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

scRNA-seq dataset integration occurs in different contexts, such as the identification of cell type-specific differences in gene expression across conditions or species, or batch effect correction. We present scAlign, an unsupervised deep learning method for data integration that can incorporate partial, overlapping or a complete set of cell labels, and estimate per-cell differences in gene expression across datasets. scAlign performance is state-of-the-art and robust to cross-dataset variation in cell type-specific expression and cell type composition. We demonstrate that scAlign identifies a rare cell population likely to drive malaria transmission. Our framework is widely applicable to integration challenges in other domains.

Background

Single cell RNA sequencing (scRNA-seq) technologies 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[1–7] is becoming increasingly important in a number of contexts. For example, in case-control studies for which a pair of scRNA-seq datasets are generated from biological replicate populations before and after stimulus[8–11], functionally matched cell types across datasets must be identified and aligned in order to estimate cell type-specific response to stimulus. Also, when similar cell populations are sequenced using different technologies or laboratories, cell type-specific differences between datasets must be estimated and removed before a combined analysis of all data can be performed.

Integrative analyses are currently challenging due to several factors. First, dataset integration can be viewed as mapping one dataset onto another. In the example of case-control studies, increased cell-type specific response to stimulus requires a more complex mapping between datasets. Therefore, integrative tools must be able to freely scale up or down the flexibility of their mapping functions to successfully perform integration at varying complexities. Second, current integrative tools can be separated into two exclusive sets: those that require all cells from all datasets to have known cell type (supervised), and those that do not make use of any cell type labels (unsupervised). Thus, if only a subset of cell can be labeled with high accuracy, or if only one dataset is labeled (as is the case when reference annotated cell atlases are available[12–17]), this partial set of labels currently cannot be used in data integration. Third, measured transcriptomes even for homogeneous populations of cells occupy a continuum of cell states, for both technical[18,19] and biological[20–22] reasons. Thus, individual cells cannot be matched exactly across datasets. Therefore, downstream analysis of integrated datasets typically involve clustering cells across datasets to find matching cell types and estimating cell type-specific differences across datasets. The clustering step makes it difficult to find rare cell populations that differ between datasets.

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 cell state 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 particularly when individual cell types exhibit strong dataset-specific signatures such as heterogeneous responses to stimulus, and 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 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**. While this paper is written in the context of integrating multiple datasets representing cell populations exposed to different stimuli or control conditions, scAlign can be readily used for any data integration context discussed above. The premise of integration methods is that when similar cell populations are sequenced under different conditions, the cells will separate first by condition, then by type (**Fig. 1a**). The first component of scAlign is the construction of a joint cell state space using scRNA-seq data from all conditions, in which cells do not separate by condition (**Fig. 1b**). This 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 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 cell state space.

In the second component of scAlign, we train condition-specific deep decoder networks capable of projecting individual cells from the cell state space back to the gene expression space of each input condition, regardless of what condition the cell is originally sequenced in (**Fig. 1c**). We use these decoders to measure per-cell variation of expression across conditions, and in the case of integrating two conditions, this generates a paired-differential expression profile representing the difference in expression of the same cell state across conditions (**Fig. 1d**). 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.

Results - scAlign captures 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**). Our first comparison used CellBench[23], a benchmark consisting of three human lung adenocarcinoma cell lines (HCC827, H1975, H2228) that were sequenced using 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. While the alignment of the pure cell populations sequenced

across platforms was trivial and did not require data integration methods (**Supplementary Fig. 1**), alignment of RNA mixtures across platforms was more challenging and more clearly illustrated the performance advantage of scAlign (**Fig. 2a**). We additionally benchmarked methods using data generated by Kowalczyk et al.[24] and Mann et al.[25], each consisting of three hematopoietic cell types sequenced in young and old mice (LT-HSC, ST-HSC, MPP), though Mann et al. additionally challenged the mice with a LPS or a control condition. Similar to our results with CellBench, scAlign outperforms other approaches on both of these benchmarks (**Fig. 2b,c**).

To better understand why the relative performance of the other methods was inconsistent across benchmarks (**Fig. 2a-c**), we next characterized the difficulty of each benchmark for alignment. For each cell type in each benchmark, we computed the differentially expressed genes (DEGs) across conditions. We observed considerable overlap in the DEGs for each cell type (**Supplementary Fig. 2**), suggesting these benchmarks may be less challenging to align and therefore more difficult to distinguish other approaches from each other. We therefore constructed a novel benchmark termed Johansen et al. by combining published scRNA-seq data on hematopoietic cells measured across different studies and stimuli. This benchmark yields increased cell type specific differential gene expression across the three cell types (**Supplementary Fig. 2**), which therefore makes it more challenging to align. On the Johansen et al. benchmark, we find that scAlign's performance is robustly superior, while Seurat also outperforms the remaining methods by a large margin (**Fig. 2d**).

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[26] and trajectory inference[27–29]. 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 shared 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 (see Methods). 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 autoencoder (**Fig. 3, Supplementary Fig. 3**), suggesting scAlign's embedding space benefits from pooling cells from across all conditions.

Most of the existing approaches tested in this paper are unsupervised, in that they do not use cell type labels to aid alignment, even if available. While the focus of this paper is on scAlign performance in this unsupervised setting, a unique feature of scAlign is that its 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 one of the conditions to align has labels, or if only a subset of cells can be reliably labeled, we can use that information to guide the alignment. **Supplementary Figures 4a-d** illustrate, for each of the four benchmarks from Figure 2, that alignment performance of scAlign improves when cell type labels are available at training time, and exceeds other supervised methods such as MINT[19]. Furthermore, even 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 cell state space, suggesting the inferred cell state space is robust to adding labels during alignment (**Supplementary Fig. 5**).

