SoptSC: Similarity matrix optimization for clustering, lineage, and signaling inference

Shuxiong Wang[1], Adam L. MacLean[1] and Qing Nie[1,2]

[1] *Department of Mathematics and Center for Complex Biological Systems, University of California, Irvine, CA 92697, United States*
[2] *Department or Developmental and Cell Biology, University of California, Irvine, CA 92697, United States*
Correspondence should be addressed to A.L.M (a.maclean@uci.edu) or Q.N. (qnie@uci.edu).

We present SoptSC, a similarity matrix optimization method for single-cell data analysis, which performs clustering, pseudotemporal ordering, lineage and marker gene identification from a cell similarity matrix. SoptSC also presents a new function: cell-cell signaling network inference, enabling the reconstruction of complex lineage relationships and associated feedback/feedforward interactions. As we show by application to several datasets, SoptSC can predict the number of clusters and the initial state unsupervised, and outperforms current methods for inference of clusters and pseudotime from single-cell data.

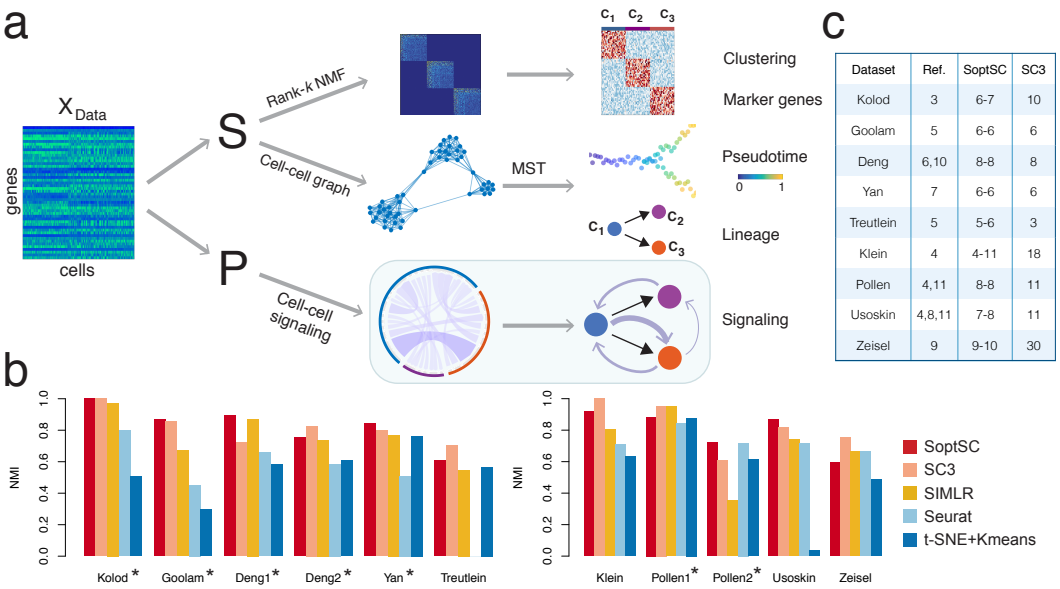
Our ability to measure the transcriptional state of a cell — and thus interrogate cell states and fates (1, 2) — has advanced dramatically in recent years (3) due in part to high-throughput single cell RNA sequencing (scRNA-seq) (4). This shift, permitting delineation of different sources of heterogeneity (1), requires appropriate dimension reduction techniques. Many approaches for cell clustering (5–7), pseudotemporal ordering (see (8) for review), and lineage inference (9–11) have been introduced, yet key challenges remain. These include parameter specification and method integration, i.e. the ability to perform multiple tasks from a single mathematical framework, allowing for consistent and coherent data analysis.

This becomes especially important in the context of multiscale analyses: below we introduce a method for inferring cell-cell communication, an inherently multiscale process involving molecular factors, cell identity/type and cell state transitions. Currently, disparate tools are used to infer low-dimensional quantities such as cell clusters, cell orderings, or lineage paths. Our goal, by providing a single framework to perform these tasks, is to be able to integrate higher-level information (on cell-cell communication) with these quantities, to reveal, for example, signaling interactions between clusters or along pseudotime.

Here we present SoptSC: similarity matrix-based optimization for single-cell analysis (**Supplementary Software** and https://github.com/WangShuxiong/SoptSC). SoptSC identifies cell subpopulations, marker genes, pseudotemporal ordering, and lineage paths between subpopulations without supervision, and introduces a new method to infer cell-to-cell signaling networks through specific ligand-receptor interactions. Coherence among multiple different tasks is achieved by centering analyses on the cell-to-cell similarity matrix S (**Fig. 1a**). To construct S, SoptSC takes as input a gene-cell data matrix and uses a low-rank representation model to capture both local and global structure of the data (12).

