Self-assembling Manifolds in Single-cell RNA Sequencing Data

Alexander J. Tarashansky[1,5], Yuan Xue[1,5], Stephen R. Quake[1,2,3], Bo Wang[1,4,*]

[1]Department of Bioengineering, Stanford University, Stanford, CA, USA. [2]Department of Applied Physics, Stanford University, Stanford, CA, USA. [3]Chan Zuckerberg Biohub, San Francisco, CA, USA. [4]Department of Developmental Biology, Stanford University School of Medicine, Stanford, CA, USA. [5]These authors contributed equally to this work.

*Correspondence should be addressed to B.W. (wangbo@stanford.edu)

Abstract

Analysis of single-cell transcriptomes remains a challenge in that subtle differences of cell types are difficult to resolve. Here we present the self-assembling manifolds (SAM) algorithm, which dynamically rescales gene expression to amplify differences between cells. We demonstrate its advantage over other methods by analyzing stem cells from *Schistosoma*, a parasite that infects >250 million people. Benchmarking on another 47 datasets, SAM consistently improves cell clustering and marker gene identification.

Text

The rise of single-cell RNA sequencing (scRNA-seq) technologies has enabled researchers to explore cell types, delineate cell developmental trajectories, and measure molecular responses to external perturbations[1-5]. Besides the rapid evolution of experimental techniques, the ever-increasing wealth of data has spawned numerous analytical methods[6-8]. These methods are often optimized for characteristics inherent to particular datasets and may require human inputs to select important genes and tune method parameters[9]. An analytical pipeline that is unsupervised and universally applicable to different datasets and organisms with little to no *a priori* knowledge remains an open challenge. In particular, robust and unsupervised detection of subtle differences in gene expression between cells in a largely homogeneous population is still not possible.

To address this general challenge, we introduce a fully unsupervised method, the Self-Assembling Manifold (SAM) algorithm. To demonstrate its utility and flexibility, we applied SAM to a difficult test dataset in which we sequenced ~370 stem cells isolated from *Schistosoma mansoni*, one of the most prevalent human parasites[10]. Testing several existing methods on this dataset, we found that they yielded poor low-dimensional embeddings that are inconsistent with known marker genes[11-13]. We reasoned that amplifying the distance between dissimilar cells may help resolve subtle differences between them. To achieve this goal, SAM iteratively rescales gene expressions and refines the nearest neighbor graph of cells. At each iteration, we assign more weight to genes that vary spatially across the current graph and feed them into the next assignment of neighbors until the graph topology converges to a stable solution.

Fig. 1a depicts the algorithm. SAM begins with a random k-nearest neighbor (kNN) graph and averages the expression of each cell with its k-nearest neighbors: $C = \frac{1}{k}NE$, where N is the

directed adjacency matrix and E is the gene expression matrix. For each gene, SAM computes a dispersion factor (**Methods**) of the averaged expressions C_i , which measures variation across neighborhoods of cells rather than individual cells. These dispersions are used to calculate the gene weights, which then rescale the gene expression matrix: $\hat{E} = EW_D$, where W_D is a diagonal matrix with gene weights along the diagonal. Using the rescaled expressions, we compute a pairwise cell distance matrix and update the assignment of each cell's k-nearest neighbors accordingly. This cycle is repeated until the distance matrix converges.

Applied to our schistosome stem cell dataset, SAM converges to a universal, stable solution independent of initial conditions and across a broad range of parameters (**Fig. 1b**, and **Supplementary Fig. 1**). **Fig. 1c** shows the iterative process, through which a kNN graph structure self-assembles. Sorting the weights by the final gene rankings demonstrates their convergence onto the final weight vector. Only a small fraction of genes (~1%) are significantly weighted and useful for separating cell clusters. These differences are too subtle to capture using other methods, which typically select a much larger percentage of features (**Fig. 2a**). As a negative control, we show that SAM cannot converge to a universal solution when applied to a randomly shuffled expression matrix (**Supplementary Fig. 2**).

Projecting the converged distance matrix onto two dimensions using t-distributed Stochastic Neighbor Embedding (t-SNE), we find that cells can be separated into four well-defined groups with orthologous gene expression patterns (**Fig. 2a**). In contrast, other commonly-used methods, including principal component analysis (PCA), Seurat[6], and SIMLR[7], fail to capture any structure. A high dimensional hierarchical representation of the final adjacency matrix is shown in **Fig. 2b**, with edge bundles connecting similar neighborhoods of cells arrayed on the periphery of the circle. Additionally, edge bundles connect subpopulations between the major clusters,

revealing additional layers of complexity that are difficult to capture in low-dimensional space.

The critical difference between SAM and other methods lies in how we assign gene weights. SAM prioritizes genes with variable expressions across neighborhoods of cells rather than individual cells as in other methods, which often use z-score-standardized dispersions to select genes for downstream analysis[6]. **Fig. 2c** reveals that many genes with high z-scores have low SAM weights, whereas high SAM weights generally correlate with high z-scores, indicating that SAM narrows the list of highly variable genes to those that are consistent with the long-range topological relationships between cells. Other methods (e.g., SC3[8]) identify marker genes based on differential gene expression between cell clusters, but this approach could suffer from poor cell cluster assignment, especially when discrete cell groups are difficult to separate or not present at all. Indeed, we observed little correlation between SC3 scores and SAM rankings for the schistosome dataset (**Fig. 2c**).

Fig. 2d and **Supplementary Table 1** list the rich panel of cell population-specific genes that SAM identifies. Furthermore, **Fig. 2e** highlights a surprising finding that suggests the current molecular definition of schistosome stem cells may also need revision. Expression of an RNA binding protein (*nanos-2*, Smp_051920), homologous fibroblast growth receptors (e.g., *fgfrB*, Smp_157300), and *eledh* (*eled*, Smp_041540) are thought to be the most important molecular signature for schistosome stem cells[11-13], but SAM reveals a novel stem cell population (arrowheads in **Fig. 2e**) that do not express any of these genes. Nevertheless, these cells still express argonaute2-1 (*ago2-1*, Smp_179320), *cyclin B* (Smp_082490) and other cell cycle regulators that are ubiquitous stem cells markers[11-13]. They also express another set of cluster-specific genes including a calcium binding protein (*cabp*, Smp_005350), an actin protein (Smp_161920), an annexin homolog (Smp_074140), a helix-loop-helix transcription factor

(*dhand*, Smp_062490), and a phosphatase (*dusp10*, Smp_034500) (**Fig. 2d, e**). Characterizing the function of these novel cell subpopulations and associated genes should become a major future direction of schistosome research. Taken together, these results demonstrate that SAM can uncover novel biology in a challenging dataset with only subtle differences between cells.

To assess the general applicability of SAM, we benchmark its performance against other common scRNA-seq analysis methods, including, Seurat[6], SIMLR[7], SC3[8], and PCA on six gold standard datasets that contain cells forming well separated clusters with high-confidence annotations[8]. For these relatively simple datasets, **Supplementary Fig. 3** shows that SAM is among the top performers.

We then compare SAM to Seurat on another 42 scRNA-seq datasets across a wide range of cell types (**Fig. 3**, **Supplementary Fig. 4-6**, **Supplementary Table 2**). Many of these datasets were previously analyzed with either manual selection of marker genes or extensive parameter optimization[2,5,9,14-18]. In contrast, SAM constructs manifolds consistent with the expressions of the identified marker genes and the underlying biological processes without any supervision or parameter changes between datasets.

Fig. 3a presents two examples to illustrate the strengths of SAM. In the natural killer T cells (NKTs) dataset[14], SAM produces tight clusters and places them in proper topological relations, whereas Seurat fails to clearly separate the cell clusters. SAM not only separates the annotated populations of the precursor (NKT0), and the mature NKT cells (NKT1, NKT2, and NKT17), but also resolves distinct expression patterns consistent with NKT assignments including *serpinb1a*, *xcl1*, *itm2a*, *il4*, and a novel pseudogene *gm15428* that is co-expressed with *il4* in the NKT2 subpopulation (**Supplementary Fig. 4a**). The proximity between NKT0 with NKT1 and NKT2 populations reflects the similarity in their transcriptional profiles, also consistent with

previous results[14]. SAM can also capture dynamic trajectories between cells. In the activated macrophage dataset[15], SAM reconstructs a circular trajectory that reflects the oscillatory nature of NF-kB activation in macrophages, with lymphoid activation gene *pbk*, chemokines such as *ccl3*, *ccl4*, and *ccl5*, and cholesterol/phospholipids transporter *abcg1* locally expressed in distinct regions around the circular projection (**Supplementary Fig. 4b**). Comparisons between SAM and Seurat manifolds on other datasets[2,16-19] highlight SAM's consistent, superior performance in discerning topological structure (**Fig. 3b**, **Supplementary Fig. 5, 6**).

