Self-assembling Manifolds in Single-cell RNA Sequencing Data

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

[1]Department of Bioengineering, [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.

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

Abstract

Single-cell RNA sequencing has spurred the development of computational methods that enable researchers to classify cell types, delineate developmental trajectories, and measure molecular responses to external perturbations. Many of these technologies rely on their ability to detect genes whose cell-to-cell variations arise from the biological processes of interest rather than transcriptional or technical noise. However, for datasets in which the biologically relevant differences between cells are subtle, identifying these genes is a challenging task. We present the self-assembling manifold (SAM) algorithm, an iterative soft feature selection strategy to quantify gene relevance and improve dimensionality reduction. We demonstrate its advantages over other state-of-the-art methods with experimental validation in identifying novel stem cell populations of *Schistosoma*, a prevalent parasite that infects hundreds of millions of people. Extending our analysis to a total of 56 datasets, we show that SAM is generalizable and consistently outperforms other methods in a variety of biological and quantitative benchmarks.

Introduction

Single-cell RNA sequencing (scRNAseq) datasets typically contain tens of thousands of genes, although many of them may not be informative for differentiating between cell types or states. Feature selection is thus commonly used to select a subset of genes prior to downstream analyses, such as manifold reconstruction and cell clustering (Crow et al., 2018; Satija et al., 2015; Vallejos et al., 2015). However, current approaches have two major limitations.

First, feature selection methods filter genes based on arbitrarily or empirically chosen thresholds, small changes in which may result in different gene sets (Vallejos et al., 2017). In addition, the selection of features typically operates under the assumption that genes with highly variable expression between individual cells capture biologically meaningful variation. Because single-cell transcriptomes are inevitably contaminated by a combination of random transcriptional and technical noise (Grün et al., 2014), the variation in biologically relevant genes may be hard to distinguish from the background noise, especially when the differences between cell populations are subtle. Resolving these differences, or "signals", is essential to study a variety of biological problems, including identifying cell subtypes (Olsson et al., 2016; Treutlein et al., 2014; Lönnberg et al., 2017; Fincher et al., 2018; Baron et al., 2016) and quantifying the effects of molecular perturbations to otherwise homogeneous populations of cells (Lane et al., 2017). In such datasets, only a small fraction of the genes, and therefore only a small fraction of the total variation, may contain the signals relevant for distinguishing cell types. Choosing these features ahead of time without *a priori* knowledge remains an unmet computational challenge.

The second limitation is that existing methods have been almost exclusively benchmarked on well-annotated, gold standard datasets with clearly distinguishable cell types (Wang et al., 2017; Kiselev et al., 2017; Duò et al., 2019; Bahlo et al., 2018). These datasets are not informative for distinguishing the performance between methods, because the differences between cell types are relatively straightforward to detect. However, evaluating the performance of feature selection and/or dimensionality reduction methods on datasets with more subtle signals is difficult as their ground truth labels are typically ambiguous or nonexistent.

To overcome the shortcomings of current feature selection approaches, here, we introduce the Self-Assembling Manifold (SAM) method, an unsupervised, "soft feature selection" algorithm that iteratively rescales gene expressions to refine a nearest neighbor graph of cells until the graph converges to a stable solution. At each iteration, SAM assigns more weight to genes that are spatially variable across the constructed graph, and this weighted gene expression is then used to improve the next nearest neighbor assignment. SAM presents two advantages: it rescales all genes according to their weights, solving the problem of thresholding, and it prioritizes genes that are variable across the intrinsic manifold of the data rather than selecting genes that are variable across individual cells.

Second, in order to better distinguish the performance between methods, we define a network sensitivity measure to identify datasets with subtle signals. With limited annotations in most high-sensitivity datasets, we introduce unsupervised graph-based metrics to quantify the degree of structure within the reconstructed manifolds for comparison between methods. In addition, we perform benchmarking using known ground truth labels on simulated datasets spanning a wide range of sensitivities by introducing increasing levels of noise to well-annotated datasets. These analyses reveal that SAM consistently improves feature selection and cell clustering.

To demonstrate the utility of SAM in practice, we provide an in-depth analysis of two datasets that are challenging to analyze using existing methods: stem cells in a human parasitic worm, *Schistosoma*, and activated macrophages (Lane et al., 2017). We show that SAM can capture novel biology undetectable by other methods and validate these results with experimental evidence.

Results

The SAM algorithm

The SAM algorithm 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 (**Figure 1a**). For each gene i, SAM computes a spatial dispersion factor of the averaged expressions C_i, which measures variation across neighborhoods of cells rather than individual cells (**Methods**). These dispersions are used to calculate the gene weights, which then rescale the expression matrix: $\hat{E} = E\sqrt{W_D}$, where W_D is a diagonal matrix with gene weights along the diagonal. Using the rescaled expressions $\hat{E}$, we compute a pairwise cell distance matrix and update the assignment of each cell's k-nearest neighbors accordingly. This cycle continues until the gene weights converge.

To demonstrate the implementation and utility of SAM, below we analyze a challenging dataset comprised of a few hundred relatively homogeneous stem cells isolated from *Schistosoma mansoni* (**Figure 1–Figure supplement 1**), a widespread human pathogen (Hoffman et al., 2014). Using a protocol we have established previously (Wang et al., 2018), these cells were collected by sorting dividing cells from juvenile parasites harvested from their mouse hosts at 2.5 weeks post infection. At this stage, the parasites use an abundant stem cell population (~15-20% of the total number of cells) for rapid organogenesis and growth (Wang et al., 2013; Wang et al., 2018). Testing several existing methods (Wang et al., 2017; Kiselev et al., 2017; Satija et al., 2015), we found that they were not able to identify distinct cell populations in this dataset. In contrast, SAM finds a stable solution independent of initial conditions (**Figure 1b**). A graph structure with clearly separated cell populations self-assembles through the iterative process (**Figure 1c**). In parallel, the gene weights converge onto the final weight vector. Eventually, only a small fraction of genes

(~1%) are strongly weighted and useful for separating cell clusters, reflecting the inherent difficulty of analyzing this dataset.

Figure 1d shows that SAM iteratively improves a series of graph characteristics, including the network-average clustering coefficient (NACC), modularity, and Euclidean norm of the spatial dispersions (**Methods**). The NACC and modularity quantify the degree of structure within the graphs – graphs with high NACC and modularity have regions of high density separated by regions of low density. The dispersion quantifies the spatial organization of gene expression – the higher the spatial dispersion the less uniformly distributed the gene expressions are along the graph. Importantly, we verified that SAM does not artificially boost these metrics in data that lack inherent structure: when applying SAM to a randomly shuffled expression matrix, none of these metrics increased from the random initial conditions.

SAM identifies novel subpopulations within schistosome stem cells

Visualizing the converged graph in two dimensions using Uniform Manifold Approximation and Projection (UMAP, Becht et al., 2019), we find that cells can be separated into three main populations, with Louvain clustering (Blondel et al., 2008) further splitting one of these clusters into two subpopulations (**Figure 2a**). In contrast, other commonly-used dimensionality reduction methods, such as principal component analysis (PCA), Seurat (Satija et al., 2015), and SIMLR (Wang et al., 2017), failed to distinguish these cell populations (see **Methods** for the selection of algorithms for comparison). **Supplementary Table 1** lists genes with high SAM weights, which includes most markers that were previously implicated to be enriched in subsets of schistosome stem cells (Wang et al., 2013; Wang et al., 2018).

Figure 2b shows that the three populations include previously characterized δ′-cells, which specifically express an RNA binding protein *nanos-2* (Smp_051920), and ε-cells, which are marked by the expression of *eledh* (*eled*, Smp_041540) (Wang et al., 2018). More importantly, SAM reveals a novel stem cell population, *μ*, comprising ~30% of all sequenced cells (*μ* denotes muscle progenitors as discussed below). While *μ*-cells express ubiquitous stem cells markers (e.g., *ago2-1*, Smp_179320; *cyclin B*, Smp_082490) and cell cycle regulators (**Figure 2–Figure supplement 1a**) (Collins et al., 2013; Wang et al., 2013; Wang et al., 2018), they are also strongly enriched for a large set of genes, with 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) as the most specific markers of this population in comparison to other stem cells (**Figure 2–Figure supplement 1b**).

Fluorescent *in-situ* hybridization (FISH) in conjunction with EdU labeling of dividing cells reveals that *μ*-cells (*cabp*$^+$EdU$^+$) are distributed near the parasite surface right beneath a layer of post-mitotic differentiated cells that also express *cabp* (**Figure 2c**). Close to the parasite surface, there are two major cell types intertwined in space: the skin-like tegumental cells and the body wall muscle cells. However, *μ*-cells express none of the recently identified markers in tegumental progenitors (Wendt et al., 2018), suggesting that they may be associated with the muscle lineage. To test this idea, we performed double FISH experiments and observed in post-mitotic *cabp*$^+$ cells the coexpression of a set of canonical muscle markers (Witchley et al., 2013), including tropomyosin (Smp_031770), myosin (Smp_045220), troponin (Smp_018250), and collagen (Smp_170340) (**Figure 2d**). These results suggest that *cabp* is a specific marker for parasite body wall muscles and *μ*-cells are the muscle progenitors. Why the juvenile parasites maintain such an active pool of muscle progenitors will be an important question for future studies.

In addition, SAM identifies two subpopulations among ε-cells: ε_α-cells that are highly enriched for an aschaete-scute transcription factor (*astf*, Smp_142120), and ε_β-cells that abundantly express another basic helix-loop-helix protein (*bhlh*, Smp_087310) (**Figure 2b**, right panels). FISH experiments confirm these cells to be in close spatial proximity but with no coexpression of *astf* and *bhlh* (**Figure 2e**). Moreover, we observed with FISH that there are fewer *astf*$^+$ cells in larger, more matured juveniles, suggesting ε_α-cells are a dynamic population during development. To verify this observation, we sequenced another ~370 stem cell from juveniles at a later developmental time point (3.5 weeks post infection). After correcting for batch effects in the combined 2.5- and 3.5-week datasets using the mutual nearest neighbors (MNN) algorithm (Haghverdi et al., 2018), we find that δ'-, μ-, and ε_β-cells remain relatively constant throughout both time points, whereas ε_α-cells comprise a significantly smaller fraction of the stem cells at 3.5 weeks (7%) compared to 21% at 2.5 weeks (**Figure 2f**). Taken together, these analyses demonstrate that SAM can identify experimentally validated stem cell populations that are previously too subtle to separate using other methods but are closely associated with the schistosome development.

The critical difference between SAM and other methods lies in how they select genes for manifold reconstruction. SAM prioritizes genes with variable expressions across neighborhoods of cells rather than individual cells as in other methods (e.g., Seurat). **Figure 2g** shows that genes with high standardized dispersion across individual cells often have low SAM weights, indicating that these highly variable genes (HVGs) are irrelevant to the topological relationships between cells. Other methods (e.g. SC3, Kiselev et al., 2017) identify marker genes based on differential gene expression between cell clusters, but this approach suffers when cell cluster assignment is poor, especially when discrete cell groups are difficult to separate or absent. Indeed, SC3 failed in the default mode as it incorrectly predicted there to be only one cluster in the schistosome dataset.

After we manually increased the number of clusters, SC3 could recover a few of the marker genes associated with only one (μ-cells, blue symbols in **Figure 2h**) of the populations detected by SAM. Furthermore, changing the number of clusters resulted in different solutions and large variability in SC3 scores for its top ranked genes.