SoptSC clusters cells into subpopulations and predicts the number of clusters (k) from the data unsupervised, by construction of a truncated consensus similarity matrix, and analysis of eigenvalue spectra of the associated graph Laplacian, which gives lower and upper bounds for the number of clusters (see Methods). Clustering is performed by rank-k non-negative matrix factorization (NMF). Also given by the rank-k NMF (in the same step as clustering) is an ordered list of marker genes for each cluster, determined by the weight associated with each gene in each cluster. This method of marker gene identification differs significantly from most current methods based on differential gene expression.

To represent relationships between cells, a graph is constructed from S where nodes represent cells and edges are determined by the nonzero elements of S. This cell-to-cell graph preserves the intrinsic structure of the data, and does not require dimensionality reduction to mea-



Dataset	Ref.	SoptSC	SC3
Kolod	3	6-7	10
Goolam	5	6-6	6
Deng	6,10	8-8	8
Yan	7	6-6	6
Treutlein	5	5-6	3
Klein	4	4-11	18
Pollen	4,11	8-8	11
Usoskin	4,8,11	7-8	11
Zeisel	9	9-10	30

Fig. 1. SoptSC framework and assessment of clustering. (**a**) Overview of the SoptSC framework and outputs generated. (**b**) Five clustering methods are applied to a range of single-cell datasets where cell cluster labels were known or validated in the original studies. Normalized mutual information (NMI) is used to quantify accuracy. Datasets marked by an asterisk are annotated 'gold-standard' for comparison purposes. (**c**) Prediction of the number of clusters by SoptSC or SC3, compared to a reference number of clusters (Ref.) from the original study; SoptSC predicts both lower and upper bounds.

sure distance: a distance is defined on the graph as the shortest path length. A weighted cluster-to-cluster graph is constructed based on distances between constituent cell pairs, and the lineage hierarchy between clusters is inferred using the minimal spanning tree of the cluster-to-cluster graph. The longest path across the cluster-to-cluster graph identifies two end states, which provide two alternative proposals for the initial state. Since these two clusters are far apart, it is usually possible to identify which is the initial state via marker gene expression. The pseudotemporal ordering of cells is inferred directly from the cell-to-cell graph by calculating distances between an initial cell and all other cells. The initial cell is inferred within the initial cluster such that the pseudotime and lineage path display the highest concordance.

We present a new method to study cell-cell signaling at single-cell resolution: whereas previous methods have focused on signaling at the level of cell clusters (13–15), SoptSC infers cell-cell signaling relationships based on single-cell gene expression for a pathway of interest. A signaling probability is defined based on weighted co-expression of signaling pathway activity in sender-receiver cell pairs (see details in Methods). As input, the user provides a ligand (or set of ligands) and cognate receptor (or set of receptors), for example, ligands from the Wnt family and Frizzled receptors. For each pathway, a set of target genes is also specified: a candidate list of genes that are known to be differentially regulated downstream of a ligand-receptor interaction, along with their sign, i.e. upregulated or downregulated. SoptSC computes the signaling probability between sender cell (expressing ligand) and receiver cell (expressing receptor and differential target gene activity). These single-pathway single-cell signaling probabilities are combined to produce summaries and determine higher-level (e.g. cluster-to-cluster consensus) signaling networks. Combining the consensus signaling networks with the lineage path allows SoptSC to infer feedback or feedforward interactions mediated by signaling factors.

To assess the methods of SoptSC, we first compare it with four current clustering methods: SC3 (5), SIMLR (6), Seurat (7), and tSNE (18) followed by k-means clustering (19) (tSNE + k-means). Nine published scRNA-seq datasets are analyzed for which the 'true' labels are known; five of these have been annotated as 'gold-standard' datasets, i.e. the clusters (cell types) have been verified (**Supplementary Table 1**). As a test statistic, we use the normalized mutual information (NMI, (20)) between predicted and verified cluster labels. We find that for the gold standard scRNA-seq data sets, SoptSC outperforms all other methods (**Fig. 1b**); for the four additional datasets, SoptSC outperforms all methods except SC3, with which it is comparable. In **Fig. 1c** a comparison is shown of the number of clusters predicted for each dataset by SoptSC, SC3, and the reference number from the original study ('Ref.'). In $6/9$ cases SoptSC is in agree-