To understand the conditions in which SAM may outperform other methods, we rank all analyzed datasets based on a network sensitivity measure, which quantifies changes in the cell-to-cell distances when randomly perturbing gene expression matrices (**Methods**). Datasets with higher sensitivity are more difficult to analyze, since changes in the selected features would strongly influence the resulting topological network. We use the network average clustering coefficient (NACC), which quantifies the degree of clustering[20], to compare graphs generated by different methods (**Methods**). Graphs characterized by regions of high density separated by regions of low density will have high NACC whereas random graphs with no structure will have low NACC[20]. We notice that Seurat produces graphs with lower clustering on datasets with high inherent sensitivity, with a Pearson correlation coefficient of $r^2 = -0.63$. In contrast, SAM's performance is insensitive to the dataset sensitivity ($r^2 = -0.09$). As a result, SAM consistently produces graphs with more structure across different types of datasets (**Fig. 3c**).

In summary, SAM improves analysis of scRNA-seq data compared to other state-of-the-art methods, supported by extensive benchmarking. While SAM performs well on simple datasets that contain well-separated cell clusters, SAM is particularly useful in analyzing datasets that contain cells in dynamic transitions or cell groups that are only distinguishable through subtle

differences in gene expression. Finally, as demonstrated by our work on the schistosome stem cells, SAM enables unsupervised analysis of scRNA-seq data from organisms with little to no *a priori* knowledge to gain novel biological insights.

Methods

Data processing

Supplementary Table 2 summarizes all datasets used in this study as well as the methods used to convert raw sequence read counts to gene expression, such as TPM (transcripts per million), CPM (counts per million), RPKM (reads per kilobase per million), or FPKM (fragments per kilobase per million). Datasets with asterisks next to their accession numbers in **Supplementary Table 2** are sourced from the *conquer* database[21]. Gene expression is measured in log space with a pseudocount of 1 (e.g., $\log_2$(TPM+1)). Genes expressed ($\log_2$(TPM+1)>1) in fewer than X = 2% or more than 100 – X = 98% of cells are excluded from downstream analysis as these genes lack statistical power. To reduce the influence of technical noise near the molecular detection limit, we set gene expression to zero when $\log_2$(TPM+1)<1. **Supplementary Fig. 1** shows that downstream analysis is robust to the data processing.

The SAM algorithm

SAM first generates a random kNN adjacency matrix and averages the expression of each cell with its k-nearest neighbors: $C = \frac{1}{k}NE$, where N is the directed adjacency matrix for the kNN graph, and E is the gene expression matrix. For each gene, SAM computes the Fano factor, $F_i = \frac{\sigma_{C_i}^2}{\mu_{C_i}}$, of the averaged expressions C_i. The Fano factor compares genes based on their variances relative to their average level of expression, which mitigates the inherent differences between gene expression distributions. Computing the Fano factors based on the kNN-averaged expressions links gene dispersion to the cellular topological structure. Genes that have highly variable expressions among individual cells but are homogeneously distributed across the topological representation should have small dispersions. We also notice that the kNN-averaging

approach reduces the effect of sequencing noise and dropout. k determines the topological length scale over which variations in gene expression are quantified. We set k by default to $\sqrt{n}$, where n is the total number of cells in the dataset and k is bounded to be at least 10. **Supplementary Fig. 1** reveals that the downstream analysis is robust to the specific choice of k. Additionally, the choice of k does not significantly affect runtime complexity or scalability.

SAM multiplies a vector of gene-specific weights calculated from the Fano factors into the original expression matrix: $\hat{E} = EW_D$, where $\hat{E}$ is the rescaled expression matrix and W_D is a diagonal matrix with $W_i = f\left(\sqrt{F_i}\right)$ along the diagonal, and f is a function that performs min-max normalization on the input vector. This matrix multiplication rescales the gene expression variances and gene-gene covariances according to their respective weights, reducing the influence of genes with low dispersions across neighborhoods. We then normalize the gene expression matrix to have unit Euclidean (L2) norm for each cell to prevent cells with large variances from dominating downstream analyses. To compute pairwise cell-cell distances, we perform PCA on $\hat{E}$. The rescaled expression matrix improves PCA robustness to variations in genes that contain little information (i.e. genes with low weights). Furthermore, this weighting strategy eliminates the typical requirement of selecting a subset of genes to feed into PCA, which often relies on arbitrary thresholds and heuristics.

Using the PC matrix (P), SAM computes a pairwise cell-cell Pearson correlation distance matrix. While typical dimension reduction approaches select a subset of the PCs, which is mostly subjective, we include all PCs and scale their variances by the corresponding normalized eigenvalues: $\hat{P} = \hat{\lambda}P$. PCs with small eigenvalues are weighted less in the distance calculation: $D_{P_iP_j} = 1 - \frac{\mathrm{Cov}(P_i,P_j)}{\sigma_{P_i}\sigma_{P_j}}$, where $D_{P_iP_j}$ is the Pearson correlation distance between PCs P_i and P_j,

$\mathrm{Cov}(P_i, P_j)$is the covariance, and σ_{P_i} is standard deviation of PC i. Using the distances to define the *k* nearest neighbors for each cell, SAM updates the kNN matrix and repeats the entire process. The algorithm continues until convergence, defined as when average correlation distance between cells' vectors of distances in adjacent iterations, $S_{j,j-1}$, is smaller than 10^{-4} or the maximum number of iterations has been reached (default 20). We define $S_{j,j-1} = \frac{1}{n}\sum D\{d_{i,j}, d_{i,j-1}\}$, where $D\{d_{ij}, d_{ik}\}$ is the Pearson correlation distance between the distances from cell *i* in distance matrices *j* and *j*-1.

Visualization

To visualize the topological structure embedded in the output cell-cell distance matrix, we fed the distances into sklearn's implementation of t-SNE using the 'precomputed' metric[22]. To directly visualize the corresponding kNN matrix (**Fig. 1c**), we used the Fruchterman-Reingold force-directed layout algorithm and drawing tools implemented by the Python package *graph-tool*[23]. The circular hierarchical graph in **Fig. 2b** was generated using the *graph_tool* package, which minimizes a nonparametric, stochastic block model to identify cell clusters and hierarchical relationships[23].

Benchmarking

To generate the convergence curves in the top panel of **Fig. 1b**, we computed the root mean square error (RMSE) between the distance matrices, kNN matrices, and weights in adjacent iterations, ensemble averaged across parallel runs. To quantify the solution stability of the SAM algorithm (bottom panel of **Fig. 1b**), we computed the RMSE for the distance matrices, kNN matrices, and gene weights at each iteration across replicates starting from different randomly generated initial graphs.

We evaluated the accuracy of each analysis method on six gold standard datasets (**Supplementary Fig. 3**) using the Adjusted Rand Index (ARI), which measures the accuracy of between two cluster assignments X and Y while accounting for randomness in the clustering:

$ARI = \frac{\Sigma\binom{n_{ij}}{2} - \left[\Sigma\binom{a_i}{2}\Sigma\binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\Sigma\binom{a_i}{2} + \Sigma\binom{b_j}{2}\right] - \left[\Sigma\binom{a_i}{2}\Sigma\binom{b_j}{2}\right]/\binom{n}{2}}$, where n is the number of cells, and n_{ij}, a_i, and b_j are elements from a contingency table that summarizes the overlap between the assignments X and Y[24]. n_{ij} denotes the number of cells assigned to X_i that are also assigned to Y_j while a_i and b_j are the sums of the *i*th row *j*th column of the contingency table, respectively. In assigning clusters, we used k-means clustering with the ground truth number of clusters set to be the unique number of labels found in the annotated datasets. SAM, SC3, and Seurat were run using default parameters.

The SIMLR package was implemented in R and run with the normalization parameter set to "True", which mean centers gene expression after normalizing them to be between 0 and 1. Seurat was implemented using the Scanpy package in Python[25]. For each gene, the expressions were mean-centered and variance normalized prior to PCA. We then performed Louvain clustering to assign clusters based on the PCA output and optimize performance by varying the number of included genes from 1% to 20% of the top Z-scoring genes with a 1% increment. ARI scores are reported based on the best performing parameter for each dataset.

To compare the quality of graphs generated by different methods, we use the NACC values to quantify the degree of structure in the computed graph topologies[20]. The NACC is the average of the local clustering coefficient for each node of a graph. The local clustering coefficient is defined as $a_i = \frac{L_i}{k_i(k_i-1)}$, where L_i is the number of edges between the k_i neighbors of node *i* and measures the degree of connectedness in a particular node's local neighborhood.

To produce the comparisons reported in **Fig. 3**, we used default parameters for SAM. For Seurat, we selected the variable genes according to their standardized dispersions using default thresholds[25] and chose the number of PCs which explain 50% of the variance (bounded between 6 and 30) for dimensionality reduction. From these PCs, we calculated a cell-cell correlation distance matrix. To keep the comparison between SAM and Seurat graphs consistent, this distance matrix was converted into a kNN matrix with the value of *k* used by SAM. The NACC was calculated for both graphs using *graph-tool*'s implementation.

