the accuracy of scAlign interpolation, for each of the three hematopoietic benchmarks, we trained decoder neural networks to map cells from the alignment space back into each of the case and control conditions. We then measured interpolation accuracy as the accuracy of a classifier trained on the original gene expression profiles of cells sequenced under one condition (e.g. stimulated), when used to classify cells that have been interpolated from the other condition (e.g. control). Comparing this interpolation accuracy to cross-validation accuracy of classifying cells in their original condition using the original measured gene expression profiles, we see that interpolation accuracy is similar to expression accuracy (**Fig. 2.6a**), suggesting that cells maintain their general type when mapped into another condition.

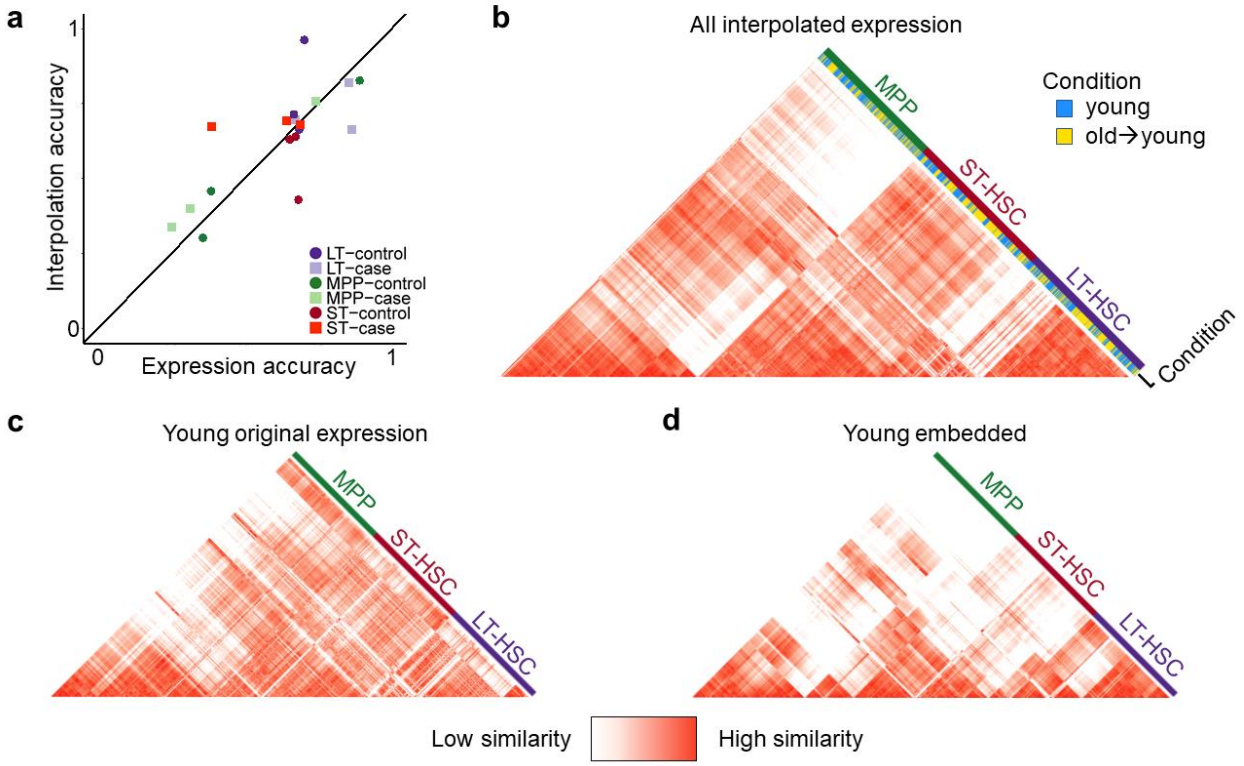


Figure 2.6: Interpolation of gene expression patterns is accurate. (**a**) Scatterplot of classifiers trained on gene expression profiles of one condition, that are subsequently used to predict labels of either measured expression profiles from the same condition in a cross-validation framework (x-axis), or used to predict labels of cells sequenced from the other condition that were then interpolated into this condition (y-axis). Similarity in accuracy represented by points near the diagonal indicates that cell type identity encoded in the gene expression profile is maintained even after interpolation. (**b**) The pairwise cell-cell similarity matrix for all cells projected into the young condition, including both the old cells interpolated into the young condition (yellow) and the cells originally sequenced in the young condition (blue). Note that cells cluster largely by cell type regardless of the condition in which they were sequenced. (**c**) The pairwise cell-cell similarity matrix for all cells computed using the original expression measurements. (**d**) The pairwise cell-cell similarity matrix for all cells computed using the low-dimensional coordinates within the alignment space learned by scAlign. Similarity between (c) and (d) indicate the scAlign embedding maintains global similarity patterns between cells in the original gene expression space.

Figure 2.6b illustrates the cell-cell similarity matrix computed in gene expression space of hematopoietic cells collected in the Kowalczyk study, when including cells sequenced in the young mice, as well as cells that have been interpolated from the old mice into the young condition. We see that cells cluster largely by cell type (LT-HSC, ST-HSC, MPP) and not by their condition of origin. Furthermore, by computing a state variance map from the interpolation of all cells into both conditions, we identify differentially expressed genes that were not identified by traditional differential expression analysis (**Appendix: 3A.S13**). This demonstrates that the encoding and interpolation process maintains data fidelity, even though the encoder is trained to align data from multiple conditions and is not explicitly trained to minimize reconstruction error like typical autoencoders. **Figures 2.6c,d** further illustrate that the cell-cell similarity matrix in embedding space is faithful to the cell-cell similarity matrix in the original gene expression space.

2.4 Experiments

This section details the experiments published in Johansen and Quon 2019 and Hodge et al. 2019 which detail a broad range of applications and collaborations with scAlign for single cell RNA sequencing-based research.

2.4.1 Interpolation identifies early gametocyte markers of the engineered ap2-g-dd strain of P. falciparum

We next applied scAlign to identify genes associated with early steps of sexual differentiation in *Plasmodium falciparum*, the most widespread and virulent human malaria parasite. Briefly, the clinical symptoms of infection are the result of exponential growth of asexual parasites within red blood cells, while parasite transmission depends on the formation of the non-replicating male and female sexual stages necessary for infection of the parasite's mosquito vector. During each round

of asexual replication, a sub-population of parasites will activate expression of the *ap2-g* gene, which encodes the transcriptional master regulator of sexual differentiation, to initiate sexual differentiation. While the gene *ap2-g* is a known master regulator of sexual commitment, and its expression is necessary for sexual commitment, the events which follow *ap2-g* activation and lead to full sexual commitment are unknown[64]. Furthermore, *ap2-g* expression is restricted to a minor subset of parasites, making the identification of the precise stage of the life cycle when sexual commitment occurs a challenging task.

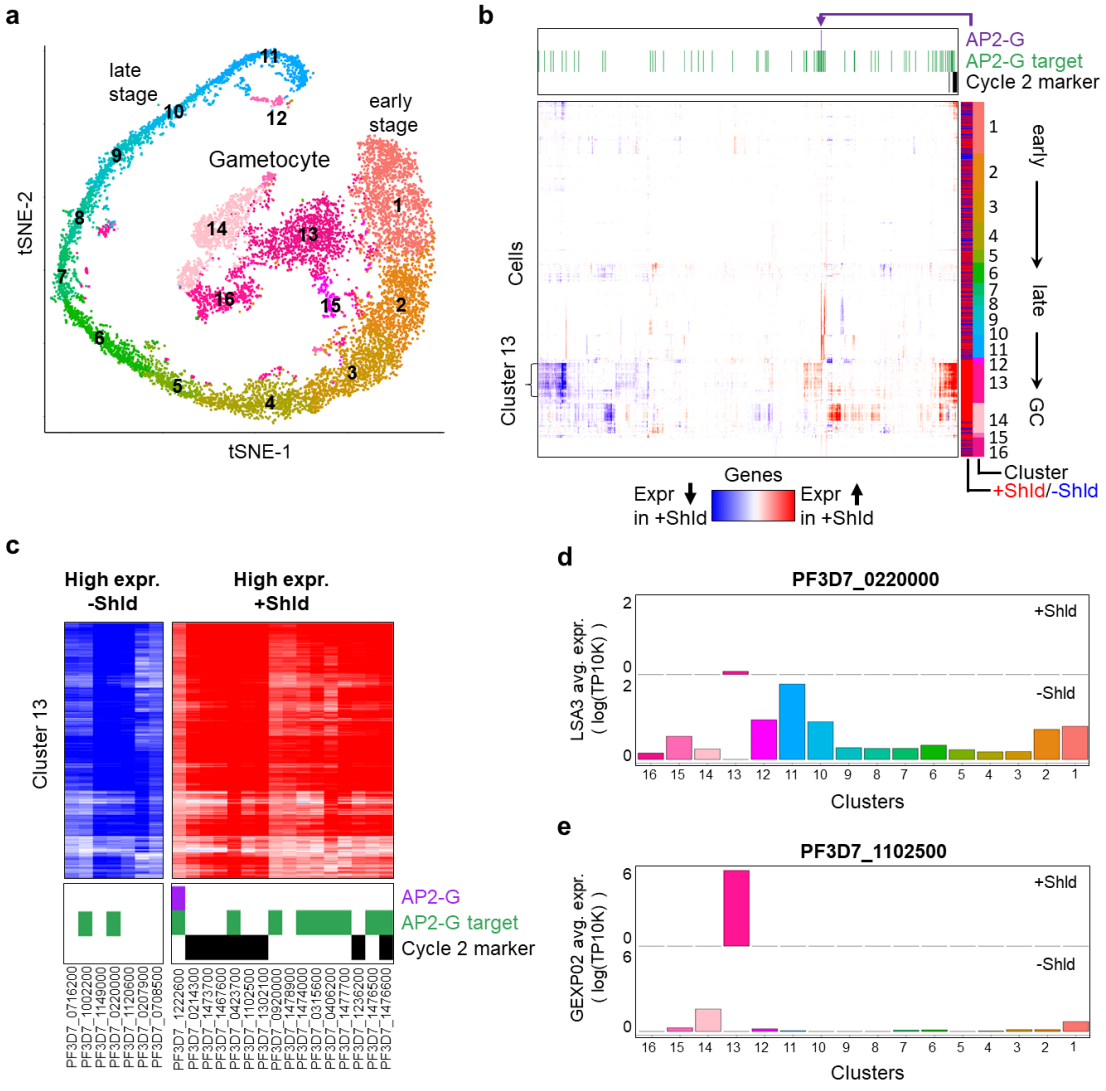


Figure 2.7: Alignment of *P. falciparum* cells sequenced from a conditional ap2-g knockdown line identifies cycle 2 gametocytes. (**a**) tSNE visualization of cells that cannot stably express *ap2-g* (-Shld) and *ap2-g* expression-capable cells (+Shld) after alignment by scAlign. Each cell is colored by its corresponding cluster identified in Poran et al., and clusters are numbered according to relative position in the parasite life cycle. (**b**) scAlign state variation map defined by projecting every cell from (a) into both the +/-Shld conditions, then taking the paired difference in interpolated expression profiles. Rows represent cells, ordered by cluster from early stage (top) to late stage and GC (bottom), and columns represent the 661 most varying genes. The state variation map reveals that cluster 13 is predicted to differ in expression the most between +/-Shld. The column annotations on top indicate which of the variable genes have been previously established as a target of *ap2-g* via ChIP-seq experiments[65] which genes have been reported as playing a role in cell cycle 2 gametocyte maturation[66] and which gene represents *ap2-g*. (**c**) The same state variation map of (c), but zoomed in on Cluster 13 and the genes predicted to be most differentially expressed between +/-Shld. (**d**) Average per-cluster expression levels of PF3D7_0220000 reported in (c), for both the +/-Shld conditions. PF3D7_0220000 is predicted to be up-regulated in -Shld relative to +Shld, which is reflected in the per-cluster expression levels. (**e**) Same as (d), but for PF3D7_1102500, a gene predicted to be up-regulated in +Shld relative to -Shld.

Figure 2.7a illustrates the alignment space of parasites which are either capable of *ap2-g* expression and will contain an *ap2-g*-expressing subpopulation in the initial stages of sexual differentiation (+Shld), or are *ap2-g* deficient and therefore all committed to continued asexual growth (-Shld). As was observed in the original paper[64], the +/-Shld cells fall into clusters that can be ordered by time points in their life cycle (**Fig. 2.7a**). scAlign alignment maintains the gametocytes from the +Shld condition as a distinct population that is not aligned to any parasite population from the -Shld condition, whereas other tested methods are unable to isolate the gametocyte population (**Appendix: 3A.S14**).

To further investigate how scAlign is able to maintain the gametocytes as a distinct population after alignment, we looked at the random walks performed by the gametocyte cells to see which cells from the -Shld condition they walked to, and found that scAlign maps a very small number of cells from similar surrounding clusters into the peripheral region of alignment space near the gametocytes. These -Shld cells in the periphery of the gametocyte cluster allows the

gametocytes to use those cells as "anchors" in their random walk and maintain their overall separation from the -Shld cells. To confirm this hypodissertation, we removed the contaminating -Shld parasites used as anchors by the +Shld gametocytes, and re-aligned the +Shld and reduced set of -Shld cells. After realignment, we found that scAlign "sacrificed" parasites from similar surrounding clusters to act as new anchors and preserve the distinct +Shld gametocytes as a distinct population (**Appendix: 3A.S15**).

Because the +Shld and -Shld cells form a set of clusters that we could order from early stage to late stage then gametocytes (+Shld), we hypothesized that the state variation map computed by scAlign could reveal where in the life cycle sexual-committing cells (a subset of +Shld cells) distinguished themselves in variation from asexual-committing cells (all -Shld cells). Using the interpolation component of scAlign, we projected each cell sequenced from each condition in the alignment space into the expression space of both of the +/-Shld conditions. By taking the difference in interpolated expression for each cell between the +Shld and -Shld transcriptomes, we computed a state variation map illustrating the predicted difference between the two conditions along the entire life cycle (**Fig. 2.7b**). From the state variation map, we observed few overall predicted differences in gene expression between the two conditions across most stages of the life cycle, except within a cluster of cells containing the gametocytes specific to the +Shld condition (**Fig. 2.7b**, cluster 13). In other words, gametocytes from cluster 13 exhibited the largest predicted differential gene expression between the +Shld gametocytes and neighboring -Shld non-gametocyte parasites. We verified that scAlign interpolation uses cells from neighboring clusters to predict -Shld expression within cluster 13 (**Fig. 2.7d,e,** see Methods).

Over all 661 highly variable genes we analyzed, we found the predicted differentially expressed genes in cluster 13 are enriched in genes previously established to play a role in

gametocyte maturation (**Fig. 2.7b**) ($p = 1.2 \times 10^{-6}$, Wilcox rank sum test), including *pfg27* (PF3D7_1302100) and *etramp4* (PF3D7_0423700)[66]. Furthermore, for the genes we predict to be upregulated in cluster 13 of the +Shld condition, we observed an enrichment of *ap2-g* targets identified via ChIP-Seq[65] ($p = 6.8 \times 10^{-7}$, Wilcox rank sum test). This upregulation of *ap2-g* targets is consistent with the fact that cells that have entered the gametocyte stage must have turned on *ap2-g* expression, but that Shld- cells cannot express *ap2-g*. Our state variation map identifies an additional eight genes not reported by Bancells and colleagues as playing a role in gametocyte maturation, but that are predicted to differ between +/-Shld (**Fig. 2.7c**). Taken in total, these results suggest the other genes we have predicted as differing between +/-Shld may also play a role in gametocyte conversion (**Fig 2.7b,c**).

2.4.2 Identification of highly variable genes in pancreatic islet cells sequenced using multiple protocols

We next tested scAlign's ability to infer an alignment space across more than two conditions by aligning pancreatic islet cells[36] derived from 8 donors and captured using four different protocols (CEL-Seq, CEL-Seq2, Smart-Seq2 and C1). The un-aligned pancreatic islet cells separate by protocol and not cell type, indicating strong protocol-specific effects which are removed after scAlign alignment (**Appendix: 3A.S16, 3A.S17a).** scAlign outperforms Seurat and scVI in terms of composite alignment accuracy on this dataset (**Appendix: 3A.S17b-c**). Interestingly, scAlign preserves the stellate, ductal and gamma cell types as separate clusters of cells, even though these three groups are represented in only a subset of the four protocols.

Having aligned the pancreatic islet cells into an alignment space, we next computed scAlign's state variance map to identify cell types and genes exhibiting high expression variation across three protocols to provide insight into how the choice of protocol affects gene expression

measurement (**Fig. 2.8a-d**). Here we excluded C1 because of the overall high gene expression specific to this protocol. We identified multiple subpopulations of cells within the alpha and beta cell types that are remarkably variable across protocols (**Fig. 2.8e**). We further show that our state variance map identifies subpopulations of alpha cells that are not consistent with the subclustering of alpha cells based on the embeddings (alignment space), illustrating that the state variance map finds unique patterns of expression variation across conditions not found by classic clustering approaches (**Fig. 2.8f**). Notably, the most highly variable genes with respect to protocol were specific to the activated stellate cells, and we confirmed these genes to be enriched in gene functions related to stellate function.

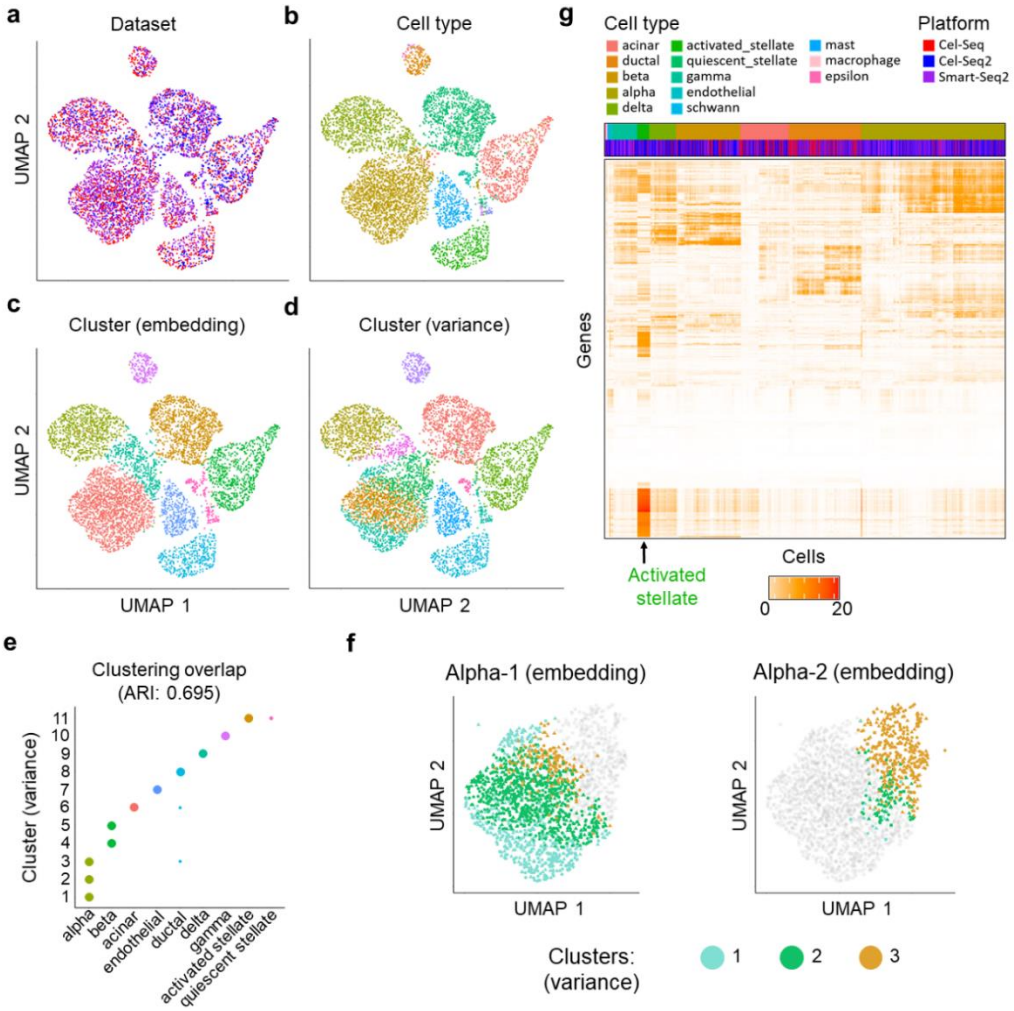


Figure 2.8: Alignment of pancreatic islet cells captured using three different protocols identifies cell type specific variation across protocols. (**a-d**) UMAP visualization of pancreatic islet cells sequenced on CEL-Seq, CEL-Seq2 and Smart-Seq2 after alignment by scAlign. colored by protocol, cell type, clustering on the alignment space or scAlign's state variance map. (**e**) Scatterplot indicating the overlap of clusters defined using the state variance map (y-axis) and based on the cell type labels as reported in Stuart et al. (**f**) Comparison of clusters identified using the embeddings, versus using the state variance map. Shown are two clusters defined in the embedding space, termed alpha-1 and alpha-2 because of their overlap with the alpha cell type. Grey points in the alpha-1 plot indicate cluster 2 cells, and grey points in the alpha-2 plot indicate cluster 1 cells. Colored points represent the three clusters identified in the state variance map. scAlign's variance map clusters (1, 2 and 3) are each found in both alpha-1 and alpha-2, indicating poor agreement. (**g**) Heatmap of the state variance map computed across the three capture protocols (CeL-Seq, CEL-Seq2 and Smart-Seq2) where red indicates high variance of expression predicted for a given gene and cell across protocols

2.4.3 Alignment of human and moues neuronal cells identifies conserved cell types and function

In collaboration with the Allen Institute, we used scAlign to perform a comparative analysis of conserved neuronal cell types between the human middle temporal gyrus (MTG) and mouse cortical regions including the primary visual cortex (VI) and the anterior lateral motor cortex (ALM). Briefly, matched cell types across species are assumed to share common expression patterns between orthologous genes which can be used to align common cell populations across species. Initial clustering of the integrated human and mouse data identified the major axis of variation to be the species-specific gene expression (**Fig. 2.9a**) which would confound any downstream analysis. To remove the primary effect of species we applied scAlign to align the human and mouse neurons with shared expression into a common representation while keeping species specific and rare cell populations such as Meis2 Adamts19 in mouse. Compared with Seurat, scAlign produced a more complete alignment indicated by an increased mixing of neurons between species (**Fig 2.9b**).

With the human and mouse neurons now clustered together, we were able to obtain cell-type homologies based on shared cluster membership along with confidence scores via bootstrapping (**Fig 2.9c**). The homology analysis identified human and mouse inhibitory neurons with both 1-to-many and 1-to-1 relationships, the latter of which enabled prediction of cellular properties in the human cortex from prior mouse annotations. Notably, the rare and distinct neuronal types per species remained distinct through-out the analysis enabled by alignment with scAlign ensuring that incorrect associations were not identified. Overall, the alignment of neurons from human and mouse revealed a conservation of cellular architecture at the resolution neurons with high specialized function and distinction.

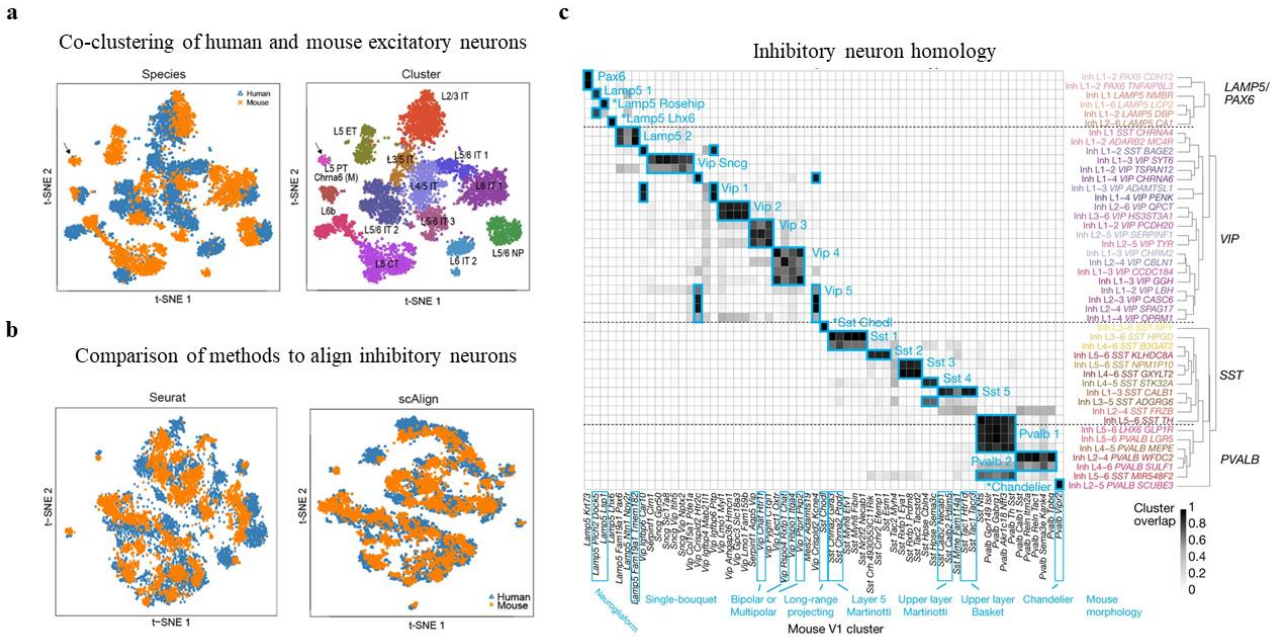


Figure 2.9: Alignment and homologies for cell types in human and mouse. (**a**) t-SNE visualization of human and mouse excitatory cell types after PCA but prior to alignment. (**b**) t-SNE visualization of human and mouse inhibitory cell types post alignment with Seurat and scAlign. The visualization are colored similarly such that human neurons are blue and mouse neurons are orange. (**c**) Human and mouse cell type homologies for the inhibitory neurons where increased overlap between species is defined by a darker color on the heatmap.

2.5 Stability of the scAlign model

Alignment methods seek to remove batch effects in a manner that adheres to a common set of goals that should not be violated in the ideal: (1) Conserve biological variation within and cross dataset, e.g. cells of the same type align together and cell types remain distinct. (2) Preserve cell populations which are unique to individual studies. (3) Produce a least effort alignment without warping the data such that biological conclusions cannot be drawn. Adherence to each of these goals is essential to ensure insights gained from aligned data are reflective of the underlying biology however many methods violate one or more during alignment. We aim to show that scAlign is robust to edge scenarios that lead to violation of these rules and can produce accurate alignments of complex data.

2.5.1 scAlign is robust to large differences in cell type representation across conditions

Besides cell type-specific responses to stimuli, we reasoned that the other factor that determines alignment difficulty is the difference in the representation (or proportion present) of each cell type across conditions. For example, cell types unique to one condition may pose challenges to alignment because there are no functionally matched cell types in the other conditions. We therefore explored the behavior of scAlign and other approaches when the relative proportion of cell types varies significantly between the conditions being aligned.

We performed a series of experiments on the Kowalczyk et al. benchmark where we measured alignment performance of all methods as we removed an increasing proportion of cells from each cell type from the old mouse condition (**Fig. 2.10**). While scAlign had superior performance across all experiments and was most robust to varying cell type proportions,

surprisingly, we found that other methods were generally robust as well. Removing even 75% of the cells of a given type only led to a median drop of 11% in accuracy across the tested methods. When we repeated these experiments on the Mann et al. benchmark, we generally found a larger decrease in performance as we removed more cells from each type compared to the Kowalczyk et al. benchmark, though scAlign still outperformed all other methods (**Appendix: 3A.S10**).

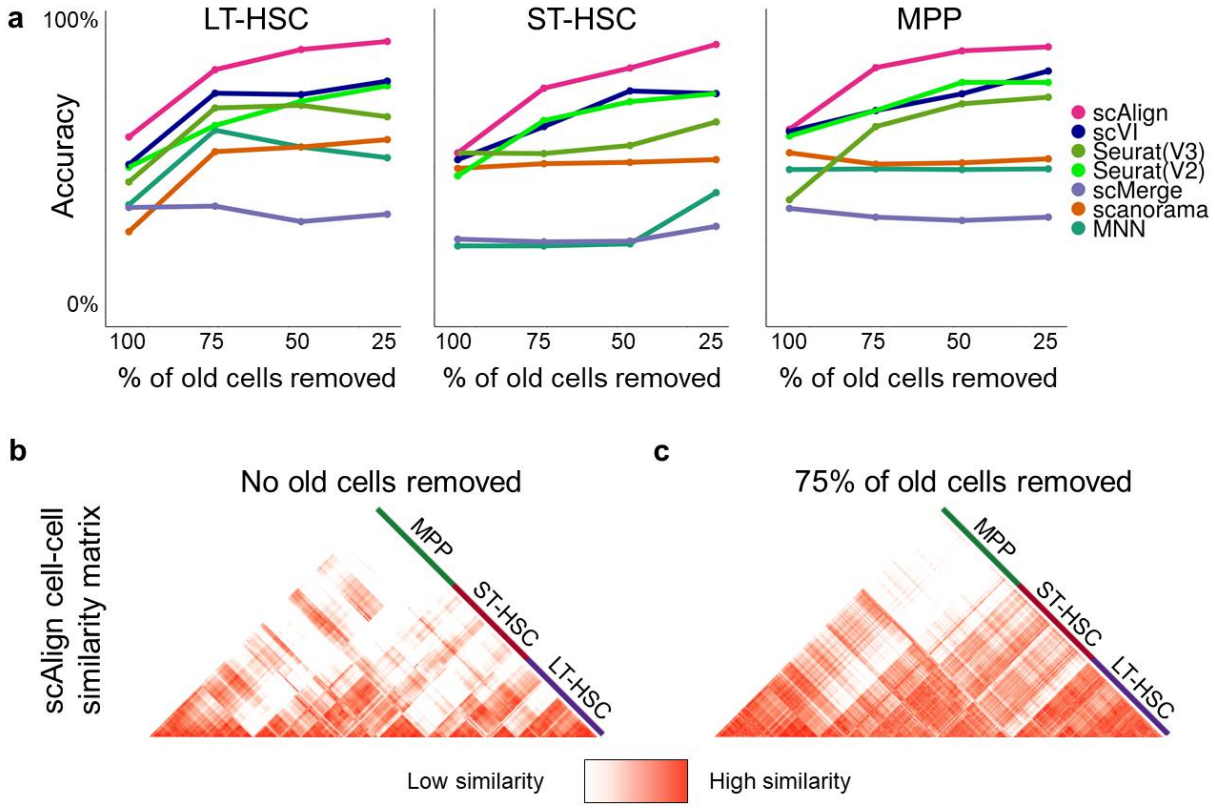


Figure 2.10: Alignment performance is robust to imbalance in cell type representation across conditions. (**a**) Accuracy of classifiers on the Kowalczyk et al. benchmark, when removing either LT-HSC, ST-HSC or MPP cells from the old condition. scAlign outperforms all other methods and exhibits minimal degradation in performance as increasing numbers of cells are removed within each cell type. (**b**) Heatmap showing the pairwise similarity matrix for the young cells from Kowalczyk et al. when no cells have been removed. (**c**) Heatmap showing the pairwise similarity matrix for the young cells from Kowalczyk et al. after removing 25% of the old mouse cells from all cell types.