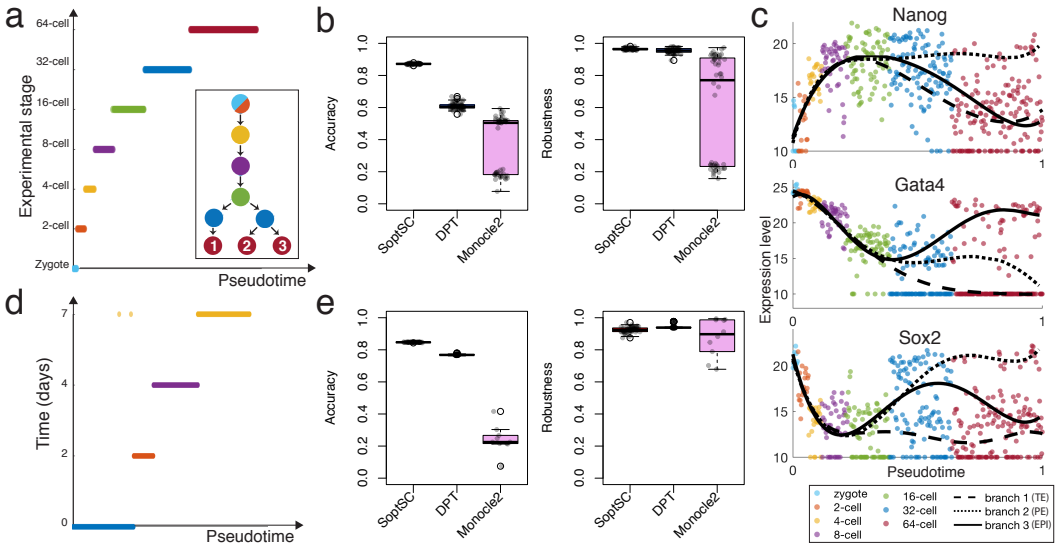


Fig. 2. Assessment of SoptSC for pseudotime inference. (**a**) Pseudotemporal ordering of single-cell data from mouse early embryo development (16) is compared with the known experimental stage. Inset shows the lineage inferred by SoptSC, colored by experimental stage of origin for each cluster. (**b**) Comparison of methods for pseudotime inference on mouse embryonic stem cell data (16) using the Kendall rank correlation between pseudotime and experimental stage as a measure of accuracy, and by subsampling 90% of cells from the data 50 times (and comparison of subsets) to measure robustness. (**c**) Marker genes identified for (16) are plotted along pseudotime; lines correspond to polynomial regression for each branch (see lineage branch numbers in part a). TE: Trophectoderm; PE: Primitive Endoderm; EPI: Epiblast. (**d**) Pseudotemporal ordering of data from (17) compared with experimental time. (**e**) Comparison of methods for (17) via accuracy and robustness measures (as for part b).

ment with the reference number of clusters ± 1. For SC3 similar agreement is seen in $4/9$ cases. SoptSC predicts lower and upper bounds for the number of clusters using the truncated graph Laplacian where the lower bound is set as the number of zero eigenvalues and the upper bound comes from the index in which the largest eigengap occurs (**Supplementary Fig. 1**). By default SoptSC uses the upper bound for inference; we have found that this is often the better estimate. We hypothesize that this may show that significantly complex structures exist within these data, which are best represented by the larger of the two estimates (the upper bound).

Assessment of pseudotemporal ordering is challenging as there are few example of data for which the true ordering is known or reliably estimated. However, correlation between pseudotime and experimental time provides means by which quantitative assessment can be made. We compared SoptSC to diffusion pseudotime (DPT) (21) and Monocle2 (22) using the Kendall rank correlation between pseudotime and experimental time as a test statistic (following the same criteria as (21)). To measure robustness, we subsample the data and evaluate the correlation for each subset of the data. For assessment we use two datasets that studied embryogenesis and embryonic stem cell differentiation, respectively (16, 17) (**Supplementary Figs. 2-3**).

SoptSC produces pseudotemporal trajectories (**Fig. 2a** and **d**) that are more strongly correlated with real experimental time that DPT or Monocle2 (**Fig. 2b** and **e**). Under subsampling, SoptSC displays robustness greater than Monocle2 and comparable to DPT, thus indicating overall that, by the criteria used, SoptSC is more accurate, and at least as robust, as the alternative methods tested. SoptSC can also resolve branch-specific marker gene dynamics along the pseudotemporal axis: three marker genes for SoptSC-identified clusters from (16) are plotted (**Fig. 2c**), showing clearly distinguishable signatures for each branch.

Following these tests of the accuracy and robustness of our method, we applied SoptSC to several other datasets (23–25) to further assess its ability to clusters cells and order them along pseudotime. We also analyze the marker genes identified for cell clusters and the predicted lineage and signaling relationships. Two studies of early embryo development (16, 24) were both found to be highly consistent with known developmental stages of early embryos. In particular, for (24) SoptSC is able to identify relevant cell clusters, lineage, and pseudotime even for a very small dataset containing only 88 cells (**Supplementary Fig. 8**). For (16), SoptSC correctly predicted two branch points in the lineage corresponding to the formation of the early blastocyst (**Supplementary**

