bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Low-rank Similarity Matrix Optimization Identifies Subpopulation Structure and Orders Single Cells in Pseudotime

Shuxiong Wang[1], Adam L MacLean[12], Qing Nie[1*]

1 Department of Mathematics, Center for Complex Biological Systems, and Center for Mathematical and Computational Biology, University of California, Irvine, CA 92697, USA

**qnie@uci.edu*

Abstract

Sequencing the transcriptomes of single cells has greatly advanced our understanding of the cellular composition of complex tissues. In many of these systems, the role of heterogeneity has risen to prominence as a determinant of cell type composition and lineage transitions. While much effort has gone into developing appropriate tools for the analysis and comprehension of single cell sequencing data, further advances are required. Optimization-based approaches are under-utilized in single cell analysis and hold much potential due to their ability to capture global properties of the system in low dimension. Here we present SoptSC: an optimization-based algorithm for the identification of subpopulation structure, transition paths, and pseudotemporal ordering within a cell population. Based on a measure of similarity between cells, SoptSC uses non-negative matrix factorization to create low dimensional representations of the data for analysis and visualization. We find that in several examples, the low-dimensional representations produced by SoptSC offer greater potential for insight than alternative methods. We tested our methods on a simulated dataset and four published single cell datasets from *Homo sapiens* and *Mus musculus*. SoptSC is able to recapitulate a simulated developmental trajectory with greater fidelity than comparable methods. Applied to two datasets on early embryonic development, SoptSC recapitulates known trajectories with high accuracy. Analysis of

[2]Orcid ID: orcid.org/0000-0003-0689-7907

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

murine epidermis reveals overall agreement with previous studies, but differs markedly regarding the composition and heterogeneity of the basal compartment. Analysis of murine myelopoiesis found that SoptSC can resolve complex hematopoietic subpopulation composition, and led to a new prediction regarding the asynchronous development of myeloid subpopulations during stem cell differentiation.

Keywords: optimization, similarity measure, non-negative matrix factorization, dimensionality reduction, single cell analysis, hematopoiesis, epidermal homeostasis, embryonic development

1. Introduction

Multicellular life can be defined by the collection of cell types present within an organism, their developmental trajectories, and potential interchange between types via cell state transitions throughout lifetime. Cell type is in turn controlled by the transcriptional state of a cell, together with interplay from proteomic and epigenetic factors. Our ability to measure the transcriptional state of a cell — and thus approach an understanding of its type or fate — has advanced dramatically within the past few years [1] due in part to high-throughput single-cell RNA sequencing (scRNA-seq) [2, 3, 4, 5]. This move away from bulk [6, 7, 8] to single-cell sequencing permits delineation of the different sources of heterogeneity from within a population, an increasingly important task given the preeminent role of biological noise in such data (e.g. [9, 10]). In addition, scRNA-seq analyses have promoted the identification of new (rare) cell types [11, 12], challenged classical models of cellular lineage hierarchies [13, 10], and deepened our knowledge of various developmental trajectories [14, 15].

scRNA-seq experiments yield measurement of $\mathcal{O}(10^4)$ genes in hundreds to thousands (and rapidly approaching $> 10^5$ [1]) cells across multiple time points and perturbations. Computational approaches are essential for the analysis of such high-dimensional datasets. Typical scRNA-seq analysis pipelines include clustering, pseudotemporal ordering of cells, and identification of marker genes, all of which require a dimensional reduction step [16, 17]. Dimensional reduction, e.g. via principal components analysis (PCA), t-distributed stochastic neighbor embedding (tSNE), etc., can also be performed directly for visualization purposes [18, 19]. Clustering — specifically the identification of functionally relevant (sub-)populations of cells and,

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ideally, the relationships between them [20] — presents a crucial challenge for the interpretation of scRNA-seq datasets. "Pseudotemporal" ordering projects cells onto a pseudotime axis that may represent (e.g.) a developmental process, or stem cell differentiation, and can deviate from real time due to the unsynchronized nature of single cells [21, 14]. Discontinuities in cellular fate transitions further complicate the analysis of pseudotime [9].

A number of factors present challenges for scRNA-seq analysis, including systematic noise, the dropout effect, sparsity, sensitivity to parameters, and non-uniqueness of outputs [22, 23]. The most significant of these, in our opinion, is appropriate consideration of the (biological and measurement) noise present in such data. Current methods for pseudotemporal ordering struggle to handle noise by selecting markers genes based on prior knowledge, projecting data into lower dimensional space and constructing diffusion distance etc [21, 24, 25].

Optimization generally seeks to find parameter values that maximize or minimize a real-valued function subject to given constraints [26, 27]. Optimization methods have found widespread use throughout computational systems biology, from early uses for the analysis of ecological or life history models [28, 29], to network inference or parameter estimation of biochemical reaction networks [30]. Optimization methods have the inherent advantages that they are able to retain global aspects of the input data through low-rank regularization, and they preserve local structure using a neighborhood projection constraint. This makes them significantly more robust to the effects of biological noise, as we will show below.