To measure the inherent sensitivity of each dataset, we randomly perturbed the gene expression matrices by scaling the expressions with a gene weight vector drawn from a uniform random distribution. PCA was applied to the perturbed data, where the number of PCs was chosen such that they explain greater than 50% of the variance in the data (bounded between 6 and 30). A correlation distance matrix was calculated from the resulting PCs and perturbations were repeated to generate distance matrix replicates. Sensitivity is then defined as the average error across all pairwise comparisons between the replicates. The error between two distance matrices *j* and *k*, S_{jk}, is defined as the average correlation distance between cells' vector of distances and their corresponding pair from the replicate matrix: $S_{jk} = \frac{1}{n}\sum D\{d_{ij}, d_{ik}\}$ where $D\{d_{ij}, d_{ik}\}$ is the Pearson correlation distance between the distances from cell *i* in distance matrices *j* and *k*.

Parameter sweeps

To illustrate the robustness of SAM to its parameters, we ran SAM across a range of values for the number of nearest neighbors *k*, the gene filtering parameter *X*, and the minimum expression value. As mentioned before, *k* determines the number of nearest neighbors to find when computing the kNN matrix. The gene filtering parameter *X* controls the genes retained in

the initial filtering step, keeping genes expressed in greater than *X*% and less than (100-*X*)% of cells. A gene is considered expressed if its expression is greater than the minimum expression value, which is in units of $log_2(TPM + 1)$. We applied SAM with *k* varied from -10 to +10 of its default value, *X* from 0 to 10%, and the minimum expression value from 0 to 6 $log_2(TPM + 1)$ over 6 trials on 8 datasets, excluding our own, with high-confidence annotations. For each parameter and dataset, we computed the average pairwise distance matrix and adjacency matrix errors. To ensure that adjacency matrices built with different values of *k* are comparable (i.e. have an equivalent number of outgoing edges), we used the output distance matrices to recompute the final adjacency matrices such that they share the same default value of *k* ($\sqrt{n}$). Additionally, we calculated the average ARI score for each dataset across all parameter values.

The error between two distance matrices is defined as before: $S_{jk} = \frac{1}{n}\sum D\{d_{ij}, d_{ik}\}$. The error between adjacency matrices *j* and *k* is defined as the proportion of edges that are different between corresponding cells: $T_{jk} = \frac{1}{2nk}\sum|N_{ij} - N_{ik}|$, where *n* is the number of cells, *k* is the number of nearest neighbors, and N_{ij} is the nearest neighbors to cell *i* in adjacency matrix *j*. For the distance matrix and adjacency matrix error metrics, we computed error bars as the standard deviation of pairwise errors from 6 trials sweeping across values for each parameter. The ARI score error bars for each dataset and parameter are the standard deviations of the scores across the different parameter values.

scRNA-seq of schistosome stem cells

Schistosoma stem cells were isolated from juvenile parasites retrieved from infected mice at 2.5 weeks post infection. At this stage, juvenile parasites undergo a massive wave of growth and develop germline primordia *de novo*. Our previous study shows that >15% of the total number of

body cells in juvenile parasites are stem cells[13]. We followed the protocol as previously described[13]. Briefly, we retrieved juvenile parasites from schistosome-infected mice (Swiss Webster NR-21963) by hepatic portal vein perfusion using 37 ℃ DMEM. Parasites were cultured at 37 ℃/5% CO_2 in Basch Medium 169 supplemented with 1X Antibiotic-Antimycotic for 24-48 hr to allow complete digestions of host blood cell in parasite intestines. Before dissociation, parasites were permeabilized in PBS containing 0.1% Triton X-100 and 0.1% NP-40 for 30 seconds, and washed thoroughly to remove the surfactants. The permeabilized parasites were dissociated in 0.25% trypsin for 20 min, and triturated with serially narrowed flamed-tip glass. Cell suspensions were passed through a 100 μm nylon mesh (Falcon Cell Strainer) and centrifuged at 150 g for 5 min. Cell pellets were gently resuspended, passed through a 30 μm nylon mesh, and stained with Vybrant DyeCycle Violet (DCV; 5 μM, Invitrogen), and TOTO-3 (0.2 μM, Invitrogen) for 30–45 min. As the stem cells comprise the only proliferative population in schistosomes, we flow-sorted cells at G_2/M phase of the cell cycle on a SONY SH800 cell sorter. Dead cells were excluded based on TOTO-3 fluorescence. Single stem cells were gated using forward scattering (FSC), side scattering (SSC), and DCV to isolate cells with doubled DNA content compared to the rest of the population. Cells that passed these gates were sorted into 384-well lysis plates containing Triton X-100, ERCC standards, oligo-dT, dNTP, and RNase inhibitor.

cDNA was reverse transcribed and amplified on 384-well plate following the Smart-Seq2 protocol[26]. For quality control, we quantified cDNA concentration using picogreen and histone *h2a* (Smp_086860) levels using qPCR, as *h2a* is a ubiquitously expressed in all schistosomes stem cell[11-13]. We picked 370 wells that had more than 0.4 ng/μL of total cDNA concentration and generated C_T values within 2.5 C_T around the most probable values (~45% of total wells)

(**Supplementary Fig. 8a-b**). cDNA was then diluted to 0.4 ng/µL for library preparation. Tagmentation and barcoding of wells were prepared using Nextera XT DNA library preparation kit. Library fragments concentration and purity were quantified by Agilent bioanalyzer and qPCR. Sequencing was performed on a NextSeq 500 using V2 150 cycles high-output kit at ~1 million reads depth per cell. Raw sequencing reads were demultiplexed and converted to fastq files using bcl2fastq. Paired-end reads were mapped to *S. mansoni* genome version WBPS9 (WormBase Parasite) using STAR. 338 cells with more than 1700 transcripts expressed at >2 TPM were used for downstream analysis (**Supplementary Fig. 8c-d**).

Acknowledgements. *S. mansoni* (strain: NMRI) was provided by the NIAID Schistosomiasis Resource Center for distribution through BEI Resources, NIH-NIAID Contract HHSN272201000005I. We thank F. Zanini, P. Li, N. Neff, and J. Okamoto for experimental help, J. Qin, F. Horns, G. Stanley, and M. Chen for conceptual input and stimulating discussions. B.W. is supported by the Burroughs Wellcome Fund through the CASI program and a Beckman Young Investigator Award.

Author contributions. A.J.T. and B.W. designed the research, A.J.T. developed the algorithm, A.J.T. and Y.X. performed the analysis, Y.X. and B.W. performed the experiments, A.J.T., Y.X. and B.W. wrote the paper and all authors commented on the manuscript, S.R.Q. and B.W. supervised the project and provided conceptual advice.

Competing Interests. The authors declare no competing financial interests.

References

1. Mohammed, H. *et al*. Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell Rep.* **20,** 1215-1228 (2017).
2. Treutlein, B. *et al*. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509,** 371-375 (2014).
3. Lönnberg, T. *et al*. Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves Th1/Tfh fate bifurcation in malaria. *Sci. Immunol.* **2,** eaal2192 (2017).
4. Segerstolpe, Å. *et al*. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* **24,** 593-607 (2016).
5. Schwalie, P.C. *et al*. A stromal cell population that inhibits adipogenesis in mammalian fat depots. *Nature*, in press.
6. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33,** 495–502 (2015).
7. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., & Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* **14,** 414-416 (2017).
8. Kiselev, V.Y. *et al*. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14,** 483-486 (2017).
9. Olsson, A. *et al*. Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature* **537,** 698–702 (2016).
10. Hoffmann, K.F., Brindley, P.J., & Berriman, M. Halting harmful helminths. *Science* **346,** 168-169 (2014).
11. Collins, J.J. *et al*. Adult somatic stem cells in the human parasite, Schistosoma mansoni. *Nature* **494,** 476-479 (2013).
12. Wang, B., Collins, J.J., & Newmark, P.A. Functional genomic characterization of neoblast-like stem cells in larval Schistosoma mansoni. *eLife* **2,** e00768 (2013).
13. Wang, B. *et al*. Stem cell heterogeneity drives the parasitic life cycle of Schistosoma mansoni. *eLife*, in press.
14. Engel, I. *et al*. Innate-like functions of natural killer T cell subsets result from highly divergent gene programs. *Nat. Immunol.* **17,** 728-739 (2016).
15. Lane, K. *et al*. Measuring Signaling and RNA-Seq in the Same Cell Links Gene Expression to Dynamic Patterns of NF-κB Activation. *Cell Syst.* **4,** 458-469 (2017).
16. Trapnell, C. *et al*. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32,** 381-386 (2014).
17. Mathys, H. *et al*. Temporal tracking of microglia activation in neurodegeneration at single-cell resolution. *Cell Rep.* **21,** 366-380 (2017).
18. Mi, D. *et al*. Early emergence of cortical interneuron diversity in the mouse embryo. *Science* **360,** 81-85 (2018).

19. Zanini, F., Pu S.Y., Bekerman E., Einav S. & Quake, S.R. Single-cell transcriptional dynamics of flavivirus infection. *eLife* **7,** e32942 (2018).

