scAlign: a tool for alignment, integration and rare cell identification from scRNA-seq data

Nelson Johansen[1,2,*], Gerald Quon[1,2,3,*]
[1]Graduate Group in Computer Science, [2]Genome Center, [3]Department of Molecular and Cellular Biology, University of California, Davis, Davis, CA

*To whom correspondence should be addressed: njjohansen@ucdavis.edu, gquon@ucdavis.edu

Abstract

Single cell RNA sequencing (scRNA-seq) experiments are now routinely being conducted on similar sets of cell types under different conditions and species, in order to characterize how gene regulation and cell identity respond to different stimuli and biological factors. The goal of scRNA-seq alignment is to compare transcriptomes collected under multiple conditions to identify cell state and type-specific response to stimuli.

Here we present scAlign, a deep learning method for aligning and integrating scRNA-seq data collected across multiple conditions into a common, low-dimensional cell state space for downstream analysis such as clustering and trajectory inference across conditions. Compared to previous approaches, scAlign uses a unified unsupervised and supervised alignment framework that can incorporate partial, overlapping or a full set of cell type labels if available to guide alignment, and also creates a bidirectional map between gene expression space of each condition and the joint cell state space. This bidirectional map allows us to project single cell transcriptomes sequenced in one condition into another condition to precisely estimate the effect of conditions on individual cells without having to first identify clusters of cells, and therefore can identify rare novel subpopulations across conditions that are typically hard to identify via clustering.

We show that scAlign achieves substantial improvements in unsupervised alignment as compared to other methods on multiple benchmark scRNA-seq data sets, and performs particularly well when there is a large degree of cell type-specific responses to conditions. We furthermore show scAlign is most robust when there is large variation in cell type composition across conditions. Finally, we demonstrate the power of our bidirectional maps by identifying a rare subpopulation of cells expressing a gene regulatory program likely to drive sexual commitment in malaria. The scAlign framework is not specific to scRNA-seq and should be applicable to alignment and data integration problems in other domains.

Introduction

Single cell RNA sequencing (scRNA-seq) technologies enable the capture of high resolution snapshots of gene expression activity in individual cells. A universal observation of scRNA-seq data is that the measured transcriptomes of even homogeneous populations of cells occupy a continuum of cell states; no two cells yield the exact same transcriptome, a result of both technical[1,2] and biological[3–5] factors.

This presents a challenge for the characterization of transcriptional changes due to stimulus. Case-control experiments in which cells are measured under control and stimulated conditions are essential for interrogating the regulatory mechanisms underlying transcriptional responses. Such experiments measuring whole-transcriptome profiling as a phenotype can now be performed at high throughput both in bulk measurements[6] as well as at the single cell level[7–9]. Unfortunately, scRNA-seq protocols require lysing the cell to extract and sequence their RNA, therefore preventing the measurement of the same cell before and after a stimulus is applied.

Current approaches to characterizing response to stimulus entail measuring biological replicates of cell populations before and after stimulus. The resulting transcriptomes then typically separate by both condition and cell type (Fig. 1b), which then makes cell type specific comparisons across conditions difficult if the cells are not labeled by experimental means. To address this problem, the scRNA-seq data across conditions is typically aligned (or "integrated")[10–16], to identify groups of cells from one condition that correspond to the same cell type as cells from another condition.

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 conditions, and a joint cell state space in which cells group by function and type, regardless of the condition in which it was sequenced. This bidirectional map enables users to generate an representation of what the same cell state looks like under each individual condition, and therefore simulate a matched experiment in which the exact same cell is sequenced simultaneously under different conditions. We demonstrate that scAlign outperforms existing alignment methods particularly when individual cell types exhibit heterogeneous responses to stimulus, and that our bidirectional map enables identification of changes in rare cell types that cannot be identified from the alignment alone. We further demonstrate the utility of scAlign in identifying changes in expression associated with sexual commitment in malaria, and posit that scAlign may be used to perform alignment in domains other than single cell genomics as well.

Results

The overall framework of scAlign is illustrated in **Figure 1**. The first component of scAlign is the construction of a joint cell state space using scRNA-seq data from all conditions (**Fig. 1a**). This joint cell state 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 across all scRNA-seq data conditions. Unlike autoencoders, which share a similar architecture but use a different objective function, our low dimensional manifold is learned by training the neural network to simultaneously encourages overlap of cells in the state space from across conditions (thus performing alignment), yet also preserving the cell-cell similarity matrices within each condition (and therefore minimizing distortion of each condition's data) (see Methods for details). Optionally, scAlign can take as input a partial or full sets of cell annotations in one or more conditions, which will encourage the alignment to match cells of the same type together.

In the second component of scAlign, we train condition specific deep decoder networks to project individual cells from the joint embedding space back to the gene expression space of each condition, by minimizing the reconstruction error for the cells originally sequenced within each condition (**Fig. 1b**). Constructing a decoder for each condition allows scAlign to project a single cell from the joint embedding space back into multiple conditions simultaneously to measure variation of expression across conditions, regardless of which condition that single cell was originally measured in. This decoding process intuitively interpolates the expression profile of cells from other conditions to represent what each cell would look like if it had been sequenced under the original condition, by using scRNA-seq data from the original condition.

Results - scAlign captures cell type specific response to stimulus

We first benchmarked the alignment component of scAlign using four publicly available scRNA-seq datasets for which replicate cell populations were subject to two or more conditions, and for which the cell type labels were obtained experimentally (**Fig. 2**). Our first comparison used

CellBench[17], a dataset consisting of three human lung adenocarcinoma cell lines (HCC827, H1975, H2228) that were sequenced under different platforms (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. Our initial experiments suggested that the Cell Bench scRNA-seq data collected on individual cell types sequenced across platforms were easily aligned by most methods, and is consistent with observations from the original Cell Bench study[17] and so this data was excluded. Alignment of transcriptomes of RNA mixtures across platforms was more challenging and more clearly illustrated the performance advantage of scAlign (**Fig. 2a, Supplementary Fig. 1**). In both the Kowalczyk et al. and Mann et al. datasets, three hematopoietic cell types were sequenced in young and old mice (LT-HSC, ST-HSC, MPP), though Mann et al. additionally challenge the mice with LPS or a control condition. Similar to our results with CellBench, scAlign outperforms other approaches on both of these datasets.

While scAlign outperformed other tested methods on those three benchmark datasets, we found the remaining approaches were inconsistent with respect to their relative performance. We found that for these three datasets, there was considerable overlap in the differentially expressed genes for each cell type across condition (**Supplementary Fig. 2**), suggesting these datasets may be less challenging to align. We therefore compiled a novel benchmark dataset by combining published scRNA-seq data on hematopoietic cells measured across different studies and stimuli, termed Johansen et al. This dataset yields more heterogeneous response to stimulus across the cell types (increased cell type specific differential gene expression). On the Johansen et al. dataset, we find that scAlign's performance is robustly superior, while Seurat also outperforms the remaining methods by a large margin.

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[18] and trajectory inference[19–21]. Dimensionality reduction of cell types generally improves when more data is used to identify the embedding dimensions, and so we hypothesized that known cell types will cluster better in scAlign's embedding space in part due to the fact we are defining a single embedding space using shared data from multiple conditions. We therefore compared the clustering of known cell types in scAlign embedding space to an autoencoder neural network trained using the same architecture as scAlign, but on each condition separately (see Methods). In two of the three datasets we tested, we found that known cell types cluster more closely and are more distinct in scAlign embedding space compared to that of the autoencoder (**Fig. 3, Supplementary Fig. 3**).