SAM outperforms other state-of-the-art methods in extensive quantitative benchmarking

Below, we assess the general applicability of SAM by benchmarking its performance against state-of-the-art scRNAseq analysis methods on a large collection of datasets. We focus on three methods, i.e., Seurat, SIMLR, and SC3, as they are mostly unsupervised, have been broadly used, and were shown to outperform other methods through extensive benchmarking (Kiselev et al., 2017; Wang et al., 2017; Duò et al., 2019; Bahlo et al., 2018; Tian et al., 2019, see **Methods** for the selection of algorithms for comparison). We first benchmark against nine datasets (**Supplementary Table 2**) that have high-confidence annotations to evaluate the accuracy of SAM in assigning cell clusters. We find that SAM has the highest Adjusted Rand Index (ARI, a measure of clustering accuracy) (Hubert and Arabie, 1985) on eight out of the nine datasets and does not over cluster the data (**Figure 3a**). Furthermore, SAM converges to the same set of gene weights for all datasets analyzed (**Figure 3b**, **Figure 3-Figure supplement 1a**) and its performance is robust to the choice of parameters and random initial conditions (**Figure 3-Figure supplement 1b-c**). In contrast, applying SAM to randomly generated datasets (**Methods**), the resulting gene weights are highly dissimilar across random initial conditions (**Figure 3b**), indicating that SAM does not converge to a stable solution on datasets with no intrinsic structure. Finally, the scalability of SAM is similar to that of Seurat, capable of analyzing hundreds of thousands of cells in minutes (**Figure 3c**), whereas SIMLR and SC3 are orders of magnitudes slower and thus excluded from further benchmarking which requires the analysis of many more datasets.

Because these nine datasets are all comprised of clearly distinguishable cell types, they may not represent the performance of methods on other datasets that contain cell populations that are only subtly different. To identify such datasets, we introduce a network sensitivity metric that quantifies the changes in the cell-to-cell distances when randomly selecting a subset of features from the gene expression matrices (**Methods**). High network sensitivity indicates that changes to the selected features strongly alters the resulting topological network. Networks that are robust to the selected features correspond to datasets that have many redundant signals or genes corroborating that network structure. In the datasets we compiled (**Supplementary Table 2**), all broadly-used benchmarking datasets have lower sensitivities whereas the schistosome dataset, which we have shown to be challenging to analyze for other methods, has the highest sensitivity (**Figure 4a)**. The fraction of genes with large SAM weights (>0.5) is negatively correlated with the network sensitivity, suggesting that the biologically relevant variation in datasets with high sensitivity is captured by relatively fewer genes (**Figure 4b**). Analyzing all 56 datasets, we found that SAM improves the clustering, modularity, and spatial organization of gene expression across the graph in comparison to Seurat as the datasets become increasingly sensitive (**Figure 4c**).

Evaluating the clustering accuracy for the highly sensitive datasets, however, is challenging, because many of them have incomplete or nonexistent cell type annotations. Therefore, we use the nine well-annotated benchmarking datasets to simulate data across a wide spectrum of sensitivities. For this, we corrupt the data by randomly permuting gradually increasing fractions of the gene expressions. As illustrated by the Darmanis dataset (Darmanis et al., 2015), **Figure 5a** shows that the sensitivity increases along with the corruption. SAM's ARI scores only marginally decrease as the corruption (and thereby sensitivity) increases, whereas Seurat's performance rapidly deteriorates. A similar contrast was observed with the NACC, modularity, and spatial dispersion between SAM and Seurat. Importantly, passing the genes with high SAM weights into Seurat rescued its performance across all metrics, indicating that SAM is able to consistently capture the genes relevant to the underlying structure of the data even with increasing levels of noise and illustrating the robustness of its feature selection strategy compared to the HVG filtering approach of Seurat. These observations generalize to all nine benchmarking datasets, quantified by the area under the curves (AUC) of the metrics with respect to corruption (**Figure 5b**).

SAM clusters macrophages by their activation dynamics with proper temporal ordering

We next highlight another dataset to show that SAM can recover biologically meaningful information that other methods fail to capture. We chose this example, which contains ~600 macrophages treated with lipopolysaccharide (LPS) when individually trapped in microfluidic channels (Lane et al., 2017), because it has high network sensitivity (**Figure 4a**) and has accompanying single cell functional data of macrophage activation dynamics that may help validate the results of our analysis. Applied to this dataset, SAM initially identifies two clusters (**Figure 6a**, top). Performing gene set enrichment analysis (GSEA, **Methods**, Subramanian et al., 2005), we find that genes with high SAM weights are dominated by cell cycle-related processes, with one of the clusters heavily enriched for cell cycle genes (e.g., Top2a, Mki67, **Figure 6-Figure supplement 1a**). After removing the cell cycle effects (**Methods**), SAM identifies two different clusters in which cells are properly ordered by the time since LPS induction, with the highly weighted genes being primarily involved in immune signaling (**Figure 6a**, bottom). These observations demonstrate that, in conjunction with GSEA, the quantitative gene weights output by

SAM can be used to infer the biological pathways that drive the clustering of cells.

One of the two clusters is enriched for TNFα expression (**Figure 6b**). It is known that LPS activates two independent pathways, one through the innate immune signal transduction adaptor (Myd88) and the other through the TIR-domain-containing adapter-inducing interferon-β (TRIF) (Lee et al., 2009). While the Myd88 pathway directly activates NF-κB, the TRIF pathway first induces the secretion of TNFα, which subsequently binds to its receptor, TNFR, to prolong the activation of NF-κB (**Figure 6c**). **Figure 6d** and **Figure 6-Figure supplement 1b** show examples of genes that are highly enriched with TNFα, a number of which are inflammatory factors known to accumulate due to prolonged NF-κB activation (Lane et al., 2017). These results suggest that SAM grouped the cells based on their activated signaling pathways: one cluster is activated through both Myd88 and TRIF pathways (MT) while the other is only activated through Myd88 (M).

To further verify that the separation between the MT and M clusters truly reflects the dichotomy in cellular response to LPS induction, we noted that this dataset combines scRNAseq with live-cell imaging of NF-κB activity in single cells. This allows us to directly test if the MT and M clusters correspond to different signaling dynamics (**Methods**). We found that most of the cells with prolonged NF-κB response (i.e., cells showing broad peaks of NF-κB activation in time) are in fact in the MT cluster (**Figure 6e-f**, and **Figure 6-Figure supplement 2a**), consistent with the expectation that TNFα signaling prolongs NF-κB activation. Although our interpretation of the data matches that provided in the original study, we were able to analyze the dataset with almost no *a priori* knowledge. In contrast, the original study required extensive manual curation, analyzed only a subset of the dataset, and could not group cells by their NF-κB activation dynamics from the gene expression data alone. Similarly, Seurat and SIMLR were unable to order the cells by the time since LPS induction or group cells based on their activation dynamics after removing the cell cycle effects (**Figure 6g**, and **Figure 6-Figure supplement 2b-c**).

Discussion

Here, we introduced a scRNAseq analysis method, SAM, that uses an unsupervised, robust, and iterative strategy for feature selection and manifold reconstruction. As demonstrated by our analysis of the schistosome stem cells and activated macrophages, SAM can capture biology that is undetectable by other methods. While SAM has consistently higher clustering accuracy than other state-of-the-art methods on datasets containing clearly distinct cell types, its advantages are especially apparent on datasets in which cell states or types are only distinguishable through subtle differences in gene expression.

The strength of SAM lies in the integration of three algorithmic components: spatial dispersion to measure feature relevance, soft feature selection, and the iterative scheme. By averaging the gene expression of a cell with that of its neighbors, the spatial dispersion quantifies the variation across neighborhoods of cells rather than individual cells. Genes with high spatial dispersion are more likely to be biologically relevant as they are capable of separating cells into distinct topological locations. Soft feature selection includes all genes and weights their contribution to the manifold reconstruction by their spatial dispersions. This mitigates the shortcoming of existing approaches in which the selection of features is a binary decision: genes are either included or not depending on arbitrarily chosen thresholds.

The conceptual challenge here is that calculating the gene weights requires the manifold, but reconstructing the manifold requires the gene weights for feature selection. SAM thus uses an iterative strategy to converge onto both the gene weights and the corresponding graph topology from a random initial graph. Each successive iteration refines the gene weights and network structure until the algorithm converges. Empirically, for all datasets analyzed we have shown that SAM converges onto a stable solution and is robust to the random initial conditions. Practically, we could initialize SAM using the graph output of another method such as Seurat, but using random initial conditions avoids potential biases in the analysis and enables the evaluation of the stability of SAM.

To demonstrate the strengths of SAM in practice, we analyzed the schistosome stem cells and identified novel stem cell populations that were validated by FISH experiments (**Figure 2**). In the analysis of activated macrophages, we showed that SAM can simultaneously order cells by the time since LPS induction and group cells according to their respective activated signaling pathways. We have validated this result using the live-cell imaging data presented in the original study (**Figure 6**).

We expect the application of SAM is not limited to feature selection, cell clustering, and manifold reconstruction; it can be readily integrated with existing analytical pipelines as its gene weights and reconstructed manifolds can be used in downstream analyses. For example, we have shown how the genes ranked by their SAM weights can be used as input to GSEA to determine the biological processes enriched among the highly weighted genes (**Figure 6**), thus directly testing if the weights reflect biologically relevant genes. Additionally, the manifold reconstructed by SAM can be used as input to pseudotemporal ordering algorithms (Setty et al., 2016; Trapnell et al., 2014).

Beyond the two example case studies, we have rigorously evaluated SAM on a total of 56 datasets. While previous studies benchmarked on datasets with clearly defined cell populations, we defined a network sensitivity measure to rank the datasets based on the inherent difficulty of their analysis (**Figure 4**). Using these datasets, we showed that SAM consistently outperforms other methods in terms of both cell clustering accuracy measured by ground truth annotations, and manifold reconstruction measured by quantitative graph characteristics. These improvements can be attributed to the robust selection of features relevant for cell clustering and manifold reconstruction even in the presence of significant amounts of random noise, as shown in the corruption tests (**Figure 5**). Overall, the network sensitivity and quantitative benchmarking metrics should help in characterizing the performance of future scRNAseq analysis methods across a wider variety of datasets.

Materials and Methods

Code and data availability. The SAM source code and tutorials can be found at https://github.com/atarashansky/self-assembling-manifold. The schistosome stem cell scRNAseq data generated in this study is available through the Gene Expression Omnibus (GEO) under accession number GSE116920.

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 are sourced from the *conquer* database (Soneson and Robinson, 2018). The nine benchmarking datasets used with high-confidence labels are marked by crosses. Gene expression is measured in log space with a pseudocount of 1 (e.g., $\log_2(\mathrm{TPM}+1)$). Genes expressed ($\log_2(\mathrm{TPM}+1)>1$) in fewer than 1% or more than 99% 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 $\log_2(\mathrm{TPM}+1)<1$. We denote the resulting expression matrix as E.

In the SAM algorithm (see below), we either standardize the gene expression matrix E to have zero mean and unit variance per gene (which corrects for differences in distributions between genes) or normalize the expressions such that each cell has unit Euclidean (L2) norm (which prevents cells with large variances in gene expressions from dominating downstream analyses) prior to dimensionality reduction. In the below section, we denote the standardized or normalized expression matrix as $\bar{E}$. Empirically, we have found that standardization performs well with large, sparse datasets collected through droplet-based methods, whereas L2-normalization is more suitable for smaller datasets with higher sequencing depth such as those prepared with the Smart-Seq2 protocol (Picelli et al., 2013). This is likely due to the fact that standardization amplifies the relative expression of genes specific to small populations in large datasets, thereby making them easier to identify. In contrast, standardization decreases the relative expression of genes specific to populations comprising larger fractions of the data, as is typically the case in smaller datasets, thereby making distinct populations more difficult to identify. **Supplementary Table 2** documents the preprocessing step used for each dataset.