Here we present SoptSC *(sopt-see)*: Similarity matrix-based optimization for Single-Cell analysis; a method for reconstructing the pseudotemporal ordering of cells, *de novo* identification of subpopulations, and identification of the transition paths between these subpopulations. SoptSC constructs a cell-to-cell similarity matrix, upon which non-negative matrix factorization (NMF) is performed to find a low-rank representation of the relationships between individual cells: rank-1 factorization determines the pseudotime axis; rank-k factorization clusters the cells into k distinct subpopulations. Key advantages of this approach include: *i*) optimization guarantees that coefficients of the linear representation are nonzero only in a local neighborhood of the data point, thus preserving the intrinsic geometric structure of the manifold; *ii)* the low-rank constraint enables SoptSC to capture global properties of the data while remaining robust to biological noise and outliers [31]; *iii)* SoptSC predicts the number of subpopulations present in the data in

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

an unsupervised manner; *iv)* the method is insensitive to 'nuisance' genes, i.e. those which are not relevant to the trajectory or process currently being studied.

The remainder of the paper is organized as follows: in the next section we describe the methods and algorithmic design of SoptSC; we then test its performance by comparison to existing methods for clustering and dimensional reduction using *in silico* data. We go on to apply SoptSC to four biological datasets, and find not only that we can comprehensively recapitulate the results of previous analyses, but that SoptSC also generates new insight into the cellular relationships and developmental trajectories of adult stem cell systems.

2. Methods

Here we describe SoptSC, an optimization-based algorithm that enables the de novo detection of subpopulations, pseudotemporal ordering, and cell subpopulation transition paths from single cell gene expression datasets. SoptSC is based on the concept of similarity between cells, i.e. we find a low-rank representation of a cell (here we use 'cell' to mean the vector of gene expression values for a cell) in terms of other cells within a given neighborhood. For each cell, the similarity score is thus defined by a set of linear coefficients (the solution of the low-rank optimization model) in a subspace given by its neighboring cells. This measure is quite distinct from distance-based metrics often used to define similarity.

The SoptSC algorithm consists of two optimization steps. In the first, a square matrix is constructed that describes the cell-to-cell similarities based on the input gene expression data. In the second step, low-rank approximations of the similarity matrix are calculated to define either (*i*) cell subpopulations within the data (rank-k, where k is the number of subpopulations), or (*ii*) pseudotemporal ordering of cells (rank-1). Via the construction of a transition matrix from the similarity matrix, we identify the transition paths between cell subpopulations. To determine the number of subpopulations, k, we propose a novel algorithm that finds the consensus matrix (S^C) for range of values of k, specified by a suitable prior, and then estimates the value of k from the eigenvalue spectra of the graph Laplacian of S^C.

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.1. ***Construction of the cell-to-cell similarity matrix***

The input to SoptSC is a single cell gene expression matrix:

$$X = \begin{bmatrix} ... & ... & ... \\ ... & X_{i,j} & ... \\ ... & ... & ... \end{bmatrix} \in \mathbb{R}^{m \times n},$$

with m genes ($1 \leq i \leq m$), and n cells ($1 \leq j \leq n$), i.e. the element $X_{i,j}$ represents the expression value of the ith gene in the jth cell. SoptSC computes the coefficient matrix Z from X by the following optimization model:

$$\begin{aligned} \mathcal{P}_1 : \min_{Z} \quad & \lambda ||X - XZ||_{2,1} + ||Z||_* \\ s.t. \quad & Z^\top \mathbf{1} = \mathbf{1}, \\ & Z_{i,j} = 0, (i,j) \in \bar{G}, \end{aligned}$$

where $||\cdot||_{2,1}$ is the $L_{2,1}$ norm (the sum of the Euclidean norm of all columns); $||\cdot||_*$ is the nuclear norm; λ is a non-negative parameter and $\mathbf{1} = (1, ..., 1)^\top$ is a vector of ones of length n. $\bar{G}$ defines the complement of G, where G is the set characterizing neighbor relationships between cells, i.e. cell pairs $(i, j) \in G$ mean that cell i is in the neighborhood of cell j. G is obtained using K-nearest neighbors [32], and we choose $K = \min\{0.1m, 20\}$. The coefficient matrix Z was found to be robust to changes in K. The linear constraint $Z^\top \mathbf{1} = 1$ guarantees translational invariance of the data [33].

The optimization model $\mathcal{P}_1$ is a representation method for the construction of graphs from nonlinear manifolds [31]. Informally, this captures the 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, the model can better capture the global structure of the overall single cell gene expression data, and is more robust to noise and outliers. The optimization problem $\mathcal{P}_1$ can be solved numerically by the alternating direction method of multipliers [31]. Let Z^* be the optimal solution of $\mathcal{P}_1$, then via symmetric weights we define the similarity matrix S as

$$S = \max\left\{|Z^*|, |Z^{*\top}|\right\}. \tag{1}$$

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

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.2. ***Rank-k NMF for cell subpopulation clustering***