All of the existing approaches tested in this paper are unsupervised, in that they do not use cell type labels to aid alignment, if available. While the focus of this paper is on scAlign performance in this unsupervised setting, a unique feature of scAlign is that its unified alignment framework can naturally incorporate cell type labels as a separate component in the objective function that is maximized during training. Thus, even if only broad categories of cell type information are available but more fine resolution is desired, or if only a subset of cells can be reliably labeled, we can use that information to guide the alignment with prior knowledge. **Supplementary Figure 4a** illustrates, for each of the four datasets from Figure 2, that alignment performance generally improves when cell type labels are available at training time. Furthermore, this performance gain does not come at the cost of distorting the cell-cell similarity matrices within each condition (**Supplementary Fig. 4b**), despite the supervised and unsupervised components of the scAlign objective function having different goals.

Results - scAlign is robust to large changes in cell type composition

There are two types of changes that we generally expect to be observed when a collection of cells is subject to stimulation: either changes in gene regulation within individual cell types, changes in the relative proportion of each cell type within the entire collection of cells, or both. We therefore explored the behavior of scAlign and other approaches when the relative proportion of cell types varies significantly between the two conditions being aligned.

We performed a series of experiments on the Kowalczyk et al. dataset where we measured alignment performance of all methods as we remove an increasing number of cells from a given cell type from the old mouse condition (**Fig. 5**). 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 a given cell only led to a median drop of 11% in accuracy across the tested methods. We found most methods had a more difficult time when we performed the same experiments using the data from Mann et al., where most methods saw a general decrease in performance as we removed more cells (**Supplementary Fig. 5**).

We next investigated the factors that underlie scAlign's robustness to varying cell type composition imbalances across datasets. 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 cell state space when performing random walks of length two. The random walk starts with a cell in one condition, then moves to a cell from the other condition based on proximity in cell state space. (Fig. 4a) The walk then returns back to a cell (excluding the starting cell) in the original condition, also based on proximity in cell state space. (Fig. 4b) For every cell in each condition, we calculated the frequency that such random walks (from the other condition) pass through any given cell (**Fig. 4**). 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 dataset, another cell is automatically selected as a replacement. (**Supplementary Fig. 6**). This suggests that these round-trip probabilities may only require a small representative set of cells for the roundtrip random walks originating from each of the conditions.

In the above experiments, we have aligned conditions in which the same set of cell types are present in both conditions. Here we explore the behavior of scAlign and other approaches when there is not a 1-1 correspondence in cell types between conditions. We expect such scenarios to arise when e.g. only a subset of cells of a given type might respond to a stimulus, or when specific cell types might appear or disappear completely due to stimulus (e.g. in the case of targeted therapy, or when speciation is treated as a stimulus[22]). For each of our benchmark datasets, we removed a subset of populations from the condition, and aligned the control and stimulated conditions to determine the extent to which the unmatched population was not aligned to cells from the other condition. **Figure 6a** demonstrates that in eight out of nine cases, scAlign outperforms other alignment methods in terms of classification accuracy. Upon further inspection, even in cases where alignment accuracy was similar between methods, scAlign visually separates cell types in its embedding space much more than approaches such as scran and Seurat (**Fig. 6b**). In particular, for other approaches, separation of different cell types within the same condition gets worse when one cell type is removed (**Supplementary Fig. 7**).

Results - scAlign interpolates gene expression accurately

One of the more novel contributions of scAlign is the ability to map each cell from the joint embedding space back into the gene expression space of each of the original conditions,

regardless of which condition the cell was originally sequenced in. This mapping is performed through interpolation: for each condition, we learn a mapping from the embedding space to expression space using cells sequenced under that condition, then apply the map to all cells sequenced under all other conditions. This interpolation procedure enables measurement of variation in gene expression for the same cell state across multiple conditions.

Figure 7a illustrates how scAlign's interpolation accurately recapitulates the condition specific expression patterns when mapping cells into each condition. For each of the three hematopoietic datasets, we trained decoder neural networks to map cells from the joint space back into each of the case and control conditions, using cells sequenced under those conditions respectively. We then mapped cells originally sequenced in the stimulated condition and interpolated their expression in the control condition, and mapped cells originally sequenced in the control condition and interpolated their expression in the stimulated condition.

We then measure 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 control cells that have been interpolated into 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. 7a**), suggesting our interpolation procedure can accurately map a cell from one condition into any other.

Figure 7b 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 which condition they originated from. This demonstrates that the encoding and decoding process maintains data fidelity, despite the fact that the encoder is trained to align data from multiple conditions and is not explicitly trained to minimize reconstruction error like typical autoencoders. **Figures 7c,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.

Results – interpolation identifies rare candidate population involved in sexual commitment

We next applied scAlign to identify genes associated with sexual commitment in malaria. Briefly, the life cycle of *Plasmodium falciparum* begins with a single asexually-committed cell and ends 48 hours later with the single cell giving rise to a new set of cells that are all either committed to asexual replication or differentiation into gametocytes. While the gene *ap2-g* is a known master regulator of sexual commitment, and an increase in expression of ap2-g is necessary for sexual commitment, the events which follow *ap2-g* activation and lead to full sexual commitment are currently unknown. Furthermore, *ap2-g* expression is restricted to a minor subset of cells, making the identification of the precise stage of the cycle when sexual commitment is occurring a challenging task.

Figure 8a illustrates the joint embedding space of malaria cells which are either capable of ap2-g expression (AP2-G CTRL) or are ap2-g deficient (AP2-G OFF). . As was observed in the original paper, cells generally form a trajectory in the embedding space, corresponding to different time points in the malaria cell cycle. scAlign is able to maintain the gametocytes from the AP2-G CTRL condition as a distinct population that is not aligned to any cells from the AP2-G OFF condition, whereas other tested methods are unable to accurately isolate the gametocyte population (**Supplementary Fig. 8**).

Using the interpolation component of scAlign, we projected each cell in either condition in the joint embedding space to the expression space of the AP2-G CTRL and AP2-G OFF conditions. Taking the difference in interpolated expression we define a paired difference heatmap (**Fig. 8b**). From the paired difference heatmap, we observed few overall differences in gene expression between the two conditions, except for a rare subpopulation of cells that lie at the interface between late-stage cells and the gametocytes that are specific to the control malaria cells in the joint embedding space (**Fig. 8c**). We found that this rare subpopulation that differs between the two conditions is reproducible even when using scRNA-seq data from a different control cell line of malaria, NF54 (**Supplementary Fig. 9**). Furthermore, clustering scRNA-seq data in either condition does not identify this rare subpopulation as a candidate cluster of cells, highlighting the utility of scAlign to identify rare subpopulations that differ between conditions.

We identified the set of genes that are predicted to differ the most between cells of the two conditions in the rare subpopulation (**Fig. 8d**). Notable are several genes with previously established roles in gametocyte maturation, including P48/45[23] and Pf11-1[24,25]. We also observed an enrichment of gender-biased transcripts in these genes upregulated in AP2-G CTRL, further suggesting some of these genes may play a role in sexual commitment.

Discussion

One of the principal advantages of scAlign over existing alignment methods is that we have shown our method can identify rare cell types that differ between conditions without the need to cluster cells. For typical alignment methods, once the effect of condition is removed, cells must still be clustered into putative cell types in order to identify which cells match across condition and therefore perform an unpaired comparison of transcriptomes within each cluster between cells from different conditions. The need to cluster cells means detection of rare cell types can be highly sensitive to choice of clustering algorithm or parameters. In contrast, through interpolation, scAlign can generate paired differential expression heatmaps from which it can be evident whether there are rare cell populations emerging in one condition (**Fig. 8**). In both our constructed datasets as well as in the malaria cell lines, we showed that scAlign can match cells of the same type even when those types are present in small proportions, unlike existing approaches.

Another unique feature of scAlign is the unified alignment framework which can optionally use labels, and furthermore can use labels only available for subsets of cells, and can naturally handle cells that might have multiple labels simultaneously. We postulate these special cases may arise frequently, as cell type markers are typically only available for well established cell types such as hematopoietic cells, markers may not be unique to individual cell types and technical factors such as dropout may cause known markers that are expressed, to not appear expressed in the transcriptome. Additionally, labels can be used only for individual conditions or all conditions simultaneously, to force specific cell types to align cross conditions, through the flexible classifier framework implemented for scAlign.