The SAM algorithm. After first generating a random kNN adjacency matrix, the SAM algorithm goes through three steps that are repeated until convergence.

1) Calculate the gene weights

First, the expression of each cell is averaged with its k-nearest neighbors:

$$C = \frac{1}{k} NE \tag{1}$$

where N is the directed adjacency matrix for the kNN graph, and E is the original n x m gene expression matrix with rows as cells and columns as genes. Here, we do not use $\bar{E}$ as it may contain negative values, for which dispersion is ill-defined. For each gene i, SAM computes the Fano factor from the averaged expressions C_i:

$$\mu_{C_i} = \frac{1}{n} \sum_{j=1}^{n} C_{ji} \tag{2}$$

$$\sigma_{C_i}^2 = \frac{1}{n}\sum_{j=1}^{n}(C_{ji} - \mu_{C_i})^2 \tag{3}$$

$$F_i = \frac{\sigma_{C_i}^2}{\mu_{C_i}} \tag{4}$$

where μ_{C_i} is the mean and $\sigma_{C_i}^2$ is the variance. We use the Fano factor to measure the gene expression variance relative to the mean in order to account for the fact that genes with high mean expressions tend to have higher variability. 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. k, set by default to 20, determines the topological length scale over which variations in gene expression are quantified. **Figure 3-Figure supplement 1b** shows 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.

To compute the gene weights, we normalize the Fano factors to be between 0 and 1. First, we saturate the Fano factors to ensure that outlier genes with large spatial dispersions do not skew the distribution of weights:

$$\{F_i | F_i > z\} = z \tag{5}$$

where z is by default the mean of the largest 50 dispersions. In other words, Fano factors exceeding this number are saturated to be z. We then calculate the gene weights as:

$$W_i = \frac{F_i}{z} \tag{6}$$

2) Rescale the expression matrix

Having calculated the gene weights, SAM multiplies them into the preprocessed expression matrix:

$$\hat{E} = \bar{E}\sqrt{W_D} \tag{7}$$

where $\bar{E}$ is the standardized or normalized expression matrix and W_D is a diagonal matrix with W_i along the diagonal. This matrix multiplication linearly rescales the gene expression variances and gene-gene covariances by their respective weights, attenuating the influence of genes with low dispersions across neighborhoods.

3) Updating the kNN graph

To compute pairwise cell-cell distances, we perform PCA on the rescaled expression matrix $\hat{E}$. The variance-scaling operation in **Eq. 7** improves the robustness of PCA to variations in genes that are uniformly distributed along the current graph (i.e., genes with low weights). Furthermore, this weighting strategy eliminates the typical requirement of selecting a subset of HVGs to feed into PCA, which often relies on arbitrary thresholds and heuristics. To perform PCA, we first mean center $\hat{E}$ to form $\hat{E}_\mu$:

$$\hat{E}_\mu = \hat{E} - \frac{1}{n} e e^T \hat{E} \tag{8}$$

where e is a column vector of ones with dimension n. We then compute the Singular Value Decomposition (SVD) of $\hat{E}_\mu$:

$$\hat{E}_\mu = USV^T \tag{9}$$

with the principal components defined as

$$P = US \tag{10}$$

The eigenvalues corresponding to the eigendecomposition of the gene-gene covariance matrix are defined in terms of the singular values as

$$\Lambda = \frac{S^2}{n-1} \tag{11}$$

where S is a diagonal matrix with singular values along the diagonal. Using the PC matrix P, SAM computes a pairwise cell-cell distance matrix. While typical dimension reduction approaches select a subset of the PCs, which is often subjective or requires computationally intensive maximum-likelihood approaches, we include all PCs and scale their variances by their corresponding eigenvalues:

$$\hat{P} = P\sqrt{\Lambda} \tag{12}$$

As a result, PCs with small eigenvalues are weighted less in the calculation of the distance between cells i and j, $D_{P_iP_j}$. $D_{P_iP_j}$ is the Pearson correlation or Euclidean distance between rows P_i and P_j in the PC matrix. Pearson correlation distance is used by default, although **Figure 3-Figure supplement 1c** shows that SAM is robust to the choice of distance metric. Using the distances to define the k-nearest neighbors for each cell, SAM updates the kNN matrix and repeats steps 1-3. The algorithm continues until convergence, defined as when the RMSE between gene weights in adjacent iterations diminished as defined by:

$$\sqrt{\frac{1}{m}\sum_{j=1}^{m}(W_{i,j} - W_{i+1,j})^2} < 5 \times 10^{-3} \tag{13}$$

where m is the number of genes and $W_{i,j}$ is the weight for gene j at iteration i.

Visualization. To visualize the topological structure identified by SAM, we feed the final weighted PCA matrix, $\hat{P}$, into UMAP (Becht et al., 2019) using Pearson correlation as the distance metric by default. To directly visualize the final kNN adjacency matrix (**Figure 1c**), we used the Fruchterman-Reingold force-directed layout algorithm and drawing tools implemented in the

Python package *graph-tool* (Peixoto, 2017).

Choosing the benchmarking methods. We used three main criteria for choosing the benchmark scRNAseq analysis methods: they should be widely used, have done extensive benchmarking against other methods, and be mostly unsupervised. We found on Web of Science that among the highest cited scRNAseq analysis tools in 2017-2018 are Seurat, SC3, SIMLR, Reference Component Analysis (RCA, Li et al., 2017), Monocle (Trapnell et al., 2014), zero-inflated factor analysis (ZIFA, Pierson and Yau, 2015), and Wishbone (Setty et al., 2016), of which we chose Seurat, SC3, and SIMLR.

SC3 is a consensus clustering algorithm that has done rigorous benchmarking against other methods such as SINCERA (Guo et al., 2015), SNN-Cliq (Xu and Su, 2015) and pcaReduce (Žurauskien and Yau, 2016) on 12 datasets with ground truth labels. SIMLR, a dimensionality reduction and clustering algorithm, evaluated its clustering performance on four annotated datasets against eight other dimensionality reduction methods, including PCA, Factor Analysis (FA), t-SNE, multidimensional scaling (MDS), and (ZIFA). Both methods have demonstrated the highest clustering accuracy across most of the tested datasets. Additionally, as both methods have built-in functions to estimate the number of clusters present within the data, they are largely unsupervised. We also selected Seurat as one of the benchmarking methods, because it is arguably the most widely used tool for dimensionality reduction and clustering of scRNAseq data and has performed well in rigorous benchmarking studies against various methods including SC3, SIMLR, RCA, and pcaReduce (Duò et al., 2019; Bahlo et al., 2018).

We did not select Reference Component Analysis as it is primarily designed for cases in which an atlas of bulk, cell-type specific, reference transcriptomes is present. Additionally, we did not benchmark against Monocle and Wishbone, because they are pseudotime analysis methods and are meant for datasets with continuous branching processes such as cell differentiation. However, it is important to note that SAM can be used for dimensionality reduction upstream of pseudotime algorithms for such datasets. Finally, we did not benchmark against ZIFA as it has already been shown to have lower clustering accuracy than SIMLR.

In addition to measuring clustering accuracy, we also introduce the unsupervised NACC, modularity, and spatial dispersion metrics to quantify both the degree of structure and spatial organization of gene expression within a nearest-neighbor graph. As such, these metrics can only be applied to dimensionality reduction methods that construct a graph representation of the dataset. Consequently, we cannot use these metrics to evaluate SC3.

Although it does technically produce a graph representation of the data, SIMLR should be considered as a hybrid between a clustering and dimensionality reduction method. Because its similarity graph is assumed to have a block structure where the number of blocks is equal to the prespecified number of clusters, the resulting nearest-neighbor graph will, by construction, tend to have a higher degree of structure and therefore artificially inflated NACC and modularity.

Furthermore, the poor scalability of SC3 and SIMLR makes them difficult to run for many trials across a large number of datasets. Although SIMLR, in particular, does provide an alternative algorithm that can scale to much larger datasets, it has not been extensively benchmarked. Even so, despite the improved speed of this large-scale implementation, estimating the number of clusters using its built-in function remains a significant computational and memory bottleneck. For example, when applied to the ~10,000 planarian neoblasts, neither implementations of SIMLR could estimate the number of clusters within 2 hours. As a result, we cannot run SIMLR in an unsupervised manner on datasets significantly larger than ~3000 cells.

As there are few practical alternatives for manifold reconstruction that have been as extensively benchmarked and widely used, we primarily compare SAM to Seurat in tests involving the unsupervised, graph-based metrics to highlight the key, advantageous characteristics of SAM as a manifold reconstruction and feature selection algorithm when applied to datasets with varying sensitivities (**Figure 4a-c**).

Benchmarking. To generate the convergence curves in **Figure 1b**, we computed the root mean square error (RMSE) of the gene weights averaged across all pairwise comparisons of ten replicates starting from randomly generated initial graphs. In **Figure 3b**, we extend this analysis to all datasets analyzed and report the final error. We use randomly generated datasets of varying sizes (ranging from 200 to 5000 cells) as a negative control to show that SAM does not converge onto the same solution across initial conditions when the data has no intrinsic structure. These datasets were randomly generated by sampling gene expressions from a Poisson distribution with mean drawn from a gamma distribution. To generate the convergence curves in **Figure 3-Figure supplement 1a**, we computed the RMSEs, which are ensemble-averaged across ten replicate runs, between the gene weights in adjacent iterations. We compute the adjacency error between kNN adjacency matrices N_i and N_j (**Figure 1b**) as

$$A_{i,j} = \frac{e^T|N_i - N_j|e}{2e^T N_i e} \tag{14}$$

where e is a column vector of ones. This simply measures the fraction of total edges that are different between the two graphs.

To compute the standardized dispersion factors in **Figure 2g**, we used Seurat's methodology implemented in Scanpy (Wolf et al., 2018), which groups the genes into 20 bins based on their mean expression values and computes the z-score of each gene's Fano factor with respect to the mean and standard deviation of all Fano factors in its corresponding bin. To generate the AUROC scores in **Figure 2h**, which quantify the likelihood of genes being cluster-specific markers, we ran SC3 on the schistosome data with the number of clusters ranging from 2 to 12. We used the AUROC scores corresponding to 4 clusters for the points on the scatter plot and the standard deviations of the scores across all tested numbers of clusters for the error bars.

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

$$ARI = \frac{\sum\binom{n_{ij}}{2} - \left[\sum\binom{a_i}{2}\sum\binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum\binom{a_i}{2} + \sum\binom{b_j}{2}\right] - \left[\sum\binom{a_i}{2}\sum\binom{b_j}{2}\right]/\binom{n}{2}} \qquad (15)$$

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 (Hubert and Arabie, 1985). 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 ith row jth column of the contingency table, respectively.

Seurat was implemented using the Scanpy package in Python (Wolf et al., 2018). For Seurat, we selected the top 3000 variable genes according to their standardized dispersions and chose the number of PCs (bounded between 6 and 50) which explain 30% of the variance 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 adjacency matrix with the value of k used by SAM. To assign cluster labels for SAM and Seurat, we applied HDBSCAN (Mcinnes et al., 2017), an unsupervised, density-based clustering algorithm to their respective PCA outputs. As HDBSCAN does not cluster any cell it deems an outlier, we assign the remaining outlier cells to clusters using kNN classification.