We next investigated the factors that underlie scAlign's robustness to imbalanced cell type representation across conditions. scAlign optimizes an objective function that minimizes the difference between the pairwise cell-cell similarity matrix in gene expression space, and the pairwise cell-cell similarity matrix implied in the alignment space when performing random walks of length two (**Fig. 2.11a**). The random walk starts with a cell sequenced in one condition, then moves to a cell sequenced in the other condition based on proximity in alignment space. The walk then returns to a different cell (excluding the starting cell) in the original condition, also based on proximity in alignment space. For every cell in each condition, we calculated the frequency that such random walks (initiated from the other condition) pass through it (**Fig. 2.11b-c**). We found that a select few representatives for each cell type are visited much more frequently than others, and that even when those cells are removed from the condition, another cell is automatically selected as a replacement (**Appendix: 3A.S11**). This suggests that a given cell type in one condition only depends on a few cells of the same type in the other condition to align properly, and so scAlign alignment does not need every cell type to be represented in the same proportion across conditions.

In the above experiments, we have aligned conditions in which the same set of cell types are present in all conditions. We next explored the behavior of scAlign and other approaches when there are cell types represented in only a subset of the conditions. We expect such scenarios to arise when only a subset of cell types respond to, or are targeted by, a stimulus or condition. For each of our benchmarks, we removed one cell type from one of the conditions (e.g. the LPS condition of the Mann benchmark, or the old mouse condition of the Kowalczyk benchmark), and aligned the control and stimulated conditions to determine the extent to which the unique population maintained separation from other cell types after alignment.

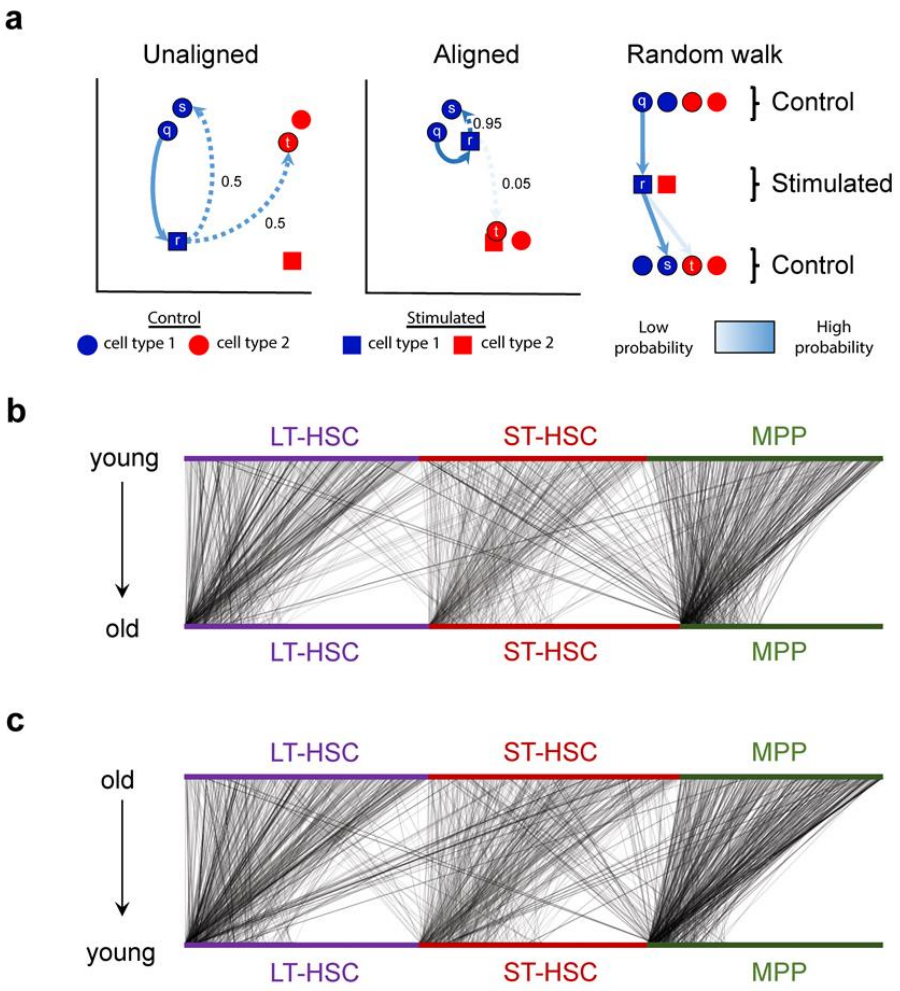


Figure 2.11. Random walks during scAlign training frequently visit a small number of hub cells. (**a**) Schematic of the cross condition round trip random walk prior to and after training of scAlign. (**b**) Visualization of the probability of a walk from each individual young cell (top) to each individual old cell (bottom) after training scAlign on the Kowalczyk et al. benchmark. Edge density represents the magnitude of the probability of a given walk. (**b**) Same as (a), except the edges represent the probability of walking from individual old cells (top) to individual young cells (bottom) in the Kowalczyk et al. benchmark.

Figure 2.12a demonstrates that in eight out of nine cases, scAlign outperforms other alignment methods in terms of classification accuracy. Even in cases where the alignment accuracy was similar between methods, scAlign visually separates cell types in its alignment space more so than

other approaches such as Scanorama and Seurat (**Fig. 2.12b**). For other approaches, the separation of different cell types within the same condition shrinks when one cell type is removed (**Appendix: 3A.S12**).

2.5.2 Robust cell type marker genes drive alignment

To gain insight into the general principles and genes used by scAlign to perform alignment, we performed a series of *in silico* expression perturbation experiments. scAlign uses the same feed-forward network to reduce the dimensionality of cells from all input conditions. We therefore hypothesized that scAlign is implicitly identifying cell type marker genes that are invariant (robust) across conditions, and using these marker genes to perform dimensionality reduction as they will naturally cause similar cell types across conditions to map to the same regions of alignment space. We tested this hypodissertation by first identifying a set of marker genes for each cell type that were robust across conditions within a given dataset. We then systematically perturbed the expression of all common marker genes across all cells, and measured the downstream effect of the perturbation on the embeddings of the cells in alignment space. Intuitively, perturbing the expression levels of genes that more strongly contribute to the alignment will yield larger deviations in the embeddings of the cells. As a control, we performed the same perturbation experiments on random control sets of genes matched for size and expression level . Perturbing the common marker genes yielded significantly larger deviations in the cell embeddings than the control sets ($P < 10^{-4}$, Permutation test), with the embeddings moving an average of 5.2 fold more than the control sets (**Appendix: 3A.S18**).

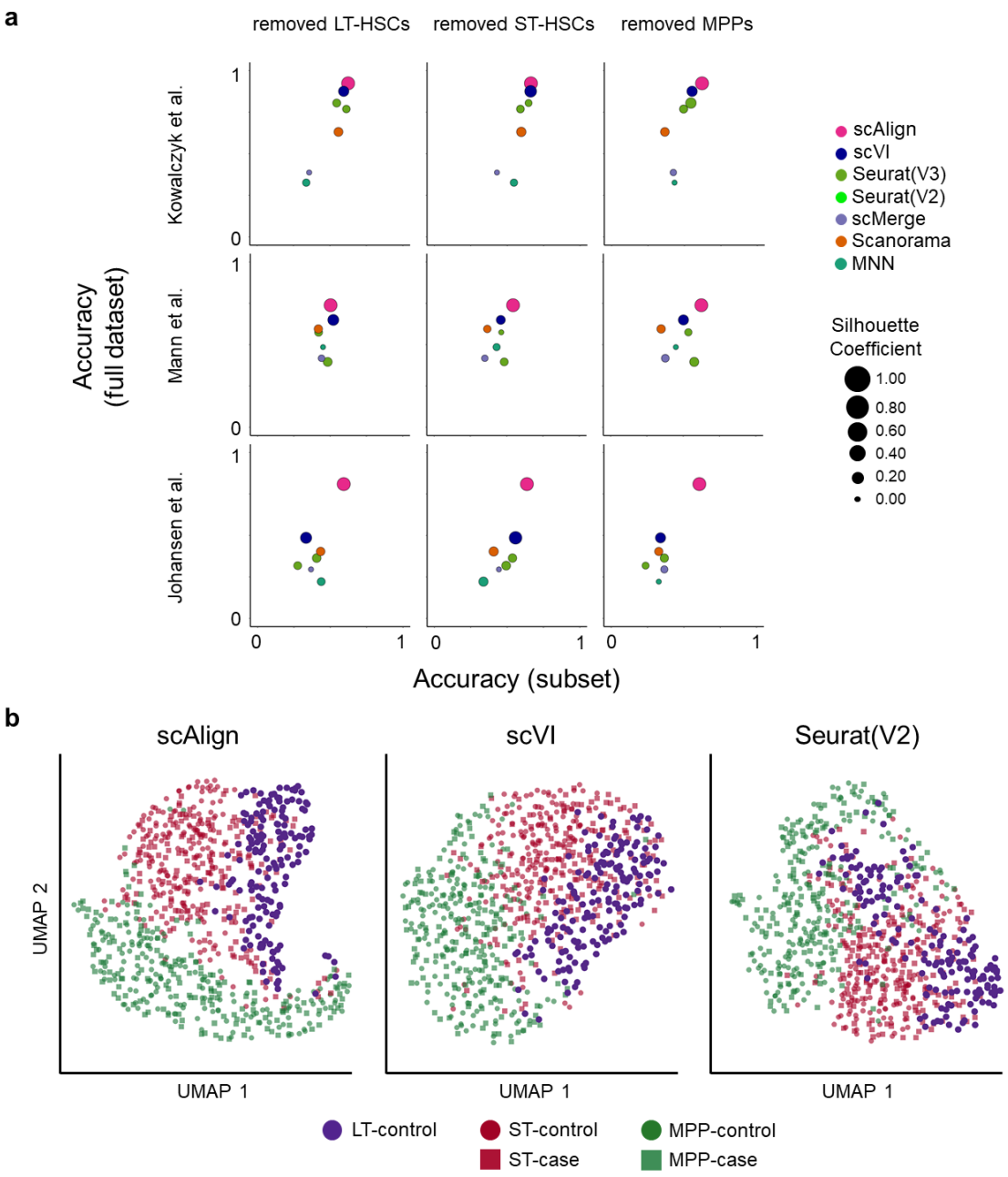


Figure 2.12: scAlign is robust to distinct cell type sets between conditions. (**a**) Scatterplot matrix of performance of each method when both conditions have the same number of cell types (y-axis), compared to when one cell type has been removed (the LPS condition of the Mann benchmark, or the old mouse condition of the Kowalczyk benchmark) (x-axis). Each point is scaled in size by the silhouette coefficient for the clustering after alignment. (**b**) tSNE plots with cells colored by cell type and condition for the top performing methods.

We next sought to evaluate the extent to which scAlign's unique random walk-based objective function contributes to its alignment accuracy. Traditional neural networks that focus on unsupervised dimensionality reduction such as autoencoders use an objective function that explicitly learn embeddings that minimize the reconstruction loss of each cell. In contrast, the scAlign objective function simultaneously encourages embeddings to maintain cell-cell similarity within condition, as well as match cells in the alignment space across conditions. We therefore evaluated the utility of scAlign's objective function by substituting scAlign's loss function for a classic reconstruction loss-based autoencoder loss function. This autoencoder shares the same number of layers and nodes per layer as scAlign, and furthermore uses a shared encoder across all conditions similar to scAlign, but unique decoders for each condition . Both the autoencoder and scAlign therefore have the same number of parameters and therefore equal model capacity, and only differ by their respective objective functions. When comparing this autoencoder to scAlign on each of our four benchmarks, we found that the autoencoder was able to achieve similar accuracy on benchmarks with minimal cell type-specific condition effects, such as Cellbench and Kowalczyk et al. (**Appendix: 3A.S19a-b**). However, on more challenging benchmarks such as Mann et al. and our HeterogeneousBenchmark, the autoencoder performed worse than scAlign (**Appendix: 3A.S19c-d**). Furthermore, the autoencoder did not maintain the cell-cell similarity matrix in embedding space as well as scAlign (**Appendix: 3A.S19, 3A.S20**), suggesting the low dimensional embeddings learned by the autoencoder may not as faithfully recapitulate the gene expression inputs.

2.6 Discussion

We have shown that scAlign outperforms other alignment approaches, particularly when there are strong cell type-specific differences across conditions, or when there is an imbalance in cell type representation across conditions. Compared to other approaches, scAlign will be particularly useful in the context where only some cell type labels are available in one or more conditions. We envision two scenarios where this may occur. First, with the increasing number of cell atlases[44–49] that are accurately labeled by domain experts and are now publicly available, scAlign can take advantage of the accurate labeling of these atlases to annotate new datasets that lack labels. Second, marker genes may be available for only a subset of cell types such as specific hematopoietic cells, in which case only a subset of cells may be reliably labeled. Even when marker genes are available, markers may not be unique to individual cell types and technical factors such as dropout may prevent truly expressed markers from being detected in the RNA. Here, scAlign can be used in conjunction with only the most confident labeled cells, or can even be used when there is overlapping labels (due to marker uncertainty).

Another advantage of scAlign over other alignment methods is the improved ability to detect rare differential expression events between conditions. For typical alignment methods, once the effect of condition is removed via alignment, cells must still be clustered into putative cell types in order to identify which cells match across condition, and then perform an unpaired differential expression test within each cluster to identify condition-specific differences. The need to cluster cells means the detection of rare cell types can be highly sensitive to the choice of clustering algorithm or parameters. In contrast, through interpolation, scAlign predicts how each individual cell within the alignment space differs in expression between any of the input

conditions, effectively performing a paired (or matched) differential expression calculation per-cell without the need to cluster. The result is scAlign can detect the presence of rare cell populations that differ in expression across conditions (**Fig. 2.7**).

scAlign implements two approaches to aligning more than two conditions simultaneously. In the reference-based alignment, a single reference condition is established and all other conditions are be aligned against the reference (**Appendix: 3A.S21**). This is expected to work well when all cell types are represented in all conditions, and has the benefit of speed. Alternatively, the all-pairs alignment mode performs an all-pairwise set of alignments simultaneously, which will be more robust to the presence of cell types only represented in a subset of the conditions.

The general design of scAlign's neural network architecture and loss function makes it agnostic to the input RNA-seq data representation. Thus, the input data can either be gene-level counts, transformations of those gene level counts or the result of a preliminary step of dimensionality reduction such as principal component scores or canonical correlation vectors. In our study, we first transformed data into a relatively large number of principal component scores before input into scAlign, as this yielded much faster run times with little to no performance degradation. The improvement in computation time due to PCA pre-processing of the input data allowed scAlign to both converge more quickly and become feasible on a CPU-based system, therefore making scAlign a broadly applicable deep learning method. More generally, the design of scAlign's neural network architecture and loss functions are general and not specific to scRNA-seq data. We therefore expect that scAlign should be applicable to any problem in which the study design consists of comparing two or more groups of unmatched samples, and where we expect there to be subpopulations of individuals within each group.

Here we have primarily compared scAlign against unsupervised alignment methods. In our supervised alignment results, scAlign compared favorably against the supervised methods MINT[3] and scmap[52] when assuming all cells are labeled. In the context of alignment, however, we reasoned that if a complete set of labels are available for all cells and conditions, then addressing the task of alignment is less useful, because cells of the same type across conditions can be directly compared via per-cell type differential expression analysis without alignment. Alternatively, in those contexts, each matching pair of cell types across conditions can be independently aligned using the unsupervised scAlign (or other unsupervised methods)_to identify matching subpopulations of cells.

The tasks of transcriptional alignment and batch correction of scRNA-seq data are intimately related, as one can view the biological condition of a cell as a batch whose effect should be removed before integrated data analysis. Compared to batch correction methods, scAlign leverages the flexibility of neural networks to perform alignment where cell states might exhibit heterogeneous responses to stimuli, yet through interpolation provides the interpretability that canonical batch correction methods enjoy.

Like all other supervised and unsupervised alignment methods, scAlign makes an underlying assumption that the two or more conditions used as input make sense biologically to align. That is, alignment methods assume that there are at least some common cell types between conditions that share some functional origin or similarity, that should be matched across conditions, even if they differ in state (e.g. expression) due to condition or stimulus. To the best of our knowledge, there is no procedure or strategy for identifying datasets that should not be aligned due to lack of matching cell types. As a result, any alignment method when applied to datasets which contain unrelated or dissimilar cell types can potentially lead to false positive matchings.

This limitation is not specific to alignment methods; scRNA-seq analysis tools designed for other purposes, such as trajectory inference, assume that a trajectory exists in the input data in the first place, and will return a trajectory regardless of whether it makes sense to do so. scRNA-seq tools in general are useful for generating hypotheses (in the case of alignment, hypotheses about which cell types match across conditions, and how they differ), but need to be used cautiously by downstream users.

A related concern is the performance of alignment methods when there exists condition-specific cell types that have no matching cell type in another condition. In our experiments, we show that scAlign outperforms other alignment methods in this scenario by choosing a small number of cells from a matching cell type, and placing those small numbers of cells in the same region of alignment space as the condition-specific cell type; in other words, scAlign purposefully misaligns a small number of cells. scAlign tends to sacrifice a small number of cells because its objective function minimizes the distortion of the cell-cell pairwise similarity matrix within each input condition, and so sacrificing many cells would lead to a large distortion of the pairwise similarity matrix.

As a neural network-based method, scAlign usage requires specification of the network architecture before training, defined by the number of layers and number of nodes per layer. In our results, we have shown scAlign is largely robust to the size of the architecture, in part because in addition to the ridge penalty we apply to the weights of the network, our objective function minimizes the difference between the similarity matrix in the original expression and alignment spaces, which also acts as a form of data driven regularization.

2.7 Chapter appendix

Measuring accuracy of pairwise alignments. Alignment performance for each method was measured as a weighted combination of cross-condition label prediction accuracy and alignment score[5]. The cross-condition label prediction was performed by training a classifier to label one condition (stimulated condition by default) using only labels from the corresponding control condition. Specifically, a K-nearest neighbors classifier from the R library 'class' was initialized with control cell embeddings after alignment, along with their corresponding cell type labels. The classifier was then used to predict labels for the stimulated cells. The predicted labels were compared against heldout labels to measure accuracy. The final score $\text{accuracy}_{\text{composite}}$ is defined by the product of the classifier accuracy and alignment score.

Measuring accuracy of multi-way (three or more) alignments. Similarly, to measure alignment performance on the alignment of three or more conditions, we measured the weighted combination of a representative-based label prediction accuracy and alignment score. The representative-based label prediction was performed by iteratively treating each condition as the representative, and training a classifier to label cells from all non-representative conditions using only labels and cells from the single representative condition. The mean accuracy was computed for all condition specific label predictions as the final accuracy. As a classifier, we chose a K-nearest neighbors classifier from the R library 'class' and initialized it with the representative condition cell embeddings after alignment, along with their corresponding cell type labels. The classifier was then used to predict labels for all the non-representative cells. The predicted labels were compared against heldout labels to measure accuracy. The final score $\text{accuracy}_{\text{composite}}$ is defined by the product of the mean accuracy and alignment score.

Measuring accuracy of transcriptional interpolation. To measure interpolation accuracy, we measured the ability of a classifier trained on the gene expression data of the cells measured under one condition to correctly label interpolated gene expression profiles of cells sequenced under the other condition (but interpolated into the current condition). A K-nearest neighbors classifier from the R library 'class' was initialized with 90% of expression data and tested on the remaining heldout set of 10% to define gene expression specific accuracy. The classifier was then used to predict the labels for cells represented by interpolated gene expression values to compute an interpolation specific accuracy. 10-fold cross validation was performed using this procedure and the average accuracy was reported.

Measuring the deviation in cell embeddings by *in silico* gene set perturbation. To determine the importance of a single gene set to scAlign's calculation of the embedded representation of each cell, we zeroed out the expression measurements of all genes in the gene set across all cells. We then measure the median and maximum change (using Euclidean distance) in cell embeddings before and after zero-ing out the expression measurements. To compute a P-value, we generated random gene sets of the same size and matched for expression levels of the genes, and calculated the number of random gene sets that yielded a deviation at least as large as what we observed for a gene set of interest.

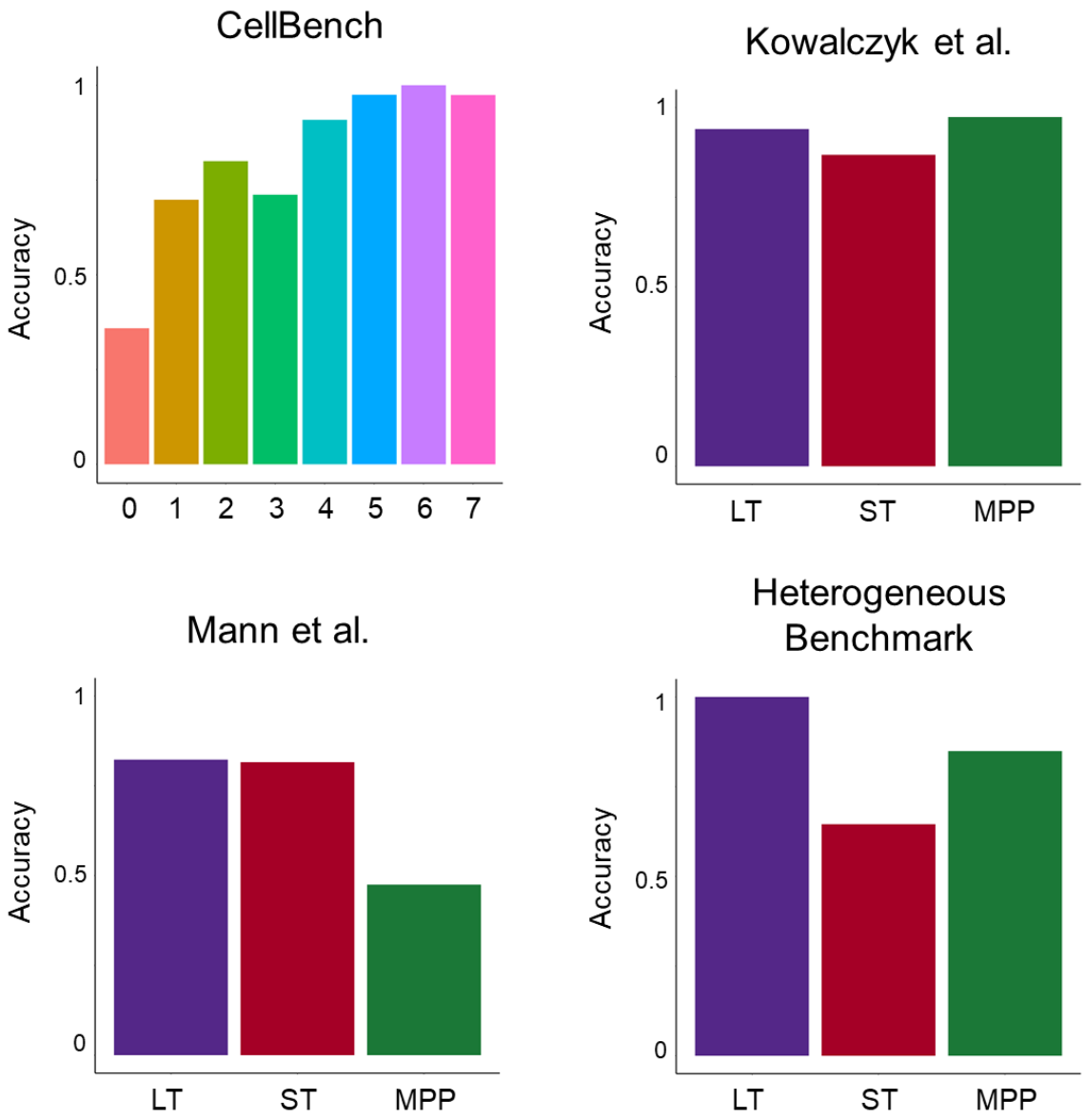


Fig. 3A.S1: Cell type-specific accuracy for each of the four benchmark datasets. Bar plots indicate the accuracy with respect to each cell type.

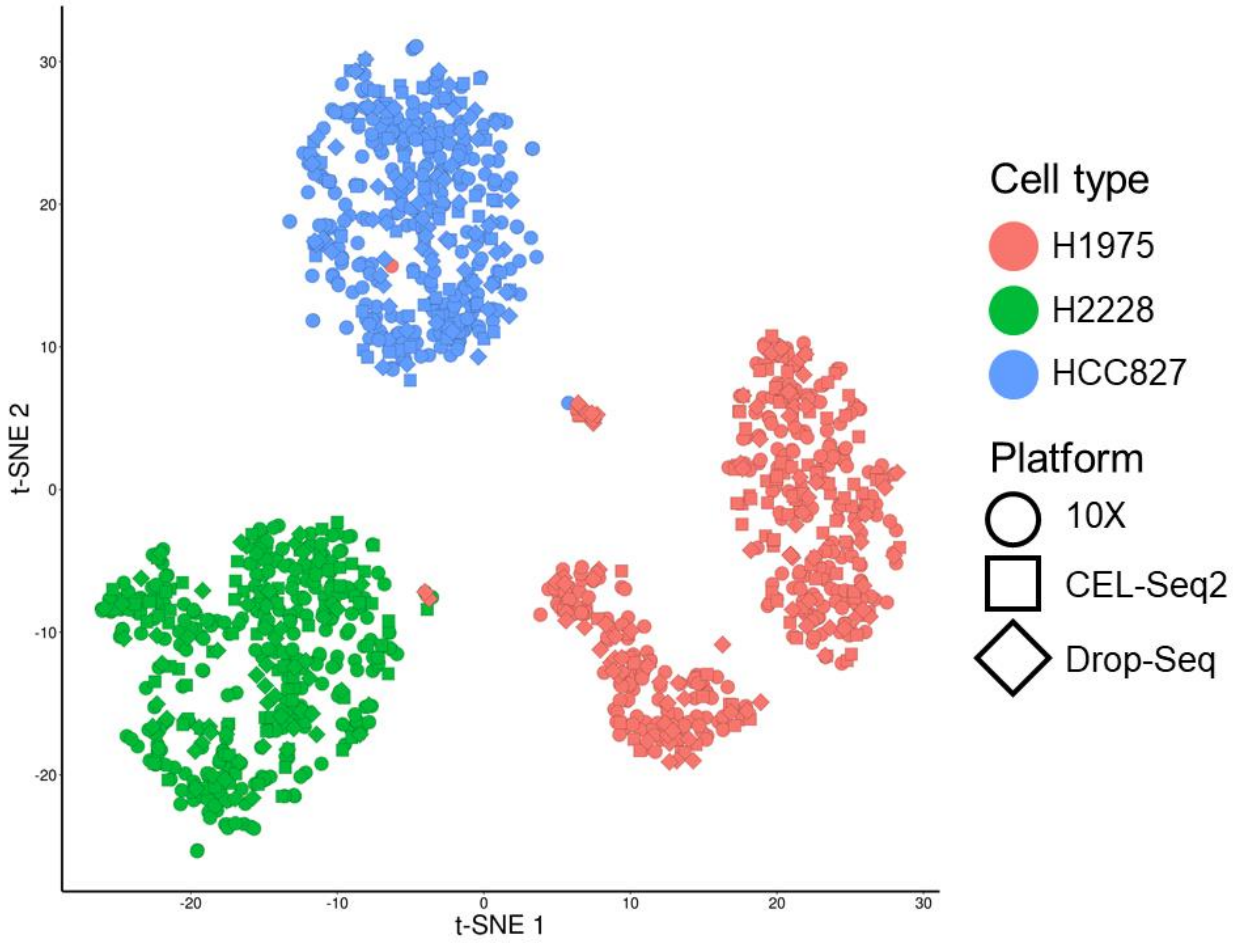


Fig. 3A.S2: Standard normalization procedures align the same cell types sequenced using different protocols by CellBench. Scatterplot of the three human lung adenocarcinoma cell lines sequenced by CellBench using either 10X Chromium, CEL-Seq2 or Drop-Seq, and normalized independently using Seurat.

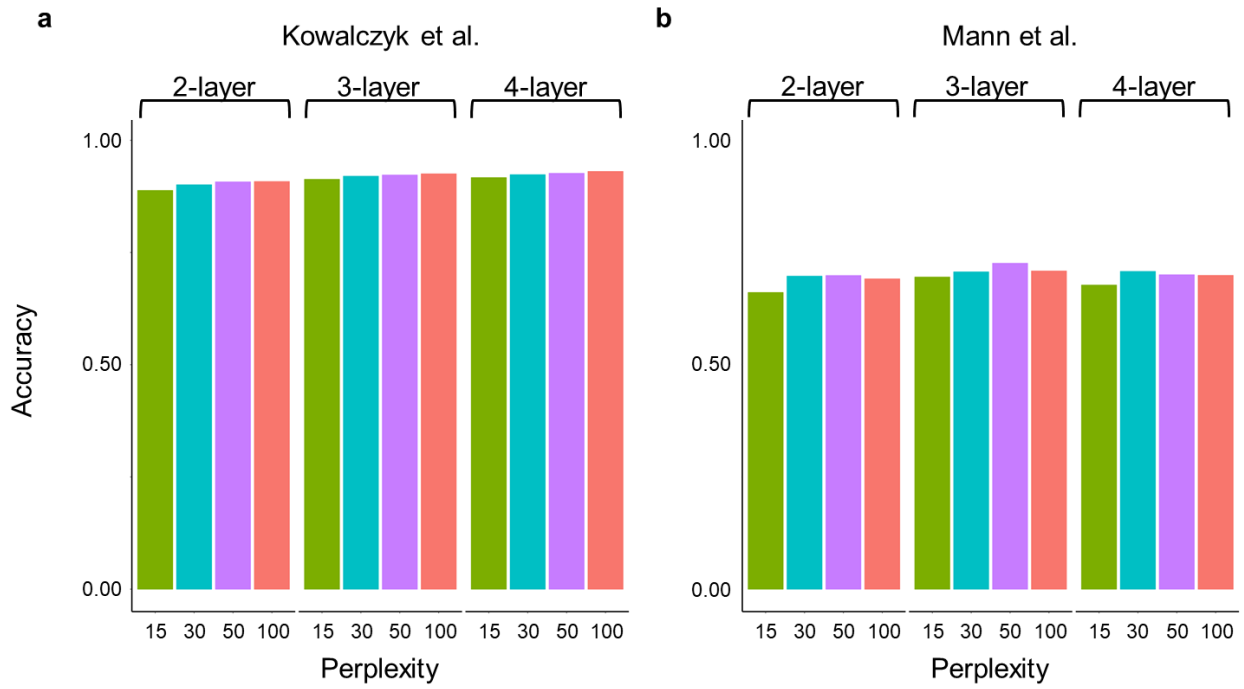


Fig. 3A.S3: Robustness of scAlign to neural network architecture and size. Barplots indicate accuracy of different network architectures, hidden layers, and perplexity parameter settings (x-axis) for scAlign trained on the Kowalcyk et al. and Mann et al. benchmarks.