Here we have mainly compared scAlign against unsupervised alignment methods. In our results, we compared favorably against MINT[19], another supervised method, when assuming all cells are labeled. In the context of alignment, however, we reasoned that if labels are available for all cells in each condition, then addressing the task of alignment is less useful, because cells of the same type across condition can be directly compared via differential expression analysis on a per-cell type level. Alternatively, if all cells are labeled but the user suspects there are subtypes of cells, each matching pair of cell types across conditions can be aligned using the unsupervised methods individually to identify subpopulations of cells.

Transcriptional alignment and batch correction of RNA-seq data share similarities in both goals and underlying methodologies, as one can think about the biological condition a cell is sequenced in as a batch whose effect should be removed in order to identify cells in one condition that represent the same cell state as cells from another condition. From a practical perspective, alignment can be viewed as a flexible batch correction method, where different corrections are applied to different cell states because the effect of the condition might be cell state-specific. A key difference, however, is that in addition to making data comparable across conditions, alignment is often performed to also characterize the differences between the same cell state across condition. scAlign uses the flexibility of neural networks to perform alignment where cell states might exhibit heterogeneous responses to stimuli, yet through interpolation provides the interpretability that linear correction methods enjoy. 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.

One of the challenges of using neural network-based approaches like scAlign is that the architecture size, defined here as the number of layers and number of nodes per layer, need to be selected before training. 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 cell state spaces, which also acts as a form of regularization. In our experiments, typical scRNA-seq datasets will require at most three layers to align multiple conditions successfully. The results of this study were robust to network depth and width (**Supplementary Figure. 10**) along with choice of hyper parameters.

One of the advantages of the general design of the scAlign neural network architecture and loss function is that by being agnostic to its RNA-seq data inputs, the input data can either be gene-level counts or even transformations of those gene level counts, such as principal component scores. 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. One of the practical disadvantages of a neural network architecture is that users who do not have access to GPU hardware do not benefit from the hardware-based acceleration of neural network training. By first reducing scRNA-seq data to principal component scores, the training of scAlign converges much quicker and is feasible on a CPU-based system, therefore making scAlign both broadly applicable and based on deep learning.

In this paper, we primarily have evaluated the accuracy of scAlign in the context of comparing scRNA-seq data across two conditions, in large part because a core component of our neural network loss function is to minimize the differences between scRNA-seq distributions of two conditions in the joint embedding space. While comparison across two conditions is currently the most prevalent study design, we envision comparison across multiple conditions becoming more frequent in the near future, as is currently done in RNA-seq when looking across different disease types[26] or looking at the same tissue across multiple organisms[27], for example. Our neural network loss function naturally scales to more than two comparisons, and can be achieved in two ways. Either a single reference condition may be established and all other conditions can be aligned against the reference, or we can maximize the overlap in embedding space between all pairs of conditions.

Acknowledgements

We thank Bjorn Kafsack for helpful discussions regarding the malaria cell line data. This project has been made possible in part by grant number 2018-182633 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation. The Titan Xp used for this research was donated by the NVIDIA Corporation.

Code and data availability

The code implementing scAlign, as well as all datasets used in this study are available at https://github.com/ucdavis/quontitativebiology.

References

1. Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostat. Oxf. Engl.* **19**, 562–578 (2018).
2. 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).
3. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363–369 (2014).
4. Eldar, A. & Elowitz, M. B. Functional roles for noise in genetic circuits. *Nature* **467**, 167–173 (2010).
5. Maamar, H., Raj, A. & Dubnau, D. Noise in gene expression determines cell fate in Bacillus subtilis. *Science* **317**, 526–529 (2007).
6. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452.e17 (2017).
7. Datlinger, P. *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
8. 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).
9. Jaitin, D. A. *et al.* Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* **167**, 1883-1896.e15 (2016).
10. 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).
11. 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).

12. 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).

13. Lin, Y. *et al.* scMerge: Integration of multiple single-cell transcriptomics datasets leveraging stable expression and pseudo-replication. *bioRxiv* 393280 (2018). doi:10.1101/393280

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

15. 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

16. 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).

17. 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

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

19. 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

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

21. 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).

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

23. van Dijk, M. R. *et al.* A central role for P48/45 in malaria parasite male gamete fertility. *Cell* **104**, 153–164 (2001).

24. Scherf, A. *et al.* Gene inactivation of Pf11-1 of Plasmodium falciparum by chromosome breakage and healing: identification of a gametocyte-specific protein with a potential role in gametogenesis. *EMBO J.* **11**, 2293–2301 (1992).

25. Ikadai, H. *et al.* Transposon mutagenesis identifies genes essential for Plasmodium falciparum gametocytogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E1676-1684 (2013).

26. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).

27. Boroviak, T. *et al.* Single cell transcriptome analysis of human, marmoset and mouse embryos reveals common and divergent features of preimplantation development. *Dev. Camb. Engl.* **145**, (2018).

28. 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).

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

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

31. 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.) **9**, 249–256 (PMLR, 2010).

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

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

34. 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).