For each outlier cell, we identify its 20 nearest neighbors among the clustered cells. Outliers are assigned to the same cluster as that of the majority of its neighbors. This minor extension to HDBSCAN is available as the built-in function *hdbknn_clustering* in SAM. SC3 was run using default parameters. The SIMLR package was implemented in R and run with the normalization parameter set to “True”, which mean-centers gene expressions after normalizing them to be between 0 and 1. Both SC3 and SIMLR provide their own functions to estimate the number of clusters and cluster assignments.

To compare the quality of graphs generated by different methods, we use the NACC, modularity, and spatial dispersion. The NACC is the average of the local clustering coefficient for each node of a graph and quantifies the degree of structure in the graph (Watts and Strogatz, 1998). The local clustering coefficient is defined as

$$a_i = \frac{L_i}{k_i(k_i - 1)} \tag{16}$$

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. We calculate the NACC using the implementation in *graph-tool* (Peixoto, 2017).

The modularity Q of a graph is defined as

$$Q = \frac{1}{4m}\sum_{i,j}^{c}\left(A_{ij} - \frac{k_i k_j}{2m}\delta_{ij}\right) \tag{17}$$

where A_{ij} is one if there is an edge from cell *i* to cell *j*, k_i is the degree of cell *i*, k_j is the degree of cell *j*, *m* is the total number of edges, and δ_{ij} is one if cells *i* and *j* are in the same cluster. High modularity indicates that clusters have on average many more edges within clusters than between clusters. To find the optimal modularity for a particular graph, we used Louvain clustering, which searches for a partition in which modularity is maximized.

To quantify the spatial organization of gene expression along the graph, we calculate the Euclidean norm of the largest 500 spatial dispersions. Spatial dispersion is defined as before in the SAM algorithm: $F_i = \frac{\sigma_{C_i}^2}{\mu_{C_i}}$, where F_i is the Fano factor of the kNN-averaged expressions, $C_i = \frac{1}{k}NE_i$. N is the directed adjacency matrix output by SAM or Seurat and E_i is a column vector of expression values for gene *i*.

To measure the inherent sensitivity of each dataset, we randomly perturbed the gene expression matrices of each dataset by randomly sampling 2000 genes and applied PCA to the subsampled data. A correlation distance matrix was calculated from the top 15 PCs and perturbations were repeated 20 times to generate distance matrix replicates. Sensitivity is then defined as the average error across all pairwise comparisons between replicates. The error between two distance matrices j and k, S_{jk}, is defined as the average correlation distance between corresponding pairs of rows in the distance matrices d_j and d_k:

$$S_{jk} = \frac{1}{n}\sum_{i=1}^{n} D\{d_{j,i}, d_{k,i}\} \tag{18}$$

where $D\{d_{j,i}, d_{k,i}\}$ is the Pearson correlation distance between the distances from cell *i* in distance matrices *j* and *k*.

We simulated datasets with increasing sensitivity by introducing increasing degrees of corruption in each of the nine annotated datasets. To corrupt a dataset, we swapped random pairs of elements in the expression matrix. The number of swaps, p, corresponds to the degree of corruption, with p varying from 0 to half of the total number of elements in the matrix. For each annotated dataset, we simulated 10 replicates per value of p. SAM and Seurat were run with default values on each corrupted dataset, clustering was performed using the *hdbknn_clustering* function in SAM, and the ARI, NACC, modularity, and spatial dispersion metrics were recorded. The Area

Under the Curve (AUC) was calculated for each metric with respect to the fraction of elements swapped, *p*, using the trapezoidal rule. Finally, to rescue the performance of Seurat, we used as input to Seurat the top 3000 genes with the highest SAM weights.

Gene set enrichment analysis (GSEA). GSEA (Subramanian et al., 2005) is typically run on a gene expression matrix with user-defined cluster assignments to quantify the differential expression for each gene. By default, differential expression is quantified using a signal-to-noise metric and the resulting scores are used to rank the genes in descending order. However, GSEA can also run in an alternative mode in which the user provides a predefined list of gene rankings. Therefore, we used the genes ranked by their SAM weights as input to GSEA to determine the biological processes enriched among the highly weighted genes. As shown in **Figure 6a**, we can directly test if SAM captures the relevant biological processes. GSEA provides a number of statistical measures to assess the significance of enriched gene sets, of which we use the False Discovery Rate (FDR). The FDR quantifies the likelihood that a highly enriched gene set represents a false positive. The significance threshold typically used with FDR is 25%, which implies that the results are likely to be valid 75% of the time.

Removal of cell cycle effects. To remove cell cycle effects from the macrophage dataset, we adopted a simpler version of the strategy used in *ccRemover* (Barron and Li, 2016), in which we subtract from the data PCs that are significantly associated with known cell cycle genes. Letting *P* represent the PCs and *L* be the gene loadings, we quantify the association between the set of cell cycle genes *G* and PC *j* as

$$A_j = \frac{1}{|G|}\sum_{i \in G} |L_{ji}| \tag{19}$$

PC j is selected if its association A_j is at least two standard deviations above the mean of the associations for all PCs. In the particular case of the macrophage data, we identified the set of PCs $S = \{P_0, P_1, P_8\}$ as being significantly associated with the cell cycle genes. We next reconstruct the data using these PCs, which thus captures the cell-cycle effects, and subtract the reconstructed data from the expression matrix E:

$$E_{removed} = E - \sum_{j \in S} P_j L_j \sqrt{W} \tag{20}$$

When reconstructing the data, we scale the gene loadings by the SAM weights W so that only the highly weighted SAM genes (which are initially dominated by cell cycle genes) contribute to the cell cycle removal, as there may be genes involved in other biological processes that could also be correlated with the PCs in S. To run SAM on the data with cell cycle effects removed, we use E as opposed to $E_{removed}$ for the calculation of spatial dispersions, because the latter may contain negative values, for which dispersion is ill-defined. This method is made available as a part of the SAM package in the functions *calculate_regression_PCs* and *regress_genes*.

Clustering the NF-κB activity time series. In the original study, Lane et al. combined imaging and transcriptomics to link NF-κB nuclear translocation dynamics to changes in gene expression within single cells. Macrophages stimulated with LPS were individually trapped in microfluidic chambers and imaged for various lengths of time (75-300 min) prior to scRNAseq library preparation. NF-κB was tagged with a fluorescent protein, and its activation was measured as the nuclear-localized fluorescence intensity. Based on the imaging data, the authors identified three main classes of NF-κB dynamics, the first with a transient initial response, the second with a prolonged initial response, and the third with a recurrent response. Because the recurrent response is found only in the 300 min time point and comprises only ~8% of these cells, we primarily focused on clustering cells based on their initial dynamics. To do this, we used the *tslearn* (Tavenard, 2017) python package to group cells based on their NF-κB activity time series. Because these time series are quite noisy, we were conservative in labeling cells as having a prolonged initial response in an effort to avoid false positives. As a result, these cells comprise only ~30% of the dataset.

For the cells sampled at 75 and 150 min after LPS stimulation, we used the time series *k*-means algorithm with the *softdtw* distance metric to cluster them each into three groups, which resulted in representative time series with transient, intermediate, and prolonged responses. Merging the cells with transient and intermediate responses into one cluster (which we simply labeled as transient response), we obtained the 75 and 150 min (black and blue, respectively) representative time series shown in **Figure 6e**. Because the cells sampled at 300 min displayed much more variability in their NF-κB time series, we clustered them into 6 groups, labeling the cluster whose representative time series had the broadest initial peak as the prolonged response cluster (red in **Figure 6e**, right). The remaining groups were labeled as the transient response cluster (red in **Figure 6e**, left).

Mapping the schistosome datasets. We used the Mutual Nearest Neighbors algorithm (Haghverdi et al., 2018) with default values to generate an expression matrix $E_{corrected}$ in which the batch effects between the 2.5-week and 3.5-week datasets were corrected for. To run SAM on the batch-corrected data, we use E for the calculation of spatial dispersions as opposed to $E_{corrected}$.

scRNAseq of schistosome stem cells. Schistosome stem cells were isolated from juvenile parasites retrieved from infected mice at 2.5 and 3.5 weeks post infection. We followed the protocol as previously described (Wang et al., 2018). Briefly, we retrieved juvenile parasites from schistosome-infected mice (Swiss Webster NR-21963) by hepatic portal vein perfusion. Parasites were cultured at 37°C/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. In adherence to the Animal Welfare Act and the Public Health Service Policy on Humane Care and Use of Laboratory Animals, all experiments with and care of mice were performed in accordance with protocols approved by the Institutional Animal Care and Use Committees (IACUC) of Stanford University (protocol approval number 30366).

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. 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 (Wang et al., 2018). 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 (Picelli et al., 2013). For quality control, we quantified the histone *h2a* (Smp_086860) levels using qPCR, as *h2a* is a ubiquitously expressed in all schistosomes stem cell (Collins et al., 2013; Wang et al., 2013; Wang et al., 2018). We picked wells that generated C_T values within 2.5 C_T around the most probable values (~45% of total wells, **Figure 1-Figure supplement 1**). 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. In 2.5-week dataset, 338 cells with more than 1700 transcripts expressed at >2 TPM were used for downstream analysis. In the 3.5-weeks dataset, 338 cells with more than 1350 transcripts expressed at >2 TPM were used for downstream analysis (**Figure 1-Figure supplement 1**).

In situ hybridization and EdU labeling. RNA FISH experiments were performed as detailed in previous publications (Collins et al., 2013; Wang et al., 2013; Wang et al., 2018). Clones used for riboprobe synthesis were generated as described previously, with oligonucleotide primers listed in **Supplementary Table 3**. Juvenile parasites were cultured with 10 µM EdU overnight, killed in 6 M $MgCl_2$ for 30s, and then fixed in 4% formaldehyde with 0.2% Triton X-100 and 1% NP-40. Fixed parasites were sequentially dehydrated in methanol, bleached in 3% H_2O_2 for 30 min, and rehydrated. Parasites were permeabilized by 10 µg/mL proteinase K for 15 min and post fixed with 4% formaldehyde. The hybridization was performed at 52°C with riboprobes labeled with either digoxigenin-12-UTP (Roche) or fluorescein-12-UTP (Roche). For detection, samples were blocked with 5% horse serum and 0.5% of Roche Western Blocking Reagent, and then incubated with anti-digoxigenin-peroxidase (1:1000; Roche) or anti-fluorescein peroxidase (1:1500; Roche) overnight at 4°C for tyramide signal amplification (TSA). For double FISH, the first peroxidase was quenched for 30 min in 0.1% sodium azide solution before the detection of the second gene. After FISH, EdU detection was performed by click reaction with 25µM Cy5-azide conjugates (Click Chemistry Tools). Samples were mounted in *scale* solution (30% glycerol, 0.1% Triton X-100, 4 M urea in PBS supplemented with 2 mg/mL sodium ascorbate) and imaged on a Zeiss LSM 800 confocal microscope.

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, N. Neff, J. Okamoto, and D. Nanes-Sarfati for experimental help, K. Lane, M. Covert, J. Qin, F. Horns, G. Stanley, J. Rink, 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.

References

Bahlo, M., Tian, L., Lönnstedt, I., Ng, M. & Hicks, S. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing. *F1000Research* **7,** 1–26 (2018).

Baron, M. *et al*. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell syst.* **3,** 346–360 (2016).

Barron, M. & Li, J. Identifying and removing the cell- cycle effect from single-cell RNA-Sequencing data. *Sci. Rep.* **6,** 33892 (2016).

Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37,** 38–44 (2019).

Blondel, V. D., Guillaume, J., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008,** P10008 (2008).

Collins, J. J. *et al.* Adult somatic stem cells in the human parasite Schistosoma mansoni. *Nature* **494,** 476–479 (2013).