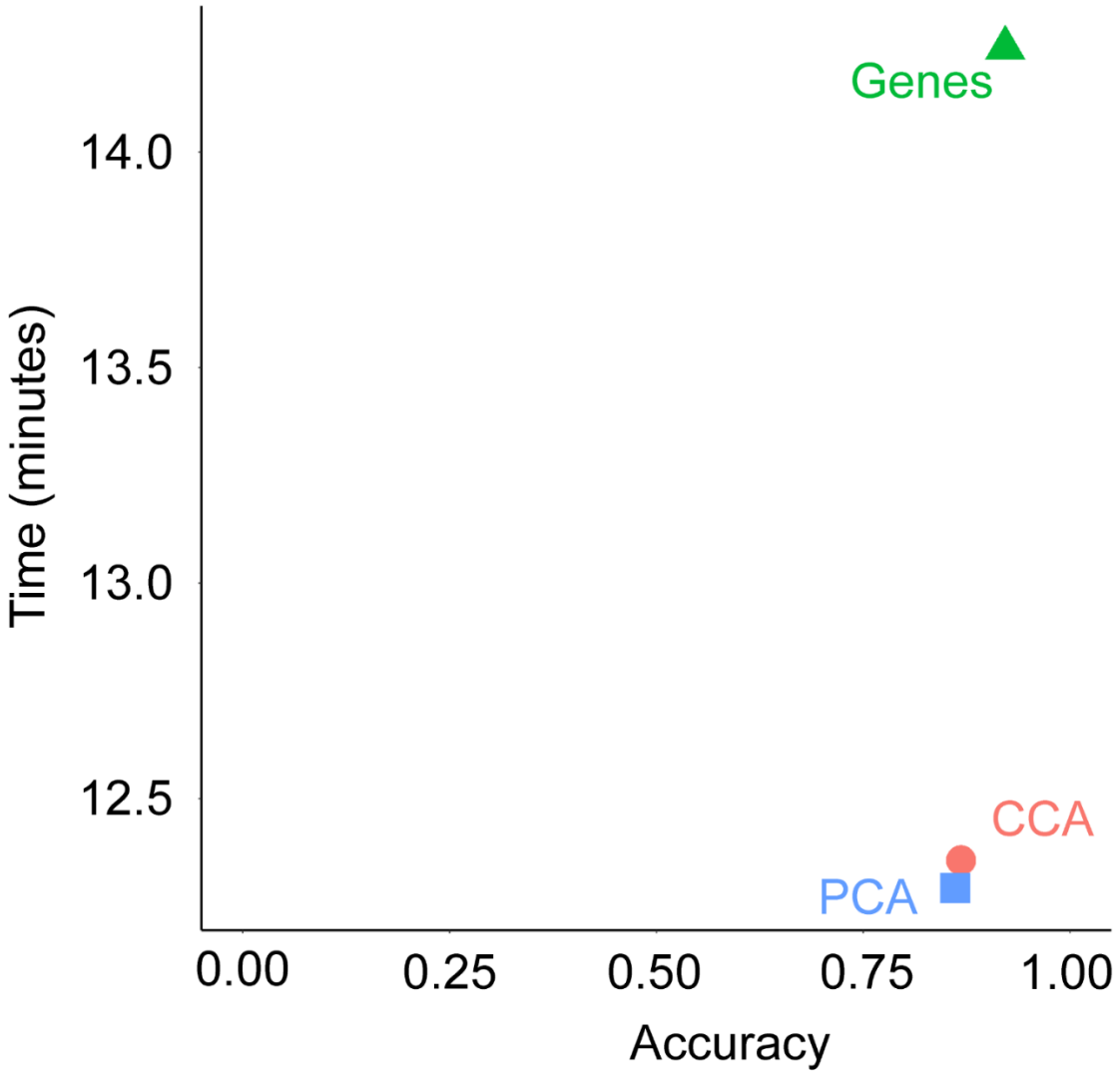


Fig. 3A.S4: Initial data preprocessing step of dimensionality reduction increases computation speed without degrading accuracy or alignment. Scatterplot indicates the computation time on the y-axis and cross condition classification accuracy on the x-axis after training scAlign with the top 3,000 variant genes, 10 CCs or 20 PCs.

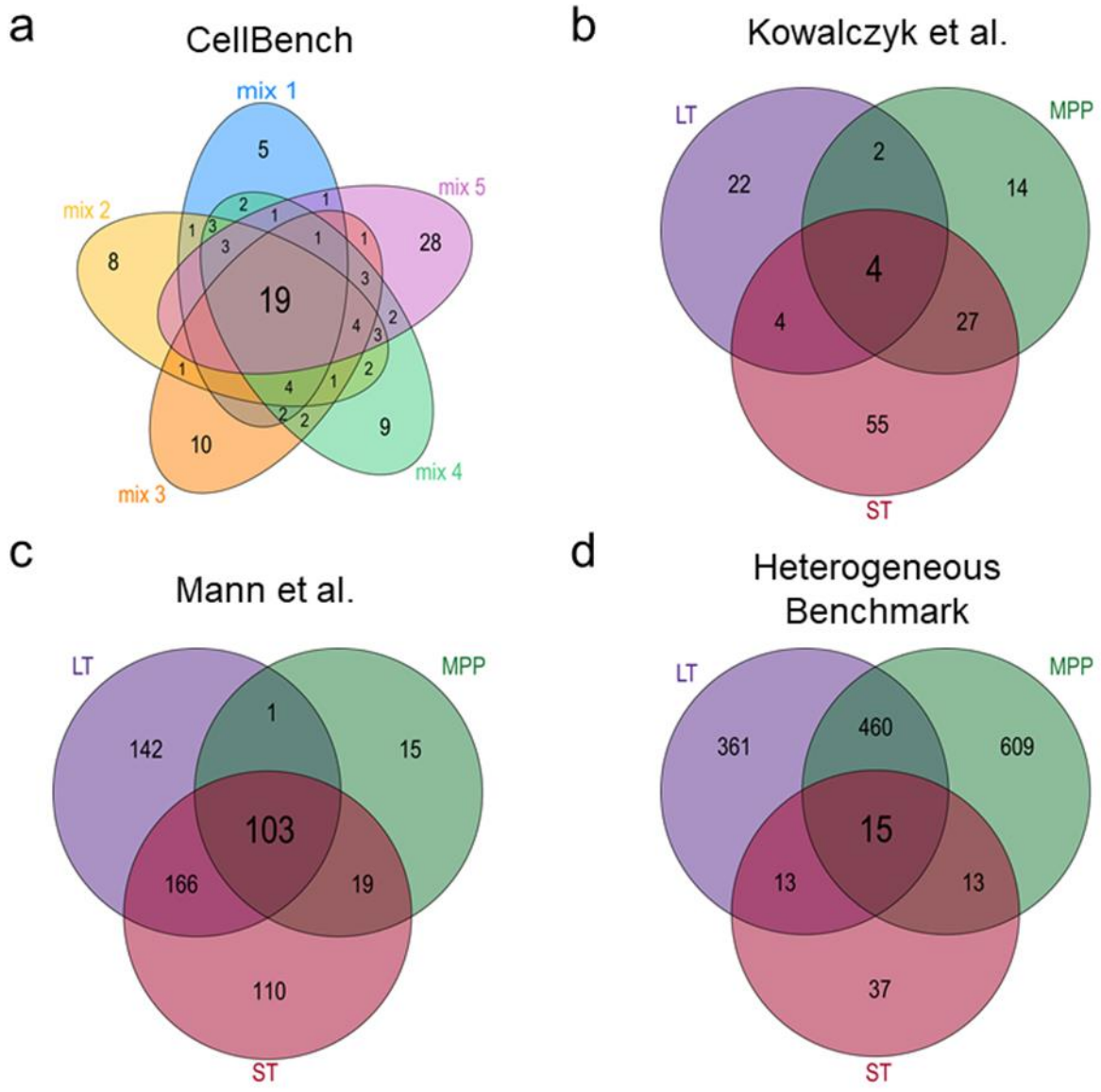


Fig. 3A.S5: Overlap of differentially expressed genes for the four benchmark datasets. (a-d) Venn diagram indicating the number of overlapping DEGs (measured across condition) for each cell type in the four benchmarks.

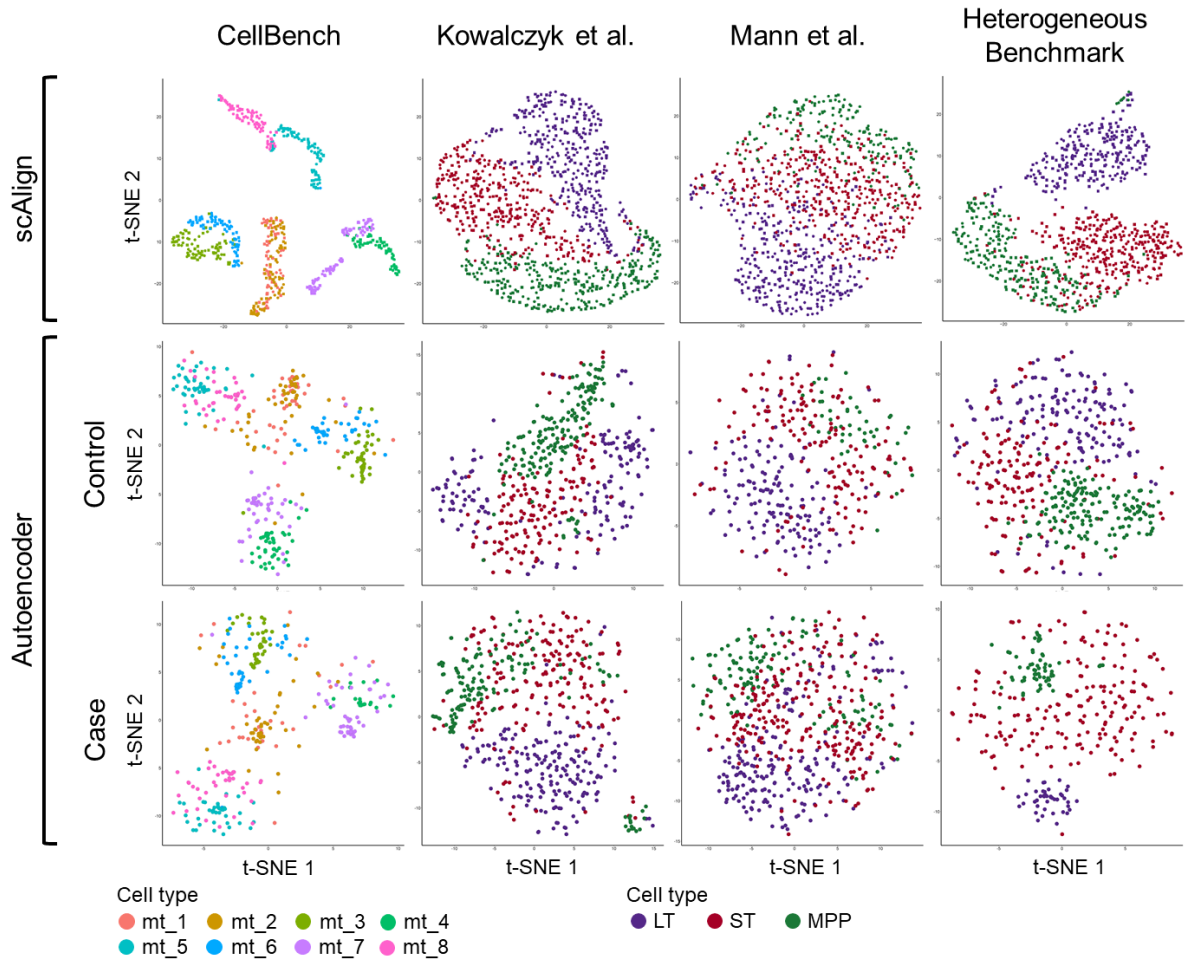


Fig. 3A.S6: Comparison of the embeddings learned by scAlign and an autoencoder. t-SNE visualization of the embeddings after training either scAlign or a condition-specific autoencoder on each of the four benchmarks.

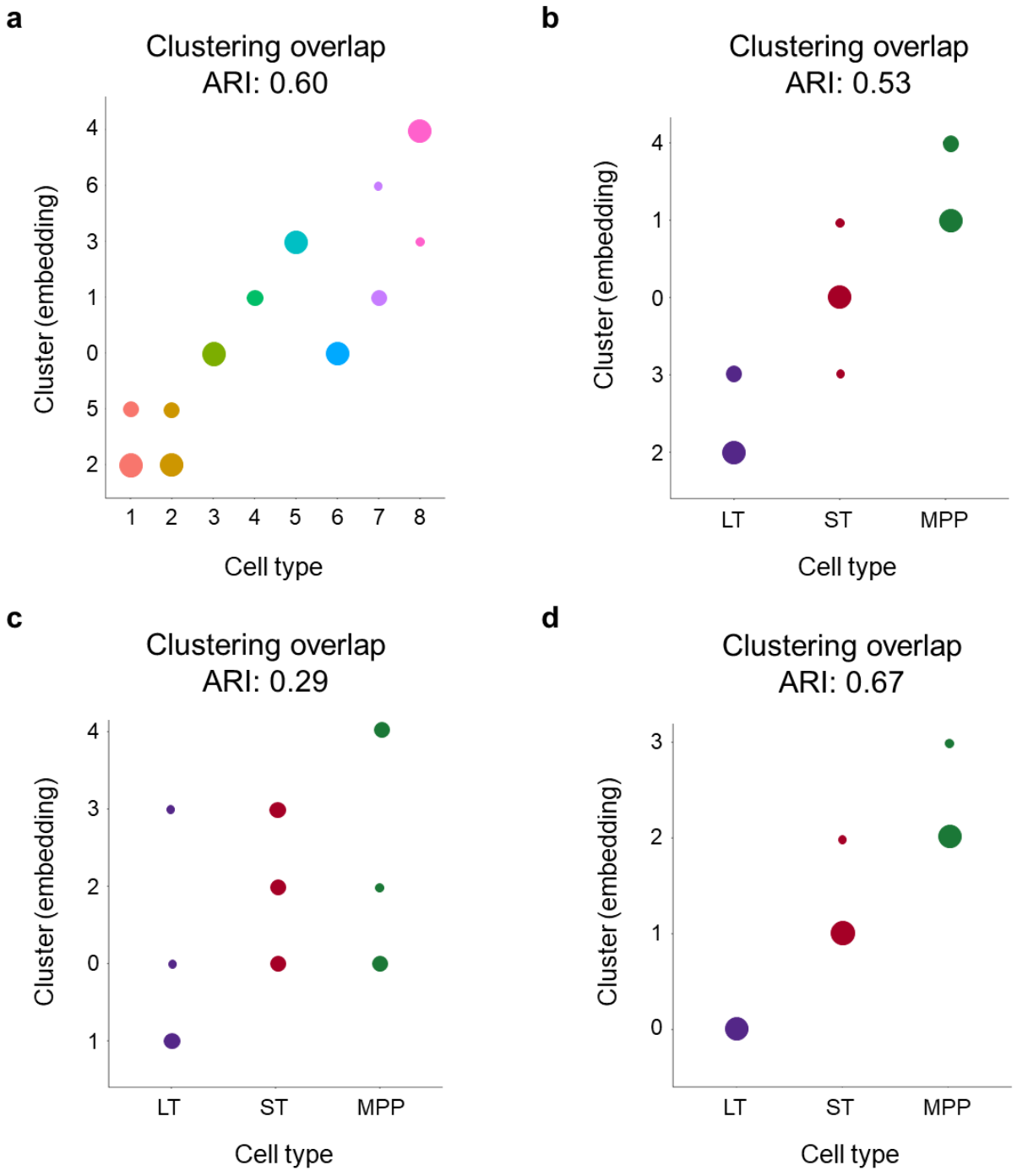


Fig. 3A.S7: Comparison of clustering in scAlign's alignment space and known cell type labels. (**a**) Dot plot visualizes the amount of overlap between de novo clustering and cell type annotations for CellBench. (b-d) Similarly to (a) but for Kowalczyk et al., Mann et al. and HeterogeneousBenchmark, respectively.

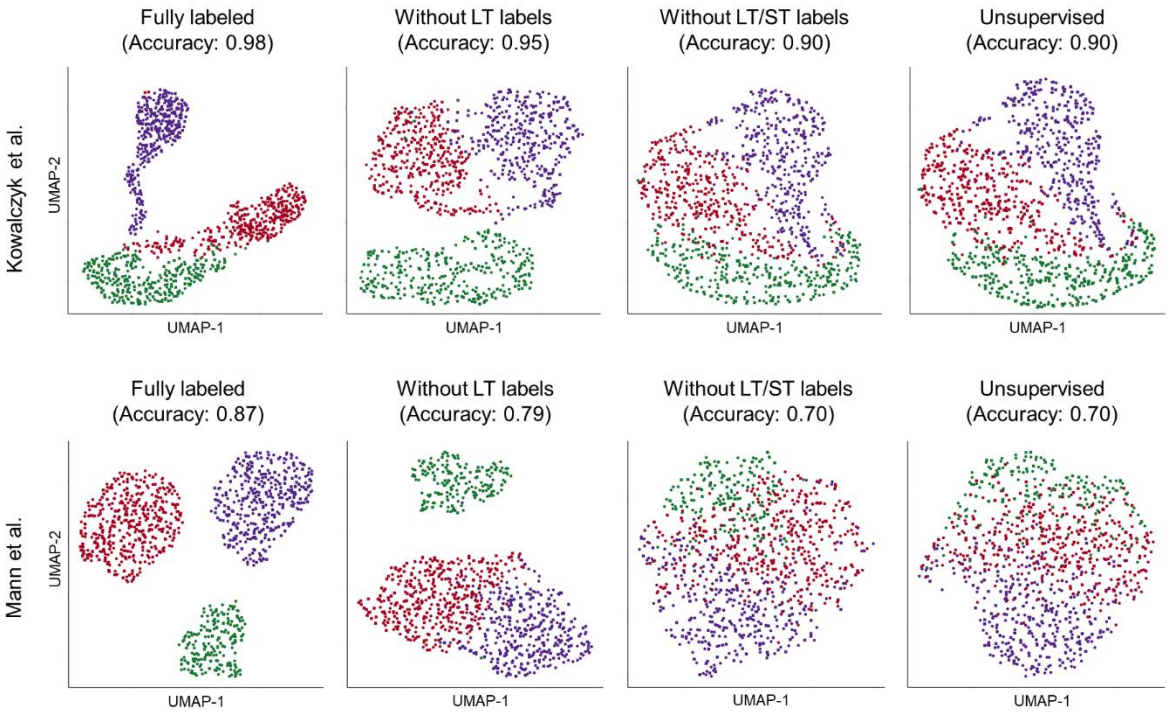


Fig. 3A.S8: Partial labels yield performance between fully labeled and no-label data. UMAP visualization of scAlign alignment of Kowalczyk et al. and Mann et al. after training with complete cell label information or after removing either the LT or LT and ST condition specific cells.

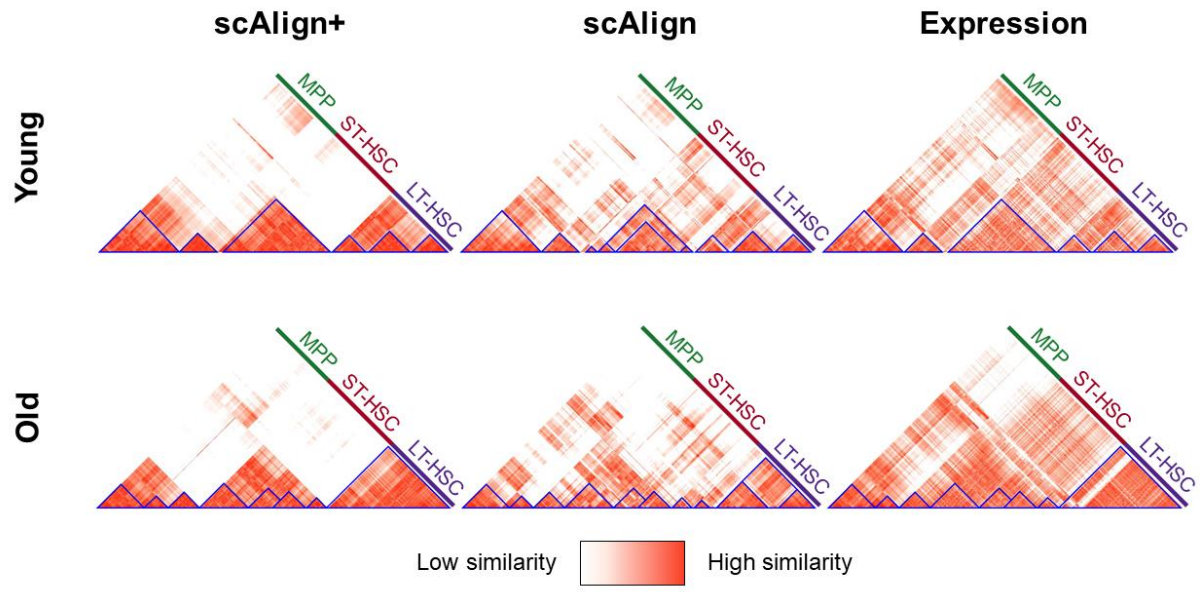


Fig. 3A.S9: Cell-cell similarity matrix after supervised or unsupervised training of scAlign compared to the gene expression-based similarity matrix. Heatmaps of cell-cell similarity illustrate the agreement between training scAlign+ with all cell type labels (supervised) or scAlign without any cell type labels (unsupervised) for both the young and old cells in Kowalczyk et al., or when directly measuring similarity in gene expression space. Clusters of cells are highlighted within and across each heatmap by blue triangles.

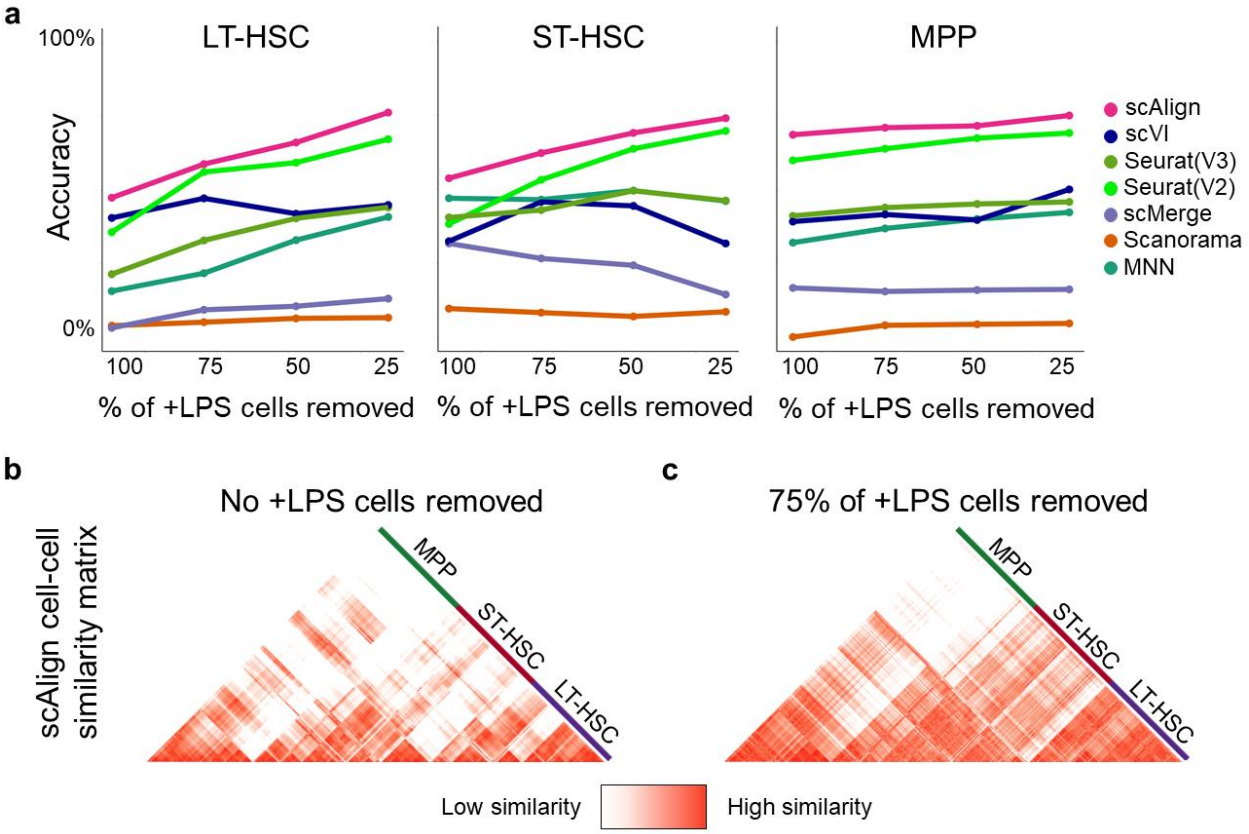


Fig. 3A.S10: Performance of alignment methods on the Mann et al. benchmark after removing cells. (**a**) Accuracy of classifiers on the Mann et al. benchmark, when removing a percentage of either LT-HSC, ST-HSC or MPP cells from the old condition. scAlign outperforms all other methods robustly on the full dataset, and most methods perform worse as more cells are removed from the old condition. (**b**) Heatmap showing the pairwise similarity matrix for the stimulated (+LPS) cells from Mann et al. when no cells have been removed. (**c**) Heatmap showing the pairwise similarity matrix for the stimulated (+LPS) cells from Mann et al. after keeping only 75% of the target cells from all cell types.

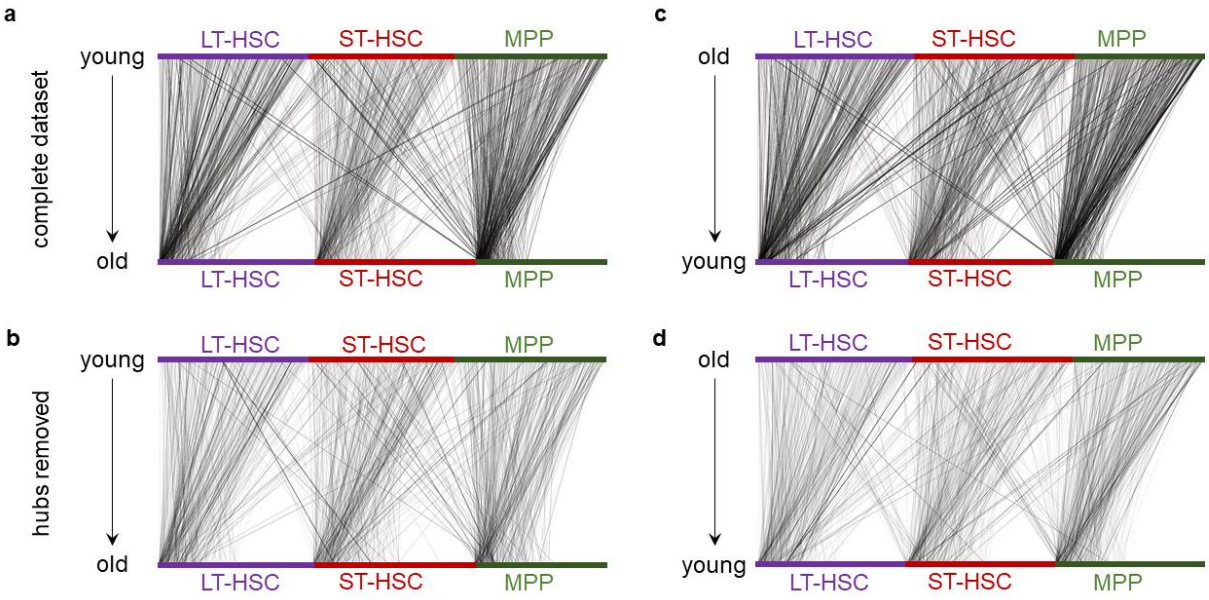


Fig. 3A.S11: Random walk probabilities measured before and after removing cells frequently visited on random walks during scAlign training on Kowalczyk et al. (**a**) Top layer of nodes represent cells in the young condition, while the bottom layer of nodes represents cells in the old condition. Nodes are sorted by degree such that hubs (cells visited frequently on the random walks) are grouped on the left of each cell type in the old condition. An edge represents a high probability walk from a young cell to an old cell, where thicker edges indicate more frequent walks. (**b**) Same as (a), but after removing the hubs identified by selecting nodes with degree above the 90^{th} quantile. (**c**) Same as (a), but the top layer now represents old cells, while the bottom layer represents young cells, and edges represent random walks from old to young cells. (**d**) Same as (c), but after removing the hubs.