In order to classify cells into subpopulations based on their similarity, we use symmetric non-negative matrix factorization (NMF) [34, 35], which can be regarded as a graph-based clustering method. The (non-negative) similarity matrix S is decomposed into a product of a non-negative low rank matrix $H \in \mathbb{R}_+^{n\times k}$ and its transpose $H^\top$ via the optimization problem:

$$\mathcal{P}_2 : \min_{H\in\mathbb{R}^{n\times k}} ||S - HH^\top||_F^2$$
$$\text{s.t.} \quad H \geq 0,$$

where k is the number of subpopulations of cells and $||\cdot||_F$ is the Frobenius norm. The low rank condition for H is ideally suited for capturing the clustered nature of the cell subpopulations, i.e. by reordering S according to the columns of H, a block-diagonal or near-block-diagonal structure can be obtained (see Fig. 1A). We denote the reordered similarity matrix S^B. The structure of S^B is such that cells within a block have high similarity to each other and low similarity to cells from other blocks. It can be shown that the solution of $\mathcal{P}_1$ is strictly block-diagonal when the data are clean and sampled from independent subspaces [36]. Due to this observation, the similarity matrix S can be approximated by a sum of rank one matrices $H^iH^{i\top}, i = 1, 2, ..., k$, where $H = [H^1, H^2, ..., H^n]$, which can be obtained by solving the NMF problem $\mathcal{P}_2$. Singular value decomposition is used to find H_0, an initial low-rank non-negative matrix required as an input for $\mathcal{P}_2$ [37]. If we now let $S = [S^1, S^2, ..., S^n]$ represent the columns of S, then the columns of S can be approximated by the space spanned by the columns of H as:

$$S^i \approx \sum_{j=1}^{k} H_{i,j}H^j.$$

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 their corresponding columns of S in the 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.

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

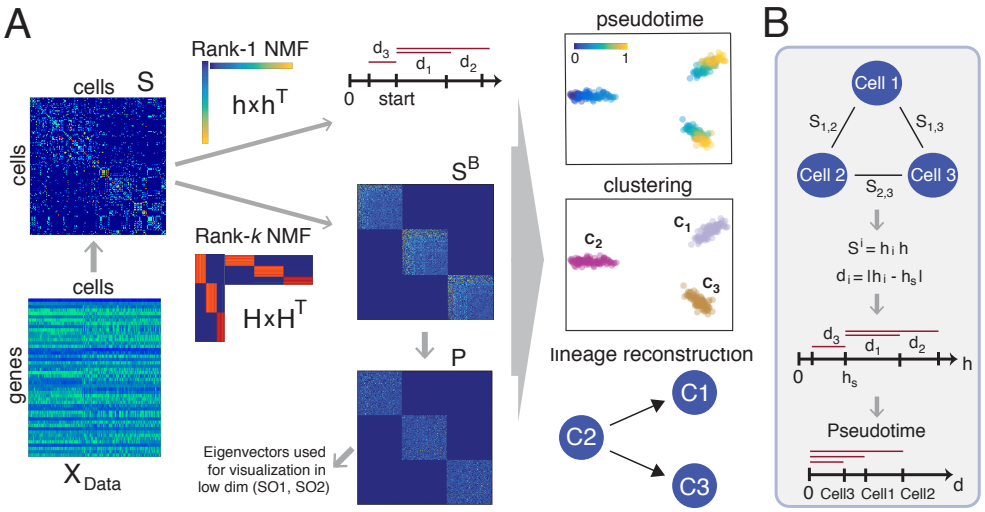


Figure 1: **Overview of SoptSC.** (A) Pipeline of the SoptSC algorithm. An input matrix $X_{Data} \in \mathbb{R}^{m \times n}$ (measuring m genes in n cells) is used to construct the cell-to-cell similarity matrix S by solving an optimization problem on the coefficients of X. Rank-k non-negative matrix factorization (NMF) is then performed (where for pseudotime $k = 1$, and for clustering $k > 1$) to find low-rank representations of S. These are used to order cells in pseudotime or cluster cells into subpopulations. From the transition matrix P lineage relationships are inferred, and eigenvectors of P are used to visualize the data in low dimension. (B) Schematic of the pseudotime algorithm. $S_{i,j}$ is the similarity between cells i and j. Then h characterizes the rank-1 NMF used to decompose S. from h, distances for each cell i in pseudotime are calculated (d_i) by choosing an initial start point h_s and ranking cells accordingly. See Methods for full details.

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.3. ***Identification of k from eigenspectra of the graph Laplacian***

Determining the number of clusters in a dataset is a fundamental problem extending far beyond the identification of cell populations from single-cell data; many clustering algorithms still require the user to specify the number of clusters. We propose a method to automatically identify the number of clusters within a dataset based on properties of the graph Laplacian (L) and the consensus similarity matrix [38], which is similar to [39]. We consider a range of values: $k_i = \{k_1, k_2, ..., k_q\}$, which can be viewed as a prior distribution for the number of cell subpopulations.

It has been shown that the number of eigenvalues of L equal to 0 is equivalent to the number of diagonal blocks of L [38].

The steps required to determine the number of clusters k are as follows:

1. Given the inputs S and $k_i, i \in (1, 2, ..., q)$, partition the cells into k_i subpopulations by solving the NMF problem $\mathcal{P}_2$.
2. Find the consensus matrix [40, 41], S^C. For each $j \in (1, 2, ..., q)$, define a matrix M^j by

$$M^j_{p,q} = \begin{cases} 1 & \text{if p and q belong to the same cluster} \\ 0 & \text{otherwise.} \end{cases}$$

   The consensus matrix S^C is then defined by

$$S^C = \sum_{j=1}^{q} M^j$$

3. Prune the consensus matrix as follows: set a tolerance $\tau \in [0, 0.5]$, and let $S^C_{i,j} = 0$ if $S_{i,j} \leq \tau q$. This increases the robustness of consensus clustering to biological noise.
4. Compute the graph Laplacian L and its eigenvalues, given the identity matrix I and a diagonal matrix D such that:

$$L = I - D^{-1/2} S^C D^{-1/2}$$

$$\text{with } D_{ii} = \sum_{j=1}^{n} S^C_{i,j}.$$

5. Find (i) the number of eigenvalues that are close to zero, and (ii) the index at which the largest eigenvalue gap occurs [38].