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

Figures

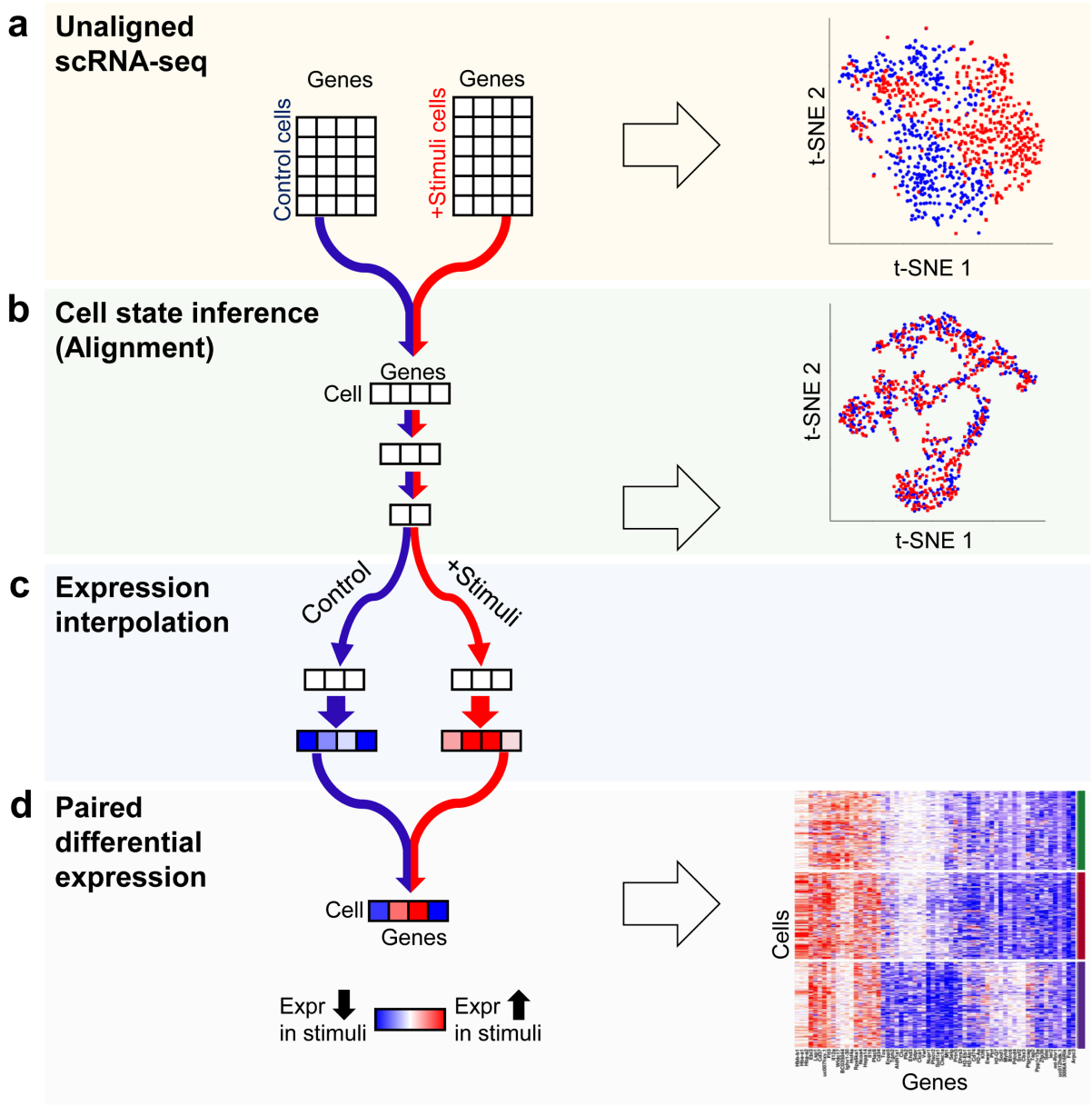

Figure 1. Schematic of unsupervised alignment and paired differential expression within the scAlign framework. (**1**) Unaligned scRNA-seq datasets from multiple conditions are used as input to scAlign. Expression can be represented as either gene-level expression, or reduced dimensions such as from PCA or CCA. (**2**) Deep encoding network learns a low-dimensional joint cell state space that simultaneously aligns cells from all conditions. (**3**) Paired decoders project cells from the joint embedding space back into the original expression space of each condition, and can be used to interpolate the expression profile of cells sequenced from any condition into any other condition. (**4**) For a single cell sequenced under any condition, we can calculate its interpolated expression profile in both conditions, then take the difference to calculate a paired differential expression for the same cell state under different conditions to identify cell state-specific changes due to stimulus.

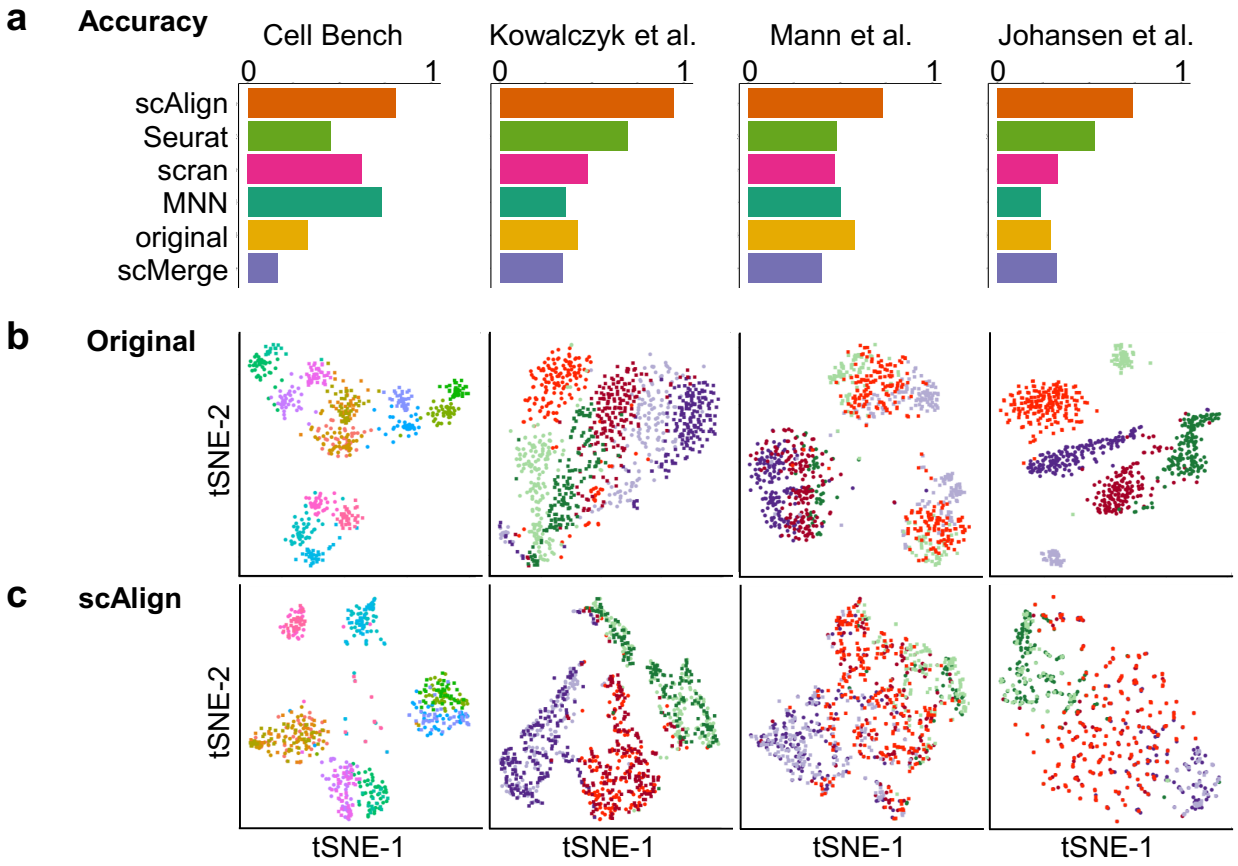

Figure 2. scAlign outperforms existing alignment approaches on benchmark datasets. (**a**) Bar plots indicate the accuracy of a classifier trained on labeled cells from one condition and used to predict cell labels from another condition. Datasets, from left to right: (1) Cell Bench, a dataset of mixtures of RNA from three cancer cell lines sequenced under multiple platforms; (2) hematopoietic cells sequenced from young and old mice; (3) hematopoietic cells sequenced from young and old mice, challenged with LPS; and (4) a composite dataset consisting of hematopoietic cells responding to different stimuli. (**b**) tSNE plots generated using the original expression profiles from each of the four datasets; each point represents a cell, and cells are colored based on their experimentally-derived cell type labels. (**c**) tSNE plots for each dataset after alignment with scAlign.