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

Besides cell type-specific responses to stimuli, we reasoned that the other factor that determines alignment difficulty is the difference in the proportion of each cell type between conditions. That is, cell types that are abundant in one condition but rare or absent in another may pose difficulty during alignment. 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. 4**). 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. When we repeated these experiments on the Mann et al. benchmark, we found most methods decreased in performance as we removed more cells from each type, though scAlign still outperformed all other methods (**Supplementary Fig. 6**).

We next investigated the factors that underlie scAlign's robustness to varying cell type proportion imbalances 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 cell state space when performing random walks of length two (**Fig. 5a**). 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 cell state space. The walk then returns to a different cell (excluding the starting cell) in the original condition, also based on proximity in cell state space. For every cell in each condition, we calculated the frequency that such random walks (initiated from the other condition) pass through it (**Fig. 5b-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 (**Supplementary Fig. 7**). 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.

In the above experiments, we have aligned conditions in which the same set of cell types are present in both conditions. We next explored the behavior of scAlign and other approaches when there are cell types unique to one of the conditions. We expect such scenarios to arise when only a subset of cells of a given type might respond to, or be targeted by, a stimulus (where a stimulus could be a targeted therapy, or even speciation[30]). 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 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. Even in cases where the alignment accuracy was similar between methods, scAlign visually separates cell types in its cell state space more so than other approaches such as scran and Seurat (**Fig. 6b**). For other approaches, the separation of different cell types within the same condition shrinks when one cell type is removed (**Supplementary Fig. 8**).

Results - scAlign interpolates gene expression accurately

One of the more novel features of scAlign is the ability to map each cell from the cell state 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 cell state 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 cell state 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 control 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. 7a**), suggesting that cells of one type sequenced in one condition, maintain their cell identity when mapped into another condition.

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 their condition of origin. 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 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 cell population potentially driving sexual commitment

We next applied scAlign to identify genes associated with sexual commitment in *Plasmodium falciparum* (malaria). Briefly, the life cycle of malaria 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 either all committed to asexual replication or differentiation into gametocytes. 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. 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 occurs a challenging task.

Figure 8a illustrates the cell state 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[31], cells generally form a trajectory in the state space, corresponding to different time points in the malaria cell cycle. scAlign is able to maintain most gametocytes from the AP2-G CTRL condition as a distinct population that is not aligned to any cell population from the AP2-G OFF condition, whereas other tested methods are unable to accurately isolate the gametocyte population (**Supplementary Fig. 9**).

Using the interpolation component of scAlign, we projected each cell from each condition in the cell state space to the expression space of the AP2-G CTRL and AP2-G OFF conditions. Taking the difference in interpolated expression between AP2-G CTRL and AP2-G OFF, we computed 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 subcluster of cells that lie at the interface between late-stage cells and the gametocytes that are specific to AP2-G CTRL (**Fig. 8c**). 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 for identifying 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 subcluster (**Fig. 8d**). Notable are several genes with previously established roles in gametocyte maturation, including P48/45[32] and Pf11-1[33,34]. 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

Here we have shown that scAlign outperforms other integration approaches, particularly when there is strong cell type-specific response to stimuli, or when there is an imbalance in cell type representation across conditions. scAlign will be particularly useful in the context where some cell type labels may be available. More specifically, the unified alignment framework of scAlign allows for the optional use of cell labels that may be available for one or more conditions, and furthermore can use labels only available for subsets of cells. Partially annotated datasets may arise frequently, as cell type markers are typically only available for well established cell types such as the hematopoietic cells. Furthermore, 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. With the increasing number of cell atlases[12–17] that are accurately labeled by domain experts and are now available, scAlign can take advantage of the accurate labeling of these atlases to annotate new datasets that lack labels.

One of the principal advantages of scAlign over existing integration methods is we can identify rare cell types that differ in expression 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 then perform an unpaired differential expression test within each cluster to identify condition-specific differences. The need to cluster cells means detection of rare cell types can be highly sensitive to the choice of clustering algorithm or parameters. In contrast, through interpolation, scAlign can generate paired differential expression heatmaps which can make evident the presence of rare cell populations that differ in expression across condition (**Fig. 8**). Additionally, on our benchmarks as well as the malaria dataset, we showed that scAlign can match cells of the same type even when the proportions of those types is imbalanced.

Here we have primarily compared scAlign against unsupervised alignment methods. In our supervised alignment results, scAlign compared favorably against MINT[19], another supervised method, 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 to identify 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. 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.

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 cell state spaces, which also acts as a form of data driven regularization. In our experiments, typical scRNA-seq data will require at most three layers to align. The results of this study were robust to network depth and width (**Supplementary Fig. 10**) along with choice of hyper parameters.

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 a preliminary step of dimensionality reduction 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. 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 (**Supplementary Fig. 11**).

In this paper, we have primarily evaluated the accuracy of scAlign in the context of comparing scRNA-seq data across two conditions. 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 comparing different disease types[35], characterizing the same tissue across multiple organisms[36], or integrating data from multiple patients for the same disease[37] 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.

Declarations

Availability of data and material. The code implementing scAlign, as well as the datasets generated and/or analysed during the current study are available in the GitHub repository, at https://github.com/ucdavis/quontitativebiology.

Funding. 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 V used for this research was donated by the NVIDIA Corporation.

Acknowledgements. We thank Bjorn Kafsack for helpful discussions regarding the malaria cell line data.

References

1. 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).
2. 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).
3. 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).
4. 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
5. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).
6. 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
7. 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).
8. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452.e17 (2017).
9. Datlinger, P. *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
10. 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).

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

12. 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).

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

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

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

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

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

18. 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).

19. 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).

20. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363–369 (2014).