20. Watts, D.J. & Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature* **393,** 440-442 (1998).

Figure legends

Fig. 1. The SAM algorithm. (**a**) SAM starts with a randomly initialized kNN matrix and iterates to refine the kNN matrix and weight vector until convergence. (**b**) Normalized root mean square error (RMSE) between adjacent iterations within a single run (top) and between multiple runs at the same iteration (bottom) to show that SAM converges to a universal, stable solution regardless of initial conditions. (**c**) Graph structures and weights converging to the final output over the course of 15 iterations (i denotes iteration number). Top: nodes are cells and edges connect neighbors. Nodes are color-coded according to the final clusters. Bottom: weights are sorted according to the final gene rankings.

Fig. 2. SAM identifies subpopulations within schistosome stem cells. (**a**) t-SNE projections of schistosome stem cells comparing SAM, Seurat, PCA, and SIMLR. Cells are color-coded based on stem cell subpopulation assignments. (**b**) Hierarchical connectivity graph between cells based on the final kNN matrix. Edge bundles connecting cells arrayed along the periphery indicate the similarity between cells. (**c**) Normalized z-score (left) and normalized SC3 AUROC, which measures relative significance of differential gene expressions (right), plotted vs. the SAM weights, with linear fits and correlation coefficients shown. The top 30 genes specific to each subpopulation are colored according to the color scheme used in (a) and (b). (**d**) Heatmap of average gene expression in the four assigned clusters, with each gene's expression normalized by its maximum value. (**e**) t-SNE projection heatmaps to show a subpopulation (arrowheads) that express none of the canonical schistosome stem cell markers, *nanos-2*, *fgfrB*, or *eled*, but express cell-cycle related genes (*ago2-1*, *cyclin B*) and another panel of genes that are specific to this population (bottom row).

Fig. 3. SAM improves analyses of a wide range of single-cell datasets. (**a**) t-SNE projections of two example datasets. Top: natural killer T-cells (NKTs), with subtypes specified by colors; bottom: macrophages activated *in vitro*. Examples of highly ranked gene expression patterns are overlaid. (**b**) Five example datasets that exemplify the superior performance of SAM over Seurat in reconstructing cellular manifolds with no supervision. The corresponding NACCs are shown in the upper-left corners. (**c**) Comparison of SAM and Seurat performance over 48 datasets, measured by the ratio of NACCs between graphs produced by SAM and Seurat. Inset: NACC of graphs produced by Seurat shows a negative correlation with dataset network sensitivity, whereas SAM is robust to sensitivity.

Fig. 1.

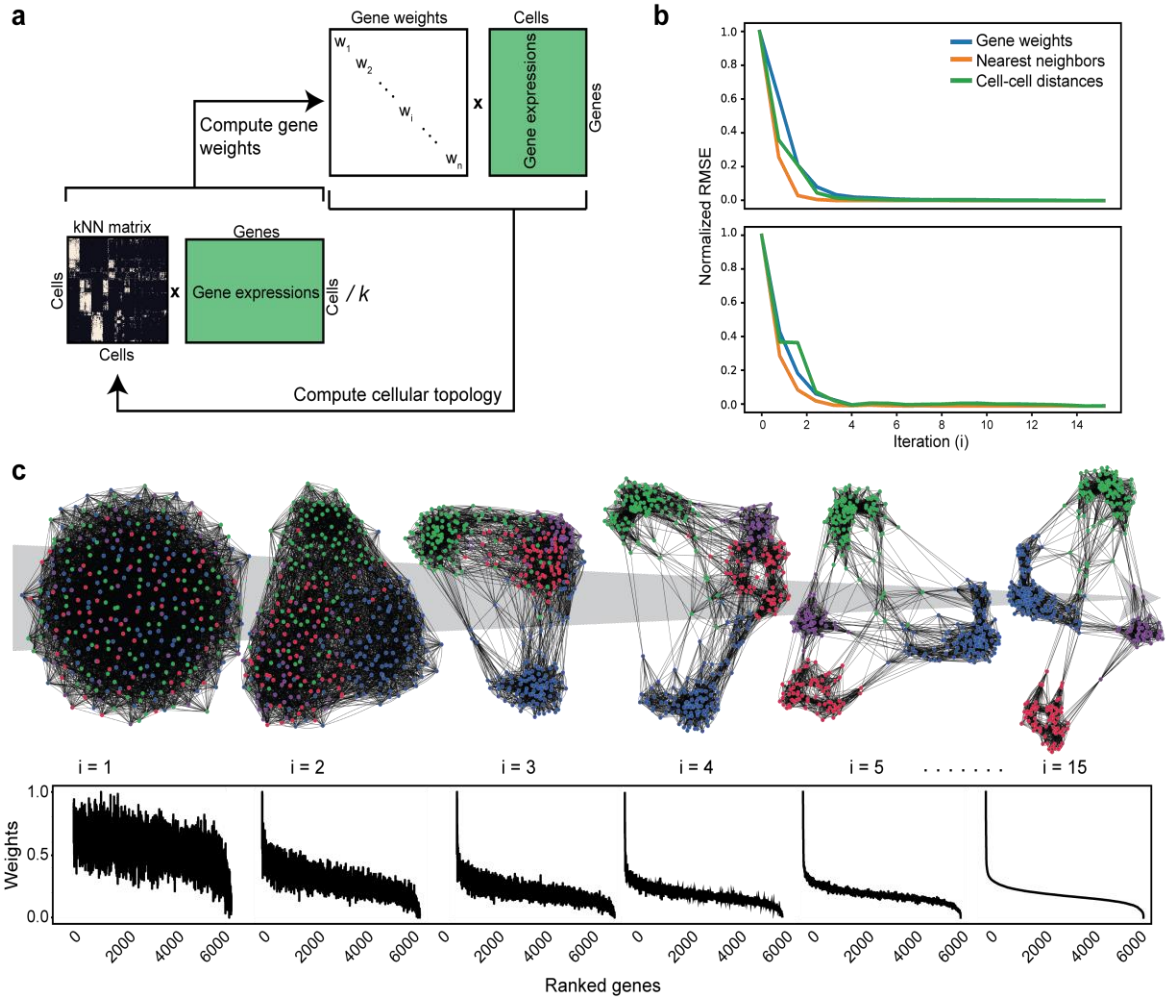


Fig. 2.