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

An initial estimate for k is given by (i). In cases where there may be significant sources of noise in the data, or where other uncertainties exist, we can use (ii) instead as an estimate of k. Especially for cases displaying a prominent largest eigenvalue gap (see e.g. Fig 5G below), (ii) can provide a better estimate of the subpopulation structure present in the data. For all analyses performed below, we choose a prior k_i so the number of clusters ranges from 1 to 25, and we set the tolerance $\tau = 0.3$.

2.4. ***Rank-1 NMF to determine pseudotemporal ordering***

Most algorithms for pseudotemporal ordering of cells proceed via graph-based methods. Here we propose an alternative method to rank cells in pseudotime based on similarity properties, rather than a distance measure. Specifically, we use the rank-1 non-negative matrix factorization of S to characterize the pseudotemporal ordering of cells. The rank-1 NMF matrix $h = [h_1, h_2, ..., h_n]$ can be obtained by setting $k = 1$ in the optimization problem $\mathcal{P}_2$. In this case, the overall structure of S can be approximated by h, i.e.,

$$S \approx hh^\top.$$

It follows immediately that each of the columns of S, denoted S^i, can be represented by a single vector h, and a non-negative coefficient h_i, i.e. $S^i \approx h_i h$. Then the non-negative coefficient h_i can be used to measure the similarity of S^i to h. If the starting cell is denoted the sth cell, then a new vector $d = [d_1, d_2, ..., d_n]$ can be defined where $d_i = |h_i - h_s|, i = 1, 2, ..., n$. Each element of d represents the relative distance from cell i to the initial cell s. The temporal order of cells is then obtained by sorting d in ascending order.

High levels of noise lead to challenges for estimation of pseudotime in the sense that the obtained similarity matrix might not fully capture similarity among all cells. In an effort to combat this, instead of using the similarity matrix directly, we propose to alternatively use the transition matrix P (defined below), and we compute pseudotemporal ordering through two steps: 1) compute the first J largest eigenvectors of P (the default value is set as $J = 6$); 2) use these eigenvectors as input and perform rank-1 NMF to obtain the pseudotemporal ordering.

2.5. ***Identification of cell subpopulation transition paths***

Following the identification of the number of cell populations present, and assigning cells to their relevant subpopulations, here we infer the transition

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

paths between these subpopulations in order to identify the cellular hierarchy present in the data. As above, we use similarity measures to determine the probabilities of transition between cell subpopulations, by first defining a transition matrix, and then using this to compute a minimum spanning tree between cell subpopulations.

The transition matrix P is defined by:

$$P = D^{-1}S,$$

where D is a diagonal matrix defined by $D_{ii} = \sum_{j=1}^{n} S_{i,j}$. For visualization purposes below, we calculate the eigenvectors of P and use the second two eigenvectors as components for visualization (SO1 and SO2). We do not use the first eigenvector as it is trivially defined as $\mathbf{1} = (1, 1, ..., 1)$. We then project P into low (three) dimensional space via principal components analysis (PCA) [42], and construct a complete weighted graph between the cell subpopulations using the centroids of the subpopulations as vertices and the distances between centroids as the weights. Then by setting a root node (the initial cell population), we can construct the minimum spanning tree for this graph: determining the order of transitions between the cell subpopulations.

2.6. ***Generation of* in silico *data for performance assessment***

To assess the performance of SoptSC , we construct a dataset for which the cell subpopulations and transition paths are known. The expression levels of genes were set as a function of a parameter t which can be regarded as cellular "differentiation time". Three distinct functions are used for simulation, where the expression values of genes generated by a functions are analogous to responses from a common biological mechanism [24]. The functions used (two nonlinear and one constant) are:

$$\begin{aligned} f_1(t) &= c_1 \cos(t/3) + 1 + \epsilon_1, \\ f_2(t) &= c_2 \sin(t/3) + 1 + \epsilon_2, \\ f_3(t) &= 1 + \epsilon_3, \end{aligned}$$

where $c_i \sim N(1, \sigma^2)$ $(i = 1, 2)$ and $\epsilon_i \sim N(0, \sigma^2)$ $(i = 1, 2, 3)$. For each function, the expression levels of genes were simulated by sampling the random variables c_i and ϵ_i. We chose 170 values of t as input to the first two functions $f_1(t), f_2(t)$ from the interval $[2\pi, 4\pi]$ to simulate two distinct cell

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

subpopulations. In order to simulate 'trajectory like' data in 2-dimensional space, we introduce the third function $f_3(t)$ on the interval $[-0.5, -0.1]$ to generate a third cell subpopulation containing 100 cells. In order to obfuscate this trajectory, we then add genes unrelated to the process being studied; we can vary this number of "nuisance" genes to test the performance of each method.