21. Eldar, A. & Elowitz, M. B. Functional roles for noise in genetic circuits. *Nature* **467**, 167–173 (2010).

22. Maamar, H., Raj, A. & Dubnau, D. Noise in gene expression determines cell fate in Bacillus subtilis. *Science* **317**, 526–529 (2007).

23. 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

24. 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).

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

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

27. 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

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

29. 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).

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

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

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

33. 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).

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

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

36. 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).

37. Chung, W. *et al.* Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* **8**, 15081 (2017).

38. 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).

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

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

41. 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).

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

Figures

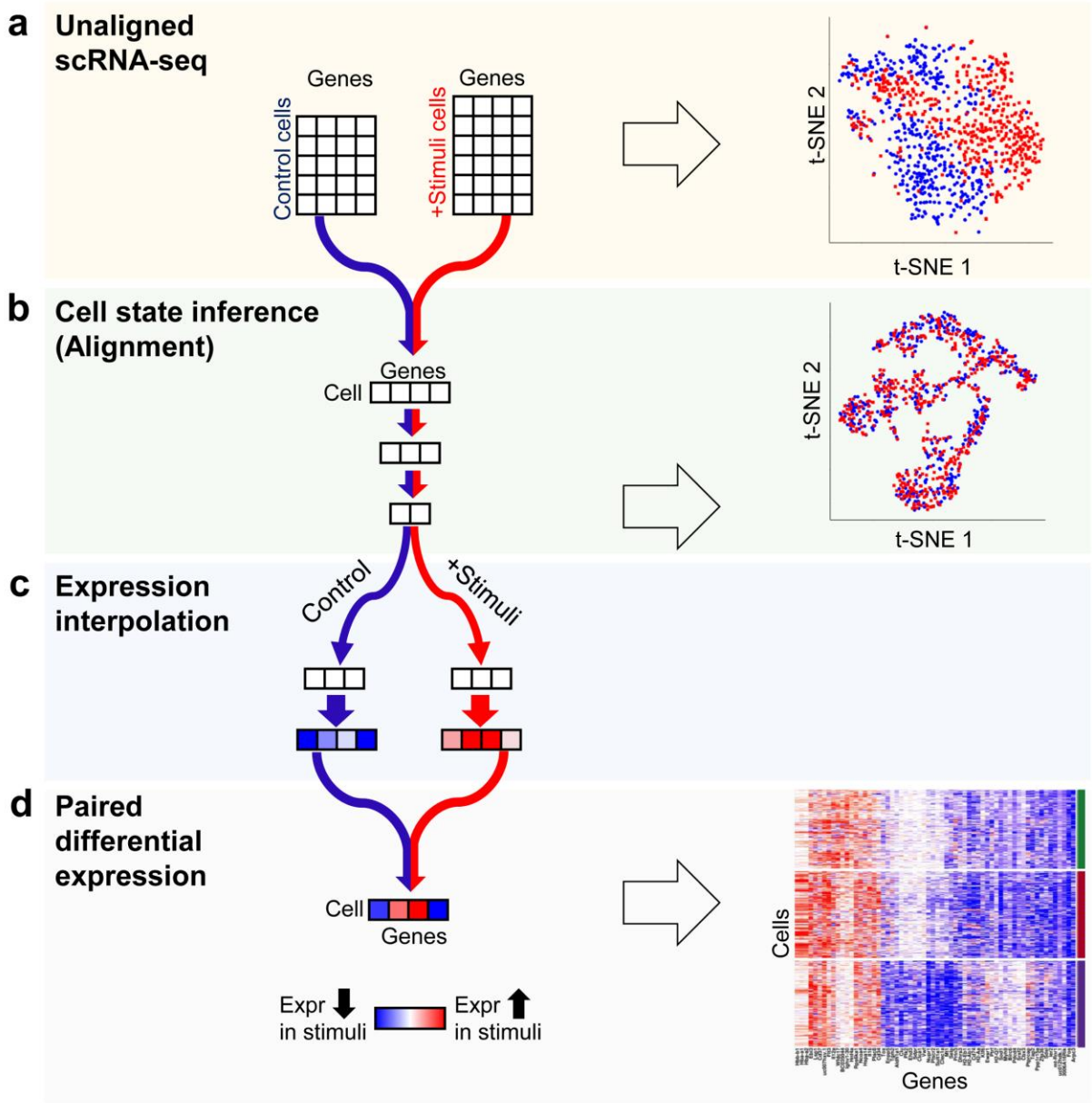

Figure 1. Schematic of unsupervised alignment and paired differential expression with scAlign. (**a**) Input to scAlign consists of multiple scRNA-seq datasets, each representing a condition for example. Expression can be represented as either gene-level expression, or embedding coordinates from dimensionality reduction techniques such as PCA or CCA. (**b**) A deep encoding network learns a low-dimensional cell state space that simultaneously aligns cells from all conditions. (**c**) Paired decoders project cells from the cell state 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 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.