Crow, M. *et al*. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9,** 884 (2018).

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

Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research* **7,** 1–21 (2019).

Fincher, C. T. *et al*. Cell type transcriptome atlas for the planarian Schmidtea mediterranea. *Science* **360,** eaaq1736 (2018).

Guo, M. *et al*. SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. *PLOS Comput. Biol.* **11,** e1004575 (2015).

Grün, D., Kester, L. & Oudenaarden, A. V. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11,** 637–640 (2014).

Haghverdi, L. *et al*. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36,** 421–427 (2018).

Hoffmann, K. F., Brindley, P. J. & Berriman, M. Halting harmful helminths. *Nature* **168,** 168–169 (2014).

Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2,** 193–218 (1985).

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

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).

Lee, T. K. *et al.* A Noisy Paracrine Signal Determines the Cellular NF- k B Response to Lipopolysaccharide. *Sci. Immunol.* **2,** ra65 (2009).

Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **49,** 708–718 (2017).

Lönnberg, T. *et al.* Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves TH1/TFH fate bifurcation in malaria. *Sci. Immunol.* **2,** eaal2192 (2017).

Mcinnes, L., Healy, J. & Astels, S. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* **2,** 205 (2017).

Olsson, A. *et al.* Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature* **537,** 698–702 (2016).

Peixoto, T. P. The graph-tool python library. (2017). doi:https://doi.org/10.6084/m9.figshare.1164194.v14

Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10,** 1096–1100 (2013).

Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16,** 241 (2015).

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

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).

Schwalie, P. C. *et al.* A stromal cell population that inhibits adipogenesis in mammalian fat depots. *Nature* **559,** 103–108 (2018).

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

Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15,** 255–261 (2018).

Subramanian, A. *et al*. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 15545–15550 (2005).

Tavenard, R. tslearn: A machine learning toolkit dedicated to time-series data. (2017).

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

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).

Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single cell RNA-seq. *Nature* **509,** 371–375 (2014).

Vallejos, C. A. *et al*. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLOS Comput. Biol.* **11.6,** e1004333 (2015).

Vallejos, C. A. *et al*. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* **14,** 565–571 (2017).

Wang, B., Collins, J. J. & Newmark, P. A. Functional genomic characterization of neoblast-like stem cells in larval Schistosoma mansoni. *eLife* **2,** e00768 (2013).

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).

Wang, B. *et al.* Stem cell heterogeneity drives the parasitic life cycle of Schistosoma mansoni. *eLife* **7,** e35449 (2018).

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

Wendt, G.R. *et al.* Flatworm-specific transcriptional regulators promote the specification of tegumental progenitors in Schistosoma mansoni. *eLife* **7,** e33221 (2018).

Witchley, J. N., Mayer, M., Wagner, D. E., Owen, J. H. & Reddien, P. W. Muscle cells provide instructions for planarian regeneration. *Cell Rep.* **4,** 633–641 (2013).

Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19,** 1–5 (2018).

WormBase, https://www.wormbase.org, release WS268.

Xu, C. & Su, Z. Gene expression Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31,** 1974–1980 (2015).

Žurauskien, J. & Yau, C. pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinform.* **17,** 140 (2016).

Figure legends

Figure 1. The SAM algorithm. (**a**) SAM starts with a randomly initialized kNN adjacency matrix and iterates to refine the adjacency matrix and gene weight vector until convergence. (**b**) Root mean square error (RMSE) of the gene weights (top) and the fraction of different edges of the nearest-neighbor adjacency matrices (bottom) between adjacent iterations (blue) and between independent runs at the same iteration (orange) to show that SAM converges to the same solution regardless of initial conditions. The differences between the gene weights and nearest-neighbor graphs from independent runs are relatively small, indicating that SAM converges to the same solution through similar paths. (**c**) Graph structures and gene weights of the schistosome stem cell data converging to the final output over the course of 10 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. (**d**) Network properties iteratively improve for the graphs reconstructed from the original data (red) but not on the randomly shuffled data (blue). Dashed lines: metrics measured from the Seurat-reconstructed graphs.

Figure 2. SAM identifies novel subpopulations within schistosome stem cells. (**a**) UMAP projections of the manifolds reconstructed by SAM, PCA, and Seurat. SIMLR outputs its own 2D projection based on its constructed similarity matrix using a modified version of t-SNE. The schistosome cells are color-coded by the stem cell subpopulations μ, δ’, ε_α, and ε_β determined by Louvain clustering. (**b**) UMAP projections with gene expressions of subpopulation-specific markers (*eledh*, *nanos-2 cabp*, *astf*, *bhlh*,) and a ubiquitous stem cell marker, *ago2-1*, overlaid. Insets: magnified views of the expressing populations. (**c**) FISH of *cabp* and EdU labeling of

dividing stem cells in juvenile parasites at 2.5 weeks post infection show that μ-cells (*cabp*$^+$EdU$^+$, arrowheads) are close to the parasite surface and beneath a layer of post-mitotic *cabp*$^+$ cells. Dashed outline: parasite surface. Right: magnified views of the boxed region. (**d**) FISH of *cabp* and a set of canonical muscle markers, *troponin*, *myosin*, *tropomyosin*, and *collagen*, shows complete colocalization in post-mitotic *cabp*$^+$ cells. Images in (**c-d**) are single confocal slices. (**e**) FISH of *astf* and *bhlh* shows their orthogonal expression in adjacent EdU$^+$ cells (arrowheads). Bottom: magnified views of the boxed region. Image is a maximum intensity projection of a confocal stack with a thickness of 12 μm. (**f**) UMAP projection of stem cells isolated from juveniles at 2.5 and 3.5 weeks post infection. Cell subpopulation assignments based on marker gene expressions are specified. Right: a magnified view to show the mapping of ε_α- and ε_β-cells. (**g**) Standardized dispersions as calculated by Seurat plotted vs. the SAM gene weights. (**h**) SC3 AUROC scores plotted vs. the SAM gene weights. Error bars indicate the standard deviation of SC3 AUROC scores between trials using different chosen numbers of clusters. In (**g**) and (**h**), the top 20 genes specific to each subpopulation are colored according to the color scheme used in (**a**).

Figure 3. SAM improves clustering accuracy and runtime performance. (**a**) Accuracy of cluster assignment quantified by adjusted rand index (ARI) on nine annotated datasets (left). Right: differences between the number of clusters found by each method and the number of annotated clusters. Smaller differences indicate more accurate clustering. (**b**) RMSE of gene weights output by SAM averaged across ten replicate runs with random initial conditions for 56 datasets (blue) and simulated datasets with no intrinsic structure (green, **Methods**). (**c**) Runtime of SAM, SC3, SIMLR, and Seurat as a function of the number of cells in each dataset. SC3 and SIMLR were not run on datasets with >3000 cells as the run time exceeds 20 minutes.

Figure 4. SAM improves the analysis of datasets with varying network sensitivities. (**a**) Network sensitivity of all 56 datasets ranked in descending order. Blue: the nine benchmarking datasets used in **Figure 3a**. Sensitivity measures the robustness of a dataset to changes in the selected features (**Methods**). (**b**) The network sensitivity plotted against the fraction of genes with SAM weight greater than 0.5 (in log scale) with Spearman correlation coefficient specified in the upper-right corner. (**c**) Fold improvement of SAM over Seurat for NACC, modularity, and spatial dispersion with respect to sensitivity for all 56 datasets. These ratios are linearly correlated with network sensitivity with Pearson correlations (r^2) specified in the upper-left corner of each plot.

Figure 5. Robust feature selection improves cell clustering and manifold reconstruction. (**a**) Network sensitivity, ARI, NACC, modularity, and spatial dispersion with respect to corruption of the Darmanis dataset, in which we swap random pairs of gene expressions with the number of swaps ranging from 0-50% of the total number of elements in the data (**Methods**). Performance is compared between SAM (blue), Seurat (red), and Seurat rescued with the top-ranked SAM genes (indigo). Error bars indicate the standard deviations across 10 replicate runs. The errors for points with no bars are too small to be seen. (**b**) Comparison of the area under curve (AUC) of the metrics in (**a**) with respect to data corruption for all nine datasets. Error bars indicate the standard deviations across 10 replicate runs. The errors for data with no error bars are too small to be seen.

Figure 6. SAM captures the cellular activation dynamics in a stimulated macrophage dataset. (**a**) GSEA analysis (left) and SAM projections (right) of the activated macrophages[7] before (top) and after (bottom) removing cell cycle effects. Teal: significantly enriched gene sets determined

by the significance threshold of 0.25 for the False Discovery Rate (FDR, dashed lines). Bottom: the two clusters are denoted as MT and M with colors representing the time since LPS induction. Arrows: evolution of time. (**b**) TNFα is enriched in the MT cluster. (**c**) Diagram of NF-κB activation in response to LPS stimulation via the Myd88 and TRIF signaling pathways. (**d**) Log_2 fold changes of the average expressions of selected inflammatory genes in the MT cluster vs. the M cluster. All genes are significantly differentially expressed between the two clusters according to the Welch's two-sample t-test ($p < 5 \cdot 10^{-3}$). (**e**) Representative traces for transient (left) and prolonged (right) NF-κB activation (**Methods**). (**f**) Cells with prolonged NF-κB response (denoted as P) are primarily in the MT population. (**g**) Seurat and SIMLR projections show that they fail to order the cells by time since LPS induction and do not identify cell clusters representing the different modes of NF-κB activation.

Figure 1 – Figure supplement 1. Quality control of library preparation and sequencing of the schistosome stem cells. (**a**) Histograms of *h2a* qPCR measurements in 2.5- (left) and 3.5- (right) week samples. (**b**) Scatter plot of gene count (>2 TPM) vs mapped read count of individual sequenced cells. Cells with low gene count or *h2a* expression are discarded and filtered from analysis (red) and the remaining cells are analyzed (blue). The number of final cells kept for analysis is annotated on the top left corner of each plot.

Figure 2 – Figure supplement 1. μ-cells express ubiquitous stem cell marker and population specific genes. UMAP projections with gene expressions of (**a**) stem cell markers and (**b**) μ-cell-specific genes overlaid.