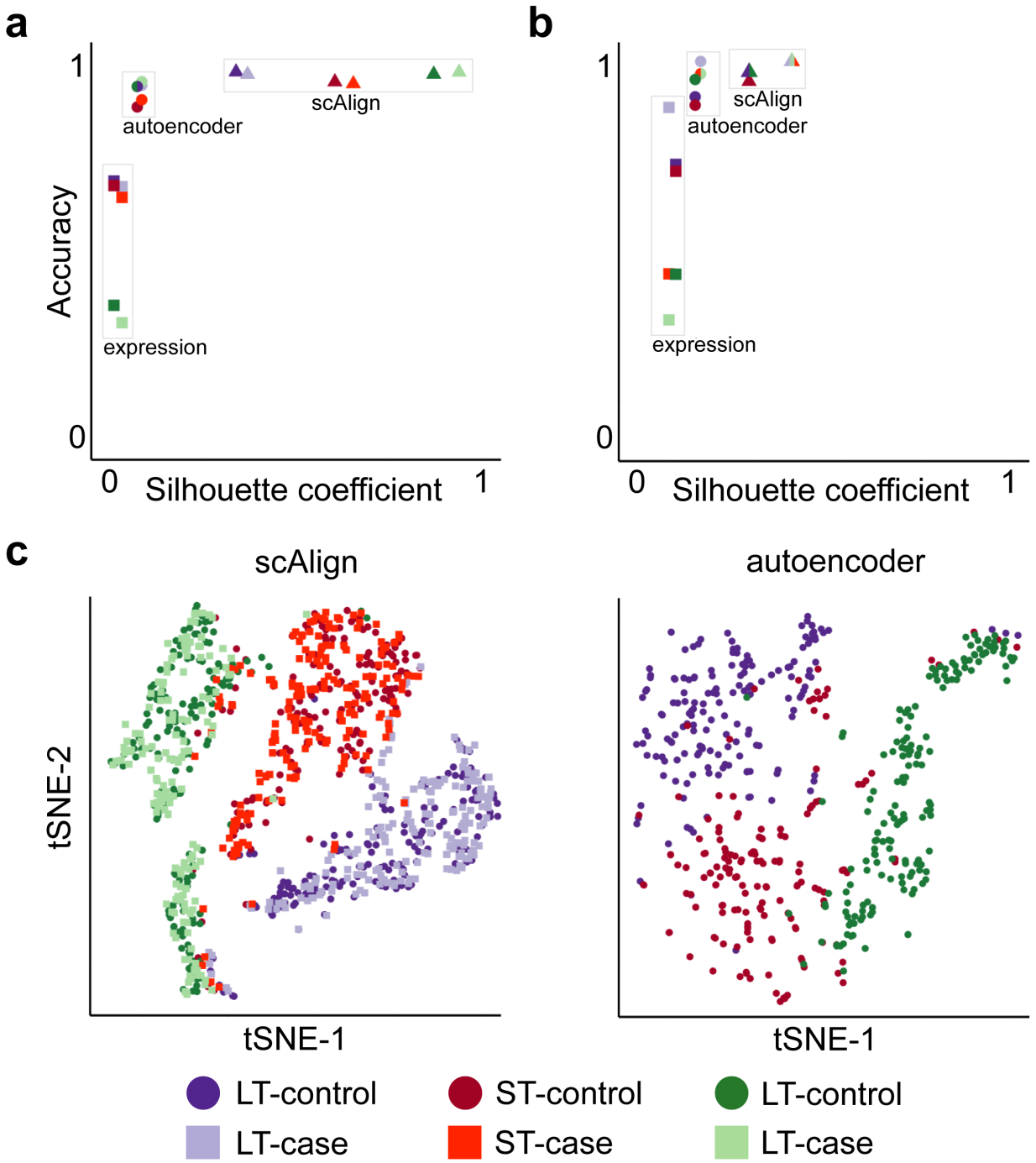

Figure 3. Joint analysis of cells from all conditions leads to more accurate clustering of cell types compared to analysis of individual conditions. (**a**) Scatterplot illustrating the quality of clustering of individual cell types from individual conditions from the Mann et al. dataset. Each point represents one cell type in one condition, when the embedding is computed using either the original expression data (expression), or the embedding dimensions of scAlign or an autoencoder with the same neural network architecture as scAlign. Y-axis represents classification accuracy while the x-axis represents the silhouette coefficient. (**b**) Same as (a), but for Johansen et al. (**c**) tSNE plots visualizing the embedding space of scAlign trained on both conditions and (**d**) an autoencoder trained on a single condition.

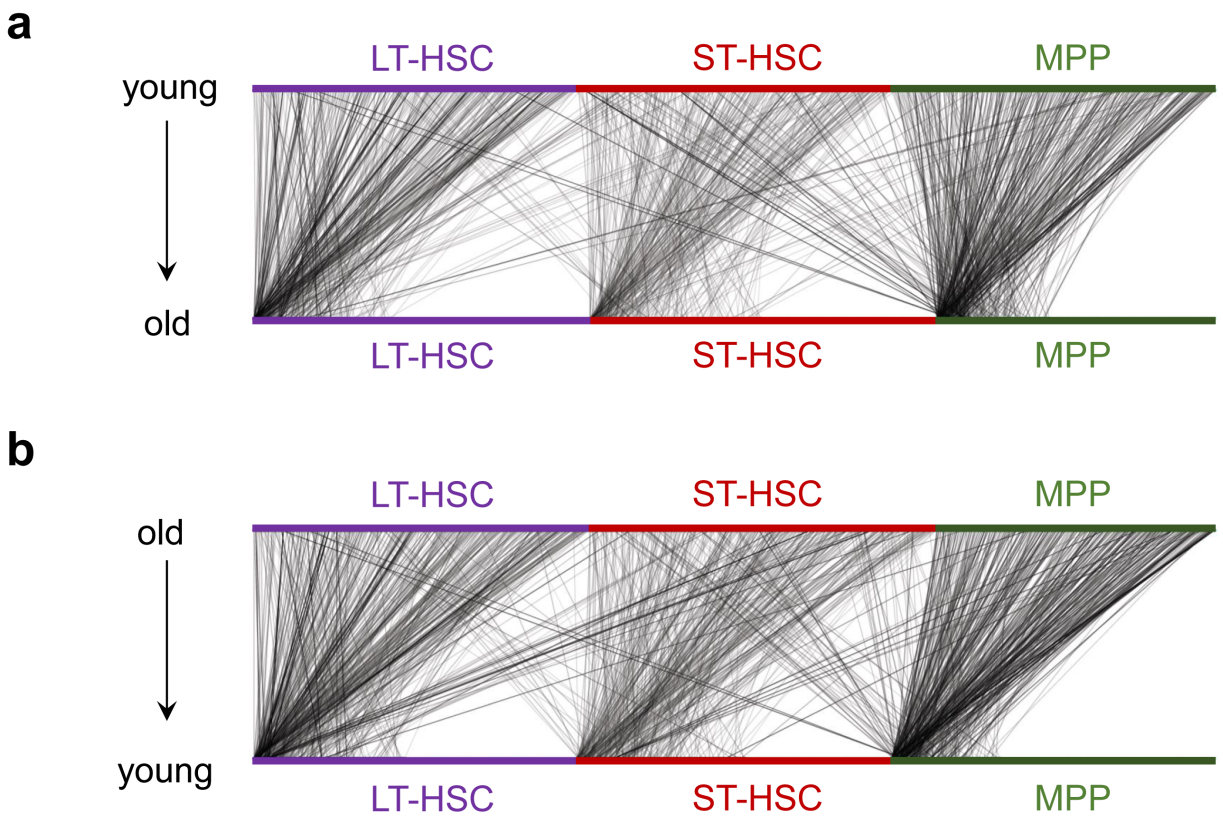

Figure 4. Random walks of scAlign tend to frequently visit a small number of hub cells. (**a**) Random walk probabilities from the top layer representing cells in the young condition from Kowalczyk et al. to the bottom layer represents cells in the old condition. Edges represent the magnitude of the probability for each random walk that initiates from the young cells and transition to the old cells, where darker edges indicate more frequent walks. (**b**) Same as (a), except the top layer represents cells in the old condition and the bottom layer represents young cells, where edges represent random walks from old to young cells.