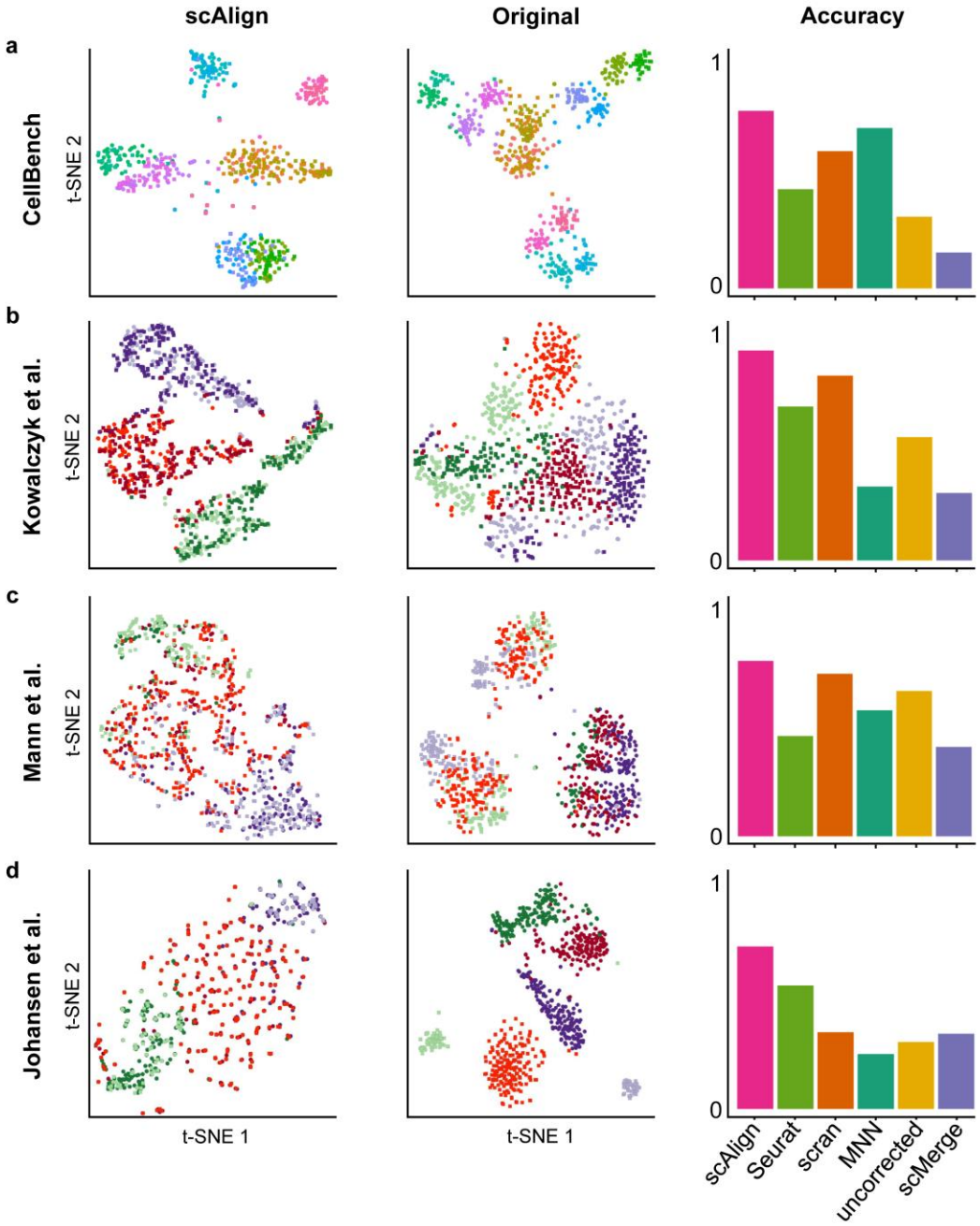

Figure 2. scAlign outperforms existing alignment approaches on four benchmarks. (**a**) CellBench, a benchmark consisting of mixtures of RNA from three cancer cell lines sequenced using multiple platforms. Plots from left to right: (1) tSNE plot of embeddings after alignment with scAlign. (2) tSNE plots generated using the original expression profiles. (3) Bar plot indicating the accuracy of a classifier trained on labeled cells from one condition and used to predict cell type labels in another condition. (**b**) Same as (a), but with the Kowalczyk et al. benchmark consisting of hematopoietic cells sequenced from young and old mice. (**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 Johansen et al. benchmark consisting of hematopoietic cells responding to different stimuli.

