Figures and figure supplements

Self-assembling manifolds in single-cell RNA sequencing data

Alexander J Tarashansky et al
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 Figure 1 continued on next page
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). The network properties converge to the same values when initializing SAM with the Seurat-reconstructed graph instead of a random graph (yellow). Dashed lines: metrics measured from the Seurat-reconstructed graphs.

DOI: https://doi.org/10.7554/eLife.48994.003
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 post infection 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 specified on the top left corner of each plot.

DOI: https://doi.org/10.7554/eLife.48994.004
Figure 1—figure supplement 2. A user interface for interactively exploring single-cell data using SAM. An interactive Jupyter notebook interface provided by the SAM package facilitates convenient visualization of single cell data (upper left) and changing of SAM parameters using various control panels (upper right, and bottom). This interface allows clustering, subclustering, visualizing of gene expression, and many other applications.
DOI: https://doi.org/10.7554/eLife.48994.005
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', and ε' 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...
Figure 2 continued

(cabp^+\text{EdU}^+, \text{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 colocalization in post-mitotic cabp^+ cells. Images in (c–d) are single confocal slices. (e) FISH of astf and bhlih 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_eA- and ε_eB-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).

DOI: https://doi.org/10.7554/eLife.48994.006
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.

DOI: https://doi.org/10.7554/eLife.48994.007
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 (N) and the number of annotated clusters (NTRUE). Smaller differences indicate more accurate clustering. Seurat* denotes Seurat analysis using parameters that maximize ARI. (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, Materials and 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 min.

DOI: https://doi.org/10.7554/eLife.48994.008
**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 for all datasets. (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.

DOI: https://doi.org/10.7554/eLife.48994.009
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 which features are selected (Materials and 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.

DOI: https://doi.org/10.7554/eLife.48994.010
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 randomly permute fractions of the data ranging from 0 to 100% of the total number of elements (Materials and methods). Performance is compared between SAM (blue), Seurat (red), Seurat with optimal parameters (black), 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.

DOI: https://doi.org/10.7554/eLife.48994.011
Figure 6. SAM captures the cellular activation dynamics in a stimulated macrophage dataset. (a) GSEA analysis (left) and UMAP projections (right) of the activated macrophages 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) log2 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 ($p<5\times10^{-15}$). (e) Representative traces for transient (left) and prolonged (right) NF-κB activation (Materials and 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.

DOI: https://doi.org/10.7554/eLife.48994.012
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.

DOI: https://doi.org/10.7554/eLife.48994.013
**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 for Seurat and SIMLR with three MT-specific marker gene expressions overlaid.

DOI: https://doi.org/10.7554/eLife.48994.014