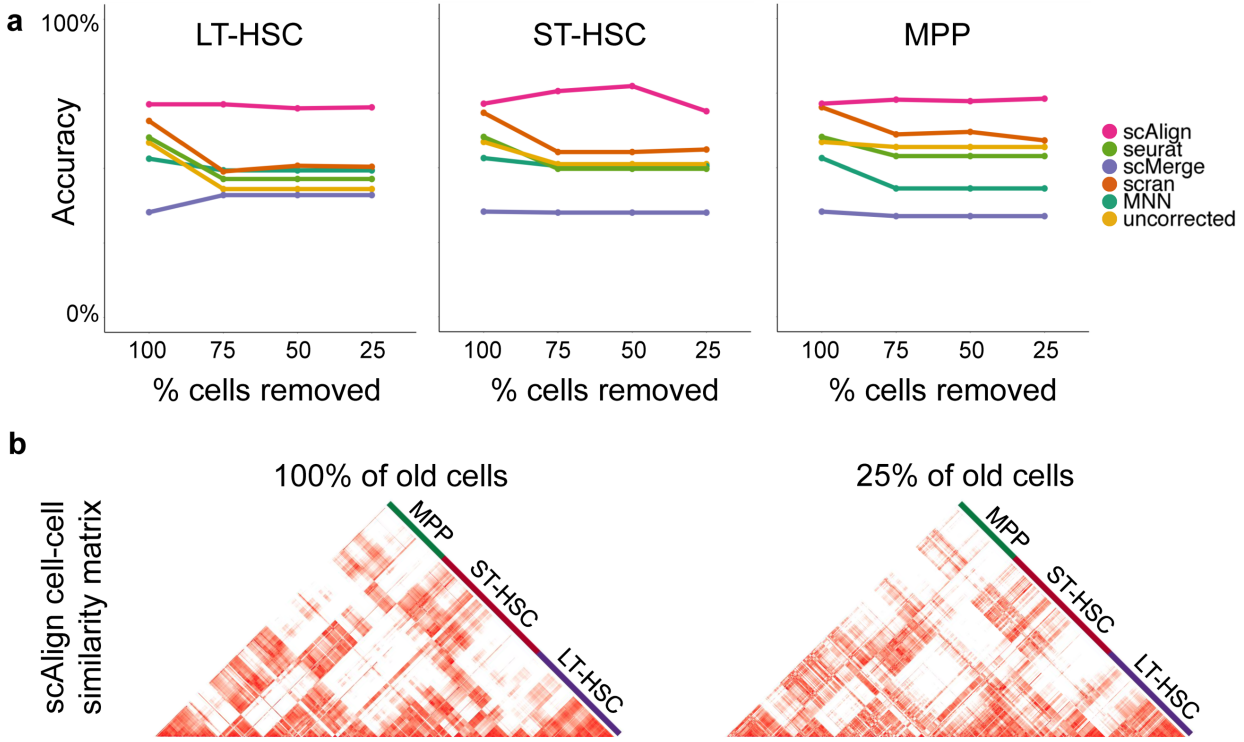

Figure 5. Alignment performance is robust to dataset imbalance in cell type composition. (**a**) Accuracy of classifiers on the Kowalczyk et al. dataset, when removing either LT, ST or MPP cells from the old condition. scAlign initially outperforms all other methods on the full dataset (100%) and exhibits almost no degradation in performance as increasing amounts 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 75% of the target cells from all cell types.

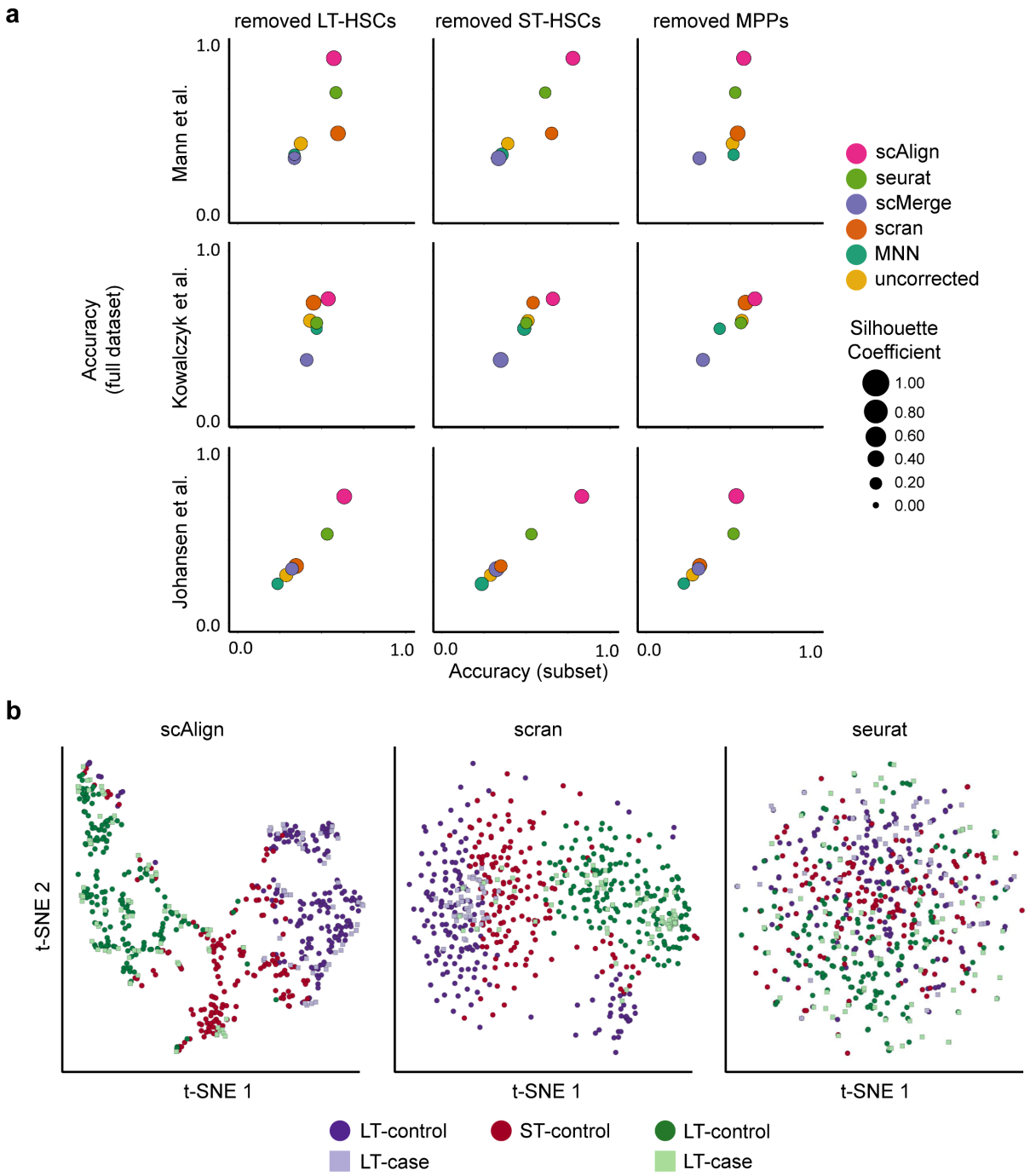

Figure 6. Alignments are robust to mismatching cell types between conditions. (**a**) Scatterplot matrix of performance of each method when both conditions have the same number of cells (y-axis), compared to when one cell type has been removed from one condition (x-axis). Each point is scaled in size 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.