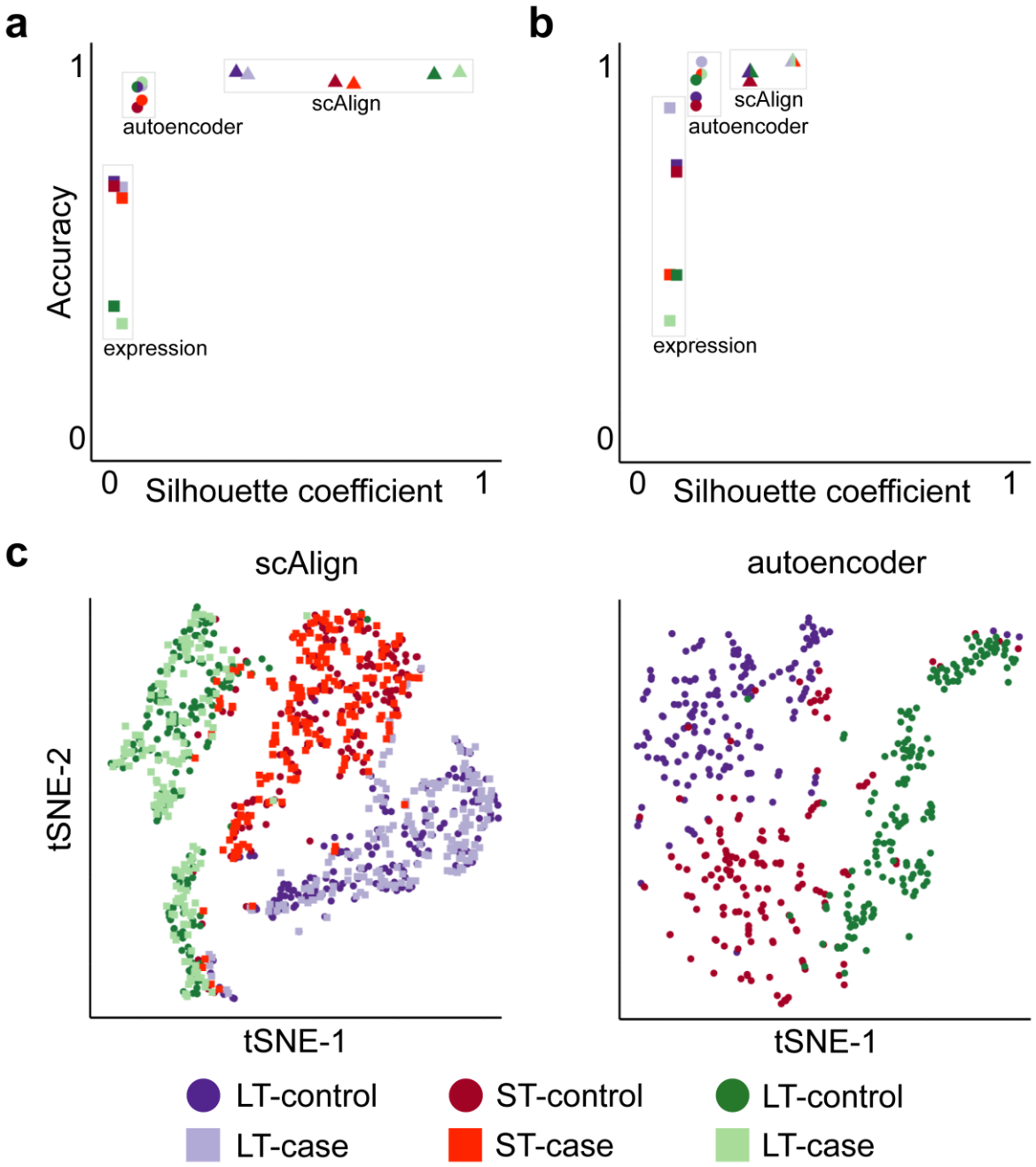

Figure 3. Joint analysis of cells from all conditions leads to more accurate clustering of cell types compared to analysis of individual conditions. (**a**) Scatterplot illustrating the quality of clustering of cell types within each condition from the Mann et al. benchmark. Each point represents one 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 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.

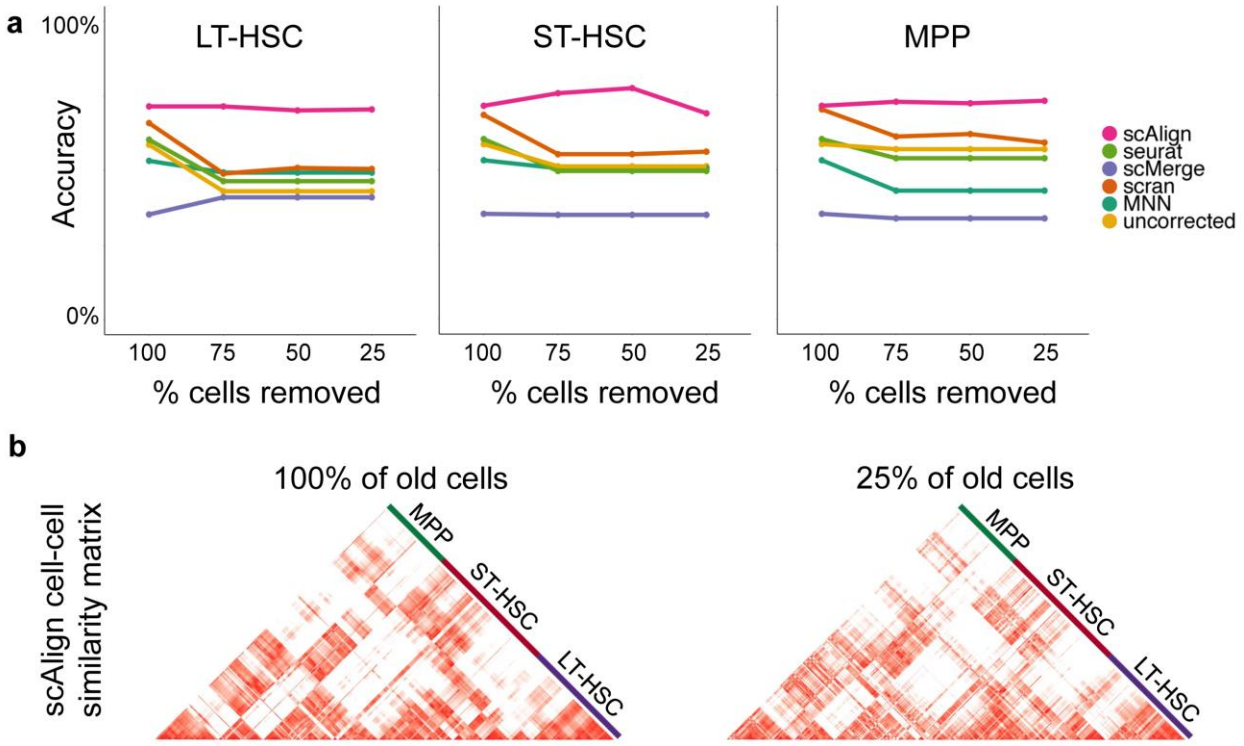

Figure 4. Alignment performance is robust to imbalance in cell type representation in each condition. (**a**) Accuracy of classifiers on the Kowalczyk et al. benchmark, when removing either LT-HSC, ST-HSC or MPP cells from the old condition. scAlign outperforms all other methods on the full dataset (100%) and exhibits almost no 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 75% of the old mouse cells from all cell types.