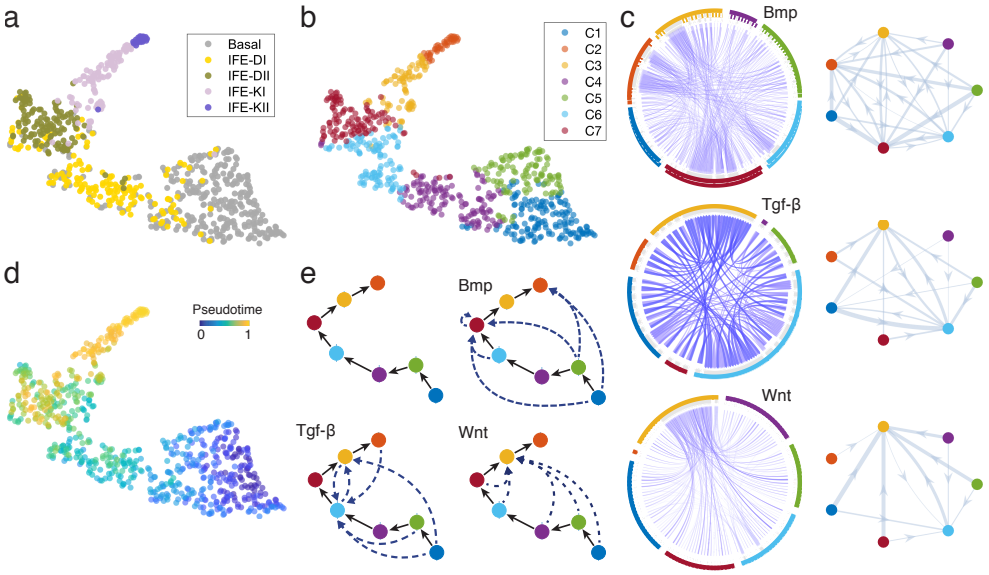


Fig. 3. Inference of epidermal lineage and signaling networks. (**a**) Cells from (23) projected into low dimension by SoptSC and colored by the cluster labels from the original study. (**b**) Cells colored by cluster labels identified by SoptSC. (**c**) Single-cell signaling networks predicted for three pathways. Left: samples from full networks where edge weights represent the probability of signaling between cells. Right: cluster-to-cluster signaling interactions where edge weights represent sums over within-cluster interactions. Colors correspond to cluster labels from part b. (**d**) Pseudotemporal ordering of cells. (**e**) SoptSC infers a linear lineage from basal to differentiated epidermal cells (top left). Summaries of the cluster-to-cluster signaling interactions with highest probability are given for the Bmp, Tgf-β, and Wnt pathways.

Fig. 5).

Applied to adult mouse skin homeostasis (23), SoptSC predicts seven subpopulations, three of which are basal, and infers a linear lineage between the basal and subsequent differentiated populations of the epidermis (**Fig. 3)**, corresponding to decreases in potency (26). We found that the marker genes identified for each subpopulation overlapped significantly with known markers of epidermal differentiation (**Supplementary Fig. 7**). We have also analyzed marker gene expression for several other datasets (**Supplementary Figs. 6F, 8G, 9F**) and find that marker genes to distinguish the various clusters can be consistently detected; these include previously known markers of cell lineages as well as novel markers.

We next study single-cell signaling networks defined by ligand-receptor pairs, to infer cell-cell communication. Using data from Joost et al. (23), the single-cell signaling networks for Bmp, Tgf-β, and Wnt show distinct temporal patterns (**Fig. 3C and Supplementary Figs. 10-13**). Given that we have prior knowledge of the structure and locations of cell types within the epidermis, we are also able to infer spatial patterns from signaling networks. We find that the strongest Bmp signals activate cells in the mid-differentiated epidermal population (C_7) marked by Krt77 and Ptgs1. The Tgf-β pathway is activated in populations C_6 and C_3, which mark late basal/early differentiated cells, and late differentiated cells, respectively. Wnt signaling is predicted to be active specifically in C_3. Thus this signaling network inference predicts temporally constrained signaling pathway activation: Tgf-β is activated earliest during epidermal differentiation, followed by Bmp. Wnt signaling activity is predicted to be important only subsequent to signals from these other two pathways, thus acting farthest from the basal membrane.

SoptSC was also applied to a dataset on myelopoiesis that displays significant complexity, containing hematopoietic subpopulations that are heterogeneous, mixed, and novel (25). SoptSC identifies 8 subpopulations within these data (**Supplementary Fig. 9**) and infers a hierarchical lineage with four branches (**Supplementary Fig. 9D**). This lineage displays finer grain detail than previous analyses of these data (22), where only two branches (corresponding to granulocytes and monocytes) were identified. The lineage identified by SoptSC contains stem/progenitor subpopulations constituting their own branch; thus acting as a pool of stem cells (that may be quiescent); distinct branch signatures can be resolved along pseudotime (**Supplementary Fig. 9E**). The heterogeneous nature of these data is reflected in the gene expression heatmap (**Supplementary Fig. 9F**), where certain clusters are hard to distinguish. However even for these data, we see that several known hematopoietic markers are

identified as top markers. Inference of the single-cell signaling networks predicts subpopulation-specific activation of Tgf-β in hematopoietic stem/progenitor cells (**Supplementary Figs. 14-16**).