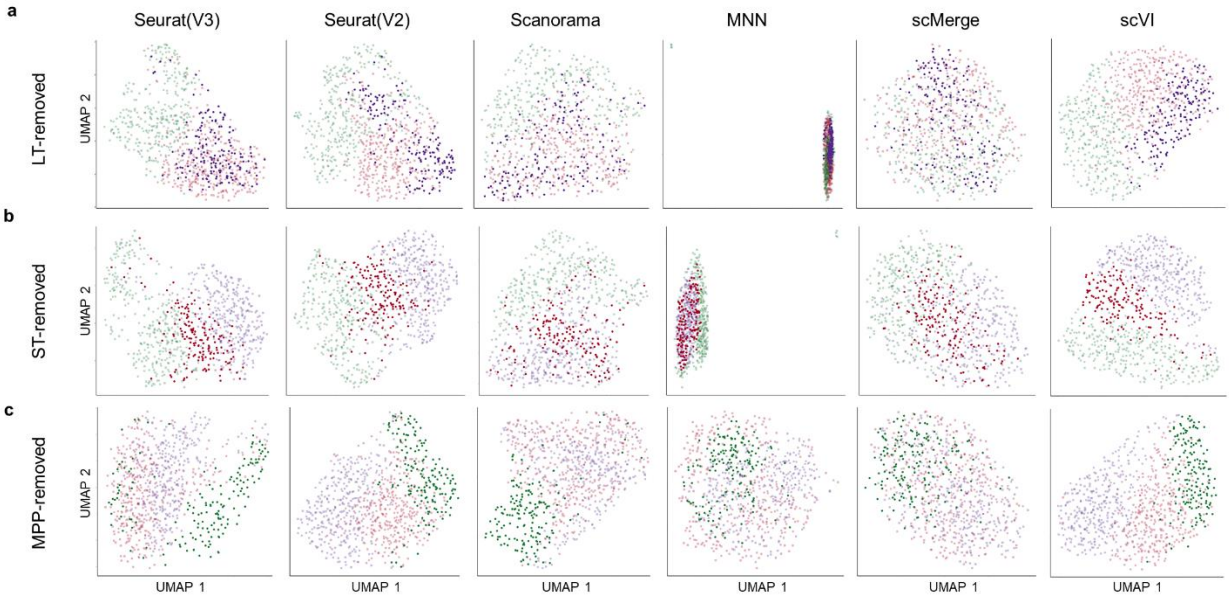


Fig. 3A.S12: Comparison of alignment methods when cell types are removed from the Kowalczyk et al. benchmark. (a) t-SNE visualization of alignments after removing LT-HSCs from the stimulated cells. The control LT-HSCs are highlighted to show the amount of incorrect overlap with either the ST-HSC or MPP cells. **(b-c)** Similar to (a), but with ST-HSCs or MPPs removed, respectively.

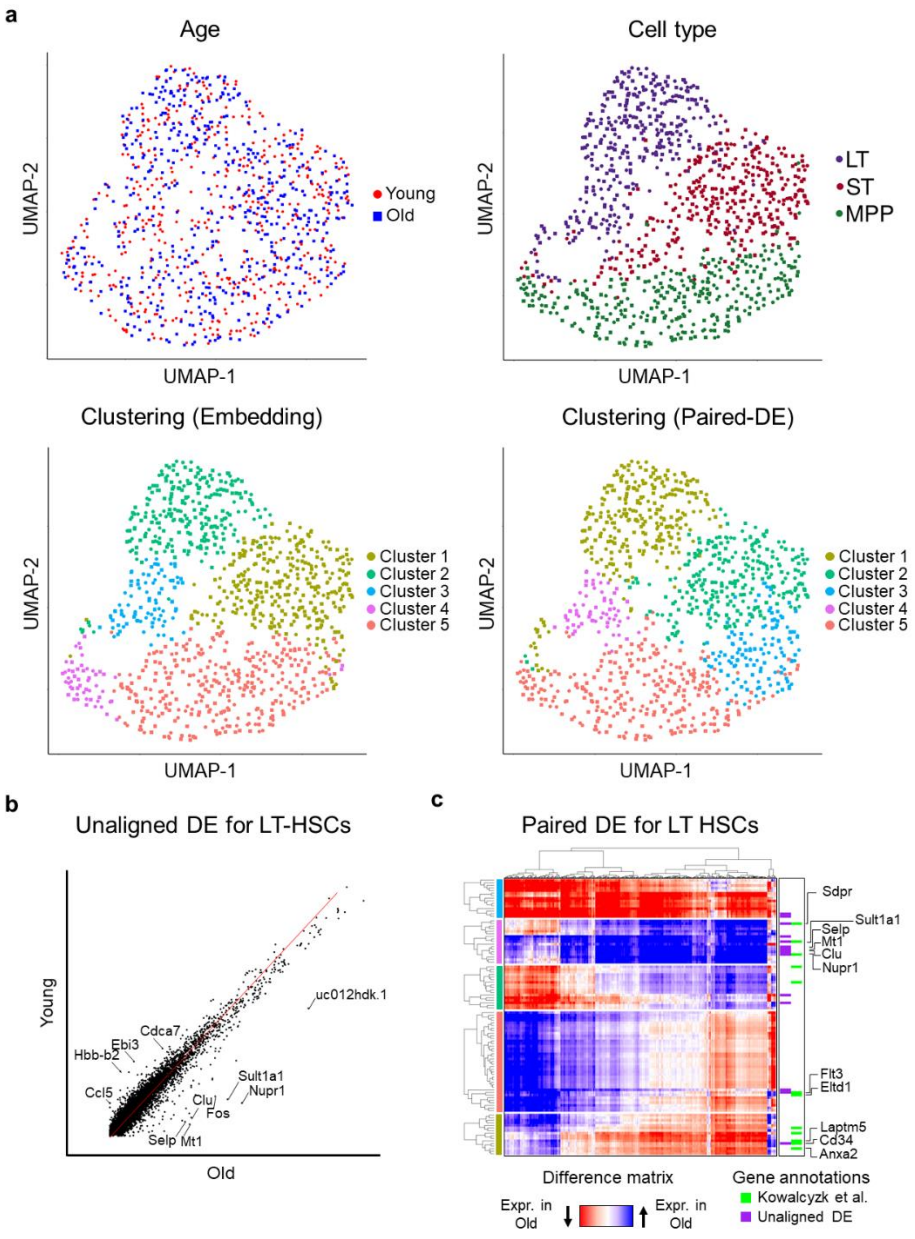


Fig. 3A.S13: Alignment of Kowalczyk et al. identifies subpopulations of LT-HSC(s) with unique response to age. (**a**) UMAP visualization of young and old mouse cells after alignment by scAlign. Each cell is colored by age, cell type, clustering in alignment space or clustering on scAlign's state variance map. (**b**) Scatterplot of significantly differentially expressed genes identified by comparing cluster averages. (**c**) Heatmap of the state variance map (paired difference for each cell interpolated into the young and old condition). Each gene is annotated based on the differential expression results from Kowalczyk et al. and comparison of cluster averages.

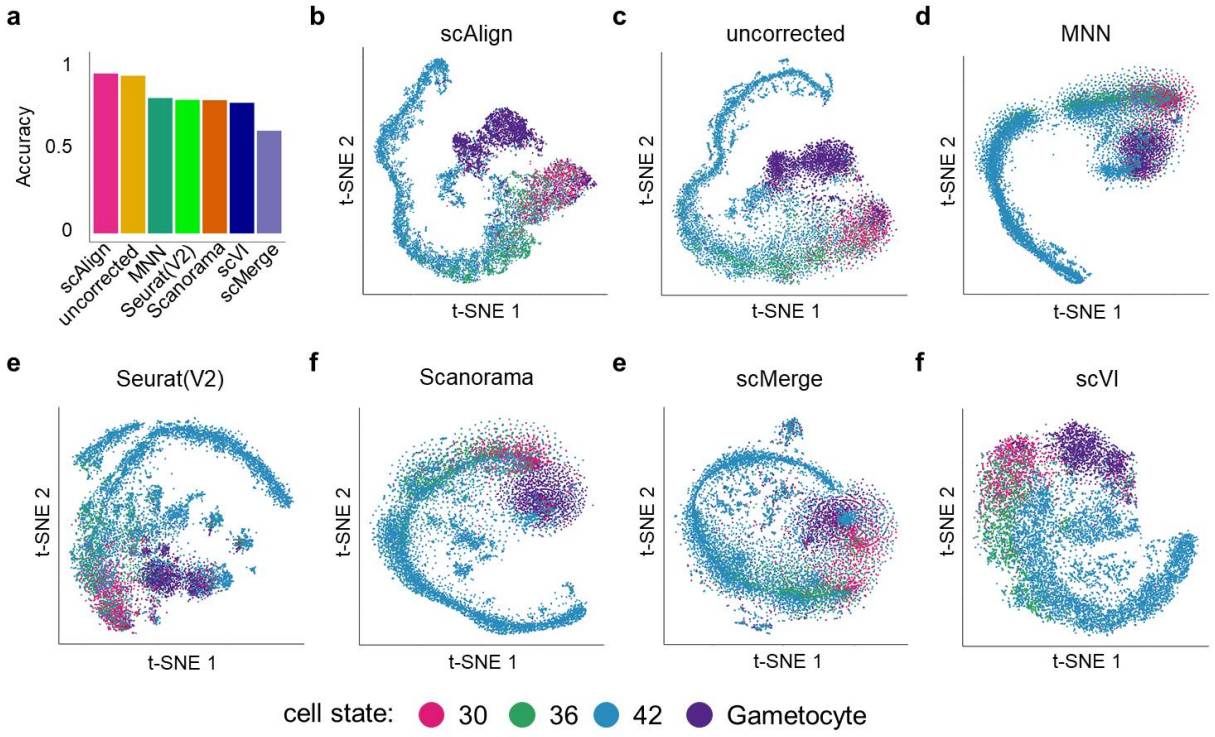


Fig. 3A.S14: Alignment of +/-Shld parasites. (**a**) Accuracy of different alignment methods after aligning +/-Shld parasites, where accuracy is based on the notion that gametocyte cells from the +Shld condition should not be aligned to any parasites in the -Shld condition. (**b-f**) tSNE visualizations of +/-Shld parasites aligned together, using different alignment methods.

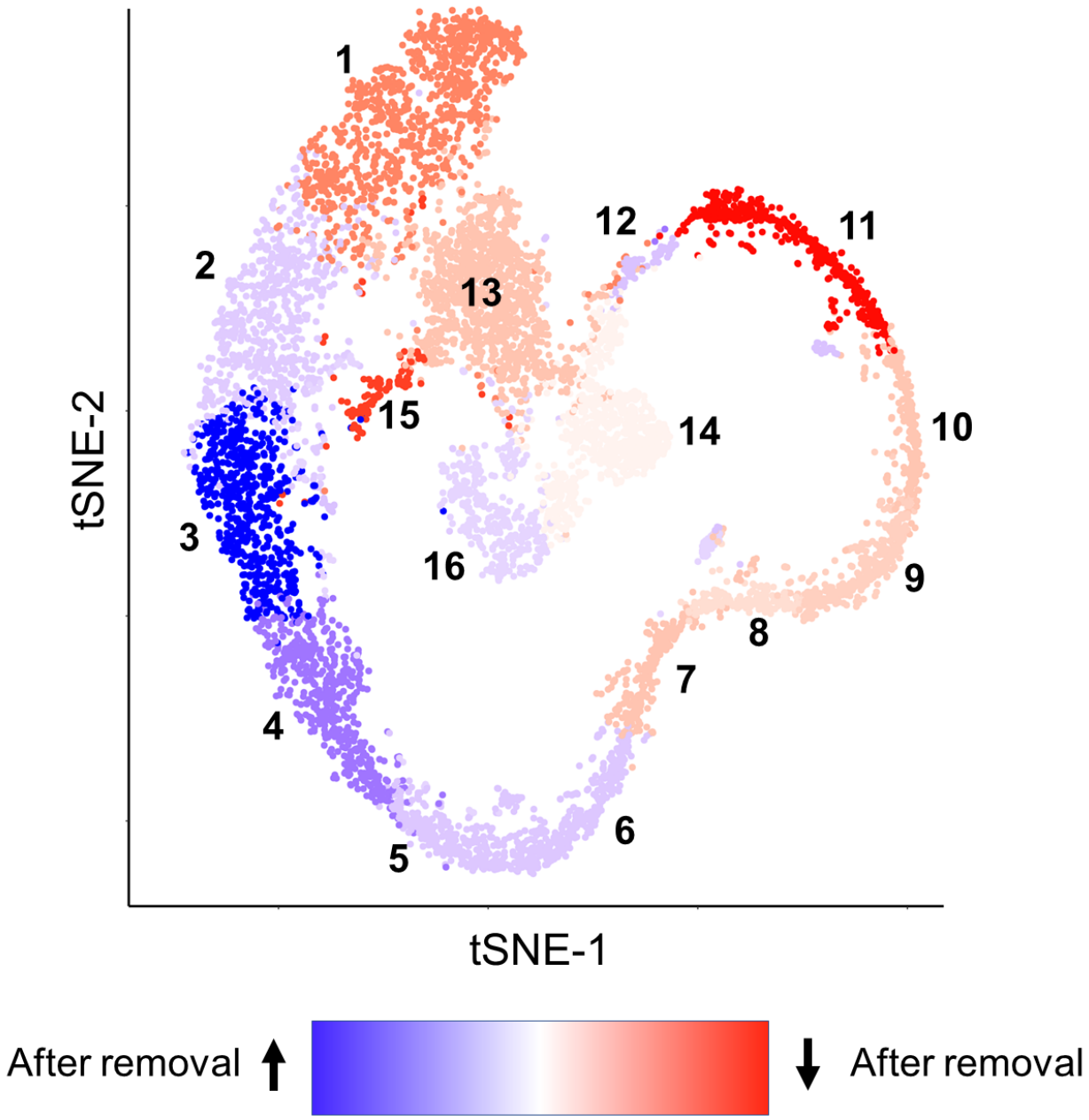


Fig. 3A.S15: Removal of -Shld parasites from gametocyte clusters leads to more diffuse round trip probabilities from cluster 12. The difference heatmap for round trip probabilities for parasites in cluster 13 (rows) in the presence of cells from -Shld grouping with +Shld gametocytes and after removal of the -Shld cells grouping with +Shld gametocytes. Dark blue indicates an increase in round trip probability after removal and red indicates a decrease in round trip probability after removal.

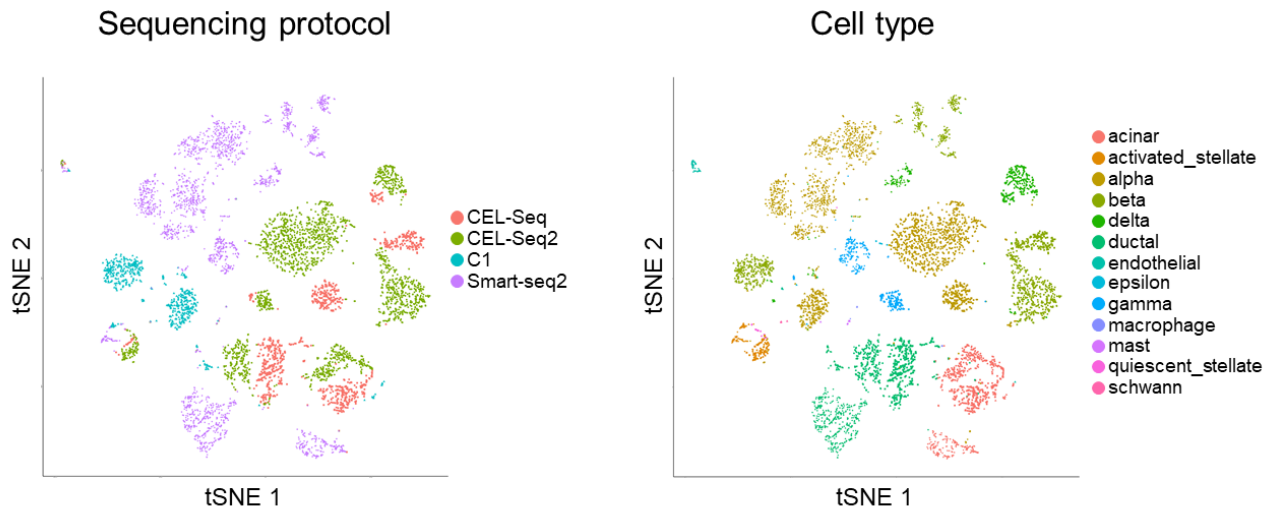


Fig. 3A.S16: Unaligned pancreatic datasets exhibit protocol specific effect. tSNE visualization of the unaligned highly variable genes, cells are colored by sequencing protocol and cell type.

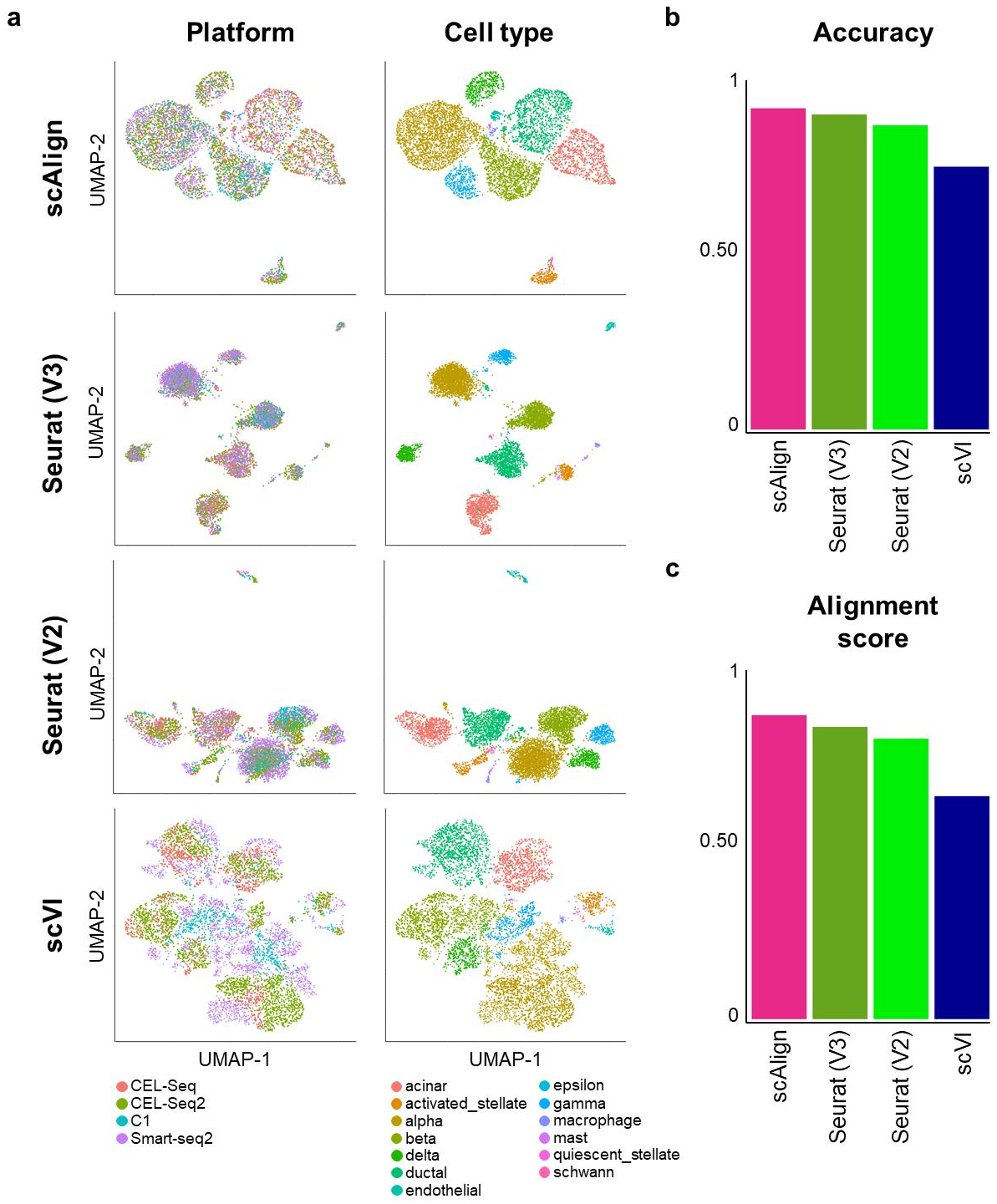


Fig. 3A.S17: Alignment of all four pancreatic islet datasets. (**a**) UMAP visualization of the aligned embeddings produced by scAlign colored by platform or cell type. (**b**) Barplot visualizing the composite accuracy for each method. (**c**) Barplot visualizing the alignment scores for each method.

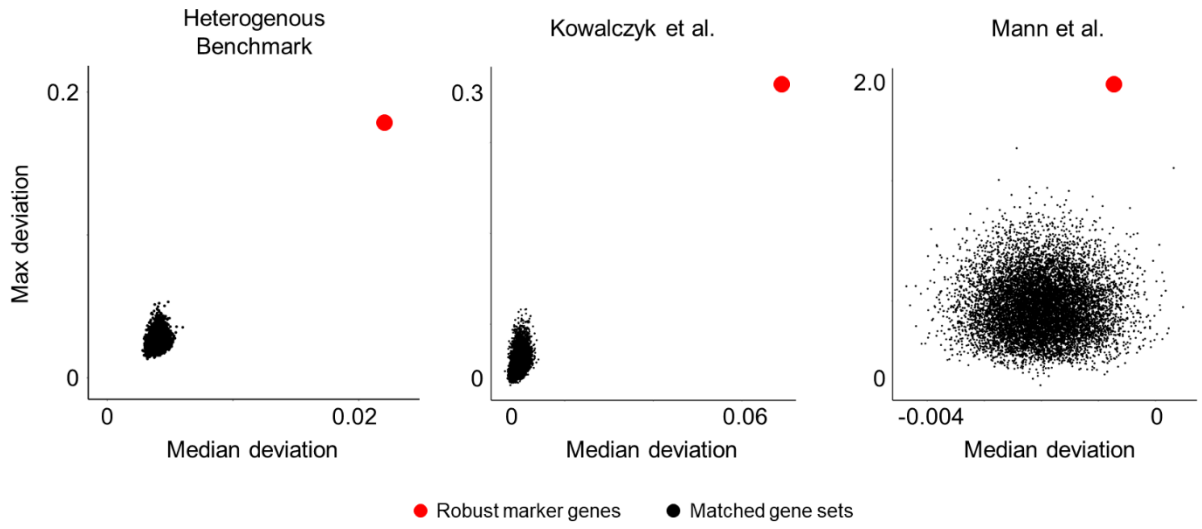


Fig. 3A.S18: Robust cell type marker genes drive alignment. For the set of cell type marker genes robustly identified across conditions, we perturbed their expression in cells and measured the corresponding deviation in cell state space embeddings. We repeated this experiment for gene sets matched for size and relative expression. Perturbation of the robust cell type marker genes led to systematically larger deviations in cell state embeddings compared to control gene sets, indicating that robust cell type marker genes contribute more to alignment than expected by chance.

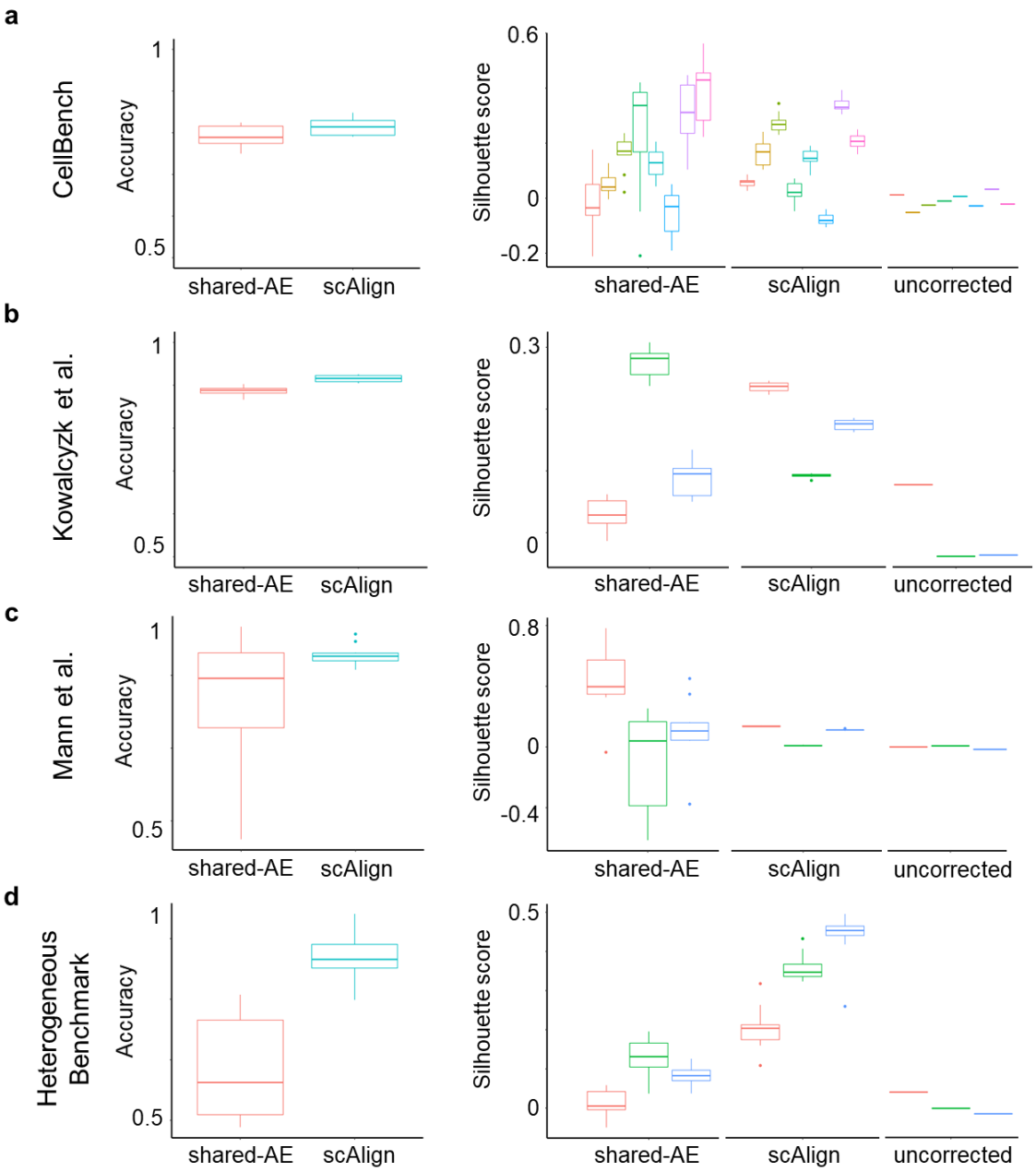


Fig. 3A.S19: scAlign outperforms a shared autoencoder (shared-AE) with similar network architecture, in terms of alignment accuracy and maintaining fidelity of the cell-cell similarity matrix. (a-d) Box and whisker plots of alignment quality metric after 10-fold CV for both the shared autoencoder and scAlign. Silhouette score is illustrated for unaligned data, cell embeddings for scAlign and for the shared autoencoder.

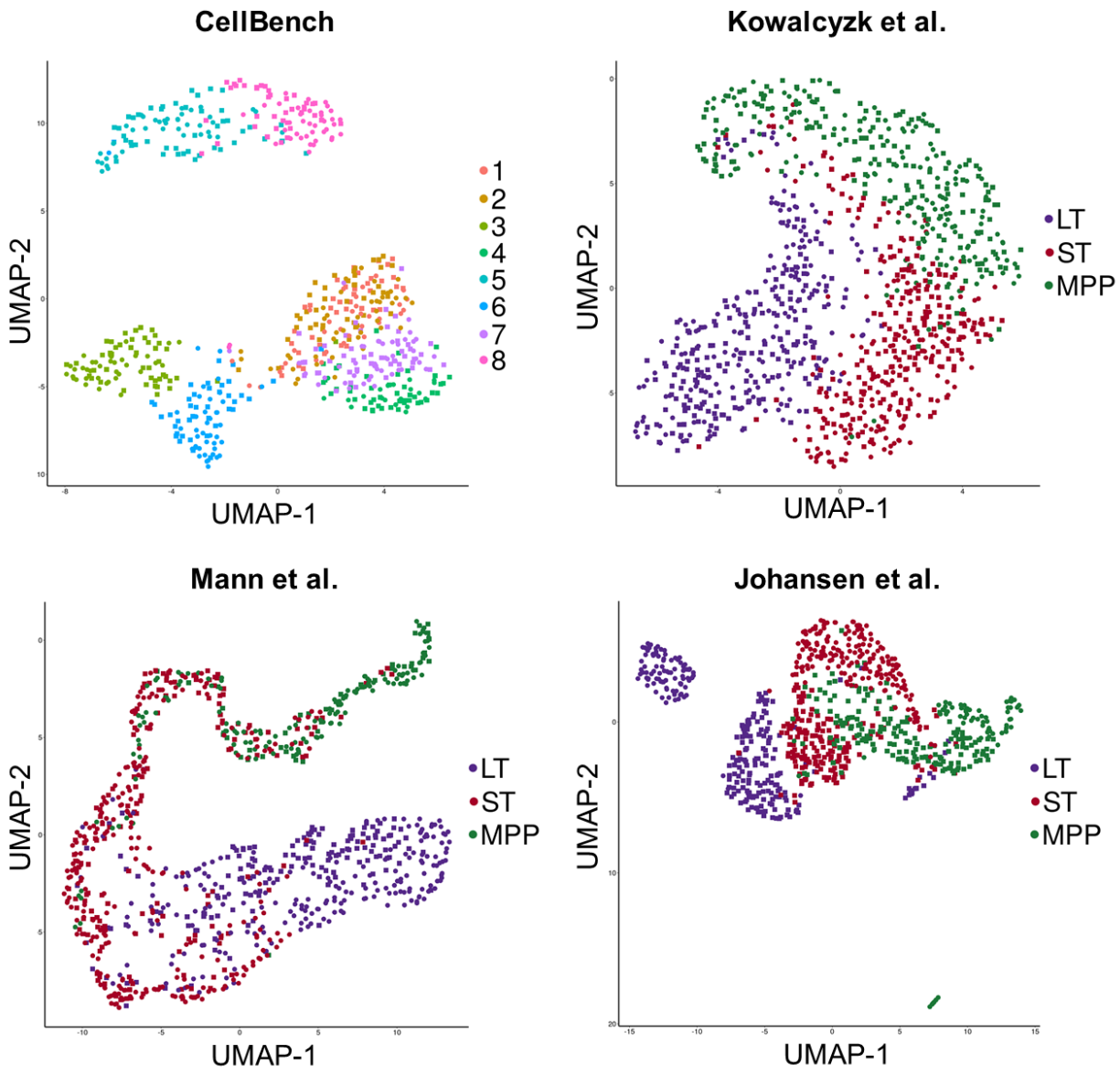


Fig. 3A.S20: Comparison of shared autoencoder cell embeddings after alignment of the four benchmark datasets. UMAP visualization shows the cell embeddings colored by cell type for each of the four benchmark datasets.