3. Results

3.1. ***SoptSC captures salient features of single cell data and outperforms other methods when data are noisy***

We applied SoptSC to an *in silico* dataset that contains three subpopulations located close to one another in gene space, which follow a nonlinear developmental trajectory, with $X_{Data} = 52 \times 270$ (see Methods for full details). As a first step, we studied low-dimensional projections of the data in order to assess, at a general level, which features of the data are captured by SoptSC in comparison with principal component analysis (PCA) or t-distributed stochastic neighbor embedding (tSNE), two widely-used dimensional reduction techniques for single cell analysis [19].

We varied the noise level by manipulating the relative standard deviation, defined as the ratio of σ to the mean of data, from 10-30% by choosing different values of σ, and visualized the known subpopulation structure of the data (three subpopulations) using the first two components of each method (Fig. 2). We can thus assess each method as we study the projections produced under increasing noise. At 10% noise (Fig. 2A) we observe that all three methods capture three distinct subpopulations, but tSNE loses the distinct shape of the developmental trajectory. At 20% noise (Fig. 2B), tSNE can distinguish neither the trajectory nor the subpopulation structure. PCA and SoptSC both retain the subpopulation structure, but PCA loses representation of the trajectory, which SoptSC retains. At 30% noise (Fig. 2C), clear distinction between subpopulations or the shape of the developmental trajectory is lost for all three methods. We note that a 30% noise level may be low in comparison with some real datasets. These results highlight the general challenge of meaningful low-dimensional data representations when "true" biological subpopulations are located close to one another in high dimensional space.

Clustering methods rely crucially on identifying how many subpopulations are present in a given sample. This is in general a difficult problem,

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

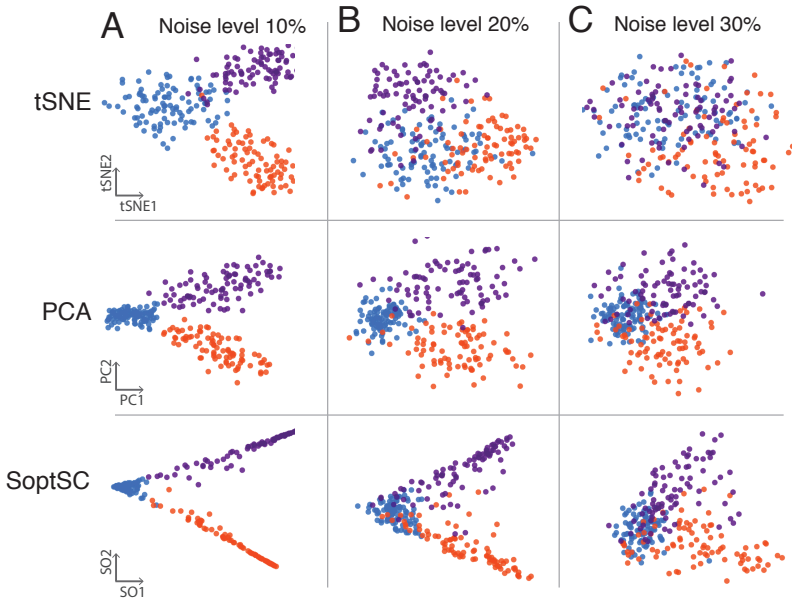


Figure 2: **Comparison of low-dimensional data representations by tSNE, PCA, and SoptSC.** The first two dominant components of the transition matrix (SO1 and SO2) are used to project the *in silico* data into 2D space, with varying levels of Gaussian noise: (A) 10%; (B) 20%; and (C) 30%. Cells are colored according to their true subpopulation labels.

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

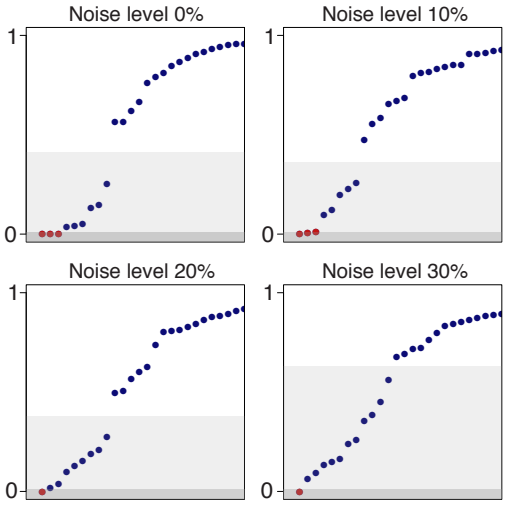


Figure 3: **Number of subpopulations within a dataset as identified by SoptSC.** The first 25 eigenvalues of the graph Laplacian of S^C, the consensus similarity matrix, are shown at different noise levels. Number of eigenvalues approximately zero (threshold = 0.01; dark gray region) predicts the number of subpopulations (marked in red). Number of eigenvalues below the largest eigengap (light gray region) provides secondary prediction of the number of subpopulations.