The optimization framework of SoptSC naturally leads itself to extensions by the inclusion of other input data, for example on methylation patterns (bisulfite sequencing) or chromatin accessibility (ATAC-seq). In summary, SoptSC performs multiple single-cell analysis tasks — identification of clusters, marker genes, pseudotime, and lineage path — from a single input (the similarity matrix), with minimal parameter specification. This enables meaningful comparisons between data representations that were not previously available, and inference of cell-cell communication networks from single-cell data.

Methods

Cell-to-cell similarity matrix construction

The input to SoptSC is a single cell gene expression matrix X with m rows (associated with genes), and n columns (associated with cells). SoptSC computes the coefficient matrix Z from X by the following optimization model:

$$\begin{aligned} \min_{Z} \quad & ||Z||_* + \lambda ||X - XZ||_{2,1} \\ s.t. \quad & Z^\top \mathbf{1} = \mathbf{1}, \\ & Z_{i,j} = 0 \quad \text{for} \quad j \notin N_i, \end{aligned} \tag{1}$$

where $||\cdot||_*$ is the nuclear norm; $||\cdot||_{2,1}$ is the $l_{2,1}$ norm (the sum of the Euclidean norm of all columns); λ is a non-negative parameter, $\mathbf{1} = (1, ..., 1)^\top$ is a vector of length n and N_i is the set of neighbors of cell i. To compute N_i, cells are projected into low dimensional space using t-distributed stochastic neighbor embedding (t-SNE) (27), and a K-nearest neighbors algorithm (28) is then applied to the low-dimensional data. The linear constraint $Z^\top \mathbf{1} = 1$ guarantees translational invariance of the data (29).

The optimization model (1) is a representation method for the construction of graphs from nonlinear manifolds (12). Informally, it captures relationships between cells by representing each cell as a linear combination of all other cells. By restricting coefficients of non-neighboring cells to be zero, the model preserves the local structure of the linear representation. By imposing the low rank constraint, this model can capture the global structure of the original data input, and is more robust to noise and outliers. Problem (1) can be solved numerically by the alternating direction method of multipliers (12). Letting Z^* be the optimal solution of (1), then via symmetric weights we can define the similarity matrix S as

$$S = \max\left\{|Z^*|, |Z^{*\top}|\right\}. \tag{2}$$

The elements $S_{i,j}$ $(= S_{j,i})$ of S thus quantify the degree of similarity between cell i and cell j.

Symmetric NMF for cell clustering

To cluster cells based on their similarity, we use symmetric non-negative matrix factorization (NMF) (30, 31) where the (non-negative) similarity matrix S is decomposed into the product of a non-negative low rank matrix $H \in \mathbb{R}_+^{n \times k}$ and its transpose $H^\top$ via the optimization problem:

$$\min_{H \in \mathbb{R}^{n \times k}} ||S - HH^\top||_F^2 \qquad \text{s.t.} \quad H \geq 0, \tag{3}$$

where $||\cdot||_F$ is the Frobenius norm and k is the reduced rank of S. The solution of (3) satisfies the condition that the similarity matrix S can be represented approximately by $HH^\top$. Thus, the columns of H represent a basis for S in the (low rank) k-dimensional space, and the columns of $H^\top$ provide the coefficients for corresponding columns of S, in the k-dimensional space spanned by the columns of H. Since $H \geq 0$, each column of $H^\top$ can be viewed as a distribution for which the ith column S^i has the component in the corresponding column of H. We can use H^{T} to classify the N cells into k subpopulations by assigning the ith cell to the jth subpopulation when the largest element among all components of the ith column of H^{T} lies in the jth position. The clustering procedure based on the similarity matrix S is such that cells within a group have high similarity to each other and low similarity to cells from other groups. (Details can be found in supplementary materials).

Marker gene identification

The non-negative low rank matrix derived from (3) not only produces cell cluster labels, but also marker genes for each cluster, by combining cluster labels with the data matrix in the following way. The element $H_{i,j}$ represents the weight by which cell i belongs to the jth cluster. Then the jth column of H defines a distribution over all cells corresponding to the jth cluster. It follows that a weight for gene v corresponding to the jth $(1 \leq j \leq k)$ cluster can be defined by

$$\omega(v, j) = \sum_{i=1}^{n} X_{v,i} H_{i,j},$$

where $X_{v,i}$ represents the vth gene expression value in cell i. These weights for gene–cluster pairs measure the significance attributed to a given gene in each cluster, thus providing a means to determine how well gene v delineates cluster j from all other clusters. Marker genes are then defined as follows: gene v is a marker for cluster j, if $\omega(v, j)$ reaches its largest value in cluster j, i.e. $\omega(v, j) = \max_{1 \leq u \leq k} \{\omega(v, u)\}$.