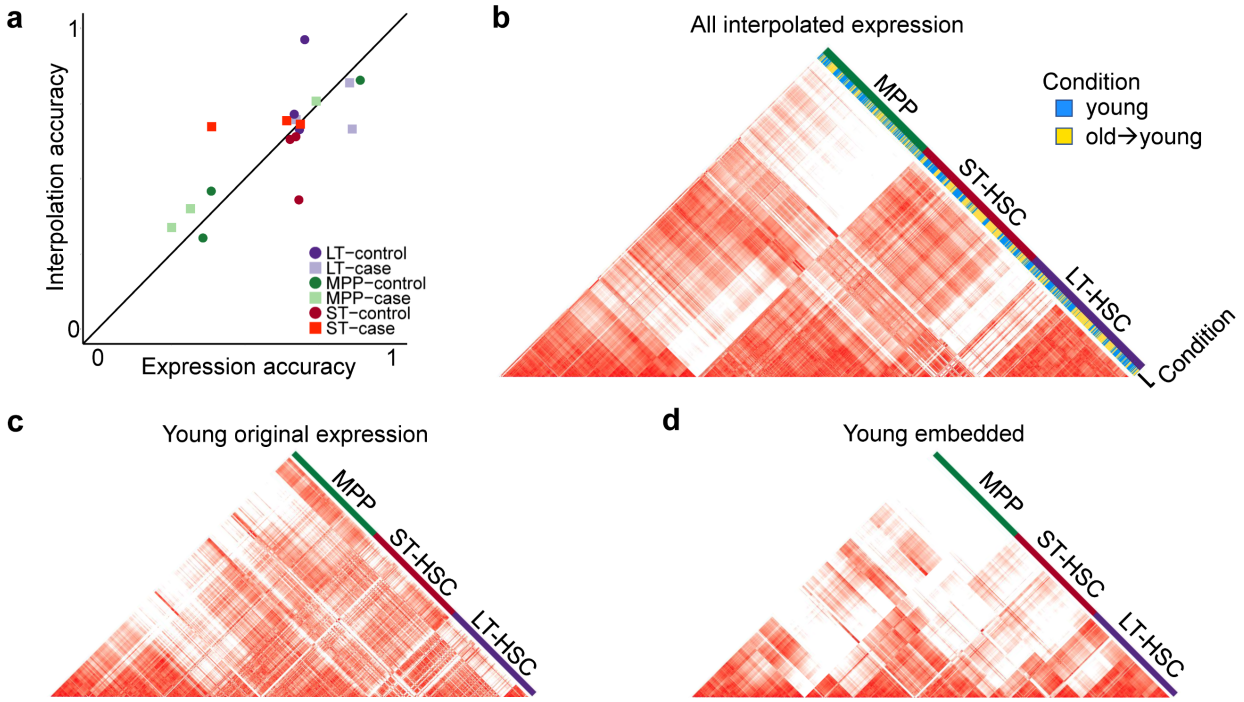

Figure 7. Interpolation of gene expression patterns is accurate. (**a**) Scatterplot of classifiers trained on gene expression profiles of one condition, then used to predict labels of either measured expression profiles from the same condition in a cross-validation framework (x-axis), or used to predict labels of cells sequenced from the other condition that were then interpolated into this condition (y-axis). Similarity in accuracy indicates that cell type information encoded in the gene expression profile is maintained even after interpolation. (**b**) Heatmap of the similarity patterns 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 cells group largely by cell type regardless of the condition in which they were sequenced. (**c**) Heatmap for the similarity patterns in original expression space as compared to (**d**) the same cells representation in the joint cell state space learned by scAlign. Similarity between (c) and (d) indicate the scAlign embedding maintains global similarity between cells in the original gene expression space.

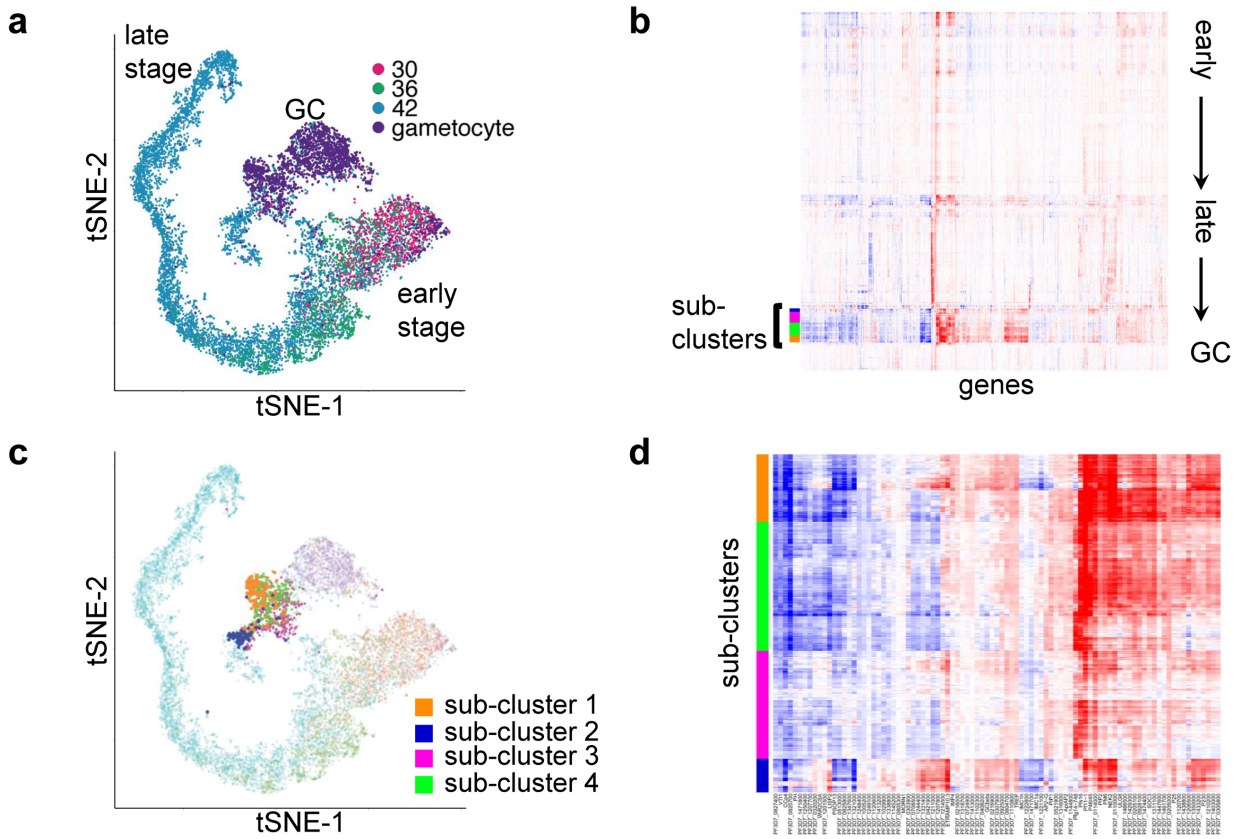

Figure 8. Alignment of malaria cells sequenced from a conditional ap2-g knockdown line reveal a rare subpopulation of cells putatively involved in sexual commitment. (**a**) tSNE visualization of cells which cannot express AP2-G (OFF) and AP2-G capable cells (ON) after alignment by scAlign. Each cell is colored by its position within three cell cycle time points represented in hours post-infection or gametocyte (GC). (**b**) Heatmap of paired differential expression when projecting every cell from (a) into both the ON and OFF condition, then taking the paired difference in expression profiles. Rows represent cells, ordered generally from early stage (top) to late stage and GC (bottom), and columns represent the 661 most varying genes. The heatmap identifies a region of large differential expression between late stage and GC cell populations, labeled as subclusters. (**c**) tSNE similar to (a), but highlighting the subcluster of cells from (b) and colored according to sub-cluster membership. (**d**) Heatmap of paired differential expression focused on the subclusters alone, and where the genes selected are the subset of the original 661 that are gender-specific.

Methods

Methods overview. The scAlign method can be broken down into two steps: (1) alignment, which learns a mapping from gene expression space of individual conditions into a common cell state space, and (2) interpolation, which learns a mapping from the common cell state space back to the gene expression space of the original conditions.

scRNA-seq alignment with scAlign. We define the alignment task as identifying a low dimensional embedding space (termed the cell state space) in which cells that perform the same function map to the same region of the cell state space. 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 cell state space as the cell before stimulus, therefore removing the effect of the stimulus.

ScAlign encodes the cell state space by extending the recent approach of learning by association for neural networks[28,29] 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. 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 cell state 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 cell state 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

$$\boldsymbol{P}^s = \boldsymbol{P}^{s\to t}\boldsymbol{P}^{t\to s}$$
$$\boldsymbol{P}^t = \boldsymbol{P}^{t\to s}\boldsymbol{P}^{s\to t}$$
$$Q_{i,k}^s = \frac{\exp\left(-0.5\|\vec{x}_i^s - \vec{x}_k^s\|^2/\sigma_i^2\right)}{\sum_{k'\neq i}\exp\left(-0.5\left\|\vec{x}_i^s - \vec{x}_{k'}^s\right\|^2/\sigma_i^2\right)}$$
$$Q_{j,k}^t = \frac{\exp\left(-0.5\left\|\vec{x}_j^t - \vec{x}_k^t\right\|^2/\sigma_j^2\right)}{\sum_{k'\neq j}\exp\left(-0.5\left\|\vec{x}_j^t - \vec{x}_{k'}^t\right\|^2/\sigma_j^2\right)}$$
$$P_{i,j}^{s\to t} = \frac{\exp\left(\vec{e}_i^{s\,T}\vec{e}_j^t\right)}{\sum_{j'}\exp\left(\vec{e}_i^{s\,T}\vec{e}_{j'}^t\right)}$$
$$P_{j,i}^{t\to s} = \frac{\exp\left(\vec{e}_i^{s\,T}\vec{e}_j^t\right)}{\sum_{i'}\exp\left(\vec{e}_{i'}^{s\,T}\vec{e}_j^t\right)}$$
$$\vec{e}_i^s = \text{encoder}(\vec{x}_i^s, \boldsymbol{W})$$
$$\vec{e}_j^t = \text{encoder}(\vec{x}_j^t, \boldsymbol{W})$$

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 cell state 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[30] for example, scAlign computes roundtrip random walks of length two that traverse the two datasets. $\boldsymbol{P^s_{i,k}}$, 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 cell state space.