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

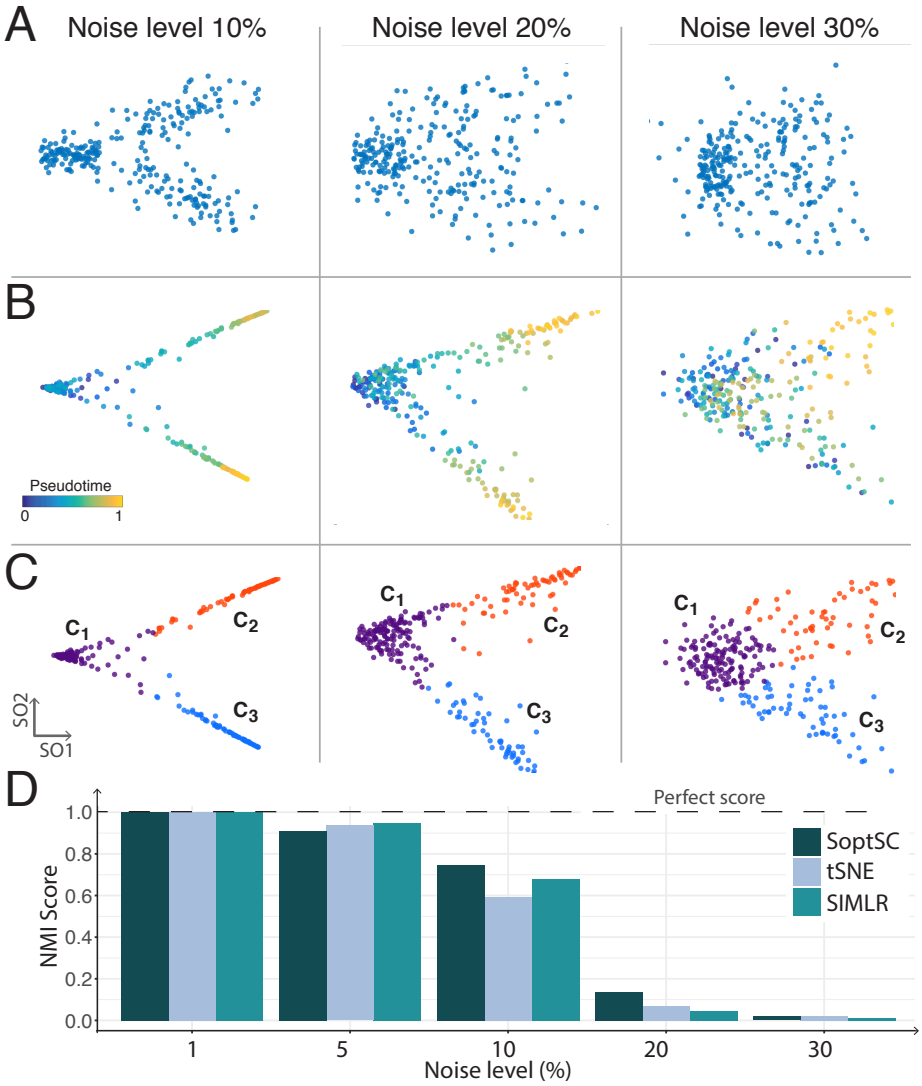


Figure 4: **Performance of SoptSC for pseudotemporal ordering and clustering.** (A) Original data used as input (true labels hidden). (B) Pseudotemporal ordering by rank-1 NMF. (C) Clustering of cells by rank-3 NMF. (D) Comparison of clustering performance of SoptSC against k-means or SIMLR clustering algorithms. The normalized mutual information (NMI) is used to assess predictions.

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and one that is compounded by the presence of noise, which is unavoidably widespread in scRNA-seq datasets. The output for the method we use to identify the number of subpopulations is illustrated in Fig. 3 (details given in Methods). We use the *in silico* dataset described above for testing; the number of eigenvalues close to zero is used to predict the number of subpopulations.

We see that at 0-10% noise, SoptSC can correctly identify that the data contain three subpopulations. At 20-30% noise, this structure can no longer be recovered by our algorithm; we note here that given 20-30% noise there is in fact little population structure left to recover (see Fig. 2). The light gray regions on Fig. 3 demarcate the number of eigenvalues below the largest gap in the eigenspectrum. This number can be used as a secondary estimate of structure, e.g. indicating the presence of hidden subpopulations, and can be used especially for clustering heterogeneous data, i.e. at 20% noise, the same number of hidden subpopulations are recovered as for 0-10% noise (± 1).

In Fig. 4 we test the ability of SoptSC to project cells in pseudotime and identify subpopulations, and we compare the results of clustering to other current methods. In the input data the true subpopulation labels are hidden and noise is added to perturb the cells in gene expression space (Fig. 4A). In Fig. 4B we plot the pseudotemporal ordering of cells at different noise levels. We see that for up to 20% noise a clear developmental trajectory through pseudotime can be obtained, but that at 30% noise our pseudotemporal ordering is no longer reliable. We then cluster the data in SoptSC via rank-3 NMF, and find that even at a 30% noise level, it is still possible to identify three subpopulations with good accuracy. In order to quantify performance, we repeat this clustering at different noise levels up to 30%. We compare the performance of SoptSC to two alternative methods: a recently published algorithm SIMLR [43]; and K-means clustering in 2D space following dimensional reduction by tSNE [44, 19]. Normalized mutual information (NMI) is the metric used for comparison [45, 46]. The results in Fig. 4D show that as the noise varies from 1-10% SoptSC performs similarly or marginally better than alternative methods (no significant differences). At the 20% noise level, SoptSC outperforms the other methods, however none of the methods attain high scores at this noise level. At the 30% noise level, none of the methods are able to cluster these data successfully. These results are in line with our central proposal: that optimization-based dimensional reduction and clustering methods are well-suited to handling biological noise.