Prediction of the number of clusters within a dataset

Prediction of the number of clusters is based on construction of a truncated consensus matrix and analysis of the spectrum of this matrix (32, 33). The number of clusters k is varied in a range $\{2, 3, ..., N\}$, and symmetric NMF is then performed for each k to identify k clusters. A consensus matrix C is defined such that $C_{i,j}$ represents the number of times that cell i and cell j are classified as belonging to the same cluster. To improve robustness of

this estimation to noise, we prune C by setting $C_{i,j} = 0$ if $C_{i,j} \leq \tau(N-1)$ where $\tau \in [0, 0.5]$ is the tolerance. The graph Laplacian $\mathbb{L} = I - D^{-1/2}CD^{-1/2}$ is computed, where I is the identity matrix, and D is a diagonal matrix of the row-sums of C (e.g., $D_{i,i} = \sum_j C_{i,j}$).

We estimate the number of clusters within a range, by giving a lower bound as well as a upper bound. The lower bound is computed as the number of zero eigenvalues of $\mathbb{L}$, and the upper bound is computed as equal to the index at which the largest eigenvalue gap of $\mathbb{L}$ occurs. It has been shown that the number of eigenvalues of $\mathbb{L}$ equal to 0 is equivalent to the number of diagonal blocks of $\mathbb{L}$ (32). By default we use the upper bound (the largest eigenvalue gap) as the estimate for the number of clusters. (Details can be found in supplementary materials).

Inference of pseudotime and cell lineage

To infer the cell lineage and a temporal ordering of cells, we construct a cell-to-cell graph $G = (V, E, A)$ where V is the vertex set of cells, and E is the edge set described by the adjacency matrix A. The adjacency matrix A is a binary matrix derived from the similarity matrix S: $A_{i,j} = 1$ if $S_{i,j} > 0$ and $A_{i,j} = 0$ otherwise. We define the distance between cells on the graph G as the length of the shortest path between two cells on the graph, i.e. $\hat{D}_{p,q}$ represents the length of the shortest path between cells p and q on the graph G.

Given a set of clusters $u = \{u_1, u_2, ..., u_k\}$ (i.e. a partitioning of cells into k distinct groups), the distance between cell clusters u_i and u_j is computed by

$$D^u_{i,j} = \frac{\sum_{p \in u_i, q \in u_j} \hat{D}_{p,q}}{|u_i||u_j|},$$

where $|u_i|$ represents the number of cells in cluster u_i. $D^u_{i,j}$ measures the average shortest path based distance between cluster u_i and u_j. A cluster-to-cluster graph $G^u = (V^u, E^u, D^u)$ is constructed where each node in V^u represents a cell cluster and E^u is the edge set described by D^u. The lineage tree, describing the cell state transition path, is inferred by computing the minimal spanning tree (MST) of graph G^u. If the initial state is provided in advance, we construct MST by setting the root as the initial state. Otherwise, we provide prediction of the initial state (u_0) by selecting the state that maximizes the path length over the MST.

Pseudotemporal ordering of cells is inferred by finding and sorting the shortest path lengths on the cell-to-cell graph G between each cell and the initial cell. An initial cell c_0, if not provided in advance, is defined such that the temporal ordering of cells and lineage tree have the highest concordance. This is achieved as follows: for each cell $c_k, k \in u_0$ (i.e. cells clustered in the initial state), we compute $\hat{D}_{c_k,i}$, the shortest path length between c_k and all other cells. By taking the mean shortest path length between c_k and $c_l, l \in u_i$ for each cluster u_i, a mean distance from the initial cell to each cluster is defined. The Kendall rank correlation between the set of mean distances and the relative positions of states according to the lineage tree is then calculated. The initial cell is defined as that which produces the highest correlation by this test statistic.

Pathway-mediated cell-cell signaling network inference

In order to study how paracrine signals are sent from and received by single cells, we implement a method to predict cell-cell signaling networks mediated by specific ligand-receptor interactions. Directed edges are drawn between two cells where a high probability of signaling is predicted by the expression of ligand in a "sender" cell, and the expression of its cognate receptor in a "receiver" cell along with appropriate expression of target genes of the pathway in the receiver cell. While such probabilities are not fully sufficient to define an interaction between a pair of cells, they represent necessary conditions for signaling, and can be indicative of spatial proximity of cells within a sample. Whereas previous works (13–15) have considered signaling activity by summing over cells within a given cluster, we seek to account for the heterogeneity between cells within the same cluster.