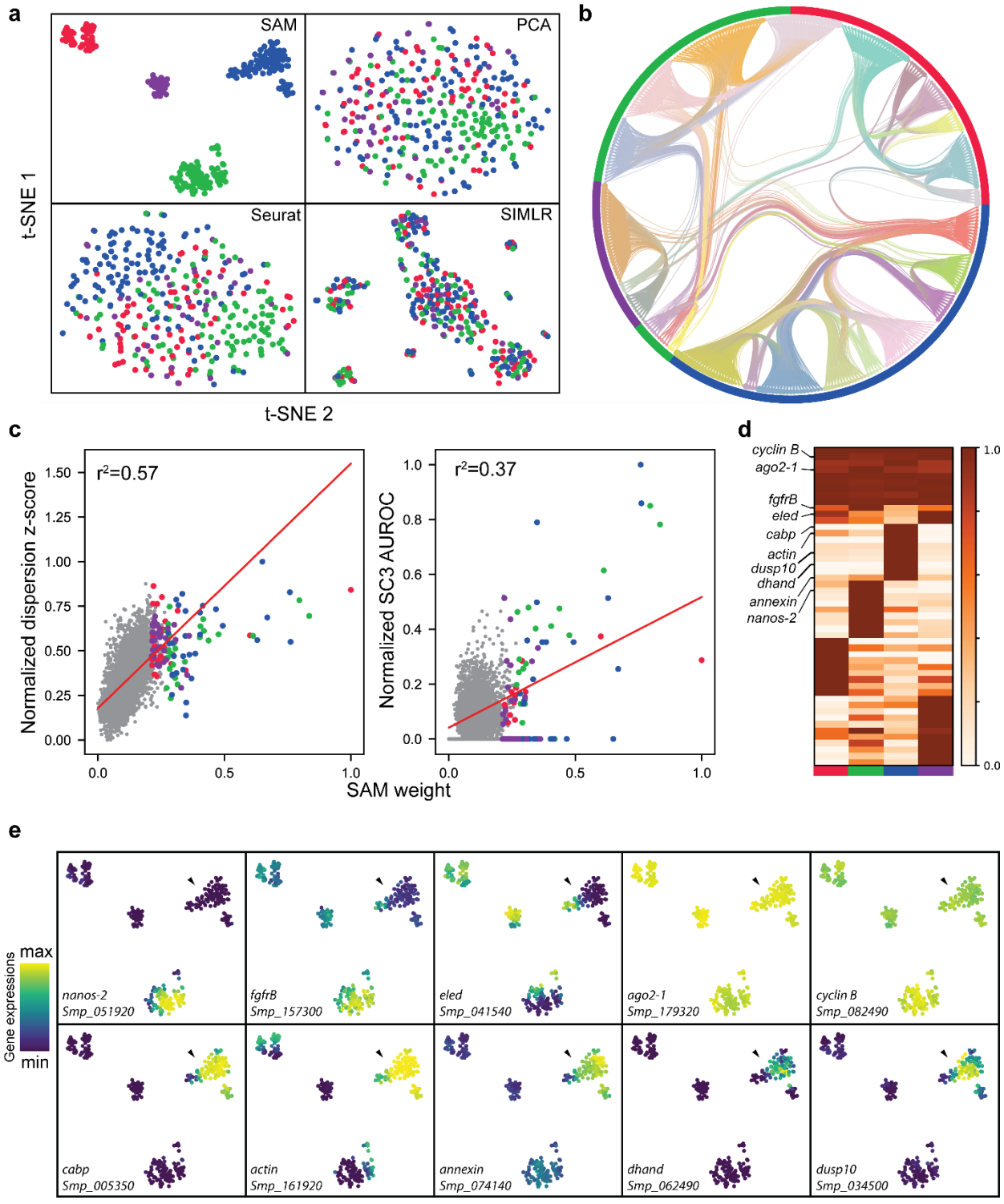
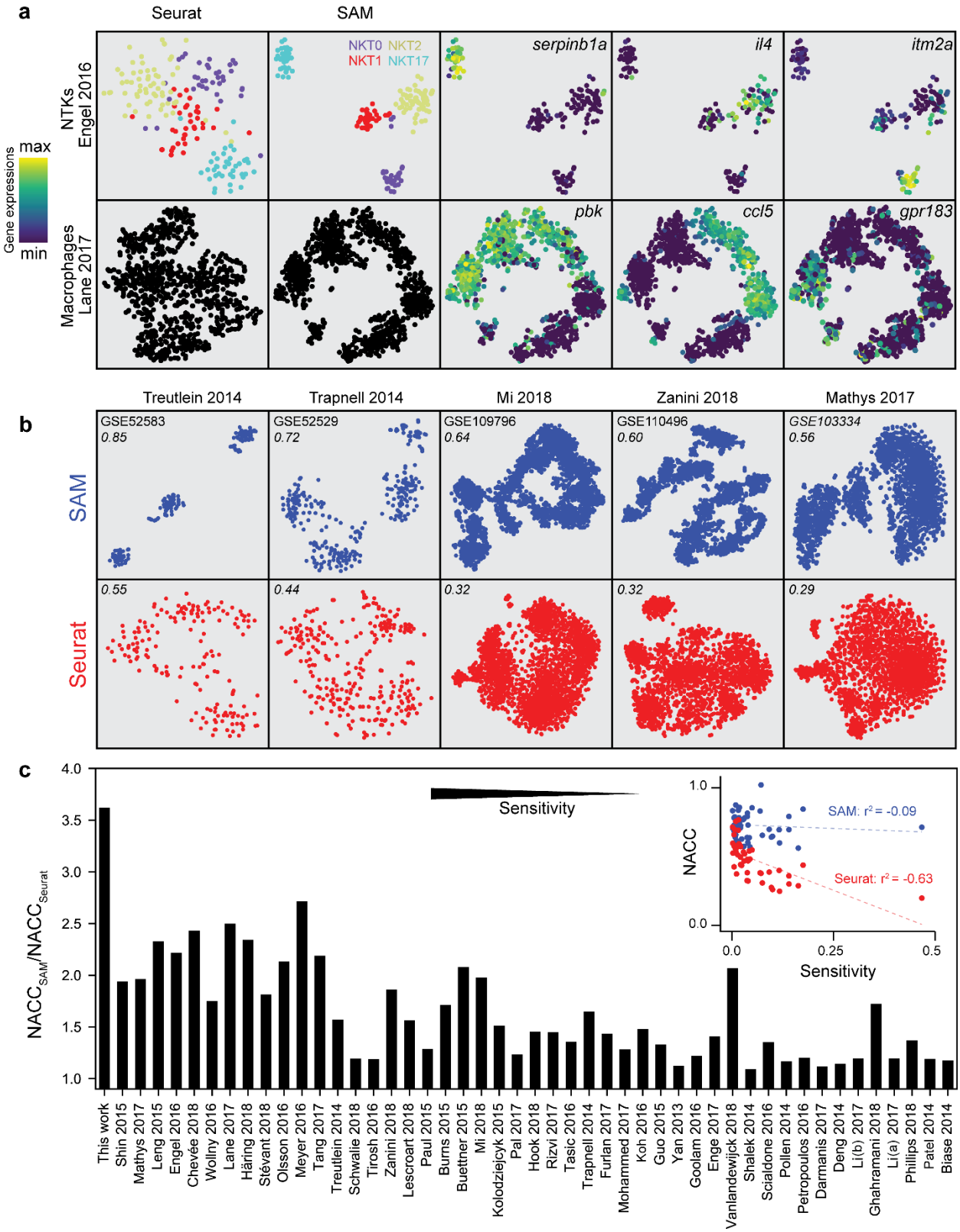


Fig. 3.