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.2. ***SoptSC recapitulates known developmental trajectories in human and mouse early embryonic single cell data***

Development of the early embryo — from oocyte to blastocyst — has been intensely studied in humans and other mammals [47, 48]. With the introduction of widespread scRNA-seq, we are now able to interrogate the beginnings of life in new detail, by characterizing the transcriptomes of single cells in these early stages [49, 50]. As a test of SoptSC, we chose to analyze two single cell embryonic datasets, from a mouse study [49] and a human study [50]. On each dataset, we ran SoptSC to extract the subpopulation structure and pseudotemporal ordering of cells and compared the results to previously characterized trajectories. These data are particularly suitable for testing our algorithm since they display clear temporal trajectories as the embryos grow. In addition, they have relatively low dimension either in number of genes (in the mouse dataset) or number of cells (in the human dataset), thus providing us with good first-step benchmarks with which to test SoptSC.

First we study 48 qPCR gene expression profiles in 438 individual cells taken from early stage mouse embryos, published by Guo et al [49]. These data describe the six cell doublings between zygote and 64-cell stage. Two well-characterized cell bifurcation/differentiation events occur during this progression, one at the 32-cell stage, and one at the 64-cell stage [51, 52]: at the 32-cell stage, totipotent cells branch into trophectoderm and inner cell mass (ICM); at the 64-cell stage, the ICM branches into primitive endoderm and epiblast. Therefore, we expect to find two distinct subpopulations emerge at the 32-cell stage, and another two distinct subpopulations to emerge at the 64-cell stage. The results of SoptSC are visualized in the 2D projection given by the first two dominant components of the transition matrix (SO1 and SO2). We begin by labelling cells with their true embryonic stage as given in [49] (Fig. 5A). The results produced by SoptSC for these data are shown in Fig. 5B-C,G.

SoptSC clusters the data into eight subpopulations (Fig. 5B). Shown in Fig. 5G is the eigenspectrum used to estimate the number of subpopulations present: the largest gap occurs after the eighth eigenvalue. Also shown in Fig. 5G are the similarity matrices constructed for all cells, and separately for the 32-cell and the 64-cell stage. Each of these displays clear structure: identifying eight subpopulations for all cells, and the two branch points that lead to two subpopulations at the 32-cell stage, and three subpopulations at the 64-cell stage.

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

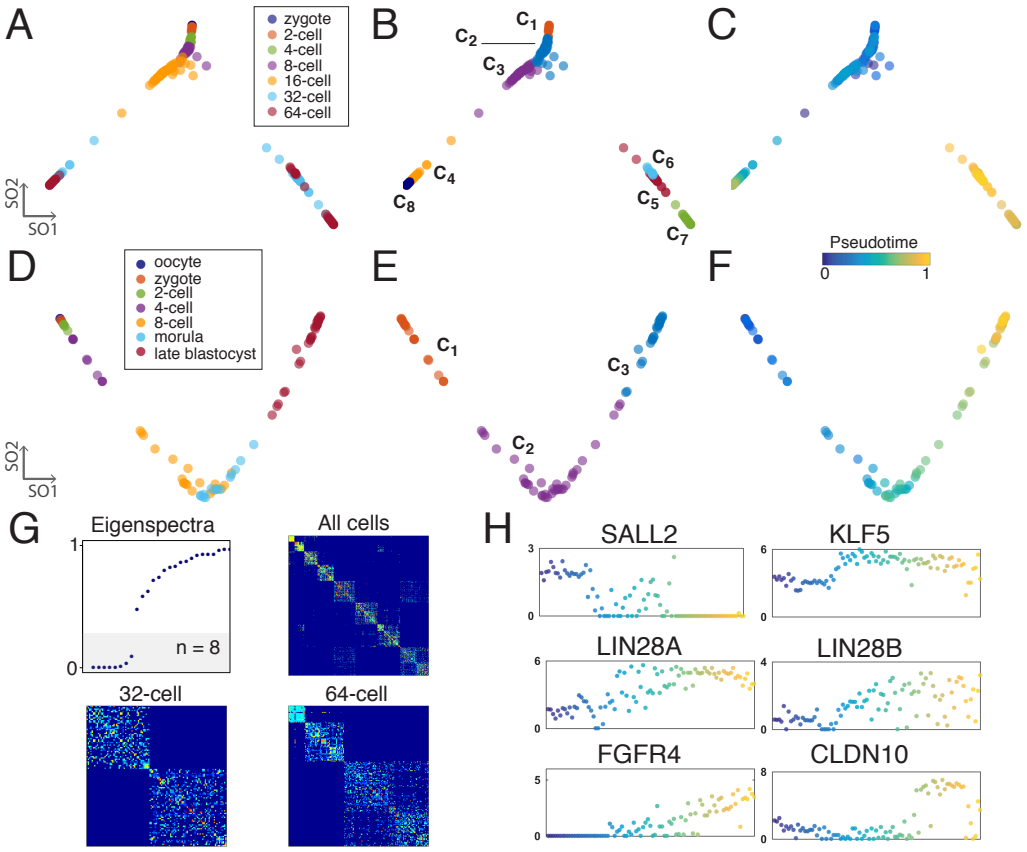


Figure 5: **SoptSC identifies cell subpopulations and developmental dynamics of single cell data from mouse and human early embryonic development [49, 50].** (A) Low dimensional projection using SoptSC of qPCR profiles from 438 single cells, developmental stages labelled according to [49]. (B) SoptSC identifies 8 clusters during mouse zygote to 64-cell stage development. (C) Pseudotemporal ordering of mouse embryonic cells by SoptSC. (D) Low dimensional projection using SoptSC of scRNA-seq profiles from 88 single human cells, developmental stages labelled according to [50]. (E) Clustering of human oocyte to blastocyst development identifies 3 populations. (F) Pseudotemporal ordering of human embryonic cells by SoptSC. (G) Eigenspectra derived from the similarity matrix for mouse embryonic data: the number of eigenvalues below the largest gap is indicated by the shaded region. The similarity matrices for the full dataset (all cells) and at two distinct cell stages are also shown. (H) Human gene expression in single cells over pseudotime.