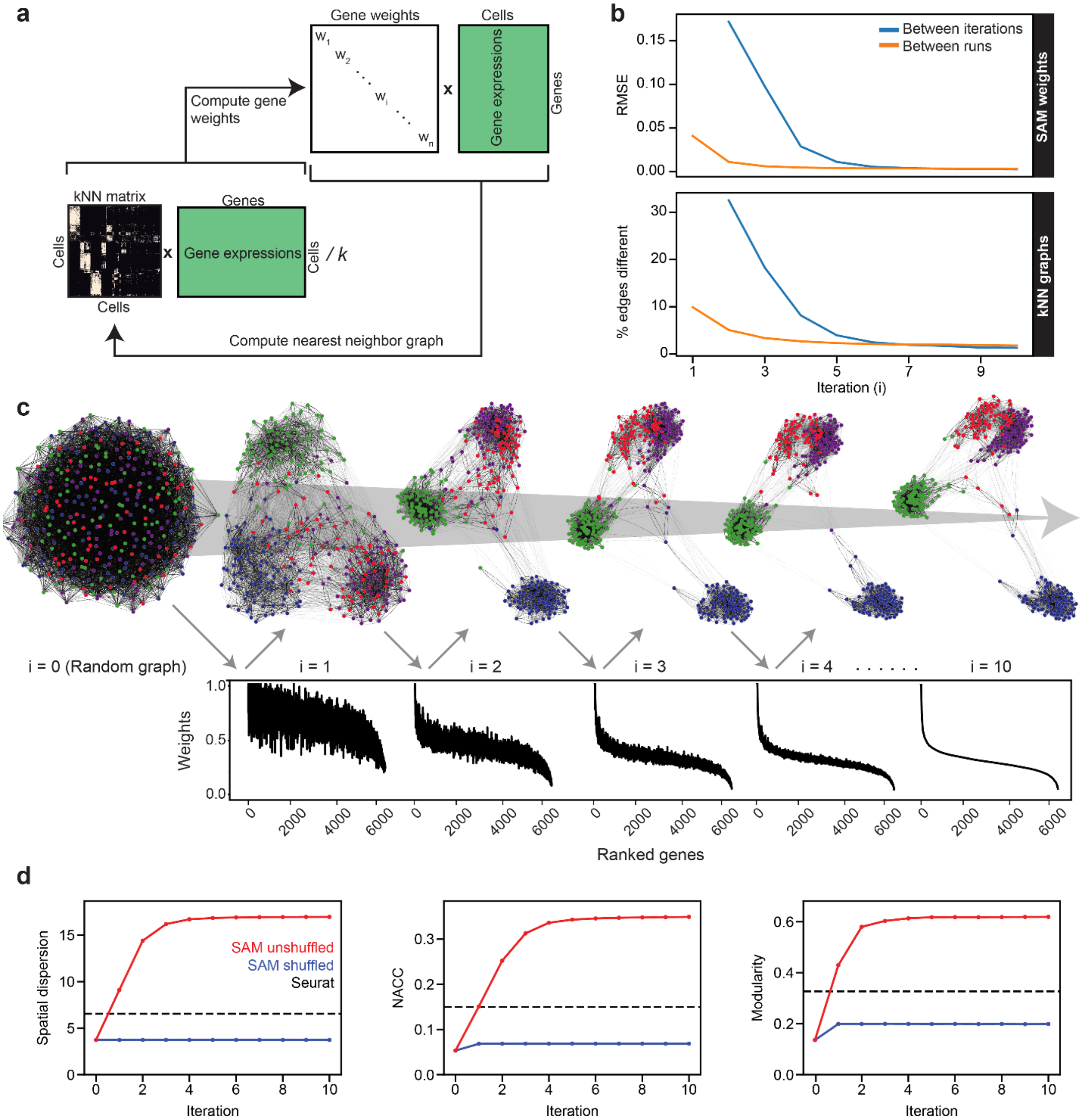
Figure 3 – Figure supplement 1. SAM converges to a stable solution independent of random initial conditions and is robust to the number of nearest neighbors and choice of distance metric. (**a**) RMSE of gene weights between adjacent iterations within a run, averaged across ten replicate runs. (**b-c**) Average ARI scores for the nine annotated benchmarking datasets when varying (**b**) the number of nearest neighbors, *k*, from 10 to 30 or (**c**) the choice of distance metric (Euclidean or Pearson correlation). Error bars indicate standard deviations of ARI scores across the different values of *k* and distance metrics. The errors for data with no error bars are too small to be seen.

Figure 6 – Figure supplement 1. Cluster-specific marker genes before and after removing cell cycle effects. UMAP projections with marker genes specific to the dividing cells (**a**) and the MT cluster (**b**) overlaid.

Figure 6 – Figure supplement 2. SAM groups cells based on NF-κB activation dynamics while other methods cannot. (**a**) UMAP projection of the macrophage cells after the removal of cell cycle effects. Cells with prolonged NF-κB dynamics are highlighted in red. (**b**) UMAP and t-SNE projections for Seurat and SIMLR, respectively, after the removal of cell cycle effects. Cells with prolonged NF-κB dynamics are highlighted in red. (**c**) UMAP projections with three MT-specific marker gene expressions overlaid.

910 **Figures**

911 **Figure 1.**



912

Figure 2.

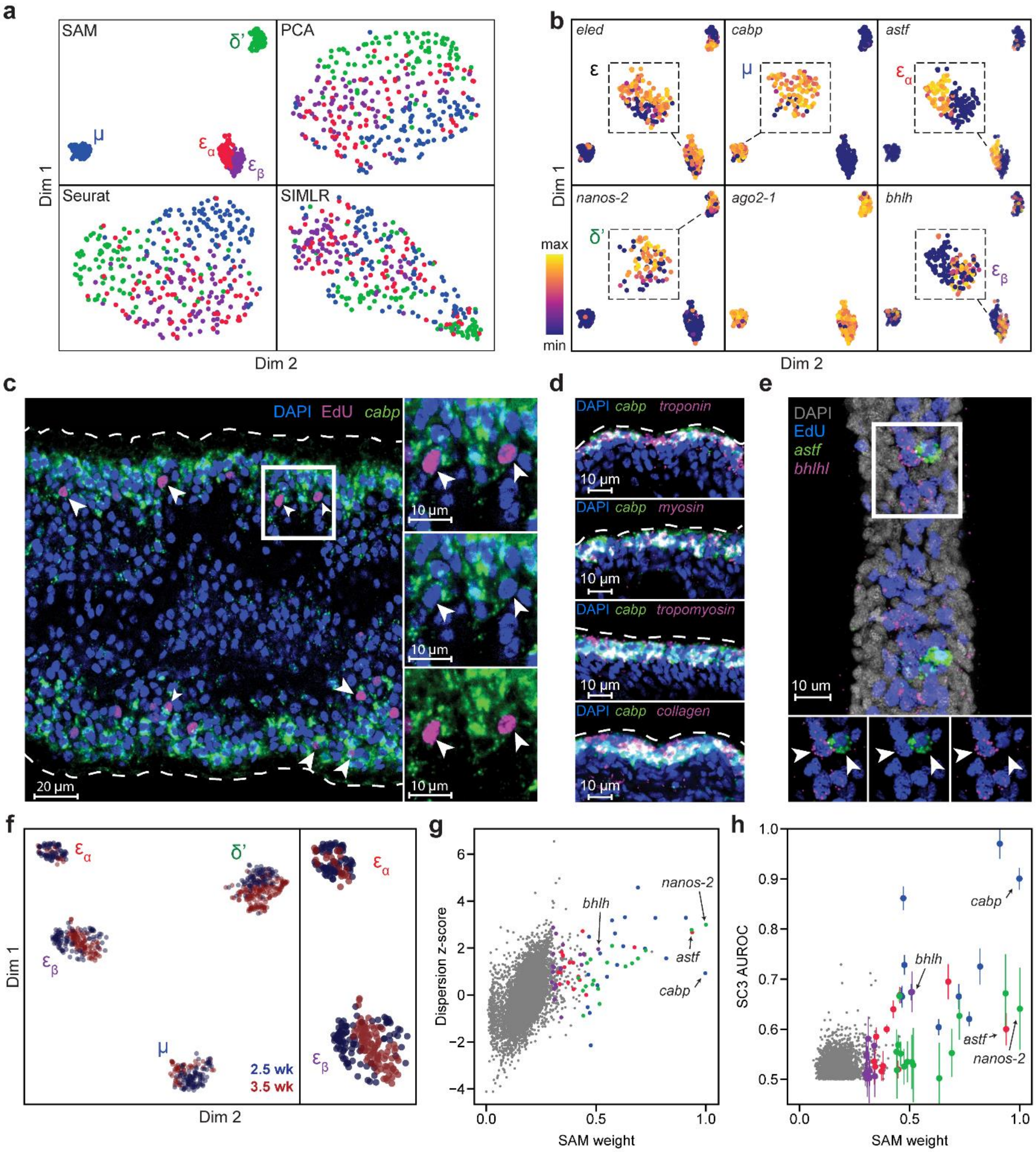
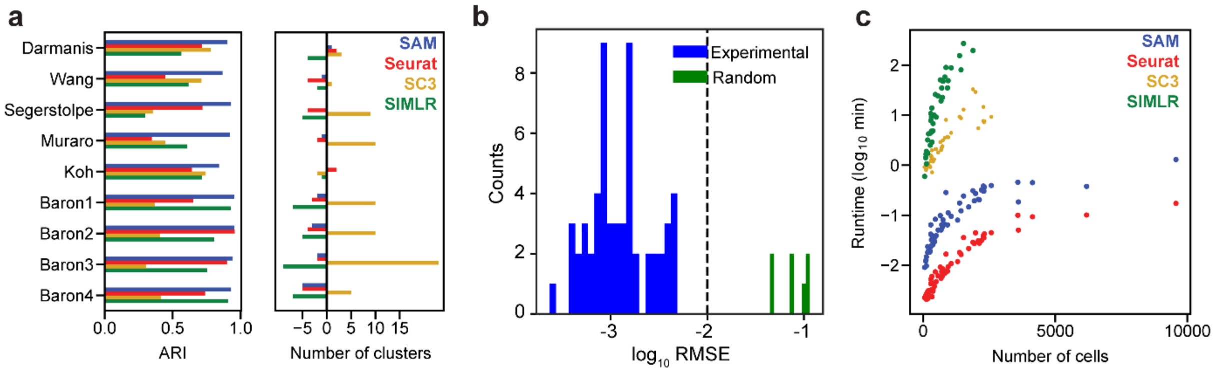
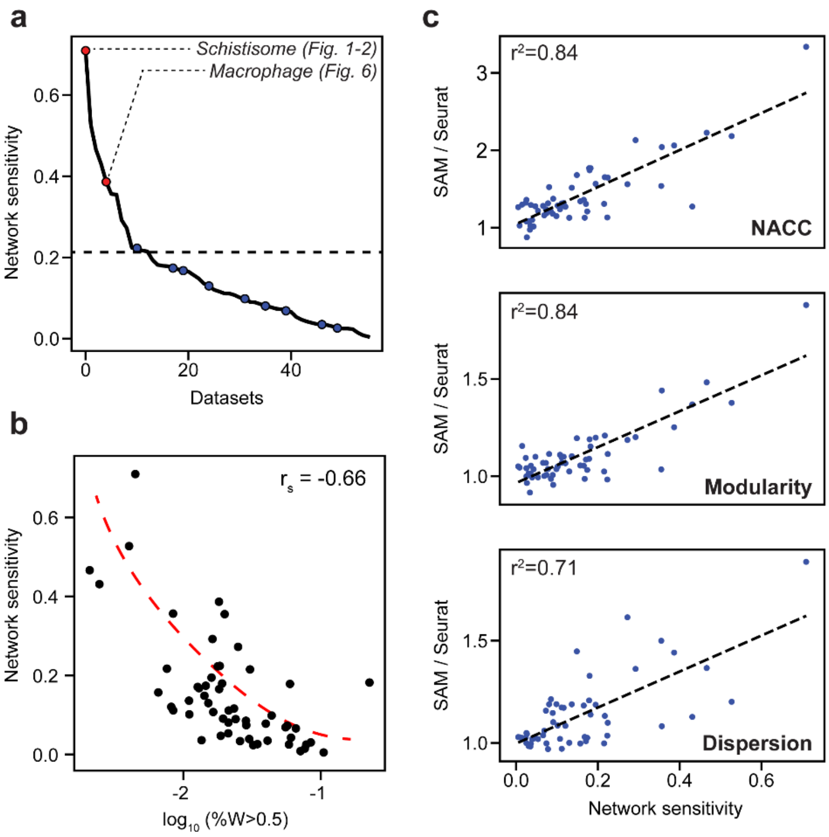

a
SAM
δ'
μ
εα
εβ
PCA
Seurat
SIMLR
Dim 1
Dim 2
b
eled
ε
cabp
μ
astf
εα
nanos-2
δ'
ago2-1
bhlh
εβ
max
min
Dim 1
Dim 2
c
DAPI EdU cabp
20 µm
10 µm
d
DAPI cabp troponin
DAPI cabp myosin
DAPI cabp tropomyosin
DAPI cabp collagen
10 µm
e
DAPI
EdU
astf
bhlhl
10 um
f
εα
δ'
εβ
μ
2.5 wk
3.5 wk
Dim 1
Dim 2
g
Dispersion z-score
SAM weight
nanos-2
bhlh
astf
cabp
h
SC3 AUROC
SAM weight
cabp
bhlh
astf
nanos-2

Figure 3.