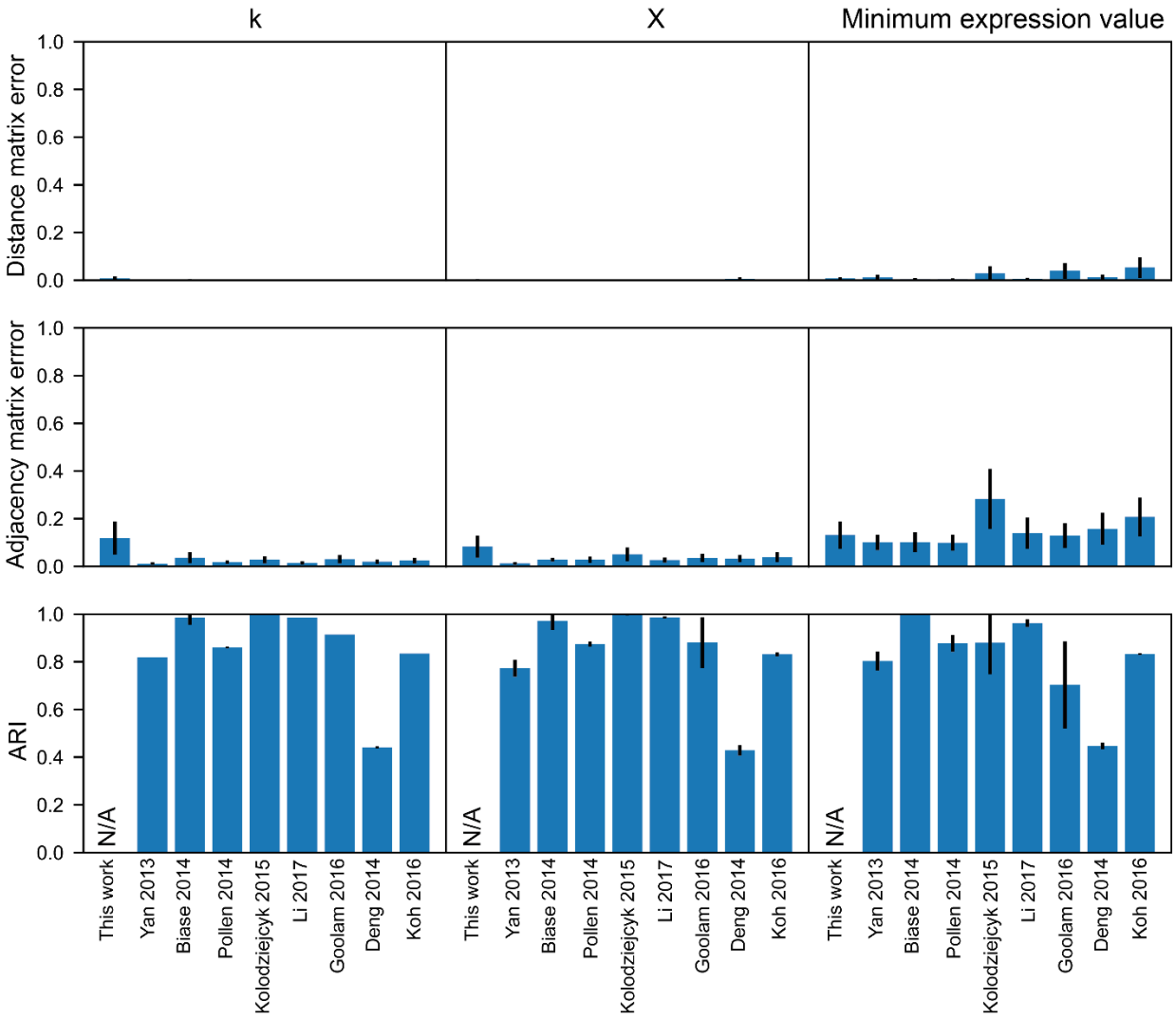


Supplementary Information

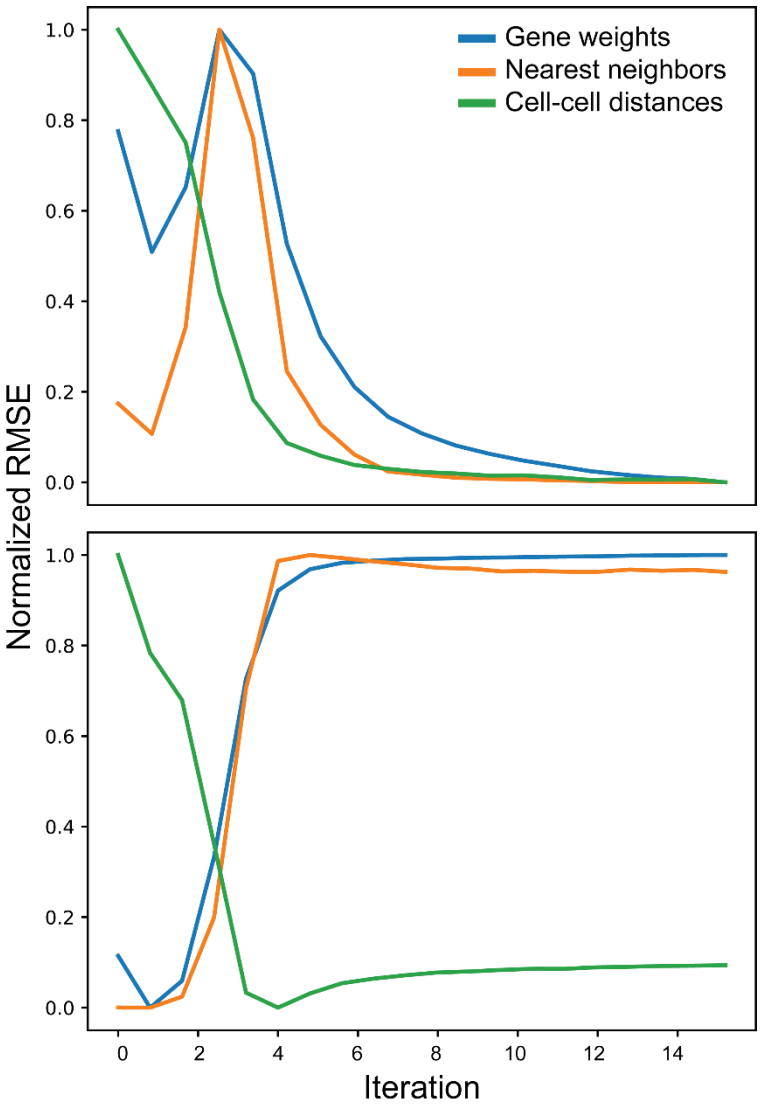
Supplementary Figure 1-7.

Supplementary Table 1-2.

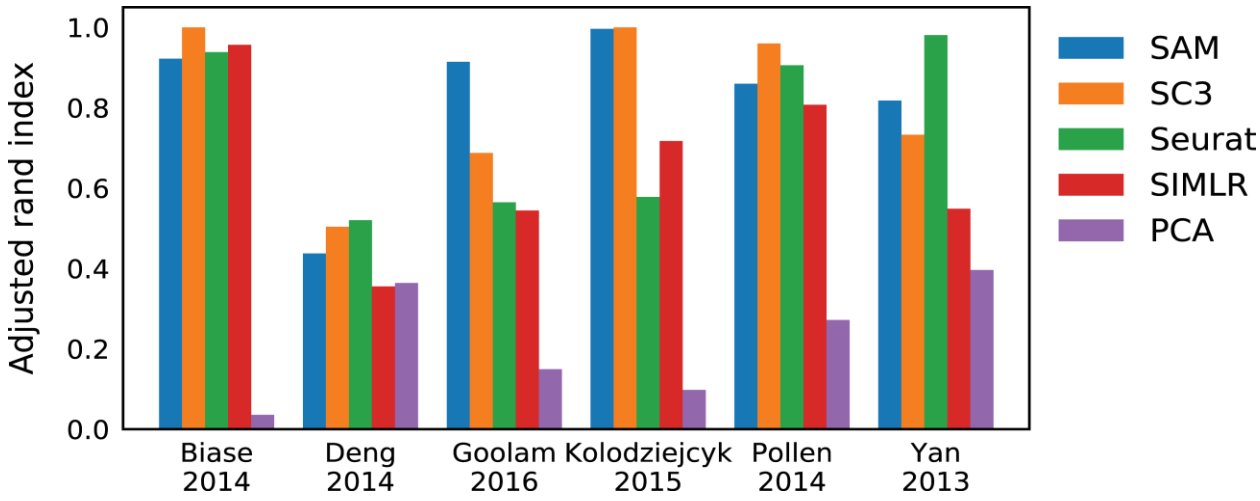
Supplementary References 21-62.



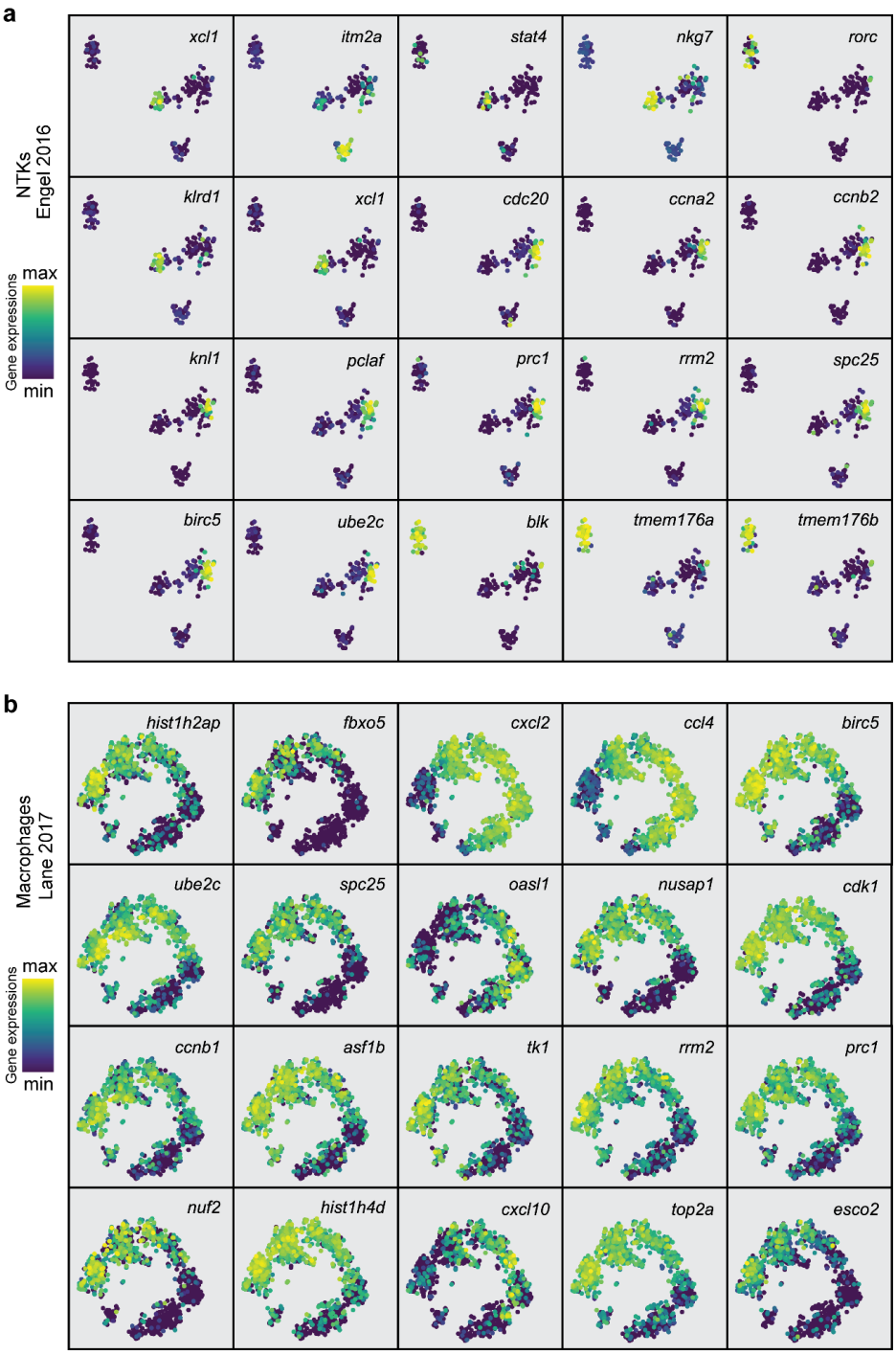
Supplementary Fig. 1: The parameter sensitivity of SAM. Top to bottom: average distance matrix errors, kNN matrix errors, and Adjusted Rand Scores (ARI), when varying the number of nearest neighbors (k, left) and the gene filtering parameters (X, middle, and minimum expression value, right) with error bars shown (**Methods**). The distance and adjacency matrices are highly robust to *k* and *X*. ARI is only available for datasets with high-confidence annotations.