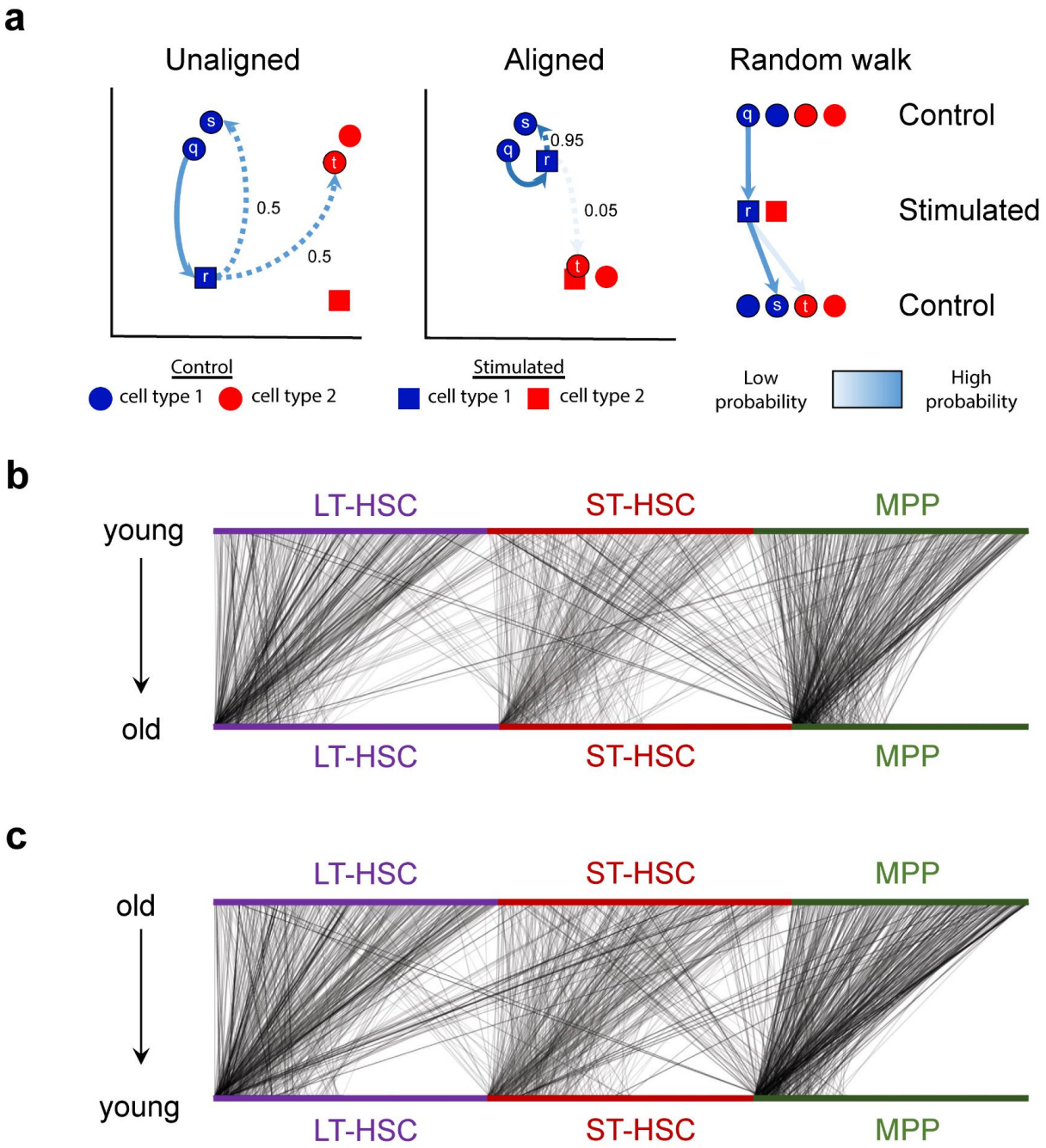

Figure 5. Random walks during scAlign training frequently visit a small number of hub cells. (**a**) Schematic of the cross domain round trip random walk prior to and after training of scAlign. (**b**) Visualization of the probability of a walk from each individual young cell (top) to each individual old cell (bottom) during training of 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.