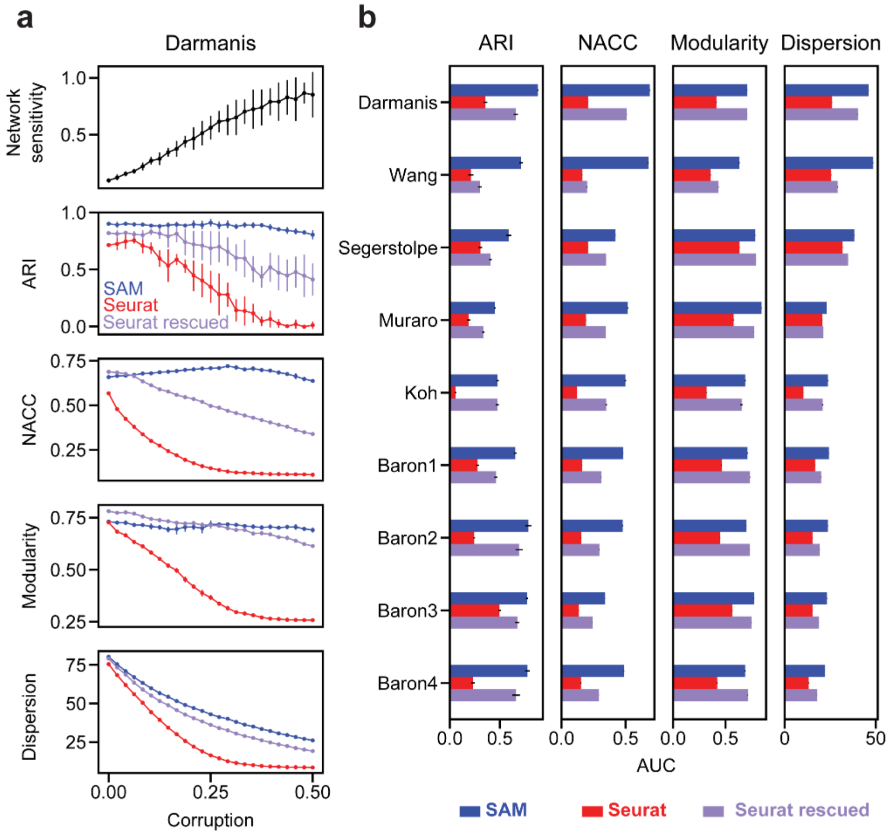

a
Darmanis
Wang
Segerstolpe
Muraro
Koh
Baron1
Baron2
Baron3
Baron4
0.0
0.5
1.0
ARI
−5
0
5
10
15
Number of clusters
SAM
Seurat
SC3
SIMLR
b
Experimental
Random
Counts
0
2
4
6
8
-3
-2
-1
$\log_{10}$ RMSE
c
SAM
Seurat
SC3
SIMLR
Runtime ($\log_{10}$ min)
2
1
0
−1
−2
0
5000
10000
Number of cells

917 **Figure 4.**

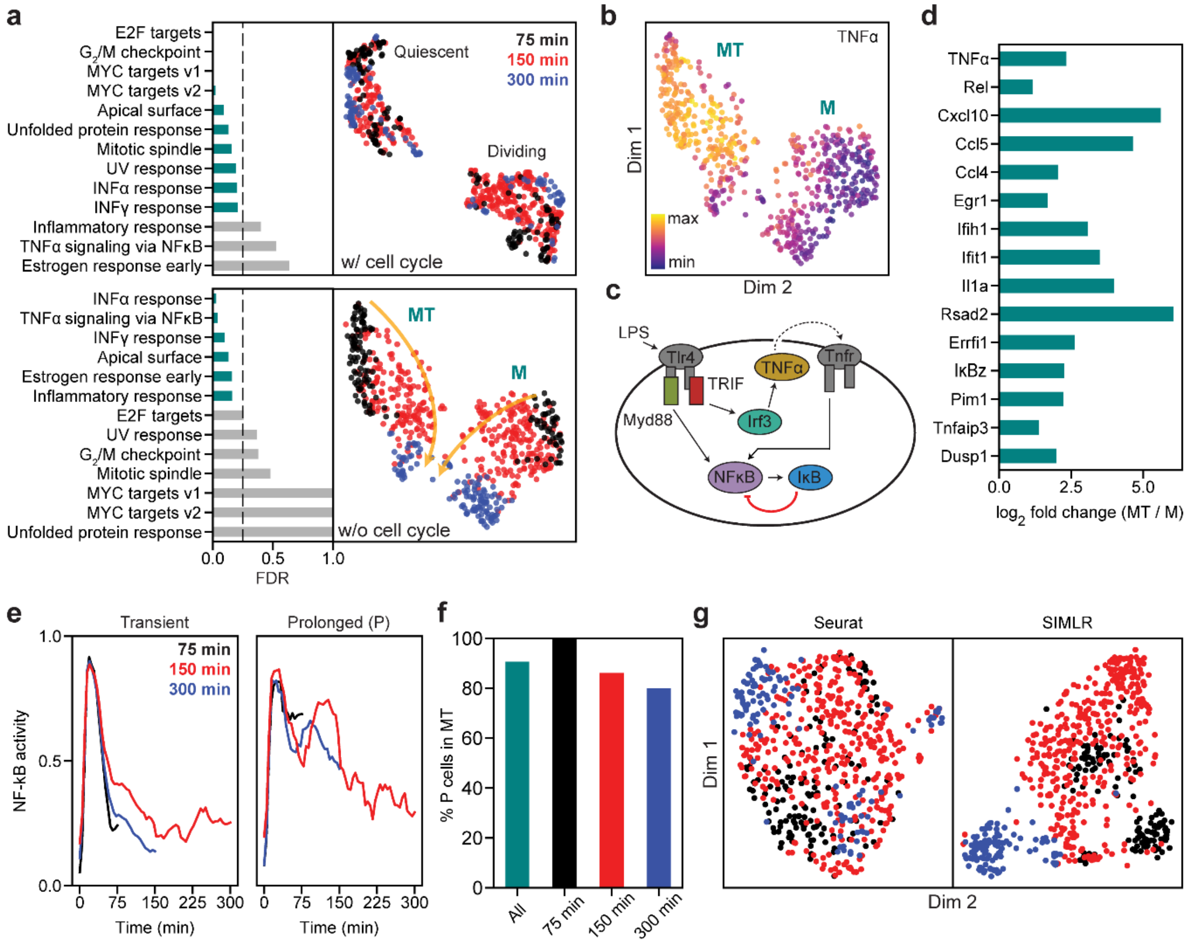


918

Figure 5.

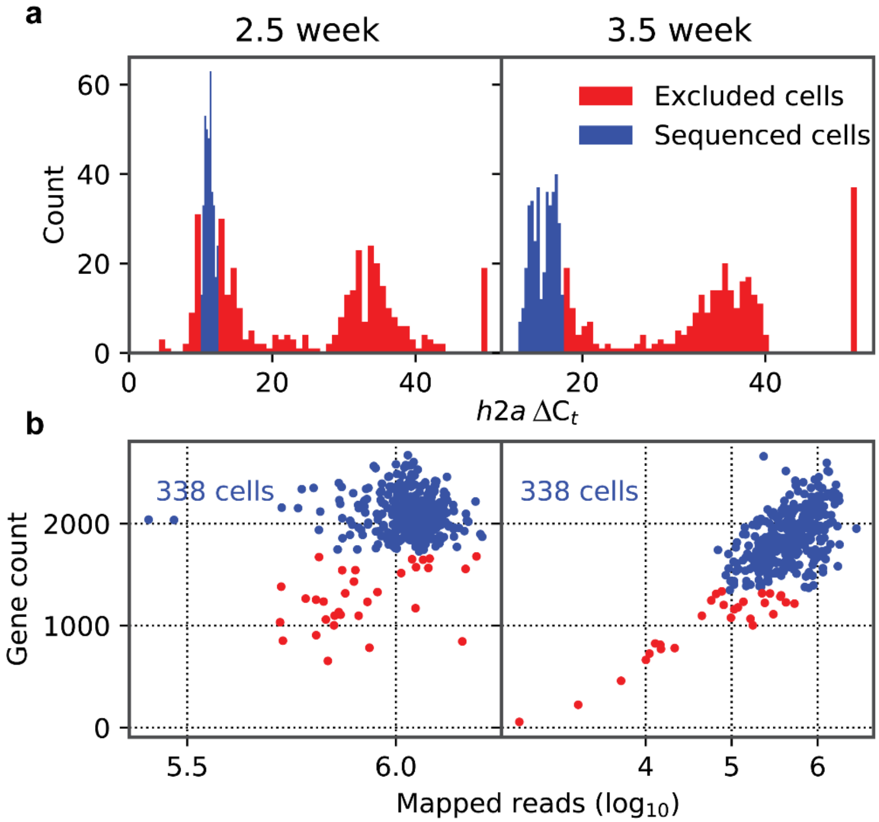

a
Darmanis
Network sensitivity
1.0
0.5
ARI
1.0
0.5
0.0
SAM
Seurat
Seurat rescued
NACC
0.75
0.50
0.25
Modularity
0.75
0.50
0.25
Dispersion
75
50
25
0.00
0.25
0.50
Corruption
b
ARI
NACC
Modularity
Dispersion
Darmanis
Wang
Segerstolpe
Muraro
Koh
Baron1
Baron2
Baron3
Baron4
0.0
0.5
0
50
AUC
SAM
Seurat
Seurat rescued