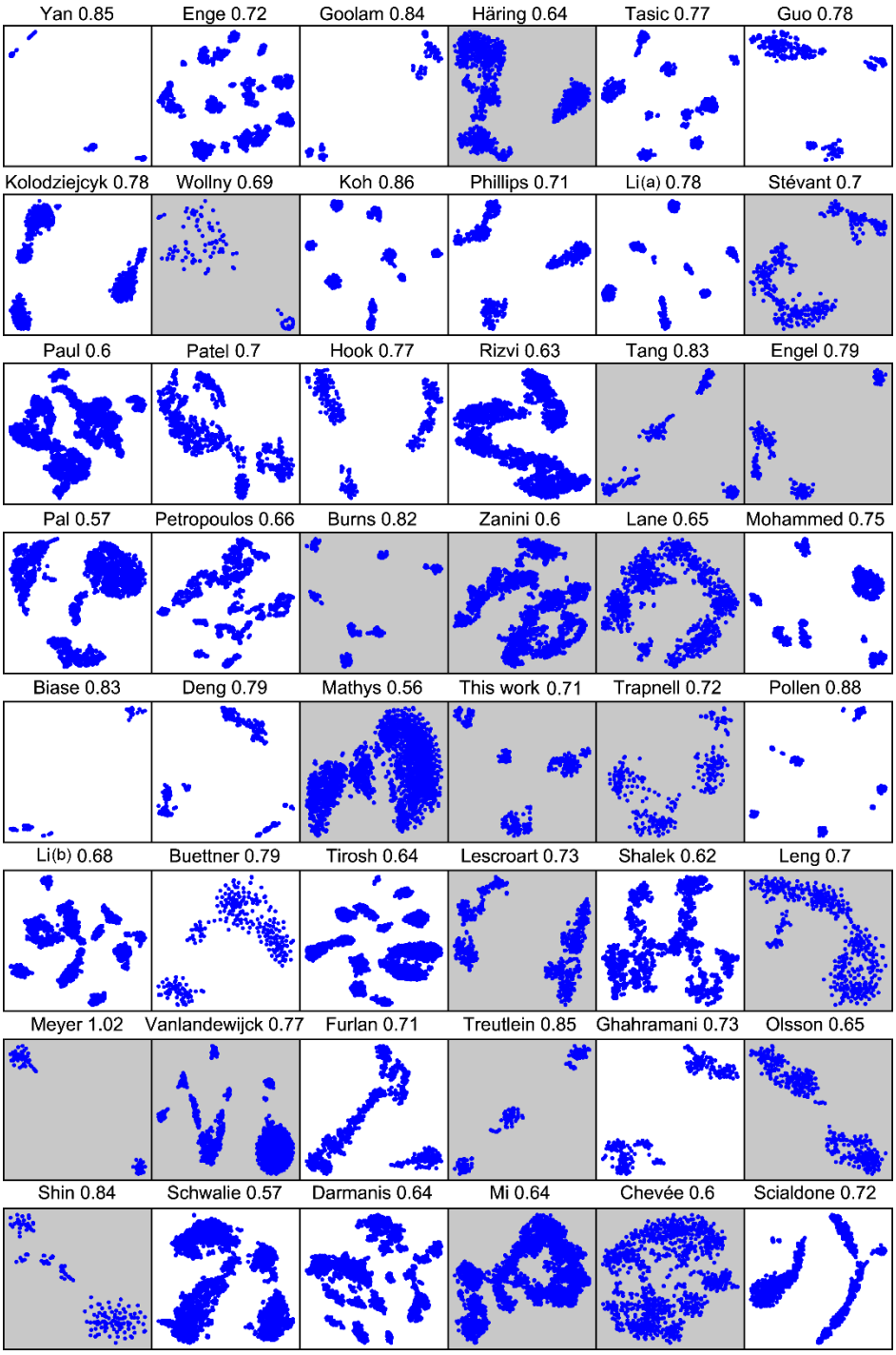
Supplementary Fig. 2: SAM does not converge to a universal solution when applied to a randomly shuffled dataset. Normalized RMSE between adjacent iterations within a single run (top) and between multiple runs at the same iteration (bottom) to show that while SAM converges to a solution within a single run, it does not converge to the same solution between runs.



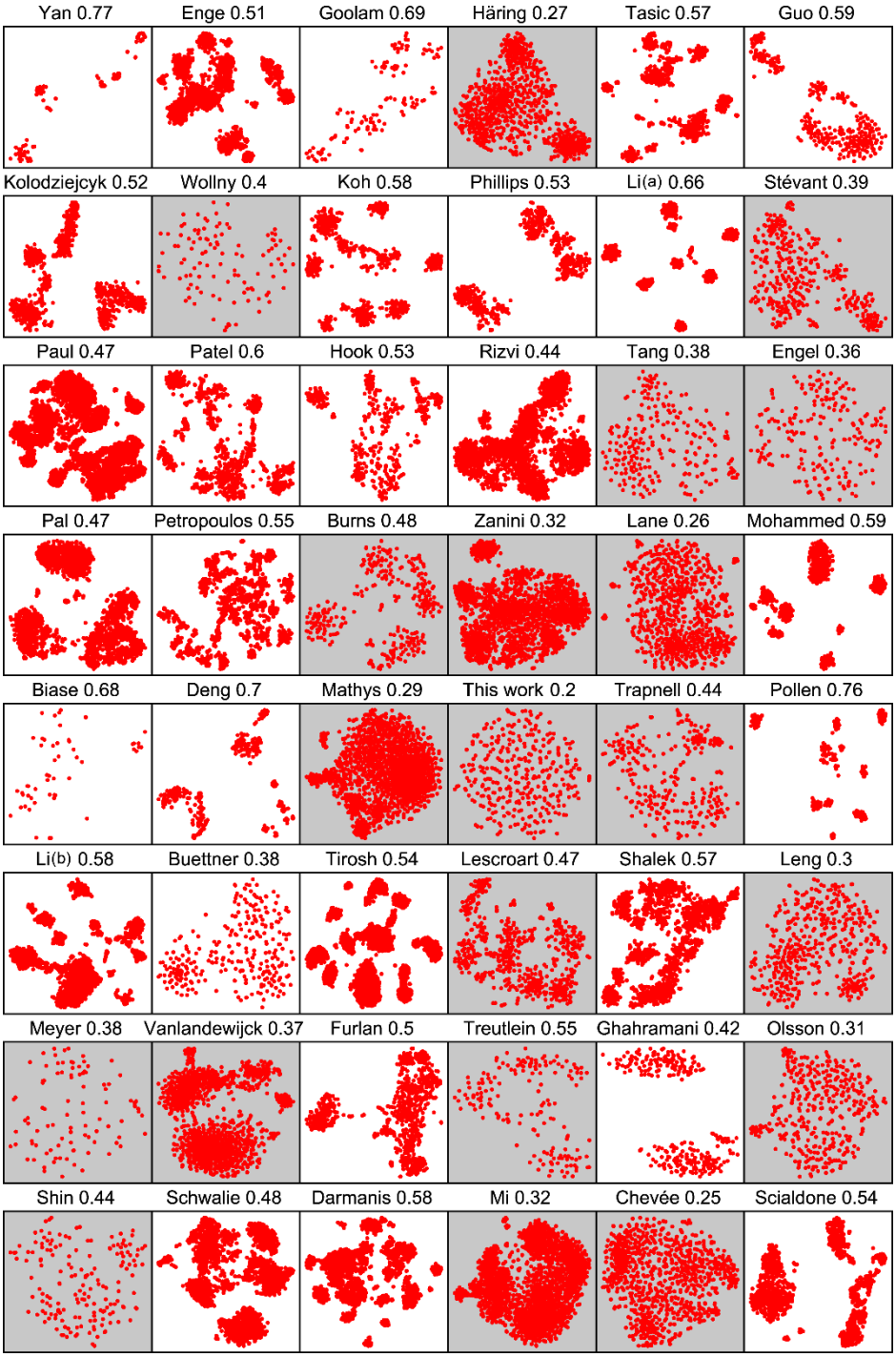
Supplementary Fig. 3: SAM performs well in benchmarks against other methods on gold standard datasets. Evaluation of clustering performance using different dimensionality reduction approaches. For Seurat and SIMLR, we use built-in clustering algorithms. For SAM, SC3, and PCA, we use k-means clustering with the same k across all methods defined as the number of known cell types from the annotations provided by the original studies.