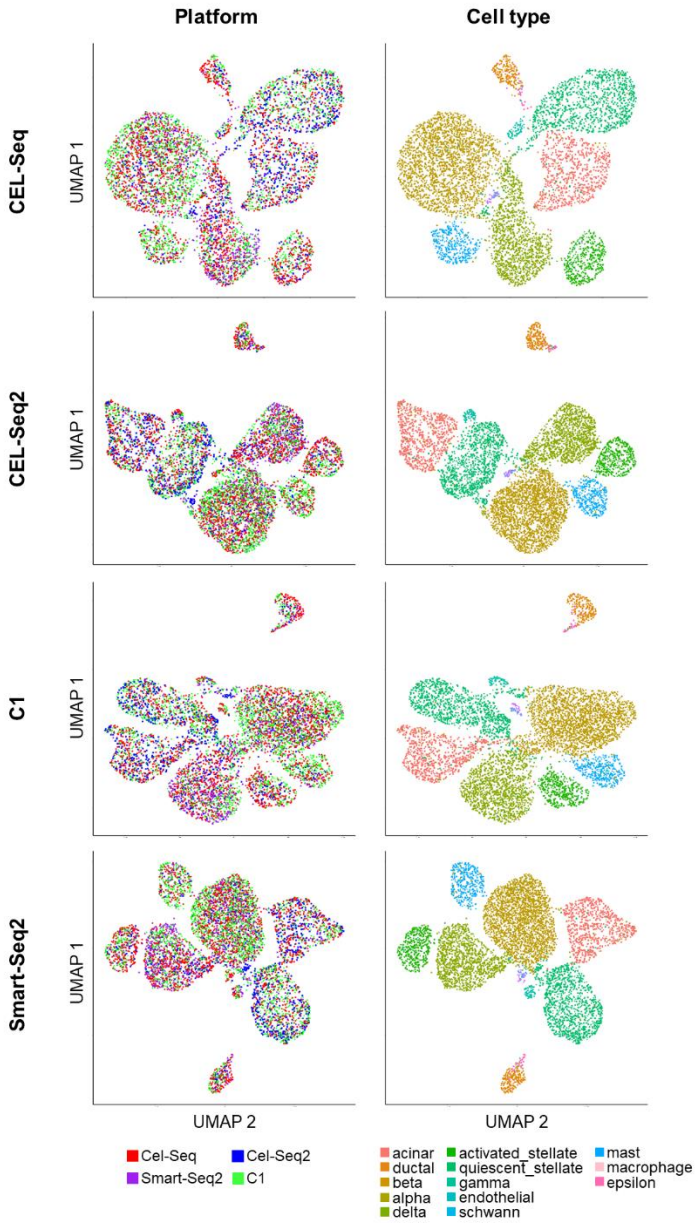


Fig. 3A.S21: Comparison of scAlign alignment of pancreatic islet cells using each protocol as a reference. UMAP visualizations after alignment where a single protocol is used as a reference (y-axis) and colored by both platform and cell type (x-axis) annotations.

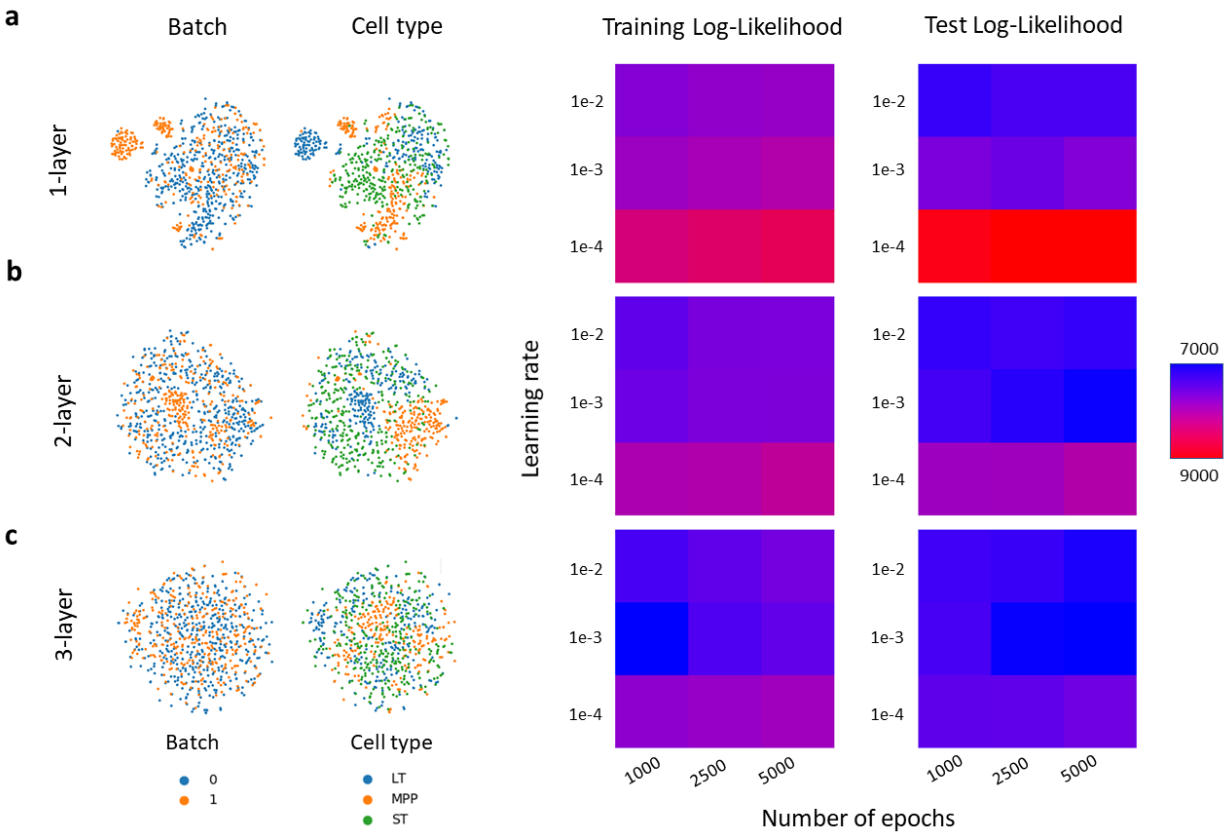


Fig. 3A.S22: scVI grid parameter search procedure identifies optimal parameterization. (**a**) (left) tSNE visualization shows the latent dimensions inferred by scVI, colored by batch (condition) and cell type, after alignment using the optimal parameters identified by grid search. (right) Parameter search results for both the training and test set with respect to log likelihood, where the x-axis is the learning rate and y-axis is the number of epochs for the respective number of network layers. (**b-c**) Same as (a), but for different numbers of hidden layers in the network.

Chapter 3

Projection and deconvolution of clumped transcriptomes

3.1 Introduction

In recent years, there has been a surge in the number and size of atlasing efforts across tissues, conditions, and species[27,45,67,68], driven by the high throughput nature of single cell- and nucleus-RNA sequencing (sc/snRNA-seq) technologies. These technologies are now routinely used to generate atlases on the scale of up to millions of cells[24,45,69,70]. Studies leveraging sc/snRNA-seq maximize the discovery of novel cell types and characterization of transcriptional heterogeneity of individual cell types within samples. One of the limitations of the sc/snRNA-seq technologies, however, is that they only capture the RNA content of each cell.

To address this limitation, there are a growing number of single cell *resolution* assays that simultaneously measure RNA content as well as other cellular annotations and modalities. For example, spatial transcriptomic sequencing assays such as Slide-seq[71] and LCM-seq[72] record both the spatial position and RNA measurements from individual spots on a sample. There are also multi-modal assays such as Patch-seq[73] that measure cellular phenotypes in addition to local RNA content, enabling the identification of connections between molecular and cellular phenotypes of neurons. Additionally, multi-modal assays such as SNARE-seq[74], sciCAR[75] and 10x Multiome simultaneously measure the DNA accessibility and gene expression in single cells.

However, single cell resolution assays have a major drawback: in exchange for collecting additional data modalities, they often trade off some precision in their RNA measurements. In the case of some spatial transcriptome sequencing assays, RNA is extracted from spots of pre-defined size and location on a tissue, leading to individual spots often capturing RNA from multiple cells. Analogously, for Patch-seq, a micropipette is used to puncture brain slices and remove RNA from a target neuron, but RNA from neighboring neuronal or glial cells can be captured as well[76]. For technologies such as MERFISH[77], in practice only a few hundred genes in the genome can be measured. This lack of true single cell RNA measurements can hinder downstream analysis of spatial gene expression patterns or inferring connections between molecular and cellular phenotypes.

3.2 scProjection

Here we present scProjection, a method for projecting single cell resolution RNA measurements onto deep single cell atlases, in order to achieve single cell precision from the original RNA measurements. First, we demonstrate our cell type-specific projections capture RNA contributions of component cells, and importantly that the gene co-expression network of the projected data is consistent with the gene co-expression network of scRNA-seq data from the same cell population. We then illustrate three use cases of scProjection. First, we show scProjection analysis of spatial transcriptomes yields substantially increased detection of cell type-specific spatial gene expression patterns across diverse tissues such as the primary motor cortex and hypothalamic regions of the brain as well as the intestinal villus. Second, we demonstrate scProjection can impute spatial genome-wide gene expression measurements when targeted sequencing of limited numbers of genes via MERFISH[78] is performed. Finally, we show scProjection can separate RNA contributions from multiple cell types when analyzing Patch-seq data, where RNA measurements

are composed of RNA from the target neuron as well as neighboring glial cells. The separation of RNA contributions leads to more accurate prediction of one data modality (electrophysiological response) from another (RNA expression levels). We conclude that integrating deep single cell atlases with single and multimodal cell resolution assays can therefore combine the advantages of both sequencing approaches to study single cells.

3.2.1 Workflow of scProjection

The scProjection model and workflow is illustrated in **Figure 1**. scProjection assumes that one or more RNA samples x_i from a single cell resolution assay are available as input (**Fig. 1a**), as well as a deeply sequenced single cell atlas that profiles the same cell types as the single cell resolution assay (**Fig. 1b**). Typical single cell resolution assays of interest include spatial transcriptome assays such as LCM-seq, Slide-seq or MERFISH, multimodal assays such as Patch-seq, or classical bulk RNA-seq. As output, scProjection simultaneously projects each RNA sample x_i onto each component cell population k within the single cell atlas to find the average cell state (expression profile) of that cell type in the sample $(y_{i,k})$ (**Fig. 1c**), as well as the relative abundance of that cell type $(\alpha_{i,k})$ (**Fig. 1d**). scProjection therefore balances selecting sets of cell states $y_{i,k}$ that help minimize reconstruction error of the original RNA measurement x_i, with the task of selecting cell states that are frequently occurring, as measured by the single cell atlas (e.g. the prior).

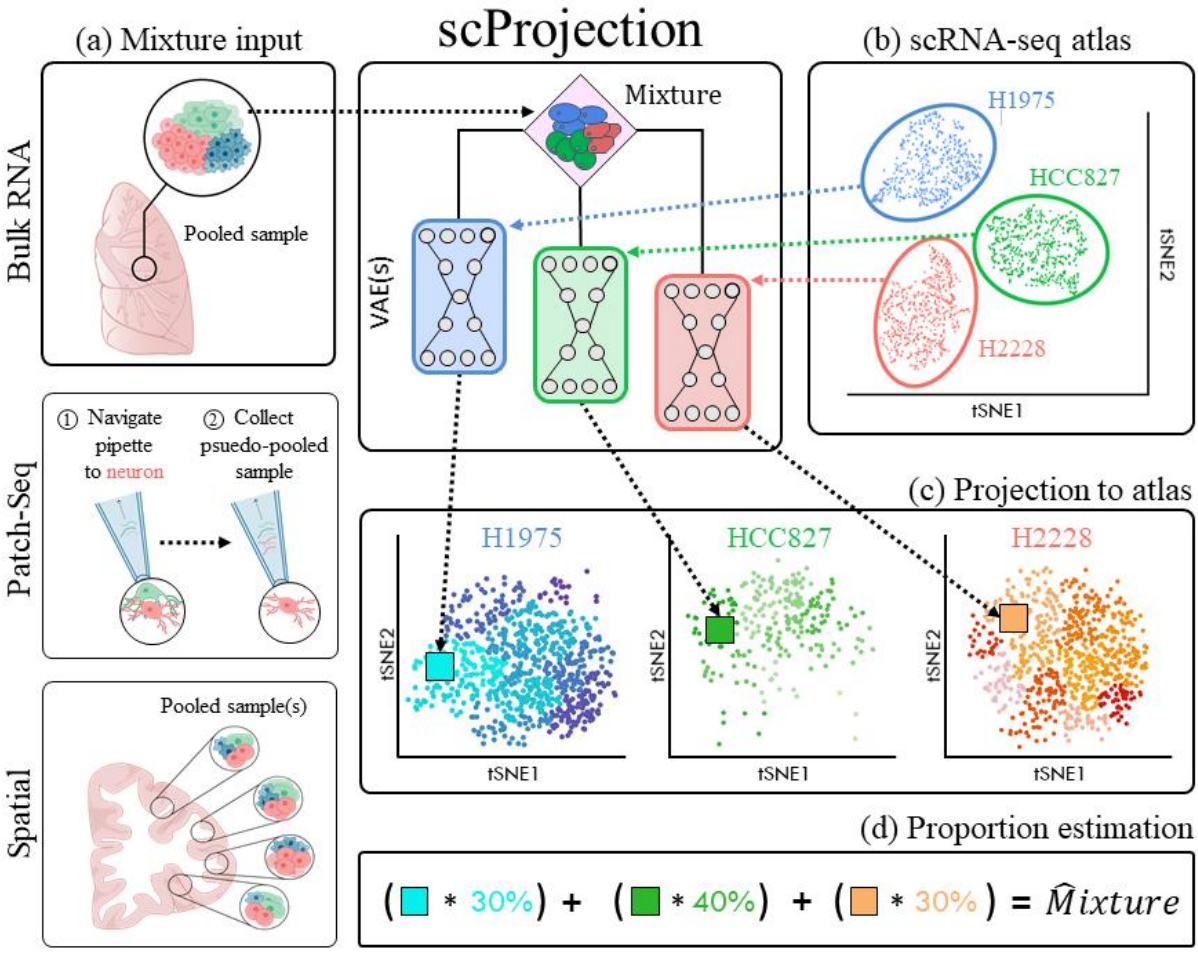


Figure 1. Schematic of cell type projection and abundance estimation with scProjection. (**a**) The primary input to scProjection consists of one or more RNA measurements originating from mixtures of cells assayed using bulk RNA-seq, multi-modal assays or spatial transcriptomics. (**b**) The secondary input to scProjection is a single cell atlas from the same region or tissue as the mixture samples, and is assumed to contain all the cell types present in the mixture samples. For each of the annotated cell types in the single cell atlas, a variational autoencoder is trained to model within-cell type variation in expression. (**c,d**) The average cell state for each cell type in a single RNA mixture, along with the relative abundances of each cell type, are estimated by balancing two objectives: (c) selection of an average cell state per cell type that is likely given the single cell measurements for each cell type (the prior), and (d) the joint selection of cell states for each cell type, and abundances, that will lead to the best reconstruction of the original mixed RNA measurements (data likelihood).

scProjection uses individual variational autoencoders[79] (VAEs) trained on each cell population within the single cell atlas to model within-cell type expression variation and delineates

the landscape of valid cell states[80], as well as their relative occurrence. Here, a valid cell state for a cell type k is defined as a genome-wide gene expression profile that has either been directly measured in a single cell atlas, or is inferred to be feasible based on the covariation of gene expression patterns observed in measured cells In practice, we ignore projections $y_{i,k}$ when the predicted cell type abundances $\alpha_{i,k}$ is small (e.g. <5%).

With scProjection, we achieve state-of-the art deconvolution performance in benchmarking with ground-truth cell type abundances for CellBench and ROSMAP [81,82] (Supplemental Note 1). scProjection most accurately estimated cell type abundances for rare neuronal and non-neuronal cell type contributions to bulk RNA samples from human Dorsolateral Prefrontal Cortex (DLPFC).

3.3 Experiments

3.3.1 Projections distinguish within-cell type variation in gene expression patterns

Initially, we established scProjection's ability to map mixed RNA samples to the correct transcriptional state for each contributing cell type. To do so, we conducted a series of simulation experiments in which a pair of cell states were selected from distinct neuron cell types, L2/3 IT and L6b, profiled in a recent human cortex cell atlas[69]. To impose a tiered difficulty, we chose these two cell types which are variable in their heterogeneity: L2/3 IT is highly variable with many cell states, and L6b is composed of five cell states (Methods). We repeatedly constructed mixed RNA samples by first selecting a random subtype, then selecting a cell state from that subtype, for each of L2/3 IT and L6b. The genes counts from this pair of randomly selected cells were added to form the final mixed RNA sample.

scProjection was then evaluated on its ability to map the mixed RNA sample back to the correct transcriptional state and subtype for each of L2/3 IT and L6b, when only providing scProjection with a cell atlas whose cells are annotated at the level of L2/3 IT and L6b (no subtype information was provided to scProjection). We found that scProjection reliably mapped each mixed RNA sample back to the correct subtype in all 10,000 mixed RNA samples. Furthermore, we found that scProjection mapped the RNA samples to the correct and higher resolution cell state in 87% of the simulations, and the projected cell state was highly correlated to the original (Spearman rho=0.99, $p < 2.2e-16$) (**Supplementary Fig. 1**). This compares favorably to CIBERSORTx, which mapped each RNA sample back to the true subtype only 61% of the time, with an average Spearman correlation of rho = 0.68 to the original cell state. These findings are consistent with experiments performed on the CellBench gold standard benchmark data (**Supplementary Note 2**).

Having demonstrated scProjection can successfully project simulated data to the correct cell state and subtype, we designed an analogous experiment using experimentally measured RNA samples from single cell resolution assays. A recent Patch-seq study[83] profiled 4,200 mouse visual cortical GABAergic interneurons from multiple layers of the mouse neocortex, of which the original study classified 1,818 of them as Sst inhibitory neurons, the most abundant class in the dataset. As we described above, Patch-seq RNA measurements typically contain RNA from the target neuron as well as neighboring non-neuronal cells, so the goal of our experiment was to perform projection to recover the cell state of the target neuron for each Patch-seq measurement (**Fig. 2a**). We first performed a sanity check by using scProjection to estimate the abundance of the Sst cell type to the Patch-seq RNA measurements from the 1,818 experimentally defined Sst

neurons and found 1764 mapped to Sst with the highest cell type abundance using two different single nucleus atlases of the brain (**Supplementary Fig. 2**).

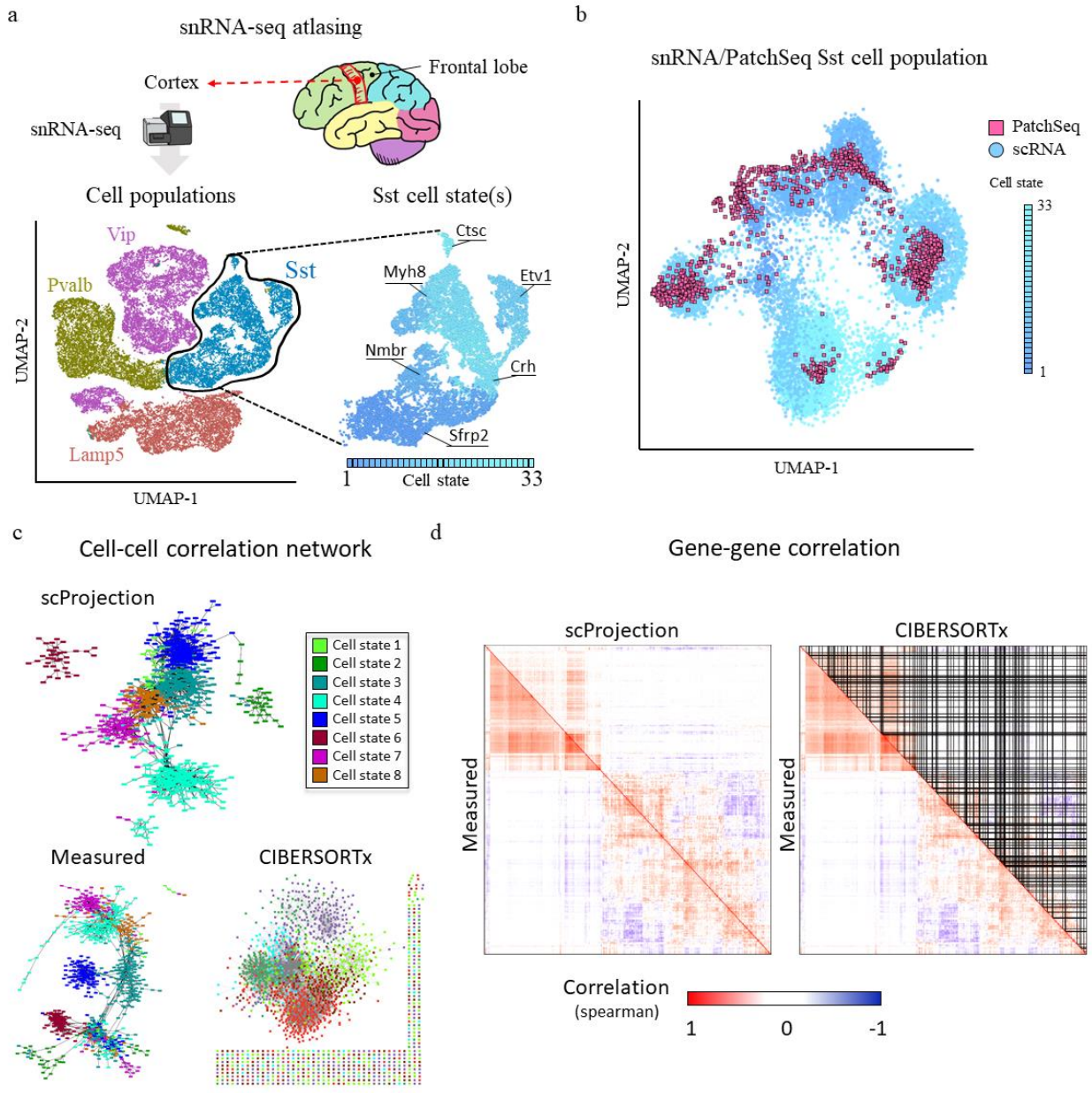


Figure 2. scProjection distinguishes within-cell type variation and maintains cell-cell and gene-gene network structure in Sst neurons. (**a**) Visualization of the snRNA-seq atlas of the mouse cortex used for projection of mouse Patch-seq data. We subsetted the data to four major cell types (Sst, Vip, Pvalb and Lamp5), of which Sst was further broken down into 33 distinct cell

states. (**b**) tSNE plot of the measured single cell (circle) Sst neurons (from (a)) alongside the mouse PatchSeq (square) measurements projected to the Sst population. snRNA-seq cells are colored according to cell state shown in (a). (**c**) Cell-cell similarity network of the measured Patch-seq Sst cells, the scProjection-based projection of Patch-seq RNA to the Sst population, and CIBERSORTx-predicted contributions of the Sst population for comparison. (**d**) Heatmaps visualizing the gene-gene covariation patterns of the measured Patch-seq RNA (lower-triangular), versus the gene-gene covariation patterns calculated from either the projections of the Patch-seq RNA to Sst via scProjection, or the CIBERSORTx-based predictions of RNA contributions by Sst. in the upper-triangular of their respective heatmaps.

We used scProjection to project the 1,818 Sst Patch-seq RNA measurements to an Sst single nucleus atlas[69] (**Fig. 2a**). Because the ground-truth cell state of the Patch-seq measurements is unknown (unlike in the simulation), we instead assessed accuracy by comparing the known Sst subtype of the Patch-seq measurement and the known Sst subtypes of the single nucleus measurements in the atlas. In 1623 of the 1,818 neurons, the cell state of the projected Sst neurons matched the annotated cell state of neighboring neurons from the single cell atlas (Methods) (**Fig. 2b**). Similarly, we projected a separate Patch-seq dataset consisting of 45 layer 1 inhibitory neurons from two electrophysiologically-defined subclasses (SBC, eNGC) onto a broad single cell atlas of inhibitory neurons. We found the SBC and eNGC neurons were better separated after projection (Acc: 0.84) compared to before (**Supplementary Fig. 3**). In total, our results on these two Patch-seq datasets suggest that scProjection distinguishes intra-cell type expression variation associated with neuronal firing patterns within the inhibitory neuron cell types.

3.3.2 High-fidelity maintenance of cell and gene network structure

One concern we had while designing scProjection was the extent to which projections altered the input RNA samples as a population. That is, if two input RNA samples are similar before

projection, we reasoned it was sensible to expect they were also similar after projection; that is, the overall similarity structure of the input samples should remain globally consistent. On the other hand, we also would expect that the co-expression behavior of individual genes after projection would be consistent with the reference single cell atlas; genes that co-vary (and therefore are more likely to co-function) in the single cell data should also do so in the projected samples, since they represent the same cells. Therefore, to measure these population level behaviors, we constructed cell-cell and gene-gene co-expression networks before and after projection to compare.

Figure 2c illustrates three cell-cell co-expression networks: that of the Patch-seq measurements before and after projection to Sst, as well as from the imputed gene expression profiles of CIBERSORTx. Overall, the structure of the cell-cell network after projection more closely resembles the before-projection measured network compared to CIBERSORTx, suggesting scProjection maintains the overall structure of a set of input samples compared to CIBERSORTx. Similarly, **Figure 2d** qualitatively compares the inferred gene co-expression network of the measured Sst scRNA-seq data, to both the projected samples from scProjection, as well as the imputed samples from CIBERSORTx. scProjection's network more closely resembles both the raw (Jaccard: 0.72) and the imputed (Jaccard: 0.76) Sst co-expression networks, in comparison to CIBERSORT (Jaccard: 0.21), which fails to impute many genes as visualized by the black lines.

3.3.3 Detection of novel spatial expression patterns of enterocytes in the intestinal epithelium

We envisioned that one primary application of scProjection is to infer single cell transcriptomes from RNA measurements produced by spatial transcriptome technologies, in order to detect spatial gene expression patterns in tissues. Technologies such as Slide-seq[71], LCM-seq[72] and Visium by

10x Genomics capture RNA from different spots of a tissue slice. Each spot potentially contains RNA contributions from more than one cell in close proximity (**Fig. 1a**). Therefore, the RNA from each spot can be viewed as a miniature bulk RNA sample composed of a small number of cells, from which we want to extract single cell transcriptomes for each contributing cell type through projection.

We initially analyzed a dataset collected by Moor et al.[84] in which they performed LCM-seq on five distinct regions, or zones, of the intestinal villus, as well as separately collected a scRNA-seq cell atlas from replicate intestinal villi. They identified spatial expression patterns in the dominant cell type, enterocytes, by (1) identifying marker (landmark) genes for each zone using the LCM-seq data, (2) assigning zone labels to the scRNA-seq cells using landmark genes, and (3) predicting zone-specific expression through zone-specific averaging of the labeled scRNA-seq data. We reasoned that identification of landmark genes from LCM-seq data could be difficult since LCM-seq captures contributions from multiple cell types thus yielding poor labeling of the single cell atlas cells. We therefore avoided this critical landmark gene selection by taking the opposite approach: we use scProjection to project the zone-specific LCM-seq samples to the enterocyte single cell atlas, to extract the enterocyte expression patterns within each zone. This approach would explicitly disregard contributions of non-enterocytes to each LCM-seq sample.

Figure 3a illustrates the projections of the LCM-seq data to the enterocyte single cell atlas, where the single cells are labeled according to Moor et al[84]. The LCM-enterocyte projections are generally proximal to the single cells assigned to the same zone by Moor et al., suggesting our approach is overall consistent with that of Moor et al. However, our approach identifies 3-fold more zone-specific spatial expression patterns compared to the genes identified by the Moor et al (**Fig. 3b**). To validate the predicted enterocyte zone-specific expression patterns, we compared our

predicted zone expression patterns to the smFISH expression quantifications and the original LCM-seq measurements provided in the original study. We found that across a small set of validated landmark genes (Ada, Slc2a2, Reg1), our spatial expression predictions were more correlated with the smFISH quantifications (**Fig. 3c**). Furthermore, our approach identified zonation patterns in genes such as Pkib, Slc2a13 and Fam120c which were not identified by the Moor et al. spatial reconstruction approach yet are clearly zone-specific according to the original LCM-seq experiments (**Fig. 3c**). These results in total suggest RNA projections improve our ability to identify zone-specific expression patterns in dominant cell types such as the enterocytes.

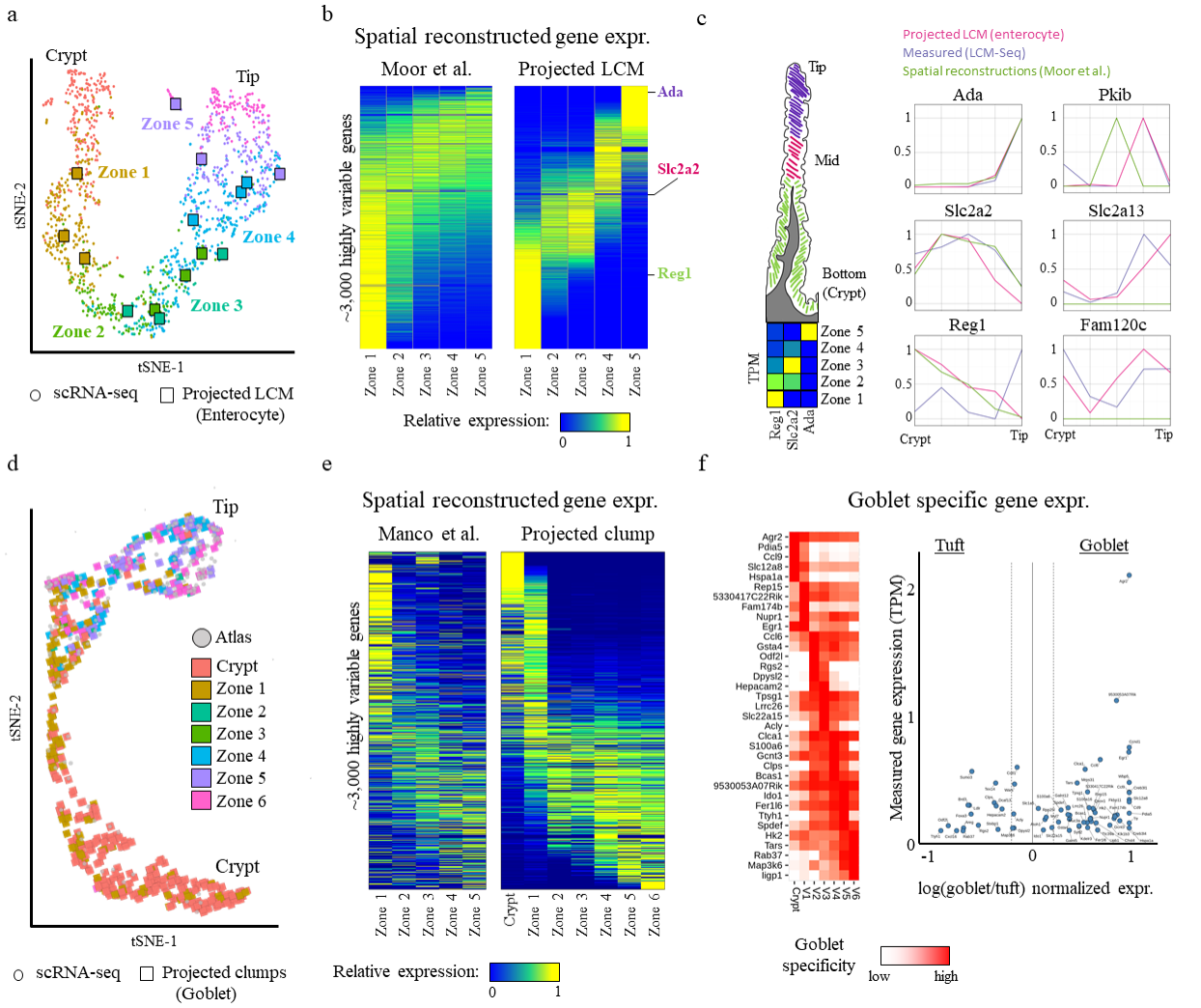


Figure 3. Projection refines spatial expression patterns in common and rare cell types of the intestinal villus. (**a**) tSNE plot of the single cell atlas (circles) and projected LCM samples (squares) across the zones of the intestinal villus. Single cells are colored based on their zone assignment by Moor et al. (**b**) Heatmap visualizing the spatial expression patterns of the top 3,000 highly variable genes using the spatial inference approach of Moor et al. on the left and after projecting the LCM samples with scProjection on the right. Three marker genes (rows) are labeled: Ada, Slc2a2 and Reg1. On the right is a schematic of a single intestinal villus, along with the expected dominant zone of expression for Ada, Slc2a2 and Reg1. Shown below the villus is the actual measured expression pattern of Ada, Slc2a2 and Reg1 in the LCM data of the five zones. (**c**) Line plots comparing the measured and projected expression of top zonated genes across the intestinal villus. (**d**) tSNE plot of the single cell atlas (circles) and projected clump-seq (squares) as annotated by the enterocyte component within each clump. (**e**) Heatmap visualizing the spatial expression patterns of the top 3,000 highly variable genes in the goblet containing clump-seq samples using the approach of Manco et al. on the left and after projecting with scProjection on the right. (**f**) Heatmap visualizing the expression of the union of the top 5 zonated genes per zone in the goblet containing clumps. The scatter plot on the right visualizes the divergence in expression of zonated genes between goblet and tuft containing clumps.