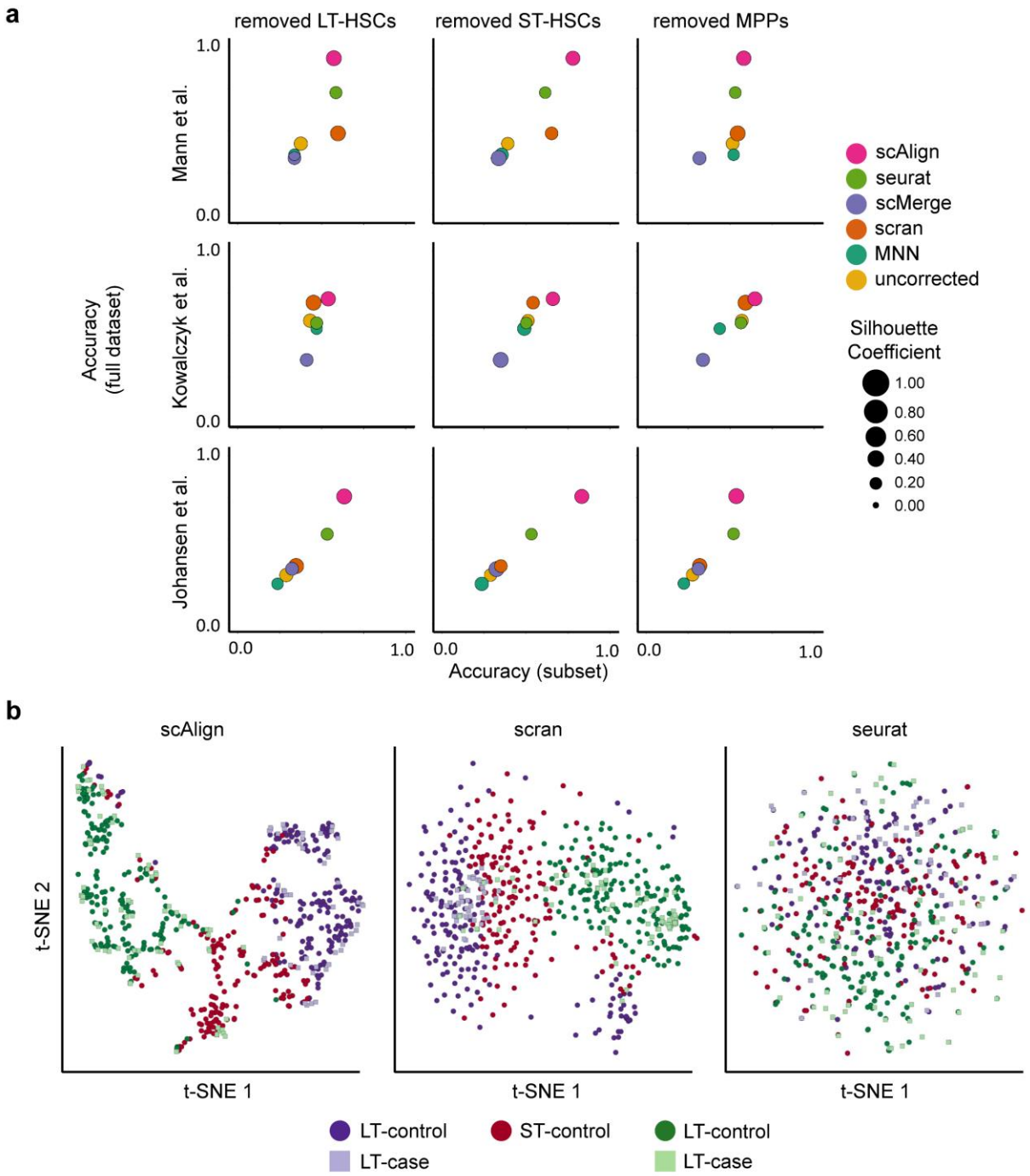

Figure 6. Alignments are robust to mismatched cell types between conditions. (**a**) Scatterplot matrix of performance of each method when both conditions have the same number of cell types (y-axis), compared to when one cell type has been removed (the LPS condition of the Mann benchmark, or the old 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.

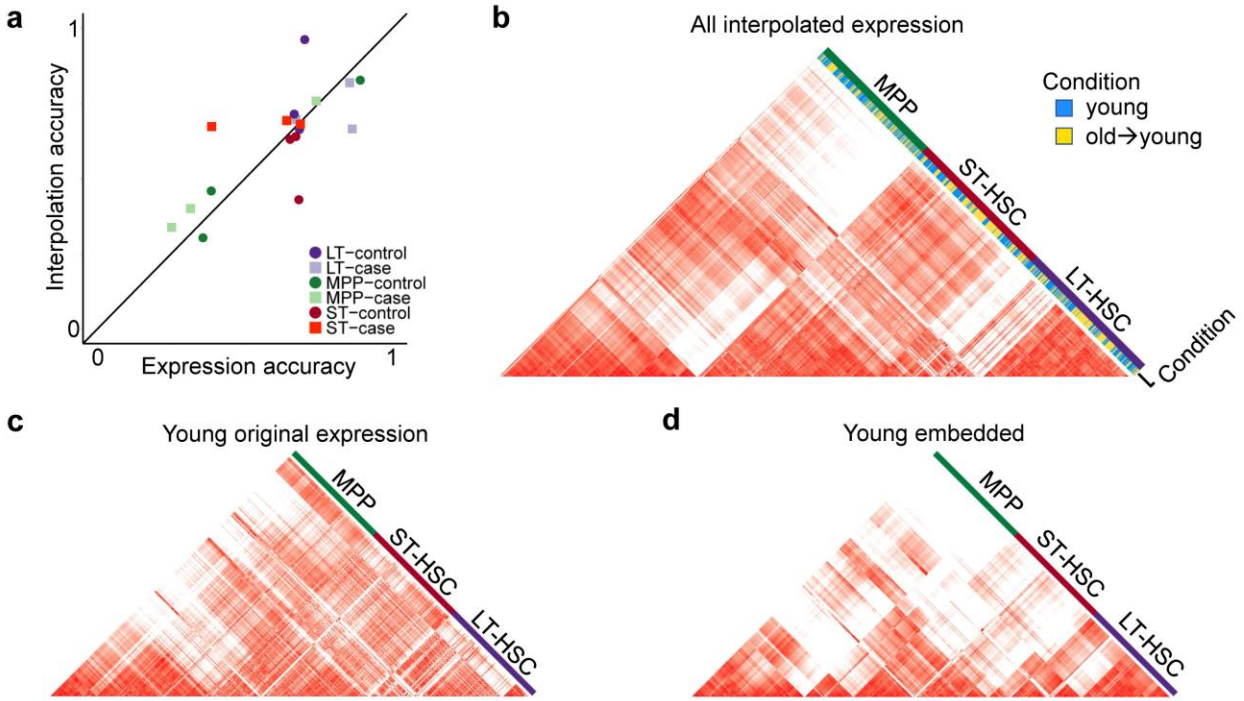

Figure 7. Interpolation of gene expression patterns is accurate. (**a**) Scatterplot of classifiers trained on gene expression profiles of one condition, that are subsequently used to predict labels of either measured expression profiles from the same condition in a cross-validation framework (x-axis), 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 cell state 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.