The network weights $\boldsymbol{W}$ are initialized by Xavier[31] and optimized via the Adam algorithm[32] with an initial learning rate of 1e-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 1e-4, and the size of the encoder network architecture.

For the tuning parameter σ_i^2, generally 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 that adjusting the magnitude of this parameter based on mini batch size was not necessary to achieve optimal results.

In our experiments, we set the size of the encoder architecture by either automatically building a network based on the dimensionality of the input data in conjunction with a complexity parameter, or from a catalog of general 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 (**Supplementary Figs.**) 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, we avoid training the neural network to perform unnecessary complex transformations on the data.

scRNA-seq interpolation with scAlign. The interpolation component of scAlign maps cells from the joint cell state space back into each of the individual condition-specific gene expression spaces. Interpolation is performed by training a condition-specific decoder neural network to perform the interpolation. 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[32] with learning rate 1e-4 and trained for at most 30,000 iterations.

Principal Component Analysis and Canonical Correlation Analysis preprocessing transformations of scRNA-seq data. The objective function that scAlign optimizes does not incorporate terms specific to RNA-seq data. To speed up the training process, we therefore tested and found that calculating 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 significantly sped up training (**Supplementary Figs.**). Note that although the encoder network would take as input PC or CC dimensions and transform them into cell state space, the decoder is still trained to transform cell state space coordinates into the original gene expression space for each condition for paired differential expression analysis.

Using partial or complete cell type annotations with scAlign. The objective function optimized by scAlign can naturally incorporate partial, overlapping or complete cell type labels on the cells, in one or both domains. We do not need to make any assumptions about the exclusivity of these labels (e.g. a cell could be assigned more than one label). Suppose there are C cell type labels available. Then define matrix $\boldsymbol{A}^s$ such that $A^s_{i,c} = 1$ if cell i in condition s has cell type label c, else $A^s_{i,c} = 0$. The scAlign objective function then becomes:

$$f = \left[\sum_i \left(\text{cross-entropy}\left(\vec{P}^s_{i,\cdot}, \vec{Q}^s_{i,\cdot}\right) + \sum_c A^s_{i,c}\text{cross-entropy}\left(\vec{P}^s_{i,\cdot}, A^s_{\cdot,c}/(\mathbf{1}^{\boldsymbol{T}} A^s_{\cdot,c})\right)\right)\right] + \left[\sum_j \left(\text{cross-entropy}\left(\vec{P}^t_{j,\cdot}, \vec{Q}^t_{j,\cdot}\right) + \sum_c A^t_{j,c}\text{cross-entropy}\left(\vec{P}^t_{j,\cdot}, A^t_{\cdot,c}/(\mathbf{1}^{\boldsymbol{T}} A^t_{\cdot,c})\right)\right)\right] + \lambda\|\boldsymbol{W}\|^2_F$$

The additional terms in f encourage random walks from a cell with known label to end at another cell with the same known label, uniformly selected at random from all cells with that label. Additionally, the classifier component is incorporated into the encoding neural network by transforming the embedding layer activations into class specific logit scores through the addition of a single hidden layer. The classifier minimizes the mean cross-entropy of the predicted and actual cell labels as defined by the second term within each summation of f.

Acquisition and preprocessing of Mann et al. dataset. We obtained the gene count matrix for HSC data generated from Mann et al.[33] from GSE100426. The provided data matrix was already filtered based on quality control metrics. We normalized the count matrix to TP10K and then removed plate specific batch effects by fitting a linear model on the scaled and centered using Seurat's NormalizeData and ScaleData functions. We retained the union of the top 3,000 variable genes between control and condition cells.

Acquisition and preprocessing of Kowalczyk et al. dataset. We obtained the gene count matrix for both C57BL6 and DBA mouse HSC data generated from Kowalczyk et al.[34] from GSE59114. Only single cell data from mouse C57BL6 was used during alignment to avoid cross mouse batch effects. We normalized the count matrix to TP10K then scaled and centered using Seurat's NormalizeData and ScaleData functions. We retained the union of the top 3,000 variable genes between young and old cells.

Acquisition and preprocessing of CellBench dataset. We obtained the gene count matrix for the RNA mixture experiments in CellBench generated by Tian et al.[17] from the R data file

mRNAmix_qc.RData available on github. We normalized the count matrix to TP10K then scaled and centered using Seurat's NormalizeData and ScaleData functions. We retained the union of the top 3,000 variable genes between mixtures profiled on CEL-seq2 and SORT-seq.

Execution of other scRNA-seq alignment methods. We compared scAlign against MNN[29], Seurat[30], scMerge[31] and scran[29], which were run based on method specific guidelines and following the workflow defined by CellBench and available on github. Prior to running each method, scran's decomposeVar function was used to identify the most variable genes for subsetting the data matrices. MNN was provided log-count data subset to the most variable genes with all parameters set to default. Seurat was provided the count level data which was normalized, then scaled and centered using the NormalizeData and ScaleData functions. Initially, 30 canonicial correlates were used for dimensionality reduction, then the MetageneBicorPlot function was used to select the optimal number of dimensions as defined by Seurat's integrated PBMC tutorial. The remaining canonicial correlates were aligned using Seurat's AlignSubspace function. scMerge was provided both count and log-count data along with a set of least variable genes identified by sorting the results of the var function in R on the normalized count matrix. The parameter kmeansK for the number of clusters was set based on cell type information. Scran was provided with log-count data subset to the most variable genes previously identified by decompseVar, and return_dense was set to TRUE.

Construction of the Johansen et al. dataset. We constructed the Johansen et al. dataset by merging multiple count matrices from the Mann et al. and Kowalczyk et al. studies. The control dataset was defined completely by young C57BL6 mouse cells. To construct the stimulated condition, we merged LT-HSCs perturbed by LPS from Mann et al., ST-HSCs from old C57BL6 mouse cells and MPPs from both young and old DBA mouse cells collected by Kowalczyk et al.

Acquisition and preprocessing of malaria dataset. We obtained the gene count matrix for the malaria data generated by Poran et al.[35] from the KafsackLab github. The data was preprocessed using the provided scripts and subset into a control and condition dataset using the provided AP2G-ON or AP2G-OFF labels in the metadata.

Identification of differentially expressed genes. Differentially expressed (DE) genes were computed using the bimod, DESeq2 and MAST methods implemented in the Seurat findMarkers function. The intersection of DE genes with p-value less than 0.01 from these three methods was used to define a final set of DE genes for each cell type. The analysis was performed on the normalized, scaled and centered data matrices computed by Seurat's preprocessing pipeline.

Measuring accuracy of transcriptional alignment. Alignment performance for each method was measured through a classifier trained to label the conditioned dataset using only labels from the corresponding control dataset. 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 conditioned cells. The predicted labels were compared against heldout labels to measure accuracy.

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.

2D tSNE visualizations of embeddings for alignment methods. By default, we use the Rtsne implementation of tSNE, which first projects input data into 50 principal components before inputing into the tSNE algorithm. All methods other than Seurat and scAlign produce corrected expression matrices, and for these we use the default 50 PCs for Rtsne. Seurat automatically selects the number of dimensions to project into for each individual dataset. scAlign was used to align scRNA-seq data into a 32-dimensional embedding space for all runs. For both Seurat and scAlign, the PCA step of Rtsne was skipped.