For a given pathway, e.g. the Wnt signaling pathway, we define a set of ligands as the protein products of the Wnt gene family, and a set of receptors as the Frizzled (Fzd) proteins that bind Wnts. Also necessary as input to this signaling inference method are a set of target genes affected by Wnt along with their sign, i.e. upregulated in response to Wnt, or downregulated in response to Wnt. Currently, the sets of target genes used are small and defined manually; while automation of this step is possible, it would introduce another level of uncertainty to be accounted for.

The probability of signal passing between two cells is then computed as follows. Suppose we have a ligand-receptor pair, the expression of which is given by the distributions $L \in \mathbb{R}^n$ and $R \in \mathbb{R}^n$ for ligand and receptor respectively, where n is the number of cells. Let $Y = [Y_{i,j}] \in \mathbb{R}^{m_1 \times n}$ define the expression of m_1 genes that are upregulated by Wnt, and $Y^* = \left[Y^*_{i,j}\right] \in \mathbb{R}^{m_2 \times n}$ define the expression of m_2 genes that are downregulated by Wnt. The probability that a signal is sent from cell i to cell j via this pathway is then given by:

$$P_{i,j} = \frac{\exp(-\frac{1}{L_i R_i}) K_{i,j} \exp(-\frac{m_1}{\sum_{v=1}^{m_1} Y_{v,j}}) \Lambda_{i,j} \exp(-\frac{\sum_{v=1}^{m_2} Y^*_{v,j}}{m_2})}{\sum_k \alpha_{i,k} K_{i,k} \beta_k \Lambda_{i,k} \gamma_k} \tag{4}$$

where:

$$K_{i,j} = \frac{\alpha_{i,j}}{\alpha_{i,j} + \beta_j}, \quad \Lambda_{i,j} = \frac{\gamma_j}{\alpha_{i,j} + \gamma_j}, \tag{5}$$
$$\alpha_{i,j} = \exp\left(-\frac{1}{L_i R_j}\right), \quad \beta_j = \exp\left(-\frac{m_1}{\sum_{v=1}^{m_1} Y_{v,j}}\right),$$
$$\gamma_j = \exp\left(-\frac{\sum_{v=1}^{m_2} Y^*_{v,j}}{m_2}\right).$$

Summaries of signaling networks and cluster–to-cluster signaling

Given a ligand-receptor pair for a specific signaling pathway, the signaling network inferred is given by the graph $G = (V, P)$, where V is the set of all cells, and $P = [P_{i,j}] \in \mathbb{R}^{n \times n}$ defines the probability of a signal being passed from cell i to cell j (Eqns. 4 and 5). For visualization of these networks we use the *circlize* package in R (34). A number of summary statistics derived from the probability matrix P can be defined, and are useful in various contexts.

Consensus over signaling pathways. For a series of ligand-receptor pairs $\{\text{Lig}_r, \text{Rec}_r; r = 1, 2, ..., N\}$, the corresponding signaling networks $G^r = (V, P^r)$ are constructed based on each probability matrix P^r, i.e. P^r defines the probability matrix for ligand-receptor pair $\{\text{Lig}_r, \text{Rec}_r\}$. Then the overall probability of signaling summed over cells, P^{tot} is given by:

$$P^{tot} = \frac{\sum_{r=1}^{N} P^r}{N}.$$

Consensus over cells (cluster-to-cluster signaling). It is also informative to consider the cluster-to-cluster signaling networks in order to predict where feedforward/feedback interactions may occur, and to compare with previous methods for cell-cell signaling study that have focussed on cluster-level signaling (13). Let $u = \{u_1, u_2, ..., u_k\}$ give a clustering of cells by assigning each cell to one of k clusters. Then the probability of a signal passed between cluster u_l and u_m, mediated by a given ligand-receptor pair, is given by:

$$P^u_{l,m} = \frac{\sum_{p \in u_l, q \in u_m} P_{p,q}}{|u_l||u_m|}.$$

Software availability

SoptSC is available on GitHub as a MATLAB (Naticks, MA) package at: https://github.com/WangShuxiong/SoptSC.

Data availability

All the datasets used in this paper are collected from the published accession numbers provided in the original studies. The datasets used for evaluating the performance of clustering (**Figs. 1b** and **1c**) are downloaded from https://hemberg-lab.github.io/scRNA.seq.datasets/. Other datasets are available from: mouse early embryo (16) (Table S4): www.sciencedirect.com/science/article/pii/S1534580710001103; human early embryo (24): GSE36552; adult skin maintenance (23): GSE67602; and myelopoiesis (25): GSE70245.

Acknowledgements

We thank H. Singh for helpful discussions on hematopoiesis and on the data described in Olsson et al. (25). We acknowledge support from NSF grant DMS1562176, NIH grants R01GM107264, R01NS095355, R01GM123731, and joint funding from the Jayne Koskinas Ted Giovanis Foundation for Health and Policy/Breast Cancer Research Foundation (to Q.N.).