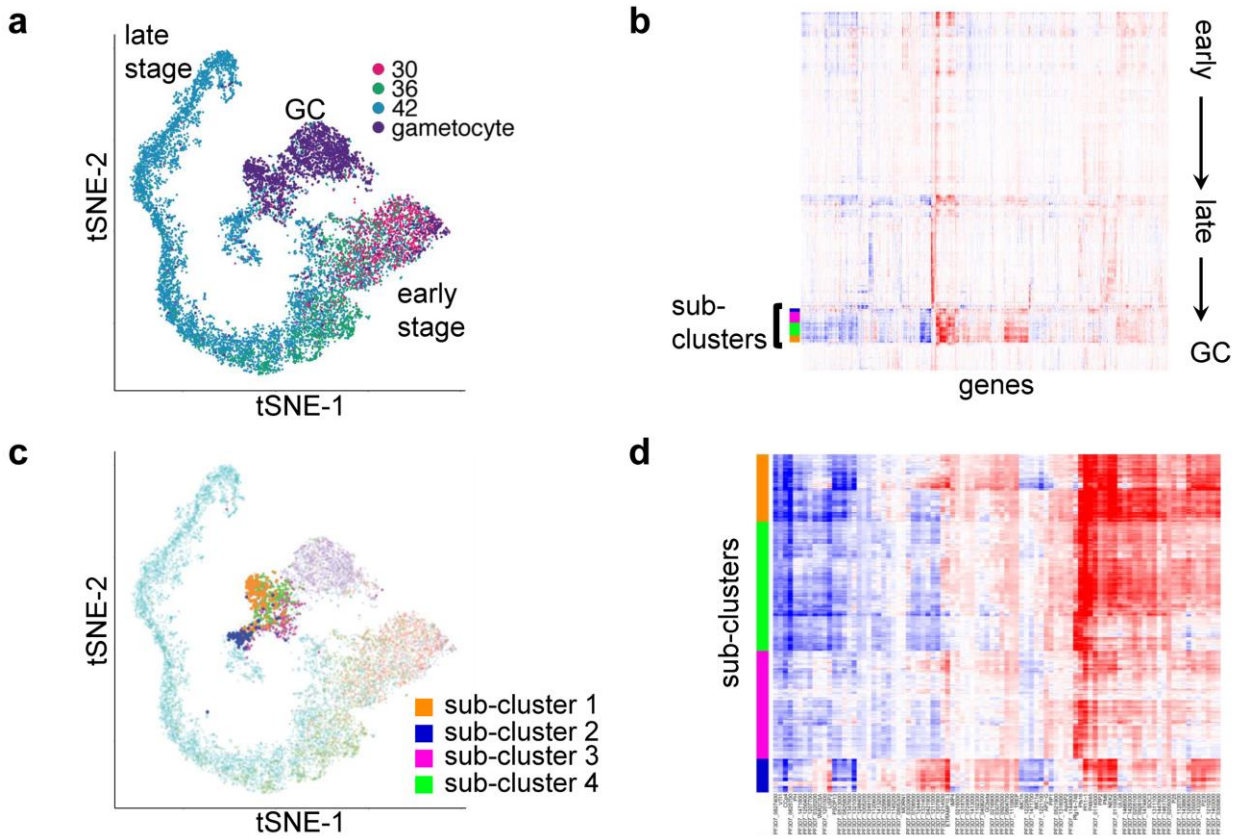

Figure 8. Alignment of malaria cells sequenced from a conditional ap2-g knockdown line reveal a rare subpopulation of cells putatively involved in sexual commitment. (**a**) tSNE visualization of cells which cannot express ap2-g (OFF) and ap2-g capable cells (CTRL) after alignment by scAlign. Each cell is colored by its position within three cell cycle time points measured in hours post-infection or gametocyte (GC). (**b**) Heatmap of paired differential expression when projecting every cell from (a) into both the CTRL and OFF conditions, 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 subcluster cells are in bold colors. (**d**) Heatmap of paired differential expression focused on the subclusters (rows) alone, and where the genes selected (columns) are the subset of the original 661 from (b) that are gender-biased in expression.

Methods

Methods overview. The scAlign method consists of 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 functionally similar cells 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[38,39] 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[40] for example, scAlign computes roundtrip random walks of length two that traverse the two conditions. $\boldsymbol{P^s_{i,k}}$, 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[41] and optimized via the Adam algorithm[42] 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, 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 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 (**Supplementary Figs. 10**) 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, we avoid training the neural network to perform unnecessary complex transformations on the data.

scRNA-seq interpolation with scAlign. The interpolation component of scAlign trains a condition-specific decoder to map cells from the joint cell state 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[42] 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 such as a negative binomial observation model. To speed up the training process, we therefore tested and 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 significantly sped up training (**Supplementary Figs. 11**). Note that even when the encoder network is given as input PC or CC dimensions instead of gene expression measurements, the decoder is still trained to transform cell state space coordinates into the original gene expression space.

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:

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

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. 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.

Acquisition and preprocessing of Mann et al. benchmark. We obtained the gene count matrix for HSC data generated from Mann et al.[40] 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. benchmark. We obtained the gene count matrix for both C57BL6 and DBA mouse HSC data generated from Kowalczyk et al.[40] 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 benchmark. We obtained the gene count matrix for the RNA mixture experiments in CellBench generated by Tian et al.[23] 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. benchmark. We constructed the Johansen et al. benchmark by merging multiple count matrices from the Mann et al. and Kowalczyk et al. studies. The control condition 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 data. We obtained the gene count matrix for the malaria data generated by Poran et al.[41] from the KafsackLab github. The data was preprocessed using the provided scripts and subset into a control and condition condition 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 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.

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 condition. 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.