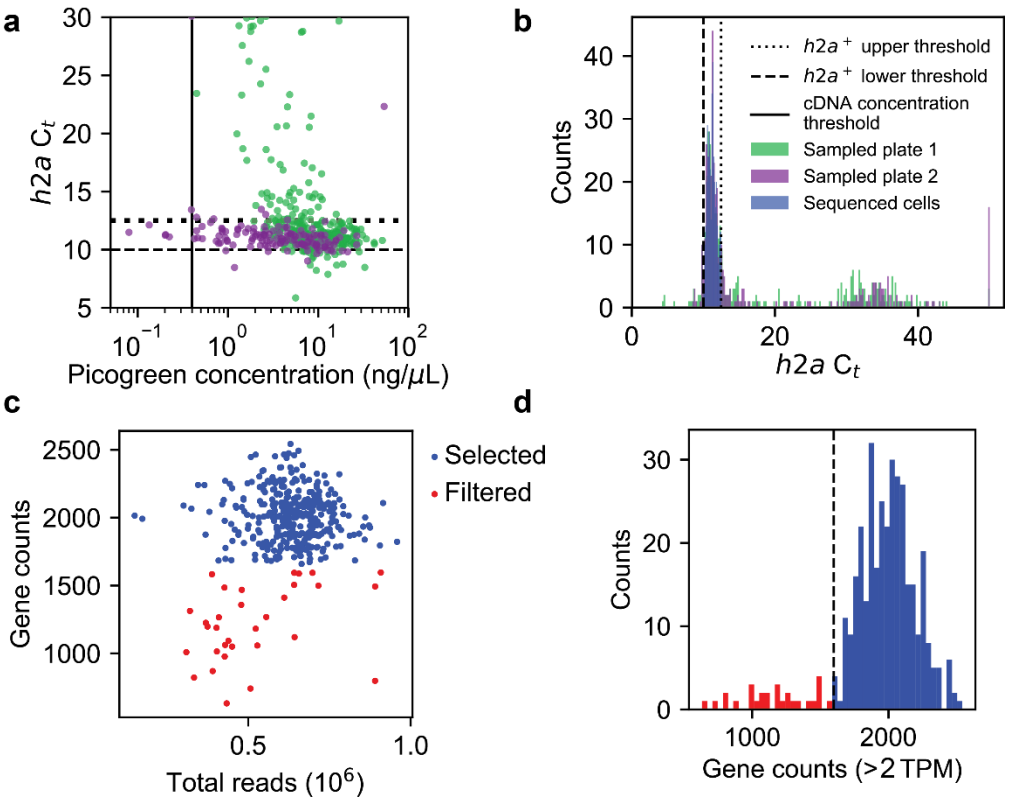
Supplementary Fig. 4: Expression patterns of highly ranked genes for NTKs and activated macrophages datasets. t-SNE projection of SAM output on (**a**) NTKs dataset (Engel 2016)[14], and (**b**) activated macrophages dataset (Lane 2017)[15] colorcoded by the expression levels of top ranked genes.



Supplementary Fig. 5: SAM t-SNE projections for 48 datasets. t-SNE projections are shown for all 48 datasets analyzed. Gray boxes highlight datasets in which SAM produces manifolds with greater topological structure than those produced by Seurat (**Supplementary Fig. 6**). Numbers indicate NACC values.



Supplementary Fig. 6: Seurat t-SNE projections for 48 datasets. t-SNE projections are shown for all 48 datasets analyzed. Gray boxes highlight datasets in which Seurat produces manifolds with less topological structure than those produced by SAM (**Supplementary Fig. 5**). Numbers indicate NACC values.



Supplementary Fig. 7: Quality control of library preparation and sequencing. (**a**) qPCR quantification of histone *h2a* (Smp_086869) expression and picogreen measurement of total cDNA after reverse-transcription and PCR amplification. (**b**) Histogram of *h2a* qPCR measurement. In (**a**) and (**b**), green and purple separate two batches (plate 1 and 2), respectively. Blue colored bars represent cells selected for downstream library preparation. (**c**) Detected gene counts and total reads of individual sequenced cells. (**d**) Histogram of detected gene counts. Cells with fewer than 1700 gene detected are filtered (red) and the remaining cells are kept for downstream analysis (blue).

Supplementary Table legends:

Supplementary Table 1: Schistosome stem cell population-specific genes. Gene IDs, protein product information, expressing clusters, and SAM ranks are provided. Rows highlighted in light green indicate genes that are shown in **Fig. 2e** and labeled in **Fig. 2d**. The colors highlighting the cluster numbers in the provided legend correspond to the colors of the clusters in **Fig. 2a**.

Supplementary Table 2: A list of all datasets used in this study. Accession numbers, normalization methods, and corresponding reference numbers are provided. Accession numbers with asterisks indicate datasets that are sourced from the *conquer* database[21].

Supplementary References

21. Soneson, C. & Robinson, D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15,** 255-261 (2018).
22. Pedregosa et al. Scikit-learn: Machine Learning in Python, *JMLR* **12,** 2825-2830 (2011).
23. Peixoto T.P. The graph-tool python library, DOI: 10.6084/m9.figshare.1164194 (2014).
24. Rand, W.M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66,** 846–850 (1971).
25. Wolf, F.A., Angerer P., & Theis F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19,** 15 (2018).
26. Picelli, S. *et al*. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10,** 1096-1098 (2013).
27. Yan, L. *et al*. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* **20,** 1131-1139 (2013).
28. Goolam, M. *et al*. Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. *Cell* **165,** 61-74 (2016).
29. Tasic, B. *et al*. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19,** 335-346 (2016).
30. Guo, F. *et al*. The transcriptome and DNA methylome landscapes of human primordial germ cells. *Cell* **161,** 1437-1452 (2015).
31. Kolodziejczyk, A.A. *et al*. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* **17,** 471-485 (2015).
32. Wollny, D. *et al*. Single-cell analysis uncovers clonal acinar cell heterogeneity in the adult pancreas. *Dev. Cell* **39,** 289-301 (2016).
33. Koh, P.W. *et al*. An atlas of transcriptional, chromatin accessibility, and surface marker changes in human mesoderm development. *Sci. Data* **3,** 160109 (2016).
34. Deng, Q. *et al*. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343,** 193-196 (2014).
35. Patel, A.P. *et al*. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344,** 1396-1401 (2014).
36. Rizvi, A.H. *et al*. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nat. Biotechnol.* **35,** 551-560 (2017).
37. Tang, Q. *et al*. Dissecting hematopoietic and renal cell heterogeneity in adult zebrafish at single-cell resolution using RNA sequencing. *J. Exp. Med.* **214,** 2875-2887 (2017).
38. Petropoulos, S. *et al*. Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **165,** 1012-1026 (2016).
39. Burns, J.C. *et al*. Single-cell RNA-Seq resolves cellular complexity in sensory organs from the neonatal inner ear. *Nat. Commun.* **6,** 8557 (2015).
40. Biase, F.H. *et al*. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Res.* **24,** 1787-1796 (2014).

41. Pollen, A.A. *et al*. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32,** 1053-1058 (2014).

42. Li, L. *et al*. Single-cell RNA-Seq analysis maps development of human germline cells and gonadal niche interactions. *Cell Stem Cell* **20,** 858-873 (2017).

43. 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).

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

45. Leng, N. *et al*. Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat. Methods* **12,** 947-950 (2015).

46. Meyer, S.E. *et al*. DNMT3A haploinsufficiency transforms FLT3ITD myeloproliferative disease into a rapid, spontaneous, and fully penetrant acute myeloid leukemia. *Cancer Discov*. **6,** 501-515 (2016).

47. Shin, J. *et al*. Single-cell RNA-seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell* **17,** 360-372 (2015).

48. Darmanis, S. *et al*. Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. *Cell Rep.* **21,** 1399-1410 (2017).

49. Scialdone, A. *et al*. Resolving early mesoderm diversification through single-cell expression profiling. *Nature* **535,** 289-293 (2016).

50. Enge, M. *et al*. Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell* **171,** 321-330 (2017).

51. Stévant, I. *et al*. Deciphering cell lineage specification during male sex determination with single-cell RNA sequencing. *Cell Rep.* **22,** 1589-1599 (2018).

52. Phillips M.J. *et al*. A Novel Approach to Single Cell RNA-Sequence Analysis Facilitates In Silico Gene Reporting of Human Pluripotent Stem Cell-Derived Retinal Cell Types. *Stem Cells* **36,** 313-324 (2018).

53. Vanlandewijck M. *et al*. A molecular atlas of cell types and zonation in the brain vasculature. *Nature* **554,** 475-480 (2018).

54. Furlan A. *et al*. Multipotent peripheral glial cells generate neuroendocrine cells of the adrenal medulla. *Science* **357,** eaal3753 (2017).

55. Ghahramani A. *et al*. Epidermal Wnt signalling regulates transcriptome heterogeneity and proliferative fate in neighbouring cells. *Genome Biol.* **19,** 3 (2018).

56. Lescroart F. *et al*. Defining the earliest step of cardiovascular lineage segregation by single-cell RNA-seq. *Science* **359,** 1177-1181 (2018).

57. Chevée, M. *et al*. Variation in Activity State, Axonal Projection, and Position Define the Transcriptional Identity of Individual Neocortical Projection Neurons. *Cell Rep.* **22,** 441-455 (2018).

58. Hook, P.W. *et al*. Single-Cell RNA-Seq of Mouse Dopaminergic Neurons Informs

Candidate Gene Selection for Sporadic Parkinson Disease. *Am. J. Hum. Genet.* **102,** 427-446 (2018).

59. Li H. *et al*. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **49,** 708-718 (2017).
60. Tirosh I. *et al*. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539,** 309-313 (2016).
61. Paul F. *et al*. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163,** 1663-77 (2015).
62. Pal B. et al. Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling. *Nat. Commun.* **8,** 1627 (2017).