Author Contributions

S.W., A.L.M. and Q.N. conceived the project. S.W. designed and implemented the algorithm. S.W. and A.L.M. performed analysis. A.L.M. and Q.N. supervised the research. All authors wrote the manuscript.

Competing Interests

The authors declare no competing financial interests.

References

1. Moris N, Pina C, Arias AM. Transition states and cell fate decisions in epigenetic landscapes. Nature Reviews Genetics. 2016;17(11):693–703.

2. MacLean AL, Hong T, Nie Q. Exploring intermediate cell states through the lens of single cells. Current Opinion in Systems Biology. 2018;9:32–41.

3. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nature Protocols. 2018;13(4):599–604.

4. Angerer P, Simon L, Tritschler S, Wolf FA, Fischer D, Theis FJ. Single cells make big data: New challenges and opportunities in transcriptomics. Current Opinion in Systems Biology. 2017;4:85–91.

5. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nature Methods. 2017;9(5):2579–486.

6. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nature Methods. 2017;14(4):414–416.

7. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161(5):1202–1214.

8. Babtie AC, Chan TE, Stumpf MPH. Learning regulatory models for cell development from single cell transcriptomic data. Current Opinion in Systems Biology. 2017;5:72–81.

9. Guo M, Bao EL, Wagner M, Whitsett JA, Xu Y. SLICE: determining cell differentiation and lineage based on single cell entropy. Nucleic Acids Research. 2017;45(7):e54–e54.

10. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. bioRxiv. 2017; p. 128843.

11. Parra RG, Papadopoulos N, Ahumada-Arranz L, El Kholtei J, Treutlein B, Soeding J. Reconstructing complex lineage trees from scRNA-seq data using MERLoT. bioRxiv. 2018; p. 261768.

12. Zhuang L, Wang J, Lin Z, Yang AY, Ma Y, Yu N. Locality-preserving low-rank representation for graph construction from nonlinear manifolds. Neurocomputing. 2016;175:715–722.

13. Camp JG, Sekine K, Gerber T, Loeffler-Wirth H, Binder H, Gac M, et al. Multilineage communication regulates human liver bud development from pluripotency. Nature. 2017;345:1247125.

14. Ramilowski JA, Goldberg T, Harshbarger J, Kloppmann E, Lizio M, Satagopam VP, et al. A draft network of ligand–receptor-mediated multicellular signalling in human. Nature Communications. 2015;6(1):7866.

15. Schiebinger G, Shu J, Tabaka M, Cleary B, Subramanian V, Solomon A, et al. Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming. bioRxiv. 2017; p. 191056.

16. Guo G, Huss M, Tong GQ, Wang C, Sun LL, Clarke ND, et al. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Developmental cell. 2010;18(4):675–685.

17. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161(5):1187–1201.

18. Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008;9:2579–2605.

19. Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics; 2007. p. 1027–1035.

20. Strehl A, Ghosh J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of machine learning research. 2002;3(Dec):583–617.

21. Haghverdi L, Buettner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nature Methods. 2016;13(10):845–848.

22. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. Nature Methods. 2017;14(10):979–982.

23. Joost S, Zeisel A, Jacob T, Sun X, La Manno G, Lönnerberg P, et al. Single-Cell Transcriptomics Reveals that Differentiation and Spatial Signatures Shape Epidermal and Hair Follicle Heterogeneity. Cell Systems. 2016;3(3):221–237.e9.

24. Yan L, Yang M, Guo H, Yang L, Wu J, Li R, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nature structural & molecular biology. 2013;20(9):1131–1139.

25. Olsson A, Venkatasubramanian M, Chaudhri VK, Aronow BJ, Salomonis N, Singh H, et al. Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. Nature. 2016;537(7622):698–702.

26. Teschendorff AE, Enver T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. Nature Communications. 2017;8:ncomms15599.

27. Maaten Lvd, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008;9(Nov):2579–2605.

28. Friedman JH, Bentley JL, Finkel RA. An algorithm for finding best matches in logarithmic expected time. ACM Transactions on Mathematical Software (TOMS). 1977;3(3):209–226.

29. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. science. 2000;290(5500):2323–2326.

30. Kuang D, Yun S, Park H. SymNMF: nonnegative low-rank approximation of a similarity matrix for graph clustering. Journal of Global Optimization. 2015;62(3):545–574.

31. Kuang D, Ding C, Park H. Symmetric nonnegative matrix factorization for graph clustering. In: Proceedings of the 2012 SIAM international conference on data mining. SIAM; 2012. p. 106–117.

32. Von Luxburg U. A tutorial on spectral clustering. Statistics and computing. 2007;17(4):395–416.

33. Meyer C, Race S, Valakuzhy K. Determining the number of clusters via iterative consensus clustering. In: Proceedings of the 2013 SIAM International Conference on Data Mining. SIAM; 2013. p. 94–102.

34. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. Bioinformatics. 2014;30(19):2811–2812.