921 **Figure 6.**



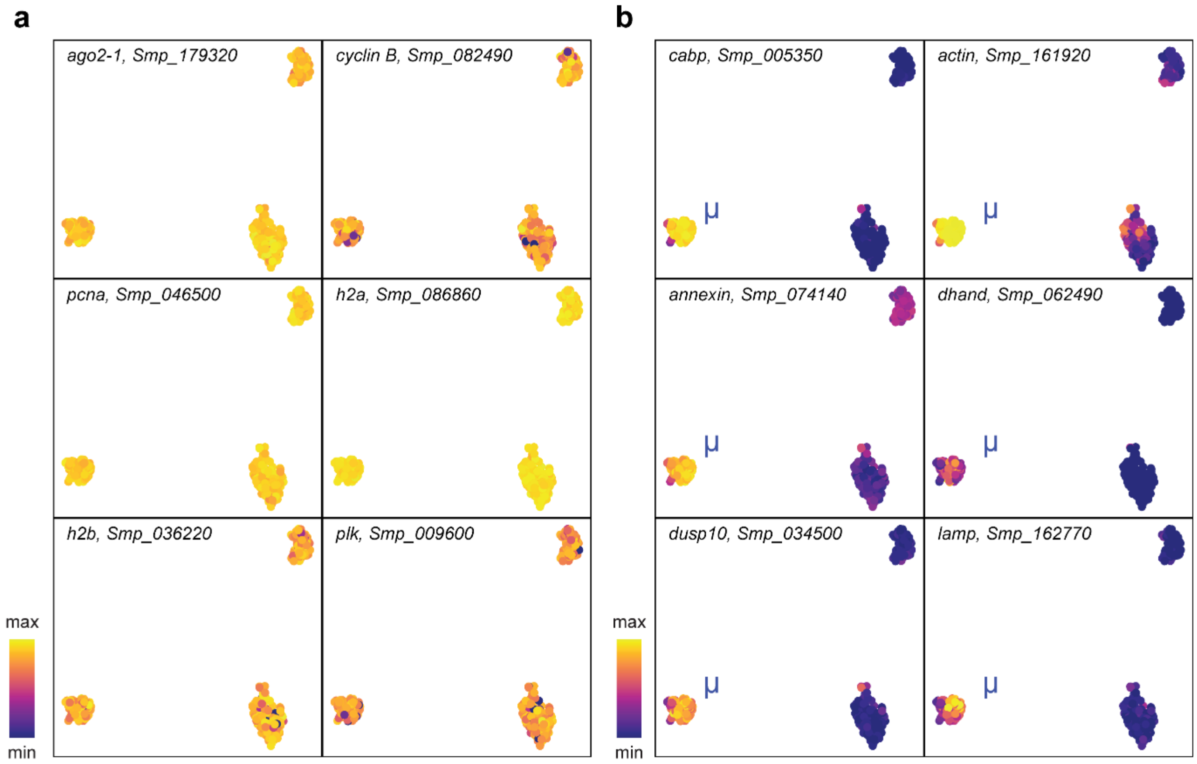
922

923 **Figure 1 – Figure supplement 1.**



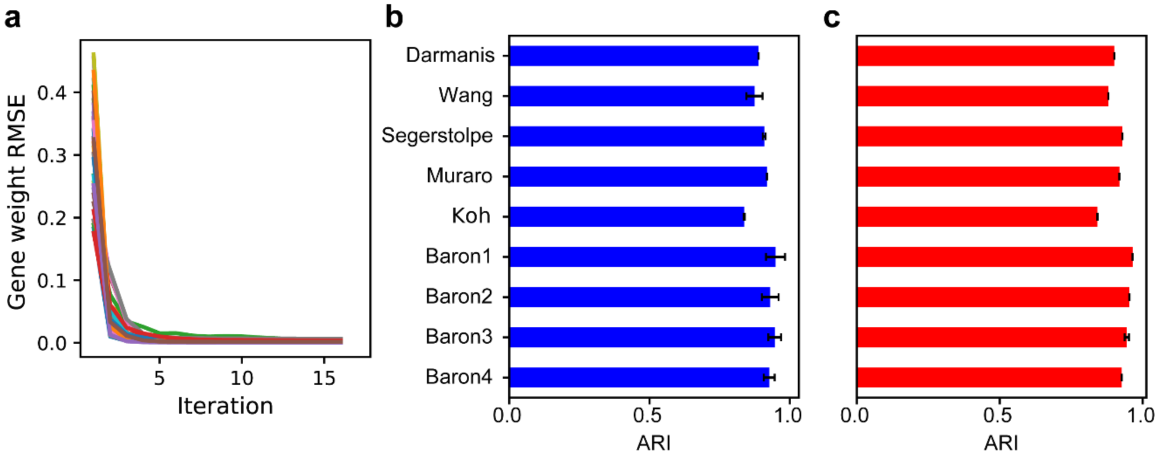
924

925 **Figure 2 – Figure supplement 1.**



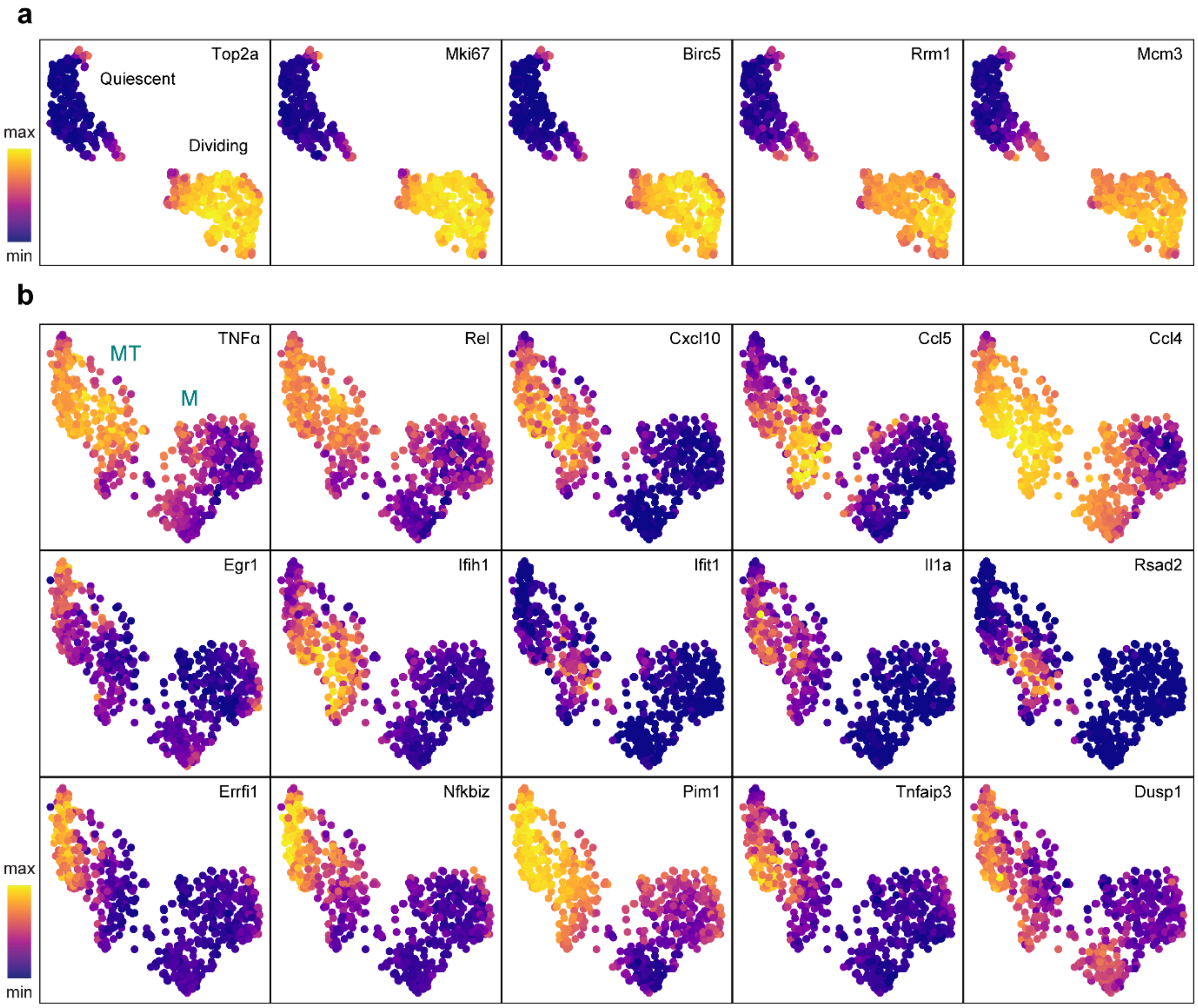
926

927 **Figure 3 – Figure supplement 1.**



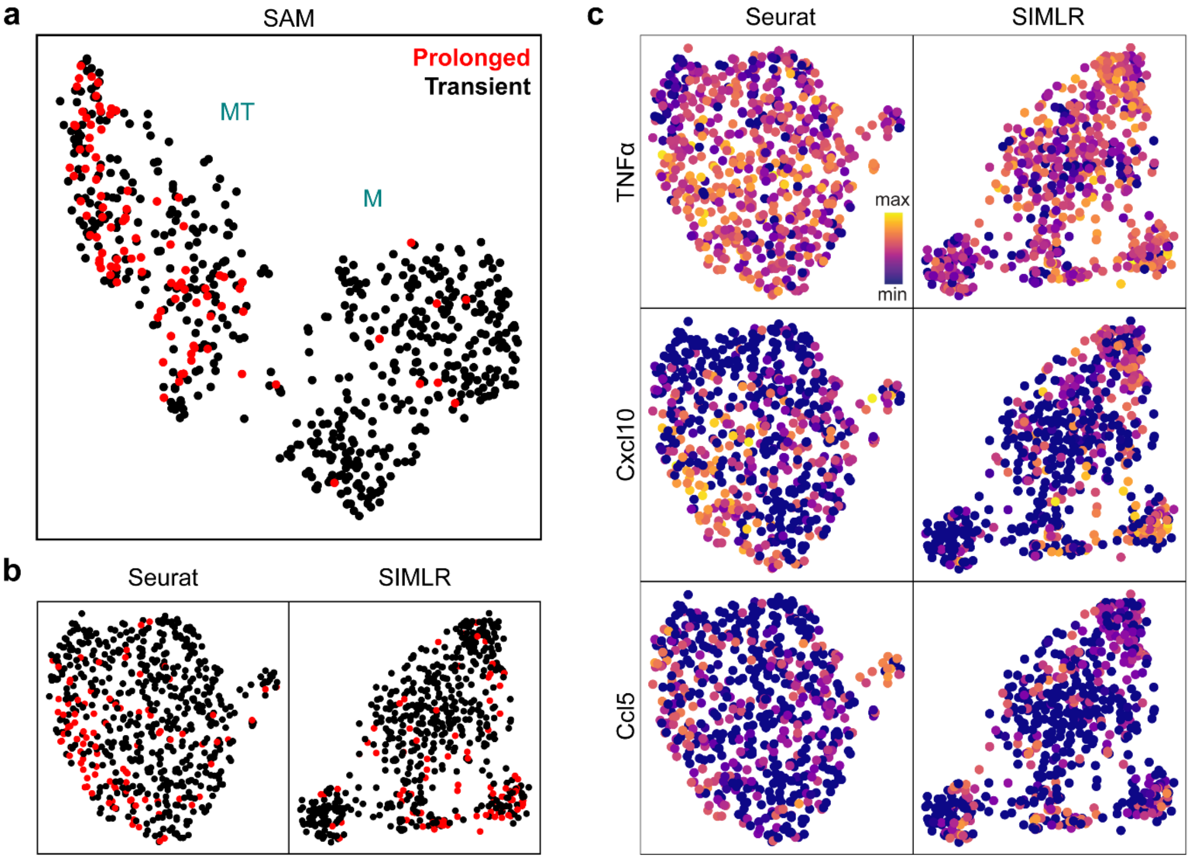
928

929 **Figure 6 – Figure supplement 1.**



930

931 **Figure 6 – Figure supplement 2.**



932

Supplementary Table legends:

Supplementary Table 1: Ranked gene list with high SAM weights in the schistosome stem cell data. Gene IDs and annotations are given in the *S. mansoni* genome version 9 (WormBase, WS268). Genes are assigned to the cluster corresponding to the marker gene, *nanos-2*, *cabp*, *astf*, or *bhlh*, that they have the highest correlation with. Genes found in our prior work[12] to be enriched in subsets of stem cells are specified.

Supplementary Table 2: Datasets used in this study. Accession numbers, library size normalization methods, data preprocessing methods, sensitivity scores, and corresponding references are provided for each dataset. Accession numbers with asterisks indicate datasets that are sourced from the *conquer* database (Soneson and Robinson, 2018). Accession numbers with crosses indicate the nine well-annotated datasets that were used for benchmarking.

Supplementary Table 3: Cloning primer sequences used for generating riboprobes for the FISH experiments. Functional annotations of the genes were given in the *S. mansoni* genome version 9 (WormBase, WS268).