3.3.4 Rare cell types of the intestinal villus can be spatially resolved

Projection of an RNA sample onto the single cell atlas of a target cell type intuitively requires sufficient abundance of the target cell type to the RNA sample in order to be successful. scProjection predicted enterocytes to contribute 90% of the LCM-seq RNA on average.In contrast, populations such as the secretory (goblets, tuft) cells are rare: for example, goblets only contribute 8% of the LCM-seq RNA on average[85], while tuft cells are only contribute 1% of the LCM-seq RNA on average (**Supplementary Fig. 4**). The mucus-producing goblet cells[86,87] and chemosensory tuft cells[88] play an important role in the protection of cells in the intestinal villus as

well as communication with other stroma cell types[89]. Manco et al.[85] captured these rare cell types in the intestinal villi by performing RNA-seq on clumps of physically-adjacent cells in the intestinal villus, by incompletely dissociating the tissue. Because of the high abundance of enterocytes and rare occurrence of goblet and tuft cells, most clumps will contain primarily enterocytes, and only occasionally contain goblet or tuft cells. To derive spatial expression patterns of the rare cell types, Manco et al. predicted the zone of the entire clump by comparing clump expression against a spatial reference from the Moor et al.[84] work described above, then assigned that zone label of the entire clump to the secretory cells within the same clump. We hypothesized that by replacing the zone-prediction step in Manco et al. with our projection approach used above for the enterocytes, we can further identify goblet and tuft specific spatial patterns of expression across the intestinal villus.

Our general strategy was to first train the scProjection VAE components on a single cell atlas of the intestinal epithelium which captured enterocytes and rare secretory types including goblet and tuft cells[85,90]. We then simultaneously project each clump to the enterocyte cell type and the secretory cell types (goblet or tuft) separately. We predicted the zone of the entire clump based on the zone-specific LCM-enterocyte projections similar to above (see Methods). We computed zone-specific expression patterns of goblet (or tuft) cells by averaging clump-goblet (or clump-tuft) projections that were predicted to land in the same zone.

We focused first on the mucus-producing goblet cells, because while rare, there were more goblet-containing clumps available to robustly estimate zone-specific expression compared to tuft cells. From an initial set of 6824 clumps, we identified 1,084 clumps that contained at least 40% cell type abundance from goblet cells. From the 1,084 goblet-containing clumps, we projected these clumps to the goblet single cell population (n=314) to identify spatial gene expression

patterns. **Figure 3d** illustrates the 1,084 clumps projected onto the goblet single cell atlas, where the clumps and single cells are labeled according to Manco et al[85] (**Supplementary Fig. 4**, Methods). The projected clumps were generally proximal to the single cells assigned to the same zone by Manco et al., suggesting our approach generally consistently captures zone-specific gene expression. Using our projections of the clumps, we predicted 2480 genes that exhibit zone-specific goblet expression patterns, compared to 972 zone-specific genes identified by Manco et al.'s approach (**Fig. 3e**). To validate the predicted goblet zone-specific expression patterns, we compared our 2480 zone-specific genes with goblet specific landmark and mucus associated genes (Methods) whose tendency for villus-tip expression was identified in Manco et al. We found that our spatial expression predictions from the clumps were correlated with reported zonated expression and smFISH quantifications (Supplementary Fig. 5) suggesting our projections can accurately capture zone-specific expression of goblet cells.

As members of the secretory cell class, the goblet and tuft cells derive from a common progenitor[88], and have previously been noted to both express common immune modulatory pathways[88]. We therefore wondered whether we could identify genes that are both zone-specific, and specific to a single lineage (goblet or tuft). We therefore identified clumps that contained at least 40% cell type abundance from the tuft cells, then projected those clumps to the tuft cell population to identify tuft zone-specific expression patterns similarly to the goblet analysis above (**Fig 3e, Supplementary Fig. 4**). To identify goblet (or tuft)-specific, zone-specific expression patterns, we computed the ratio of goblet and tuft specific expression for each gene per zone, and identified the top five genes per zone exhibiting goblet specific expression ($\log(\text{goblet}/\text{tuft}) > 0.9$) (**Fig. 3f**). The goblet-specific gene, Agr2, in the crypt zone stands out as highly expressed and goblet specific (**Fig. 3f**), and is a known landmark[91]. However, most genes that were specific to

goblet or tuft were expressed at relatively low levels (TPM<1), suggesting the differences in expression between goblet and tuft may be driven by noise.

3.3.5 Identification of spatial motifs in the primary motor cortex

The identification of spatial gene expression patterns is a task often performed at the individual gene level; many approaches have been developed to identify non-random spatial single gene expression patterns [92,93] . Here, we wondered to what extent recurring spatial patterns in cell neighborhoods could be identified. At a coarse level, the mammalian brain organizes neurons into functional neighborhoods that vary with cortical depth, as seen in recent spatial transcriptomics studies utilizing MERFISH[77,94]. We hypothesized that there might be more localized structure to cell organization in the brain, involving potentially small groups or types of cells that frequently spatially co-occur together. We term these larger groups of co-occurring cells “spatial motifs”.

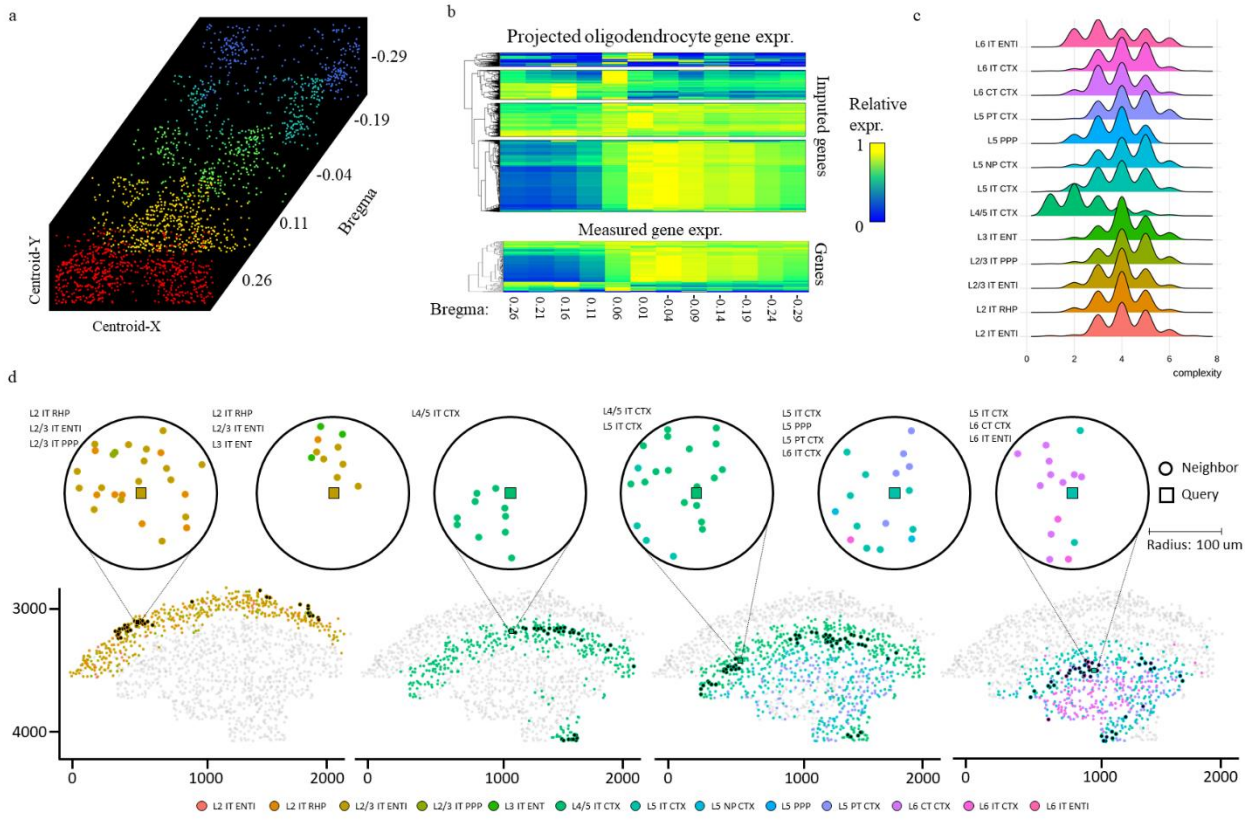


Figure 4. Imputation and high-resolution label transfer identifies spatial expression patterns in the brain. (**a**) Stacked tSNE plots of oligodendrocyte populations identified according to dominant cell type with scProjection across Bregma indices from Moffit et al. (**b**) Heatmap visualizing the spatial expression patterns within the oligodendrocytes of imputed (top) and measured (bottom) genes from the original study. (**c**) Neighborhood density plots for each cell type annotated by scProjection, where the x-axis indicates the neighborhood complexity for each cell. (**d**) tSNE plot of a single slice separated by layer type of the neurons according to the post significant neighborhoods highlighted in the circle plots for a few neurons from the mouse cortex mapped by Zhang et al. and as annotated by scProjection.

To identify spatial motifs as a function of cortical depth, we jointly analyzed a recent MERFISH study by Zhang et al.[94] and a million-neuron atlas from Yao et al.[69] of the mouse primary motor cortex (MOp). We used scProjection to infer a revised high resolution cell type cell label for each MERFISH measurement by projecting MERFISH measurements to the snRNA-seq

atlas and assigning labels based on the taxonomy of Yao et al. , which defines 129 cell types what broadly fall under the category of glutamatergic, GABAergic, and non-neuronal subtypes.

Having assigned each MERFISH measurement to one of hundreds of possible discrete high-resolution cell type labels, we first performed neighborhood analysis by quantifying, for each high-resolution label, the complexity of its physical neighborhood (within 100um radius). More specifically, we define the complexity of a cell's neighborhood as the number of distinct cell types present in a 100um radius of the cell (Methods). For each brain slice, we computed the distribution of neighborhood complexities of glutamatergic (excitatory) neurons as a function of cortical depth and high-resolution cell type annotated by scProjection. Comparing the neighborhood complexity of excitatory neurons across cortical depth revealed that at most cortical depths were comparably complex (mean complexity: 4 cell types), with the notable exception of L4/5 IT CTX neurons which were overall less complex (mean complexity: 1.5 cell types) (Fig. 4c). 24% of the L4/5 neuron cells had homogenous neighborhoods that contained no neurons from any other layer, an observation unique to the L4/5 neuron cells. This is potentially a result the fact that these neurons are a rare population in the MOp region.

Delving into the types of neighborhoods occupied by L4/5 neurons we clustered the cells by their neighborhood memberships (Methods) which identified a diverse set of neighborhood types ranging from homogenous L4/5 populations to neighborhoods which exists on the L2/3 and L6 boundaries (Figure 4d). Of note, the L4/5 IT CTX neurons were the only high-resolution cell type to form islands of neurons containing only the same type (Complexity: 1) within 100um. Even at a neighborhood radius of 500um we identified high-resolution type specific grouping of cells by neighborhood complexity indicating that each high-resolution type annotated by scProjection contain diverse neighborhood types in the MOp . By annotating higher resolution high-resolution

cell type annotations onto the MERFISH data with scProjection we can uncover neighborhood structure underlying coarser cell type spatial variation.

3.3.6 Transcriptome imputation helps infer global spatial expression patterns in the brain

Imaging-based spatial transcriptome technologies such as MERFISH and seqFISH enable imaging of individual transcripts in 2D tissue slices and therefore provide insight into spatial expression patterns at sub-cellular resolution. However, these technologies have two drawbacks: (1) it may not be practical to spatially profile all genes in the genome; for example, MERFISH experiments have profiled only hundreds of transcripts[95] to date, and (2) imaging pipelines[96] are required to segment the images into cells in order to compute single cell expression patterns, which can be an error-prone process and lead to transcripts from adjacent cells being grouped into one 'cell'[96].

To address the limitation of the smaller number of genes that can be measured by imaging-based technologies such as MERFISH and seqFISH, we modified scProjection so that even with a small, refined set of measured genes for the input RNA samples, scProjection would project those RNA samples to genome-wide expression profiles of individual cell types. Intuitively, scProjection uses direct and indirect correlation between the measured genes and missing genes (assessed from the single cell atlas) to perform non-linear imputation of gene expression measurements. In that way, scProjection could be used to simultaneously attain single cell resolution and impute the rest of the genome's expression signal.

In a study of neurons from the hypothalamic preoptic region of the mouse brain, Moffit et al. assayed 155 marker genes across millions of neurons using MERFISH and generated a paired scRNA-seq cell atlas. Using scProjection, we imputed genome-wide expression patterns for the

entire MERFISH dataset spatially profiling millions of neurons. Labeling each MERFISH sample by the cell type that contributes that most RNA. scProjection recovered the spatial organization of Oligodendrocytes across slices from the mouse brain defined by Bregma indices (**Fig. 4a**). More specifically, the oligodendrocytes spatially organize into one cluster at Brega 0.26, then eventually diverge into two populations by Bregma -0.29. To explore potential functional implications of the segmentation of oligodendrocytes from one into two spatial regions, we computed Bregma index-specific expression patterns of Oligodendrocytes between Bregma 0.26 and -0.29 and identified many genes with clear differential expression patterns across the two distal Bregma indices (**Fig. 4b**). Of particular note are Calca and Dpp10, both of whom are associated with oligodendrocyte differentiation that occurs along the bregma axis with immature and mature oligodendrocytes occupying separate compartments of the hypothalamus[95]. Neither of these markers belonged to the 155 marker gene set measured by MERFISH in the original study. scProjection therefore helps identify genes with spatially distinct expression patterns, even if they were not measured in the original spatial transcriptome assay.

3.3.7 Projection of Patch-seq RNA improves identification of connections between gene expression and neuron electrophysiology

Besides spatial transcriptome technologies, there are several other single cell resolution assays that could benefit from scProjection. For example, Patch-seq[73] is a protocol for jointly measuring the RNA, electrophysiological (ephys) and morphological properties of individual neurons, and is critical for linking the molecular and cellular properties of neurons. Patch-seq uses a micropipette to puncture a neuron in order to simultaneously measure its RNA and electrophysiological properties. When applied *in vivo* or *ex vivo* slices of brain tissue, the micropipette passes through other surrounding cells in order to reach the neuron of interest, leading to the RNA measurements

containing contributions from both the target neuron as well as surrounding glial cells[76].scProjection analysis of several Patch-seq studies indicates cell type abundances from non-neuronal cells are predicted to be as high as 30%, suggesting significant contamination of RNA (**Fig. 5a**). We therefore hypothesized that projecting Patch-seq RNA measurements to a single cell atlas of neurons would reduce the effect of contaminating RNA and improve downstream analyses such as correlating gene expression measurements to electrophysiological measurements of neurons.

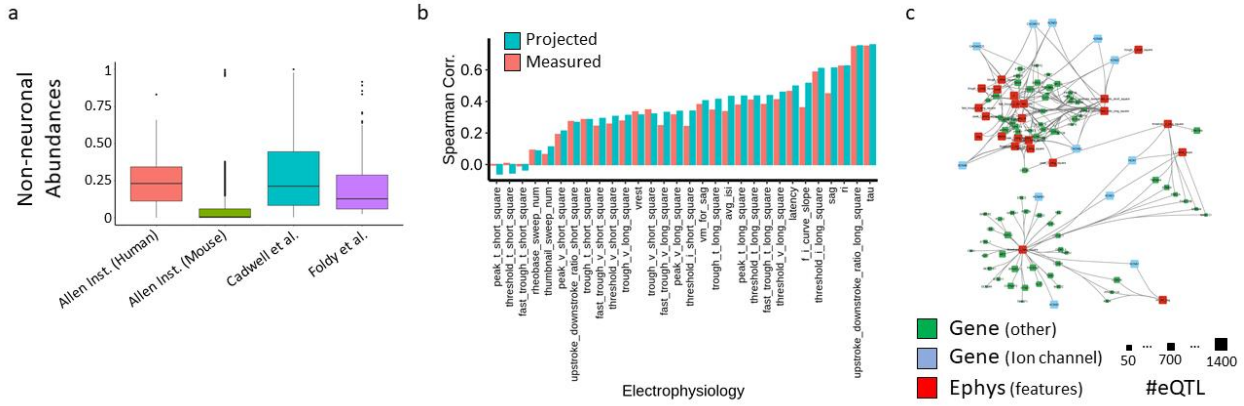


Figure 5. Projection of Patch-seq RNA links molecular measurements to electrophysiology of neurons. (**a**) Box and whisker plots visualizing the abundances of non-neuronal RNA estimated by scProjection across all samples of multiple PatchSeq studies. (**b**) Bar plot of the accuracy (based on Spearman correlation) of gene expression-based prediction of electrophysiology measurements, when predictions are made using either the original measured RNA, or the Sst projected PatchSeq samples. (**c**) Gene – electrophysiology correlation network, where edges are between significantly correlated genes and electrophysiology features. Node size is proportional to the number of eQTLs identified in the xQTL study of the ROSMAP cohort.

We applied scProjection to a set of 4,200 Patch-seq measurements targeting mouse GABAergic neurons[83], together with a reference atlas of the mouse brain[69]. Of the 4,200 measurements, scProjection predicted that 1,912 of them were primarily targeting Sst inhibitory neurons (**Supplementary Fig. 6**), consistent with the fact that these 1,912 assayed neurons were

experimentally identified as Sst before Patch-seq. We focused our experiments on the 1,912 predicted Sst inhibitory neurons because they were the best represented type of neuron and projected the Patch-seq measurements to the Sst single cells sequenced in the reference atlas.

Here we assumed that more accurate Patch-seq RNA measurements should enable better prediction of ephys properties of neurons from gene expression levels. To this end, our RNA projection enabled 27% higher prediction of two ephys features, sag and latency, from genome-wide expression profiles (spearman correlation of 0.62 compared to 0.43, p = 5e-18, rank sum test) (**Fig. 5b**). On the level of individual gene-ephys feature correlations, we found that our RNA projections led to an order of magnitude higher number of significant ($q<0.05$) correlations between Sst-projected gene expression levels and ephys properties as compared to the unprojected RNA measurements (**Fig. 5b**). Additionally, we identify cell type-specific correlations between ephys properties and ion channel genes that play a role in neuronal signaling (**Supplementary Fig. 7**). These results together suggest that RNA projections remove noise driven by the presence of non-neuronal abundances, which leads to better identification of connections between gene expression and neuron electrophysiology.

Having used scProjection to establish more gene-ephys connections than could be previously appreciated, we further hypothesized that genetic variation may drive systematic changes in some ephys features, through changes in gene expression patterns. We extracted cis-eQTLs detected in the human dorsolateral prefrontal cortex from the ROSMAP consortia[27], and found that 91 genes were both associated with genetic variation, and also correlated with ephys features of neurons. Although gene-ephys connections were identified via correlative analysis and so we cannot directly infer that these eQTLs will causally influence ephys properties in general, we looked specifically at ion channel genes because they play critical roles in establishing ephys

responses to neuron stimuli. We found 12 ion channels associated with neuronal firing and under genetic control, of which 58% of them were only identified after projection (but not with the original Patch-seq measurement). We also identified 79 genes not annotated as ion channels that are also associated with electrophysiology and eQTLs (Fig. 5c). In fact, 83% of all genes associated with the 31 ephys features are not ion channel genes. While much of the focus of interactions between genes and electrophysiology is on ion channels, our results suggest there may be many more genes that either directly influence ephys in novel ways, or indirectly interact with ion channels for example.

3.4 Discussion

In our experiments, we have demonstrated the utility of projections for the analysis of diverse single cell resolution assays such as spatial transcriptomes and Patch-seq. At its heart, projection maps RNA samples into the cell state space defined by a single cell atlas. Therefore, RNA projections can also potentially play a role in up-sampling the per-cell sequencing depth of spatial and multi-modal sequencing assays, by projecting lower depth samples into a high depth cell atlas. For example, because RNA capture is not per-cell but per-spot for technologies such as Slide-seq, the number of effective transcripts sequenced can vary spot to spot [71]. Furthermore, mRNA capture efficiencies can vary between protocols [97], and technologies such as SMART-seqv2 yield significantly high read depth per cell compared to 3' tagging technologies such as the 10x Chromium [98].scProjection can be used to project RNA samples sequenced from specialized spatial and multi-modal sequencing assays into a deeply sequenced scRNA-seq atlas for example, in order to increase the resolution of the resulting gene expression profiles. This is conceptually similar to the process of imputation that we demonstrated in our results, though imputation is typically cast

as a problem of filling in zero transcript counts rather than up sampling both non-zero and zero counts.

RNA projections are complementary to deconvolution methods. The task of deconvolution methods[99–102] is primarily to estimate the cell type abundances of a set of reference cell populations towards a single RNA sample, and is a very well-studied problem dating back several decades[103]. While scProjection also computes such cell type abundance to a set of populations, its primary goal is to distinguish intra-cell type variation by also mapping the RNA sample onto the precise cell state within each of the cell type populations that best represents the expression profile of those cell types within the RNA sample. scProjection therefore distinguishes intra-cell type variation, whereas deconvolution methods principally focus on differences in cell type abundances in a sample.

A major feature of scProjection is that it implicitly fits a probability density function (PDF) over the cell state space for each cell type. This is advantageous for several reasons. First, this enables scProjection to reason about the relative frequency of a cell state observed in the training data, where more frequently observed states have higher probability of being projected to. Second, it enables scProjection to interpolate between observed cell states when the training data is small, which can be important for training on rare cell types or on data from smaller studies. Third, scProjection can also naturally ignore outlier sequenced cells in the training data because they will not appear often in the cell atlas. In contrast, a number of other methods either average the expression profiles all cells of the same type such as CIBERSORTx that we tested here[99], or only map RNA samples to measured single cells in the atlas[104]. Methods that average cells of the same type together will be sensitive to outliers, and more importantly will be unable to account for variation within a given cell type.

One of the caveats of scProjection and related methods, is that by projecting RNA measurements to a reference single cell atlas, scProjection assumes that the single cell atlas contains accurate representations of the cell state of cell populations within the RNA sample. There could be scenarios where this is false; for example, projecting RNA from a spatial transcriptome assay of (liver) hepatocellular carcinoma samples to a normal liver atlas would miss expression variation in hepatocytes that is driven by carcinomas. Therefore, if no suitable single cell atlases are publicly available, it would make sense to collect scRNA-seq data on some biological replicate samples in addition to the spatial transcriptome datasets. This experimental design of collecting both scRNA-seq as well as spatial transcriptome data is common[71,84,105,106] so we expect this caveat to not limit the widespread applicability of scProjection.

Finally, we envision applications of RNA projections beyond what we have illustrated here. For example, databases such as the Gene Expression Omnibus (GEO) catalog gene expression data from bulk RNA samples collected since RNA sequencing was first deployed. Using the increasing number of single cell atlases derived for different tissues and cell types across organisms, scProjection can be used to re-analyze historic bulk RNA samples to extract average cell states for individual cell populations that contribute to the bulk RNA sample. Cell type-specific changes in case-control studies could then be inferred, as could cell type-specific eQTLs from genetic studies of disease, for example.

3.5 METHODS

scProjection overview. Our framework, scProjection, projects N gene expression profiles $\boldsymbol{b}_n \in B$ generated from RNA samples into each of K different cell populations represented in a reference single cell atlas, yielding a new set of gene expression profiles $\boldsymbol{x}_{n,k}$, for $k = 1, \ldots, K$. scProjection also estimates $\alpha_{n,k}$, the proportion of RNA contributed by each cell population k to sample n (**Fig. 1**). scProjection assumes that each $\boldsymbol{b}_n$ is a weighted linear combination of the cell population-specific projections $\boldsymbol{x}_{n,k}$:

$$\boldsymbol{b}_n = \sum_{k}^{K} \alpha_{n,k} \boldsymbol{x}_{n,k}$$

Only $\boldsymbol{b}_n$ is formally observed, and the goal is to estimate $\alpha_{n,k}$ and $\boldsymbol{x}_{n,k}$.

To perform estimation, scProjection leverages a separate reference single cell atlas in which single cells $\boldsymbol{s}_{j,k}$ (representing the j^{th} cell sequenced for cell population k in the atlas S) have been sequenced. In the first step, scProjection trains a deep variational autoencoder (VAE) separately for each cell population k using all single cells sequenced for cell population k ($s_{*,k}$), yielding a parameter set $\{\phi_k, \theta_k\}$ (representing the encoder and decoder parameters, respectively) for each cell population k. After training, each VAE implicitly defines the set of cell states that projections into cell population k ($x_{n,k}$) can occupy. In the second step, the VAEs with trained parameters $\left\{\hat{\phi}_k^{(0)}, \hat{\theta}_k^{(0)}\right\}$ are used to get initial projections $\hat{\boldsymbol{x}}_{n,k}^{(0)}$ by inputting each $\boldsymbol{b}_n$ into the k^{th} VAE and sampling from the output to estimate $\hat{\boldsymbol{x}}_{n,k}^{(0)}$. In the second step, we estimate the RNA proportions $\hat{\alpha}_{n,k}$ by solving the above equation by using linear regression by setting $\boldsymbol{x}_{n,k} = \hat{\boldsymbol{x}}_{n,k}^{(0)}$. Finally in the third step, we fix the mixing proportions $\hat{\alpha}_{n,k}$, and re-update all VAE parameters $\{\phi_k, \theta_k\}$ simultaneously to improve estimates of $\boldsymbol{x}_{n,k}$ by maximizing the reconstruction of each $\boldsymbol{b}_n$.

scProjection training of cell population-specific VAEs (Step 1).

scProjection uses VAEs to perform the projection of RNA samples $\boldsymbol{b}_n$ into the gene expression space of each cell population k to yield the projection $\boldsymbol{x}_{n,k}$. The set of cell population-specific VAEs are identical in network structure and are comprised of a deep encoder network parameterized by weights ϕ_k, and decoder network parameterized by weights θ_k. To train the VAEs, we optimize the following objection function with respect to the VAE parameters $\{\phi_k, \theta_k\}$:

$$L(\{\phi_k, \theta_k\}; \{\boldsymbol{s}_{j,k}\}) = \sum_{k=1}^{K}\sum_{j=1}^{J} E_{q_{\phi_k}(\boldsymbol{z}_{j,k}|\boldsymbol{s}_{j,k})}[\log p_{\theta_k}(\boldsymbol{s}_{j,k}|\boldsymbol{z}_{j,k})] - \prod_{k=1}^{K}\prod_{j=1}^{J} D_{KL}[q_{\phi_k}(\boldsymbol{z}_{j,k}|\,\boldsymbol{s}_{j,k})||p(\boldsymbol{z}_{j,k})]$$

$$q_{\phi_k}(\boldsymbol{z}_{j,k}|\boldsymbol{s}_{j,k}) = N(\boldsymbol{z}_{j,k}; \mu_{\phi_k}(s_{j,k}), \sigma^2_{\phi_k}(s_{j,k})I)$$

$$p_{\theta_k}(\boldsymbol{s}_{j,k}|\,\boldsymbol{z}_{j,k}) = N(\,\boldsymbol{s}_{j,k}; \mu_{\theta_k}(z_{j,k}), \sigma^2_{\theta_k}(z_{j,k})I)$$

The functions $\{\mu_{\phi_k}(\cdot), \sigma^2_{\phi_k}(\cdot)\}$ and $\{\mu_{\theta_k}(\cdot), \sigma^2_{\theta_k}(\cdot)\}$ represent the mean and variance of the normal distribution predicted by the encoder and decoder, respectively. The parameters of the VAEs $\{\phi_k, \theta_k\}$ are regularized through 30% dropout [13], batch normalization [14] and L2 weight regularization to ensure robust training. ADAM [15] is used for optimization with a decaying learning rate starting at 1e-3 and a smooth warmup of the KL term in the ELBO, which has been shown to produce more accurate reconstructions [107]. We denote the trained VAE parameters by $\{\hat{\phi}_k^{(0)}, \hat{\theta}_k^{(0)}\}$.

For the experiments in which we impute genome-wide expression measurements from limited sets of marker genes such as those measured by MERFISH, the structure of the VAE becomes asymmetric with the input measurements to the encoder defined by a subset of gene expression measurements $G_e \subseteq G$ (corresponding to marker genes). The decoder output is still defined by the full set of gene expression measurements G made in the single cell atlas. Only estimates of those genes G_e directly measured in mixture samples b_n are used in subsequent steps of scProjection.

scProjection estimation of cell type abudnacne of each cell population (Step 2).

Here, scProjection projects each RNA sample $\boldsymbol{b}_n$ to each cell population k via the VAE parameterized by $\left\{\hat{\phi}_k^{(0)}, \hat{\theta}_k^{(0)}\right\}$ to estimate $\hat{\boldsymbol{x}}_{n,k}^{(0)}$:

$$\hat{\boldsymbol{x}}_{n,k}^{(0)} = \mu_{\hat{\theta}_k^{(0)}}\left(\mu_{\hat{\phi}_k^{(0)}}(\boldsymbol{b}_n)\right)$$

Then, we estimate the mixture proportions $\alpha_{n,k}$ and nuisance parameters of a multi-layer perceptron f_{σ_b} (and hold all other variables fixed) by optimizing the following objective function:

$$L(b_n) = \sum_{n=1}^{N} \log N(b_n \mid \sum_{k}^{K} \hat{x}_{j,k}^{(0)} \alpha_{n,k}, f_{\sigma_b}(\sigma_{n,k}^2 \oplus \alpha_{n,k})I)$$

Optimization is performed with ADAM [15] and a learning rate of 1e-3 until convergence. The estimated mixing proportions $\hat{\alpha}_{n,k}$ are kept fixed for the remainder of the training procedure.

scProjection final estimates of RNA projections (Step 3).

In this step, scProjection re-optimizes the encoder and decoders of the individual VAEs $\{\phi_k, \theta_k\}$ by minimizing the following composite objective function, which includes the likelihood of both the single cell atlas data $s_{j,k}$ and the RNA samples b_n:

$$\text{ELBO} = \sum_{n}^{B} \log\ N(\boldsymbol{b}_n \mid \sum_{k=1}^{K} \mu_{\theta_k}\left(\mu_{\phi_k}(\boldsymbol{b}_n)\right)\hat{\alpha}_{n,k}, f_{\sigma_b}(\sigma_{n,k}^2 \oplus \hat{\alpha}_{n,k})I) +$$

$$\sum_{k}\sum_{j} E_{q_{\phi_k}(\boldsymbol{z}_{j,k}|\boldsymbol{s}_{j,k})}[\log p_{\theta_k}(\boldsymbol{s}_{j,k}|\boldsymbol{z}_{j,k})] -$$

$$\left[\sum_{k}\sum_{n} D_{KL}[q_{\phi_k}(\boldsymbol{z}_{n,k} \mid \boldsymbol{b}_n)||p(\boldsymbol{z}_{n,k})] + \sum_{k}\sum_{j} D_{KL}[q_{\phi_k}(\boldsymbol{z}_{j,k} \mid \boldsymbol{s}_{j,k})||p(\boldsymbol{z}_{j,k})\right]$$

Note in this case, the VAE parameters are initially set to $\phi_k = \hat{\phi}_k^{(0)}$ and $\theta_k = \hat{\theta}_k^{(0)}$ before optimization, and the parameters of f_{σ_b} are fixed at their values estimated at Step 2. Intuitively, we are adjusting the RNA projections $\boldsymbol{x}_{n,k} = \mu_{\theta_k}\left(\mu_{\phi_k}(\boldsymbol{b}_n)\right)$ to better predict the RNA sample $\boldsymbol{b}_n$, because the single cell reference data may be collected in a different experiment from the RNA samples. The single cell data are included in the objective function and serve as a regularization term to ensure identifiability of each VAE as specific to one cell population k. After training to obtain final VAE parameter estimates $\left\{\hat{\phi}_k^{(1)}, \hat{\theta}_k^{(1)}\right\}$, we estimate our final RNA projections $\hat{\boldsymbol{x}}_{n,k}^{(1)} = \mu_{\hat{\theta}_k^{(1)}}\left(\mu_{\hat{\phi}_k^{(1)}}(\boldsymbol{b}_n)\right)$.

Acquisition and preprocessing of the intestinal villus dataset.

We obtained the gene expression matrices for the LCM-seq, scRNA-seq and spatial reconstructions experiments described in Moor et al.[84] from GSE109413 and https://doi.org/10.5281/zenodo.1320734. We independently normalized the count matrices to TP10K, then scaled and centered using Seurat's NormalizeData (without log transform) and ScaleData functions. We retained the union of the marker genes of each cell type identified from the original study, together with the top 2,000 variable genes of both LCM-seq and scRNA-seq.

Acquisition and preprocessing of the brain MERFISH dataset.

We obtained the processed MERFISH gene luminescence matrix described in Moffitt et al.[95] from dryad.8t8s248 and the scRNA-seq count matrix from GSE113576. We independently preprocessed each data modality by normalizing to TP10K, then scaled and centered using Seurat's NormalizeData and ScaleData functions. We removed entire cell types from the scRNA-seq data that had no analog in the MERFISH experiments and are defined in Table S9. We retained the union of the marker genes of each cell type identified in the original study, together with the top 2,000 variable genes across the entire scRNA-seq atlas.

Acquisition and preprocessing of the mouse Patch-seq dataset.

We obtained the gene count matrix for the mouse Patch-seq experiments described in Berg et al.[108] from portal.brain-map.org/explore/classes/multimodal-characterization on January 2019. We discarded samples that did not pass QC as defined in the original paper in both the RNA and electrophysiology modalities. We normalized the count matrix to TP10K (without log transform), then scaled and centered using Seurat's NormalizeData and ScaleData functions. We retained the union of the marker genes of each cell type identified from the original study, together with the top 2,000 variable genes across each of the cell types defined in the snRNA-seq.

Acquisition and preprocessing of the mouse brain atlas.

We obtained the gene count matrix for the human brain atlas described in Yoa et al.[69] from the Allen Institute Cell Types database: RNA-Seq data page on the Allen Institute's webpage. We normalized the count matrix to TP10K, then scaled and centered using Seurat's NormalizeData and ScaleData functions. We retained the union of the marker genes of each cell type reported in the original study, together with the top 2,000 variable genes across each of the cell types defined in the snRNA-seq.

Acquisition and preprocessing of the Tasic et al. mouse brain atlas.

We obtained the gene count matrix for the mouse brain atlas described in Tasic et al. from the Allen Institute Cell Types database: RNA-Seq data page on the Allen Institute's webpage. We normalized the count matrix to TP10K, then scaled and centered using Seurat's NormalizeData and ScaleData functions. We retained the union of the marker genes of each cell type reported in the original study, together with the top 2,000 variable genes across each of the cell types defined in the snRNA-seq.

Acquisition and preprocessing of the CellBench benchmark.

We obtained the gene count matrix for the RNA mixture experiments in CellBench described in Tian et al.[57] from the R data file mRNAmix_qc.RData available on GitHub (https://github.com/Shians/CellBench). We normalized the count matrix to TP10K (without log transform), then scaled and centered using Seurat's NormalizeData and ScaleData functions. We retained the union of the marker genes of each cell type identified by CIBERSORTx, together with the top 3,000 variable genes computed separately on the RNA mixtures profiled on CEL-Seq2 and SORT-Seq.

Acquisition and preprocessing of the ROSMAP-IHC benchmark.

We obtained the gene count matrix for the bulk-RNA experiments and IHC measurements described in Patrick et al. from the R data files available on Github (https://github.com/ellispatrick/CortexCellDeconv). We normalized the count matrix to TP10K (without log transform), then scaled and centered using Seurat's NormalizeData and ScaleData functions. We retained the union of the marker genes of each cell type reported in Darmanis et al.[109], together with the top 2,000 variable genes.

Execution of deconvolution methods.

In the two sections below on benchmarking cell proportion estimations in different datasets, we compared scProjection against CIBERSORTx[99], MuSiC[100], NNLS, dtangle[101], DSA[102], and single gene deconvolution. Each method was run based on method-specific guidelines provided by the original authors and following the workflows defined by in tutorials for each approach. Prior to running each method, the FindVariableGenes function implemented in Seurat was used to identify the most variable genes for a consistent subsetting of the data matrices. CIBERSORTx was provided counts for all highly variable genes in the scRNA-seq data along with cell type annotations to create a signature matrix. Then counts for all highly variable genes in the mixture data were provided to CIBERSORTx which then estimates RNA proportions. MuSiC was provided counts for all highly variable genes in the scRNA-seq and mixture data

along with cell type annotations. NNLS (as implemented by us in R) was provided the TPM values for all highly variable genes in the scRNA-seq and mixture data. Proportions from NNLS for cell type k were computed by summing the learned weights across all cells annotated as cell type k; this was repeated for each cell type and each mixture sample. dtangle was provided with a mean count vector per cell type in the scRNA-seq data and the original counts from the mixture data along with cell type markers and annotations. DSA was provided with the original counts for the mixture data and cell type specific marker genes. Single gene deconvolution was performed by identifying individual marker genes of each cell type, which were used to estimate the relative proportion of each cell type with respect to the remaining markers.

Benchmarking cell population proportion estimation on the CellBench dataset.

The CellBench dataset provides gene expression profiles obtained from sequencing titrated RNA mixtures from three human lung adenocarcinoma cell lines (H1975, H2228, HCC827), as well as single cell RNA profiles from each cell line. Sequencing was performed using either plate based (CEL-Seq2 or Drop-Seq) or droplet based (10x Chromium and Drop-seq Dolomite) protocols. The proportion of RNA from each cell line was recorded for each mixture and defines a baseline for methods aiming to computationally estimate the RNA percentages. We trained scProjection using the RNA mixtures as inputs b_n and the single cell data as the atlas S. We treated the scProjection estimates $\hat{\alpha}_{n,k}$ as our predictions of abundances for each cell type. We then compared scProjection-based deconvolution against other methods as described above (Supplementary Figure 8).

Benchmarking cell population proportion estimation on the ROSMAP-IHC dataset.

To provide a more challenging and realistic deconvolution benchmark, we used the ROSMAP-IHC dataset consisting of 70 bulk RNA samples of the dorsolateral prefrontal cortex (DLPFC), an scRNA-seq atlas derived from the DLPFC, and cell population proportions estimated using IHC from adjacent samples to

those samples used for sequencing. The bulk RNA, reference single cell atlas and cell population proportions were collected and estimated in three different studies, thus introducing technical and biological variability between data modalities that does not exist in the CellBench study. We trained scProjection using the RNA mixtures as inputs b_n and the single cell data as the atlas S. We treated the scProjection estimates $\hat{\alpha}_{n,k}$ as our predictions of abundances for each cell type. We then compared scProjection-based deconvolution against other methods as described above (Supplementary Figure 9, 10). Furthermore, for each proportion estimated by scProjection we assign a confidence score indicating the certainty of the mixture being assigned to a specific cell type (Supplementary Figure 11).

Prediction of cell population using scProjection.

From scProjection's estimates of cell population specific proportions, treated as probabilistic class assignments, the class with maximal probability is assigned as the cell population label for each sample.

Cell annotation with KNN label transfer

After estimating the projection of a mixture onto a single cell atlas the projected mixture is labeled based on annotations in the single cell atlas of its 5-nearest scRNA-seq neighbors.

Zonated gene expression scoring

For each gene we compute the distance from an idealized zone-specific measurement as the difference between $gene_{ideal}$= (1,0,0,0,0) and the computed gene zonation score vector. A threshold was set based on the 75th quantile of the resulting scores to compare the number of zonated genes across methods.

Constructing a gold standard set of zonated goblet expression patterns based on clumps.

Recently, Manco et al. sequenced 'clumps' consisting of multiple physically-proximal cells from partially-dissociated intestinal villi. Using scProjection, we performed expression deconvolution and identified enterocyte-goblet clumps that contained both enterocytes and goblet cells, using a single cell atlas of enterocytes[84] and goblet cells[85]. Based on our zonated enterocyte expression patterns (Fig. 4b), we predicted the zone of each enterocyte-goblet clump based on the projection of the enterocyte-goblet clump onto the enterocyte single cell atlas. Because the goblets in the enterocyte-goblet clumps are physically proximal to the enterocytes, we then assumed the goblets in each enterocyte -goblet clump was from the same zone as the projected enterocyte. For each zone, we identify all enterocyte-goblet clumps from that zone, project the enterocyte -goblets to the goblet single cell atlas, and average across all such projections to estimate zone-specific goblet expression.

Supplemental Note 1:

As a first step, we performed experiments to determine the extent to which scProjection can identify the primary cell type of an RNA sample. Specifically, we benchmarked the deconvolution performance of scProjection utilizing recent bulk RNA-seq studies for which the proportions of each cell type in the mixed RNA samples were experimentally determined[81,82]. Our first benchmark is CellBench[81], a dataset where mixed RNA samples were experimentally constructed by mixing RNA from three human lung adenocarcinoma cell lines and varying either the relative concentrations of RNA content or the numbers of cells. With an ideally matched scRNA-seq atlas, the deconvolution of CellBench mixtures was not a challenging task and most methods estimated the proportions near perfectly (avg. acc: 0.95) (Supplementary Fig. 8). The second benchmark is the ROSMAP dataset[82], consisting of 70 bulk RNA and 80,660 scRNA-seq samples from the dorsolateral prefrontal cortex, where abundances per cell type were estimated based on immunohistochemistry (IHC). The ROSMAP dataset presented a more challenging task than CellBench due to increased technical and biological variation between the bulk RNA samples and the reference single cell atlas. For ROSMAP, scProjection clearly performs best across all tested methods (MSE: 0.04 compared to median MSE: 1.3 for other methods) with respect to estimating cell type proportions of each bulk sample.

In contrast, the remaining model based methods, methods which do not need marker gene sets per cell type, overestimated the neuronal content of each sample and underestimated the rarer non-neuronal cell types (**Supplementary Fig. 9, 10**). A novel aspect of scProjection is the ability to compute, per sample, an estimate of the likelihood for each cell type proportion that enables the identification of samples with low concordance to the atlas (**Supplementary Fig. 11**) and higher error in reconstruction of the original mixture measurement.

Supplemental Note 2:

We first used the CellBench[57] benchmark to validate that projections of mixed RNA samples to individual cell populations yield cell states that resemble the single cells used to train scProjection. CellBench is a dataset which consists of scRNA-seq datasets generated on three human lung adenocarcinoma cell lines (H1975, H2228, HCC827), as well as bulk RNA mixtures of all three cell lines combined at varying magnitudes. We used scProjection to project 636 mixed RNA samples to each of the composite H1975, H2228, and HCC827 cell populations. ScProjection estimated gene expression profiles and likelihood (Supplementary Fig. 12) for each of three cell lines per input for all 636 RNA mixtures. Projections were highly correlated (rho>=0.98) with the average measured scRNA-seq profiles for each cell line, suggesting projections globally look similar to the single cell data. We also demonstrated that our projections retain the gene co-expression networks exhibited in the measured single cell atlas after imputing for gene expression dropout (Supplementary Fig. 13) as compared to the deconvolution method CIBERSORTx[99]. We saw similar results on another benchmark, the ROSMAP-IHC dataset[82], where projections of 70 bulk RNA samples onto five cell types were also similar to the cell type averages (rho = 0.91)despite the higher degree of technical and biological variation. ScProjection therefore projects cell population-specific expression profiles that are consistent with the single cell measured profiles from the reference single cell atlas.

Supplemental Note 3:

Having validated scProjection's predictions of the dominant cell type from a single sample, we next assessed scProjection for the ability to impute genome-wide expression measurements using a limited set of marker genes. In a study of neurons from the hypothalamic preoptic region of the mouse brain, Moffit et al. assayed 155 marker genes across millions of neurons using MERFISH, and generated a paired scRNA-seq cell atlas[95]. Using the scRNA-seq cell atlas, for each individual cell population, we performed a series of experiments where we randomly sampled a cell from the atlas, extracted the expression levels of only the 155 marker genes used for MERFISH, and used scProjection to impute ~4000 genes from the 155 marker genes. We found the projected and measured gene expression patterns correlated well and the predicted cell state agreed with the original scRNA-seq measurement (Spearman rho=0.63, p=2e-8). We also found that scProjection could identify the correct cell type in 0.78 of cells (**Supplementary Fig. 14**). These results suggest scProjection can be used to impute genome-wide expression profiles based only on marker gene expression.

3.6 Supplementary Materials

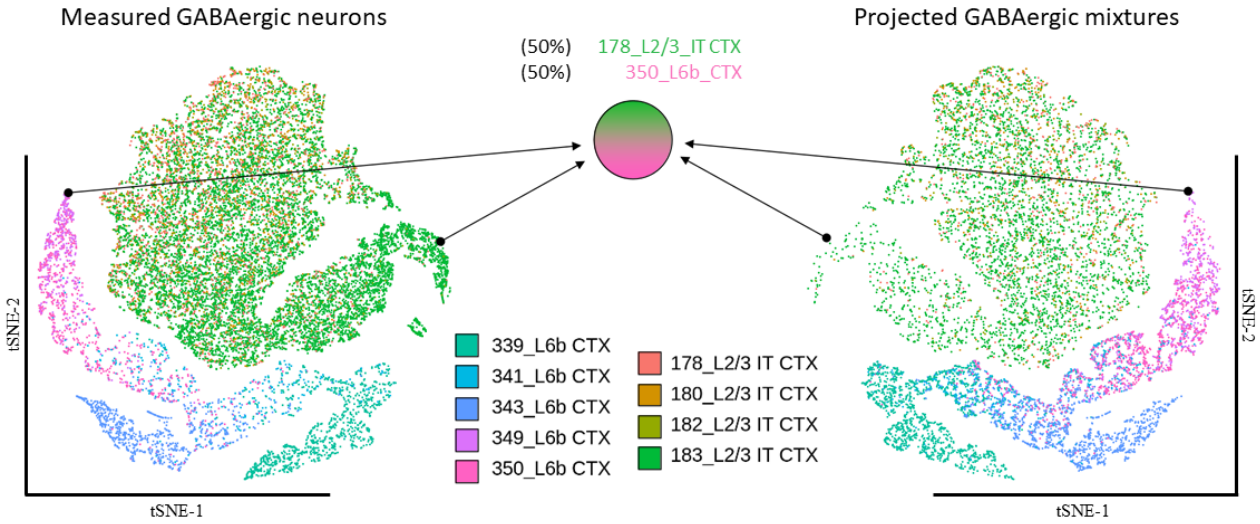


Fig S1: Simulated mixed RNA samples from L2/3 and L6b neuronal populations. Simulated samples were generated by selecting pairs of cells, one from each population, and adding them together. Samples were projected back onto the two populations to determine how close

projections were to the originally selected cells. Shown is an example of a single mixture generated by selecting two cells (from the tSNE plot on the left), and projecting them back onto those same populations (to the tSNE plot on the right). tSNE plot visualizes the measured GABAergic gene expression patterns (left) and projected mixtures (right) in the same low dimensional space (with a reversed tSNE axis on the right for visualization symmetry).

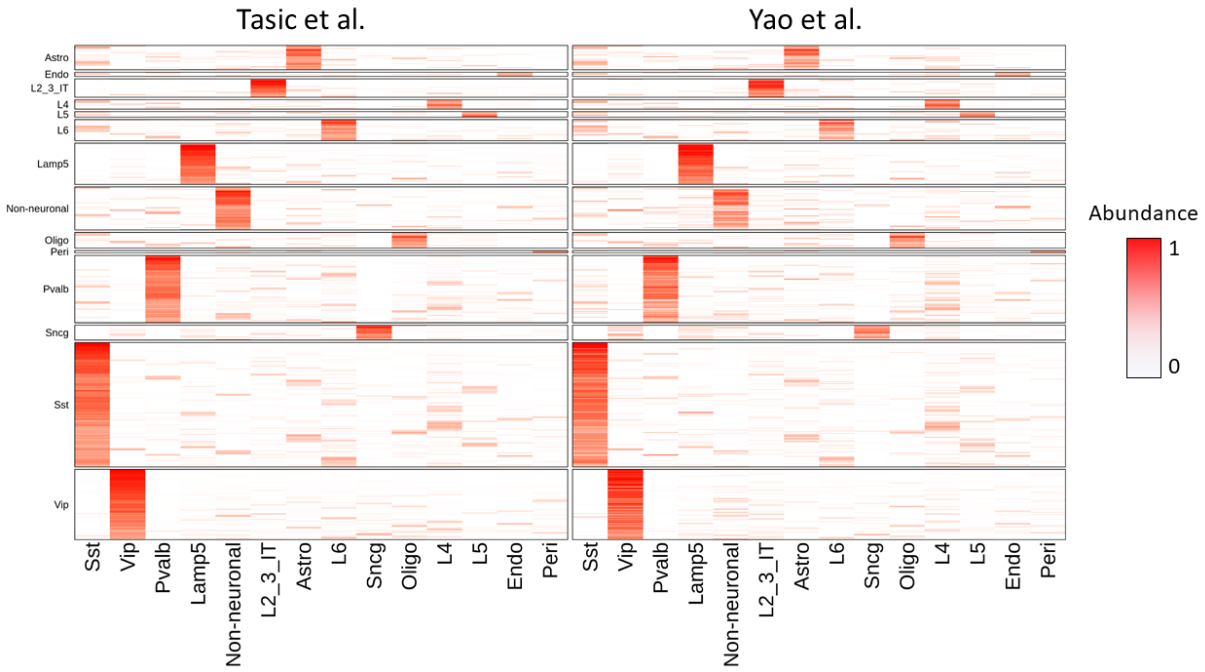


Fig S2: Estimated cell type abundances using multiple atlases of the mouse cortex. Heatmaps visualize the abundance of each cell type (columns) for each PatchSeq sample (rows) based on training scProjection using either the Tasic et al. atlas or a recent atlas from Yao et al.

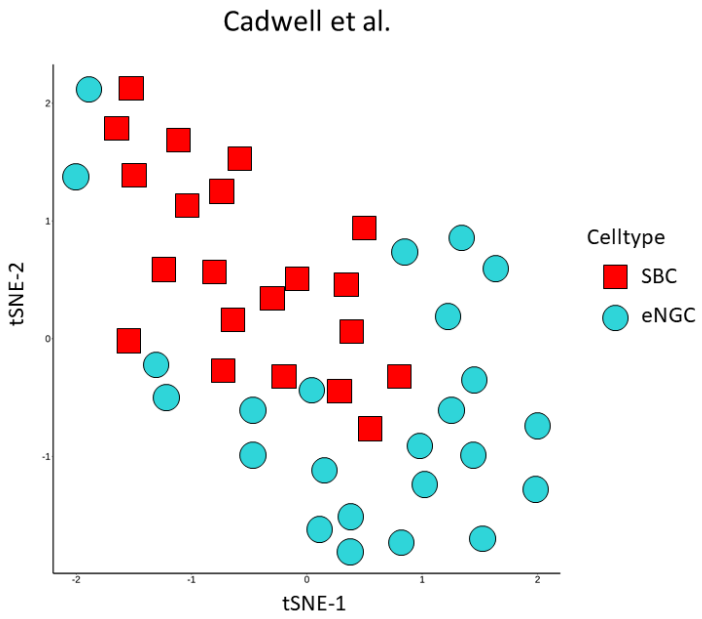


Fig S3: Embeddings of SBC and eNGC neurons. tSNE plot visualizes the separation of SBC and eNGC neurons from Cadwell et al. projection by scProjection.

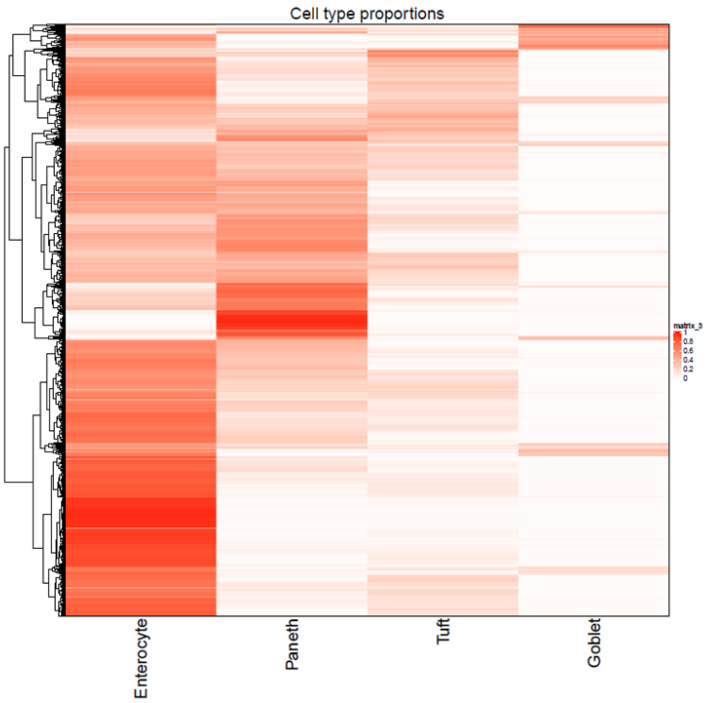


Fig S4: Estimated cell type abundances for clump-seq samples. Heatmap visualizes the cell type abundances of each cell type (columns) for each clump (rows) in Manco et al.

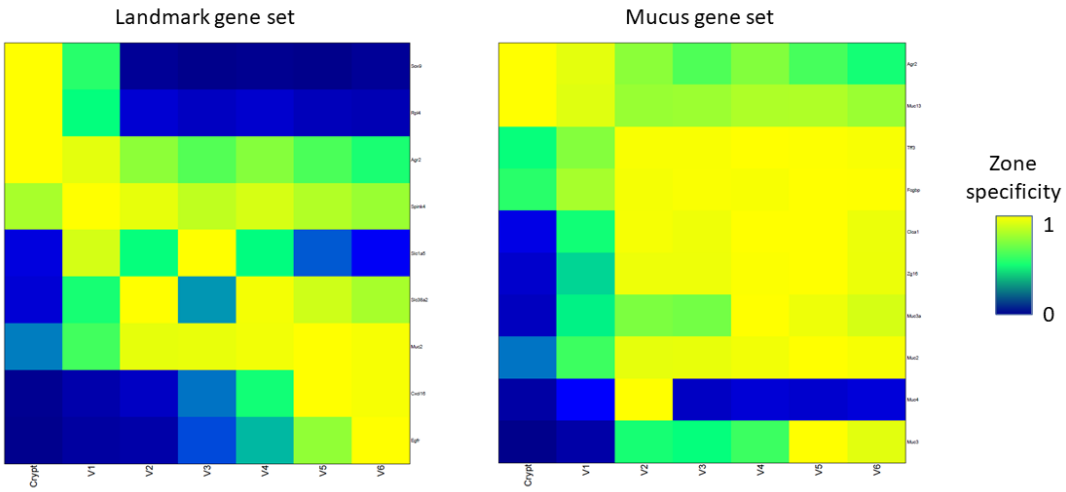


Fig S5: Zonated expression of projected goblet clumps. Heatmaps visualize the zonation pattern for goblet specific expression of landmark genes (left) and mucus genes (right), where yellow indicates increased zone specific expression.

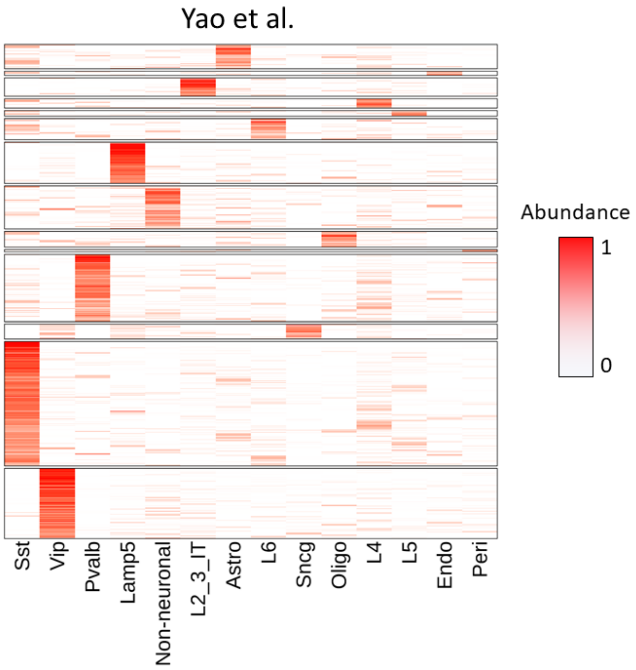


Fig S6: Estimated cell type abundances for mouse PatchSeq data. Heatmap visualizes the cell type abundances of each cell type (columns) for each neuron (rows) in the Allen mouse PatchSeq data.

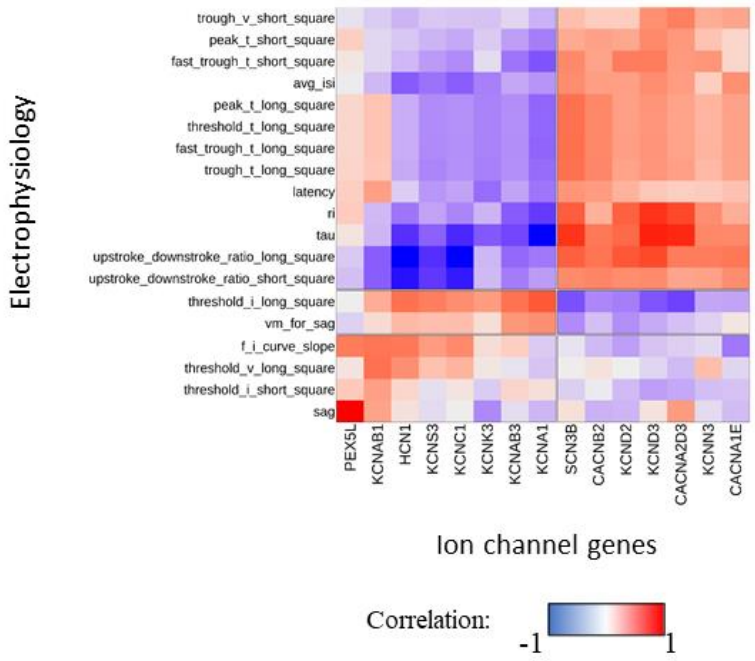


Fig S7: Correlation of ion channel genes with electrophysiology features. Heatmap visualizes the correlation of the most variable ion channel genes that play a role in neuronal signaling (columns) with electrophysiology features (rows).

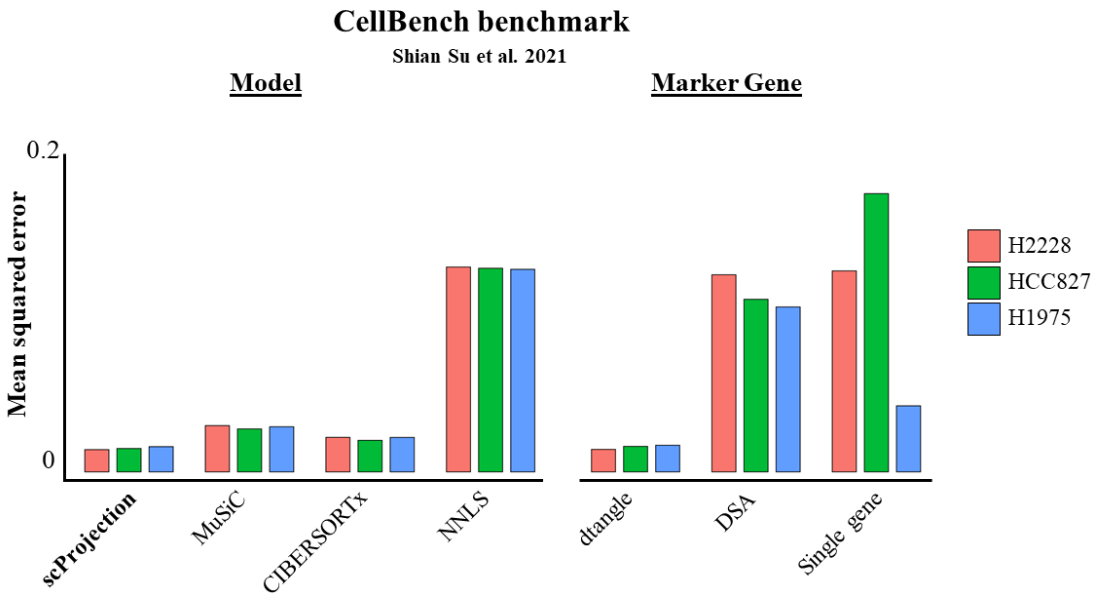


Fig S8: Benchmarking of deconvolution methods on CellBench. Barplots indicate the error in predicted cell type abundances compared against the ground truth for each deconvolution method.

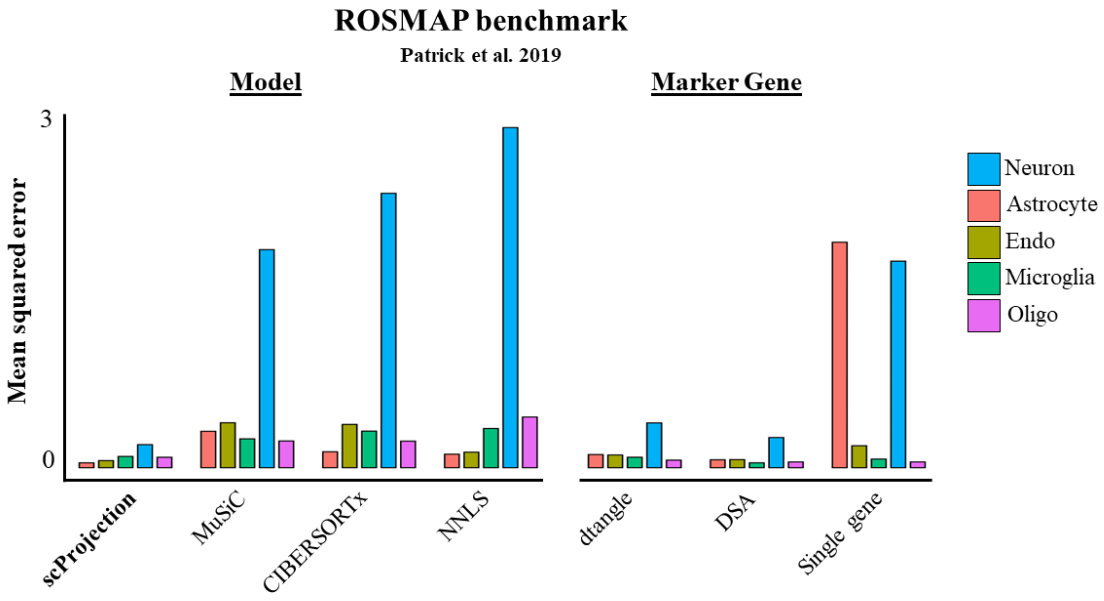


Fig S9: Benchmarking of deconvolution methods on ROSMAP. Barplots indicate the error in predicted cell type abundances compared against the ground truth for each deconvolution method.

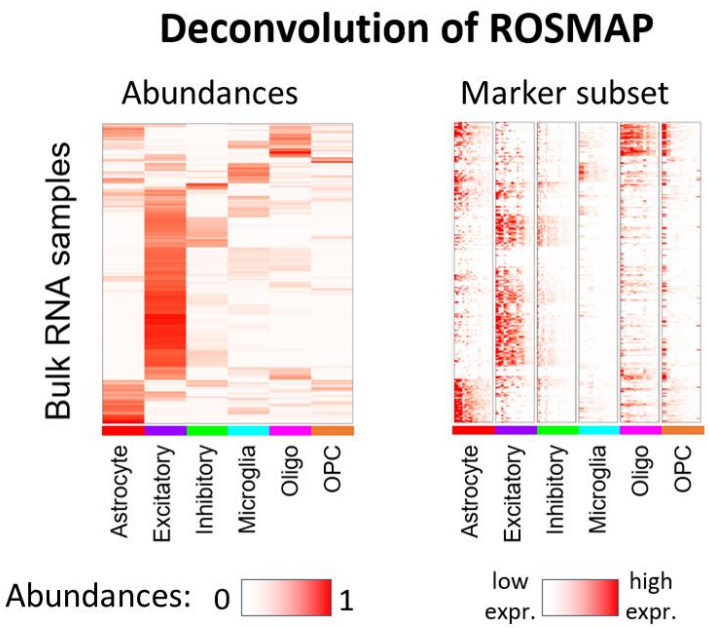


Fig S10: Cell type abundances and marker gene expression of ROSMAP bulk samples. Left heatmap indicates the estimated abundances of cell type (columns) for each bulk sample (rows), and the right heatmap shows the expression of the top marker gene expression (columns) for each cell type.

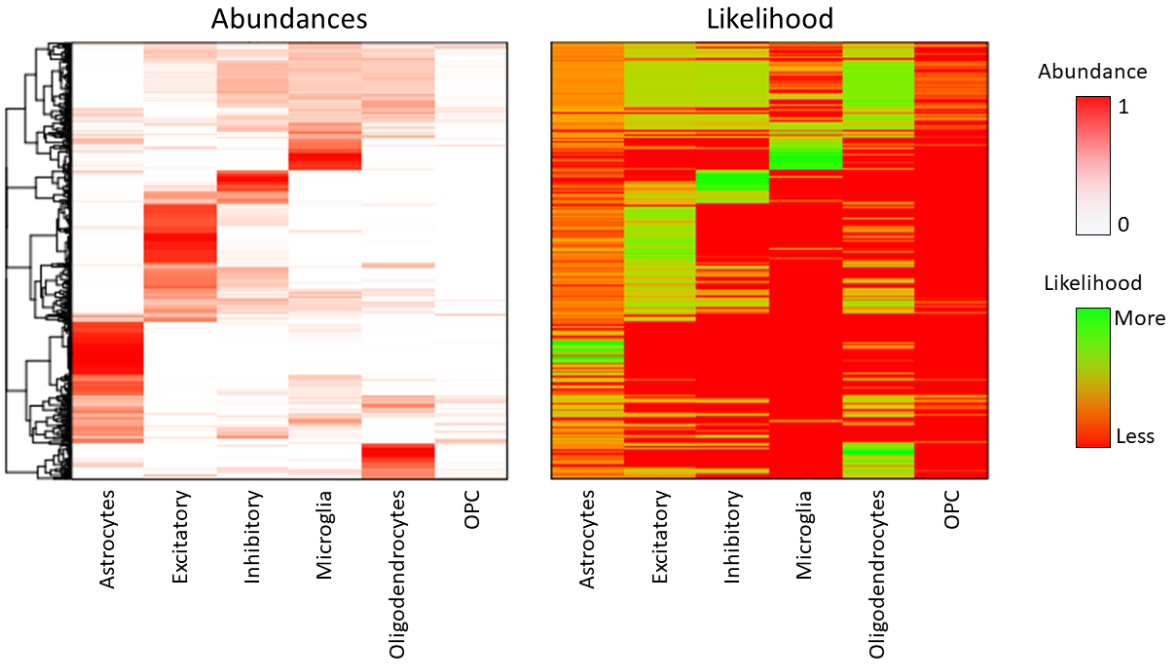


Fig S11: Estimated cell type abundances and likelihood for ROSMAP bulk samples. Left heatmap indicate the estimated abundances of each cell type (columns) for each bulk sample (rows), and the right heatmap shows the likelihood of each abundance value for each cell type (columns).

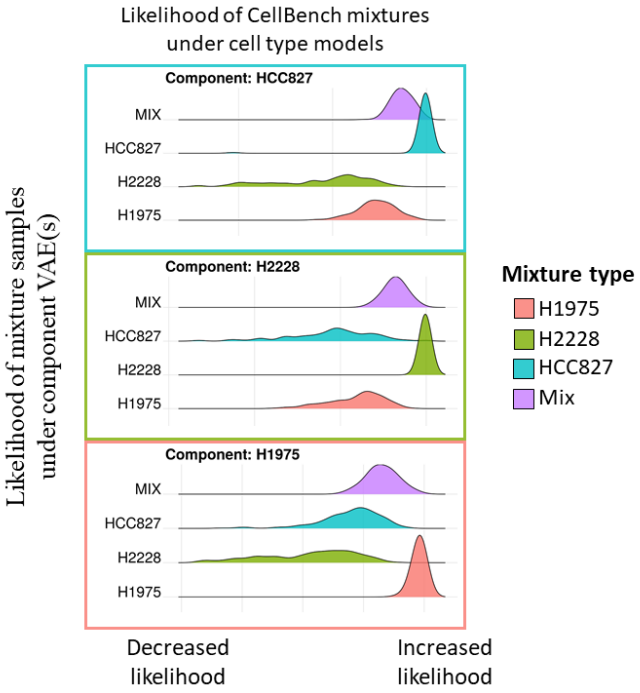


Fig S12: Likelihood of CellBench mixtures under each component VAE in scProjection. Density plots for each component VAE (bounded boxes) indicate the likelihood of each mixture type under the corresponding cell type model trained by scProjection.

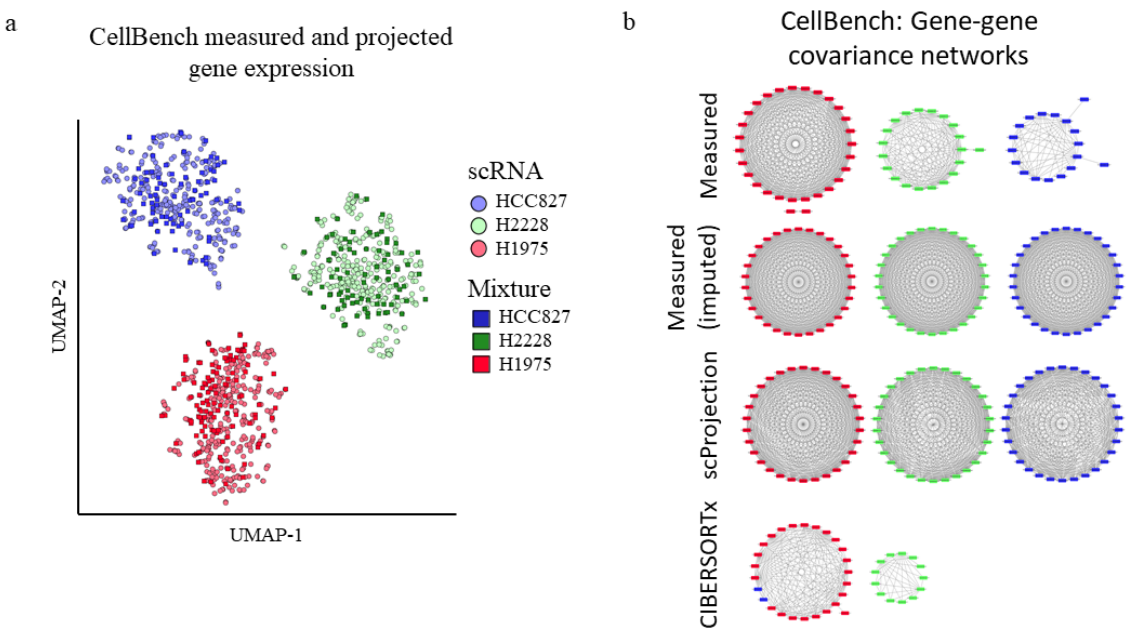


Fig S13: Validation of projected CellBench mixtures. (**a**) Left tSNE plot visualies the measured scRNA-seq data overlayed with the projected mixtures for each cell type. (**b**) Gene-gene correlation networks indicates the differences in conservation of gene structure by scProjection and CIBERSORTx.

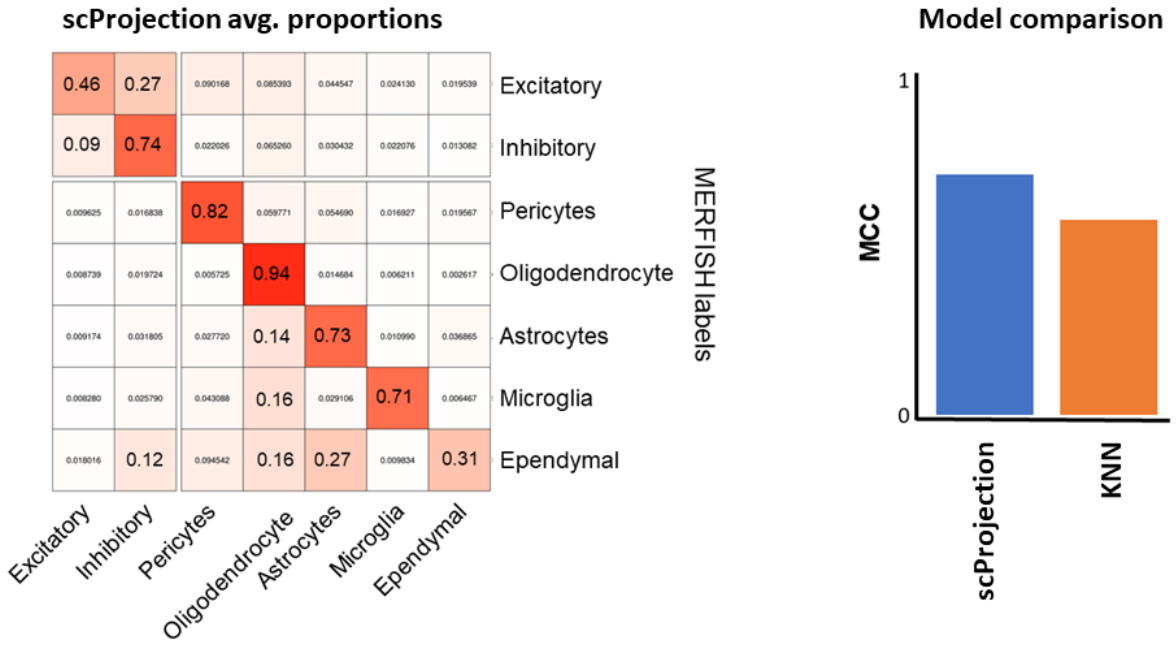


Fig S14: Validation of scProjection cell type annotation on MERFISH. Heatmap visualizes the average abundance for each cell type called by scProjection on the columns and the label annotated in the original MERFISH study. The barplot shows the classification performance of scProjection against a k-nearest neighbors approach which labeled based on neighborhood proximity of the MERFISH data mapped onto the scRNA-seq atlas.

Chapter 4

Conclusions and Future Directions

4.1 Conclusions

The single cell transcriptomics technologies that now capture highly sensitive measurements of gene expression have almost followed a Moore's Law with exponential scaling in the number of cells that can be sequenced and analyzed simultaneously. A critical task in single cell genomics is to ensure transcriptomic measurements from millions of cells sequencing under varying batch effects are comparable to enable joint analysis of variation within complex tissues. Furthermore, leveraging deeply sequenced single cell atlases with novel methods sheds new light on historic studies of disease which utilized bulk RNA sequencing, to uncover the cell-type specific variability. The aim of this thesis is to progress the tool development in computational biology towards models, neural networks, which can scale along with sequencing technologies throughput to enable rapid analysis and accurate normalization of non-biological confounders.

In this thesis I presented scAlign, a model to align multiple single cell RNA sequencing datasets into a common expression space in which the underlying biology and cell state were preserved. Previous, methods could not handle non-linear technical variation as well as cases in which the cell type distribution did not agree across datasets. Additionally, we introduced a novel form of per-cell differential expression using techniques from style transfer in computer vision to enable high-resolution mapping of gene expression patterns under specific disease states as I showed in the analysis of malaria transmission. Along with these primary contributions of scAlign, I demonstrated that neural networks are flexible models which

can scale to single cell genomics analysis of millions of cells and I believe will become the foundation for next-generation of single cell analysis.

Proving the flexibility of neural networks, we presented scProjection a hierarchical generative neural network for the deconvolution of bulk RNA, clumped spatial transcriptomics and contaminated multi-omics technologies such as PatchSeq. Prior work in deconvolution were limited in the ability to capture rare-cell types and complex variation in complex systems such as the human brain. scProjection improves not only on the ability to estimate cell type abundances as well as the provides the ability to recover the exact cell state of individual cell types underlying transcriptomic measurements from mixed RNA samples. I demonstrated that scProjection can take spatial transcriptomics where each sample contains RNA from multiple cell types and accurately recover the underlying cell states enabling higher resolution analysis of cell neighborhoods in the mouse cortex. We also found that by removing contaminating RNA from multi-omics technologies such as PatchSeq identified more significant associations between molecular and functional measurements of individual neurons. scProjection is the only deconvolution method that can accurately recover cell states from mixed samples that faithful recaptures single cell transcriptomics measurements which enables researchers to revisit historical disease studies of Alzheimer's disease, for example, to characterize cell type specific associations which were masked by bulk RNA measurements.

4.2 Future directions

The work in this thesis on scAlign and scProjection leaves opportunities for advancement of these methods to adapt the underlying network models for newly emerging sequencing technologies and analysis challenges in computational biology.

scAlign focused on the alignment of single cell genomics datasets which contain unwanted technical variation in gene expression measurements of individual cells. However, the similarity between datasets in terms of cell type composition, species divergence or expression shift could be quantified and

incorporated into the network model to guide the merging of more than two datasets in order to iterative build a latent representation without technical confounding. One could modify the all-pairs alignment procedure to be a guided alignment to ensure the predominate biological signal across all datasets is preserved while limiting the wrapping of unique biological signals. Additionally, due to the general design of scAlign in terms of data input it would be trivial to extend this model for the alignment of additional molecular measurements such as ATAC-seq data which exhibits similar technical variation albeit in a significantly higher dimensional space, the complete genome. One could modify the similarity or kernel functions used by scAlign to those of natural language processing which are common in ATAC-seq analysis to ensure the biological signal is retained while removing technical variation in the epigenetic signal of chromatin accessibility.

The development of scProjection enabled the highly accurate prediction of cell type abundances and transcriptomic projections from mixed sample. Increasing the number of cell types which scProjection needs to model increases the cross-correlation between types and can confound abundances estimates. The framework of scProjection could be expanded to perform iterative deconvolution where a single cell atlas defines varying resolution of cell type annotation. scProjection could be modified to perform deconvolution at the coarsest cell type annotation and then use the estimated abundances as a prior for the estimate of cell type abundances for higher resolution types as you move down a cell type hierarchy. In doing so the estimates would be less noisy due to separating cells first by the most distance cell types and then fine tuning the abundances for more related types with less clear marker gene expression patterns. The performance of such a modified scProjection would not only improve the abundance estimation but also the projection of mixed samples to individual expression measurements through a more exact mapping onto cell states defined in an single cell atlas.

A recent open challenge in single cell genomics is the analysis data of acquired through multi-omics technologies which measure for an individual cell multiple molecular or functional readouts. For example, 10x Multiome enables the simultaneous profiling of both gene expression and chromatin accessibility for each individual cell sequenced. The challenge of integrating the information across these

modalities requires a model to find the relationships between the epigenetic signature of individuals cell types and the role of such DNA structure on the downstream gene expression product. Intuitively, a neural network that can learn a joint embedding space that identifies the comparable features across modalities would be able to find a latent representation that captures the global language that define a cell across many genomics readouts. One could envision a similar model to scAlign, where the goal is to ensure the retention of biological variability within each data modality while aligning the modalities together into a joint embedding space. The development of such multi-modal neural network models will have the opportunity to enable new analyses and hypothesis generation to fuel research directions not yet explored. Machine learning models that exploit the measurement of multiple modalities from individual cells will enable deeper molecular and functional characterization of cell types that can enable new break throughs in precision and target medicine for rare or complex diseases.

Bibliography

1. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *bioRxiv* 065912 (2016) doi:10.1101/065912.
2. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
3. Rohart, F., Eslami, A., Matigian, N., Bougeard, S. & Lê Cao, K.-A. MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinformatics* **18**, 128 (2017).
4. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
5. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
6. Lin, Y. *et al.* scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proc. Natl. Acad. Sci.* 201820006 (2019) doi:10.1073/pnas.1820006116.
7. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).
8. Argelaguet, R. *et al.* Multi-Omics factor analysis - a framework for unsupervised integration of multi-omic data sets. *bioRxiv* 217554 (2018) doi:10.1101/217554.

9. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).

10. Hie, B. L., Bryson, B. & Berger, B. Panoramic stitching of heterogeneous single-cell transcriptomic data. *bioRxiv* (2018) doi:10.1101/371179.

11. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257 (2019).

12. Snyder, M. P. *et al.* The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).

13. Johansen, N. & Quon, G. scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data. *Genome Biol.* **20**, 166 (2019).

14. Haeusser, P., Frerix, T., Mordvintsev, A. & Cremers, D. Associative Domain Adaptation. in *IEEE International Conference on Computer Vision (ICCV)* (2017).

15. Ganin, Y. *et al.* Domain-Adversarial Training of Neural Networks. *ArXiv150507818 Cs Stat* (2015).

16. Lander, E. S. Array of hope. *Nat. Genet.* **21**, 3–4 (1999).

17. Fare, T. L. *et al.* Effects of atmospheric ozone on microarray data quality. *Anal. Chem.* **75**, 4672–4675 (2003).

18. Rhodes, D. R. *et al.* Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9309–9314 (2004).

19. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

20. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

21. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).

22. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).

23. Cannoodt, R., Saelens, W. & Saeys, Y. Computational methods for trajectory inference from single-cell transcriptomics. *Eur. J. Immunol.* **46**, 2496–2506 (2016).

24. Bakken, T. E. *et al.* Evolution of cellular diversity in primary motor cortex of human, marmoset monkey, and mouse. *bioRxiv* 2020.03.31.016972 (2020) doi:10.1101/2020.03.31.016972.

25. Hodge, R. D. *et al.* Conserved cell types with divergent features between human and mouse cortex. *bioRxiv* 384826 (2018) doi:10.1101/384826.

26. The Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).

27. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).

28. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).

29. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostat. Oxf. Engl.* **8**, 118–127 (2007).

30. Kiselev, V. Y. *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017).

31. Ji, Z. & Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* **44**, e117–e117 (2016).

32. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

33. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).

34. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).

35. Lawlor, N. *et al.* Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* **27**, 208–222 (2017).

36. Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst.* **3**, 385-394.e3 (2016).

37. Segerstolpe, Å. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.* **24**, 593–607 (2016).

38. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43 (2019).

39. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452.e17 (2017).

40. Datlinger, P. *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).

41. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e17 (2016).

42. Jaitin, D. A. *et al.* Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* **167**, 1883-1896.e15 (2016).

43. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. ZINB-WaVE: A general and flexible method for signal extraction from single-cell RNA-seq data. *bioRxiv* 125112 (2017) doi:10.1101/125112.

44. Hon, C.-C., Shin, J. W., Carninci, P. & Stubbington, M. J. T. The Human Cell Atlas: Technical approaches and challenges. *Brief. Funct. Genomics* **17**, 283–294 (2018).

45. Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).

46. Vento-Tormo, R. *et al.* Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**, 347–353 (2018).

47. Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, (2018).

48. Hodge, R. D. *et al.* Conserved cell types with divergent features between human and mouse cortex. *bioRxiv* 384826 (2018) doi:10.1101/384826.

49. Plasschaert, L. W. *et al.* A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).

50. Stuart, T. *et al.* Comprehensive integration of single cell data. *bioRxiv* 460147 (2018) doi:10.1101/460147.

51. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).

52. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).

53. Haeusser, P., Mordvintsev, A. & Cremers, D. Learning by Association - A versatile semi-supervised training method for neural networks. in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

54. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

55. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (eds. Teh, Y. W. & Titterington, M.) vol. 9 249–256 (PMLR, 2010).

56. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs* (2014).

57. Tian, L. *et al.* scRNA-seq mixology: towards better benchmarking of single cell RNA-seq protocols and analysis methods. *bioRxiv* 433102 (2018) doi:10.1101/433102.

58. Kowalczyk, M. S. *et al.* Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.* **25**, 1860–1872 (2015).

59. Mann, M. *et al.* Heterogeneous Responses of Hematopoietic Stem Cells to Inflammatory Stimuli are Altered with Age. *bioRxiv* 163402 (2017) doi:10.1101/163402.

60. Ji, Z. & Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* gkw430 (2016) doi:10.1093/nar/gkw430.

61. Setty, M. *et al.* Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **34**, 637–645 (2016).

62. Welch, J. D., Hartemink, A. J. & Prins, J. F. SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biol.* **17**, 106 (2016).

63. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).

64. Poran, A. *et al.* Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites. *Nature* **advance online publication**, (2017).

65. Josling, G. A. *et al. Regulation of sexual differentiation is linked to invasion in malaria parasites*. http://biorxiv.org/lookup/doi/10.1101/533877 (2019) doi:10.1101/533877.

66. Bancells, C. *et al.* Revisiting the initial steps of sexual development in the malaria parasite Plasmodium falciparum. *Nat. Microbiol.* **4**, 144–154 (2019).

67. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).

68. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).

69. Yao, Z. *et al.* A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell* **184**, 3222-3241.e26 (2021).

70. Travaglini, K. J. *et al.* A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).

71. Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).

72. Espina, V. *et al.* Laser-capture microdissection. *Nat. Protoc.* **1**, 586–603 (2006).

73. Cadwell, C. R. *et al.* Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nat. Biotechnol.* **34**, 199–203 (2016).

74. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).

75. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* eaau0730 (2018) doi:10.1126/science.aau0730.

76. Tripathy, S. J. *et al.* Assessing Transcriptome Quality in Patch-Seq Datasets. *Front. Mol. Neurosci.* **11**, 363 (2018).

77. Xia, C., Fan, J., Emanuel, G., Hao, J. & Zhuang, X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci.* **116**, 19490–19499 (2019).

78. Spatially resolved, highly multiplexed RNA profiling in single cells | Science. https://science.sciencemag.org/content/348/6233/aaa6090.

79. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *ArXiv13126114 Cs Stat* (2014).

80. Mohammadi, S., Davila-Velderrain, J. & Kellis, M. A multiresolution framework to characterize single-cell state landscapes. *Nat. Commun.* **11**, 5399 (2020).

81. Tian, L. *et al.* Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods* **16**, 479–487 (2019).

82. Patrick, E. *et al.* Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. *PLOS Comput. Biol.* **16**, e1008120 (2020).

83. Gouwens, N. W. *et al.* Integrated Morphoelectric and Transcriptomic Classification of Cortical GABAergic Cells. *Cell* **183**, 935-953.e19 (2020).

84. Moor, A. E. *et al.* Spatial Reconstruction of Single Enterocytes Uncovers Broad Zonation along the Intestinal Villus Axis. *Cell* **175**, 1156-1167.e15 (2018).

85. Manco, R. *et al.* Clump sequencing exposes the spatial expression programs of intestinal secretory cells. *Nat. Commun.* **12**, 3074 (2021).

86. Pelaseyed, T. *et al.* The mucus and mucins of the goblet cells and enterocytes provide the first defense line of the gastrointestinal tract and interact with the immune system. *Immunol. Rev.* **260**, 8–20 (2014).

87. A sentinel goblet cell guards the colonic crypt by triggering Nlrp6-dependent Muc2 secretion - PubMed. https://pubmed.ncbi.nlm.nih.gov/27339979/.

88. Gerbe, F., Legraverend, C. & Jay, P. The intestinal epithelium tuft cells: specification and function. *Cell. Mol. Life Sci. CMLS* **69**, 2907–2917 (2012).

89. The Intestinal Epithelium: Central Coordinator of Mucosal Immunity - PubMed. https://pubmed.ncbi.nlm.nih.gov/29716793/.

90. Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* **551**, nature24489 (2017).

91. Park, S.-W. *et al.* The protein disulfide isomerase AGR2 is essential for production of intestinal mucus. *Proc. Natl. Acad. Sci.* **106**, 6950–6955 (2009).

92. Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2021).

93. Yang, Y., Zhao, H., Wang, J. & Zhou, Y. SPOT-Seq-RNA: predicting protein-RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction. *Methods Mol. Biol. Clifton NJ* **1137**, 119–130 (2014).

94. Zhang, M. *et al. Molecular, spatial and projection diversity of neurons in primary motor cortex revealed by in situ single-cell transcriptomics*. 2020.06.04.105700 https://www.biorxiv.org/content/10.1101/2020.06.04.105700v1 (2020) doi:10.1101/2020.06.04.105700.

95. Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, (2018).

96. Perkel, J. M. Starfish enterprise: finding RNA patterns in single cells. *Nature* **572**, 549–551 (2019).

97. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 1–14 (2018).

98. Zhang, M. J., Ntranos, V. & Tse, D. Determining sequencing depth in a single-cell RNA-seq experiment. *Nat. Commun.* **11**, 774 (2020).

99. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).

100. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).

101. Hunt, G. J., Freytag, S., Bahlo, M. & Gagnon-Bartsch, J. A. dtangle: accurate and robust cell type deconvolution. *Bioinformatics* **35**, 2093–2099 (2019).

102. Zhong, Y., Wan, Y.-W., Pang, K., Chow, L. M. & Liu, Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics* **14**, 89 (2013).

103. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* **11**, 5650 (2020).

104. Biancalani, T. *et al.* Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat. Methods* **18**, 1352–1362 (2021).

105. Zhang, M. *et al.* Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature* **598**, 137–143 (2021).

106. Fawkner-Corbett, D. *et al.* Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell* **184**, 810-826.e23 (2021).

107. Burgess, C. P. *et al.* Understanding disentangling in β-VAE. *ArXiv180403599 Cs Stat* (2018).

108. Berg, J. *et al.* Human cortical expansion involves diversification and specialization of supragranular intratelencephalic-projecting neurons. *bioRxiv* 2020.03.31.018820 (2020) doi:10.1101/2020.03.31.018820.

109. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7285–7290 (2015).