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

By comparison of the known labels (Fig. 5A) with the subpopulations identified by SoptSC (Fig. 5B), we see that SoptSC clusters the zygote and the 2-cell stage into a single population, and similarly clusters the 4-cell and the 8-cell stages together. The 16-cell stage is identified as a single subpopulation. At the 32-cell stage the first bifurcation occurs (forming trophectoderm and ICM); this can be seen by the two distinct 32-cell stage subpopulations in Fig. 5A. SoptSC identifies each of these subpopulations (C_4 and C_5). Furthermore we see that a second branching event occurs during the differentiation of cluster C_5 (which we thus can identify as the ICM) and SoptSC correctly clusters two subsequent 64-cell stage subpopulations distinctly (C_6 and C_7, corresponding to primitive endoderm and epiblast). A separate 64-cell stage subpopulation (C_8) emerges following the differentiation of the trophectoderm (C_4). In Fig. 5C we order the cells along pseudotime, and see that the inferred pseudotime is consistent with the cellular developmental stages. Overall, we find that the results of SoptSC are in excellent agreement with previous analyses of these data and with the known biology [51, 53].

Next we ran SoptSC on single-cell RNA-seq data from the very early stages of human embryo development, published by Yan et al. [50] and studied further in [54, 55]. The results are shown in Fig. 5D-F,H. The data consist of 88 cells from seven stages of human early embryonic development, beginning from the oocyte and transitioning through 2-cell to 8-cell stages before differentiating into the morula and finally the late blastocyst stage (approximately 32 cells). We selected 8220 genes from a total of 20,012 based on two criteria: 1) a minimum gene expression level (FPKM > 1) must be satisfied in at least 50% of the cells; 2) the variance of $\log_2$-transformed FPKM of each gene is larger than 0.5. Fig. 5D shows the distribution of cells along the seven human early embryo developmental time points (labels from [50]). We see that SoptSC projects these subpopulations into a single developmental trajectory.

In Fig. 5E we plot the subpopulation structure as identified by SoptSC: three subpopulations were predicted based on the eigenspectra of the graph Laplacian, i.e. we predict that overlap between these early developmental stages leads to fewer functionally different (in gene expression space) subpopulations. The first four stages (oocyte to 4-cell) are clustered together, as are the 8-cell and morula stages. The late blastocyst stage is clustered alone. We note that among the cells studied, the cluster boundaries are distinguished in perfect agreement at C_1/C_2, and very good agreement (two cells mis-classified) at C_2/C_3. The pseudotemporal trajectory inferred by SoptSC

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Fig. 5F shows that the ordering cells by this method is highly consistent with the known stage of development. To further dissect the pseudotemporal ordering obtained, we plot six genes previously identified as important markers [54] along pseudotime (Fig. 5H). We find very good agreement between the dynamics predicted here and those previously studied [54].

3.3. ***SoptSC reveals new structure within epidermal cell subpopulations during telogen***

The mammalian epidermis is a well-characterized adult stem cell system [56], yet significant questions remain regarding the constituents of specific epidermal cell subpopulations and the interactions between them. Cells of the epidermis exhibit considerable heterogeneity [57], and can transition between multiple compartments (sometimes crucial for function, see e.g. the formation of hair follicles [58]). In the interfollicular epidermis (IFE), cells are highly stratified: a stem cell population in the basal layer maintains the tissue through proliferation and production of differentiated cells (populations DI and DII below) and finally keratinized cells (populations KI and KII below). The keratinized cells form the outermost layer of the skin that is eventually shed [56].

Here we analyze a recent scRNA-seq dataset of murine epidermis taken during the second telogen [59] in order to assess the effects that such epidermal heterogeneity may have on subpopulation structure and pseudotemporal ordering. Joost et al. [59] performed multi-level clustering in order to identify the various subpopulations of the epidermis, and found five subpopulations within the interfollicular epidermis (IFE) at the first level of clustering (Fig. 6A). Here we focus our studies on the IFE, as it likely represents a faithful trajectory in pseudotime, and we thus analyze 720 single cells using SoptSC. We select 1523 variable genes as input, based on the criterion that the gene expression variance > 0.8. By inspection of the eigenspectra of the graph Laplacian we predict that eight subpopulations exist within the IFE: three more than were identified in the first level of clustering by Joost et al., including one subpopulation than was not identified at all in their analysis (including at second level clustering). We visualize the results of SoptSC in Fig. 6 using the first two dominant components (SO1 and SO2).

SoptSC projects the IFE cell population onto a 2D plane that reflects the known differentiation trajectory of the IFE, and preserves the overall subpopulation structure identified by Joost et al. (Fig. 6A), although we also see that in some regions (on the right of the plot) there is overlap where

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

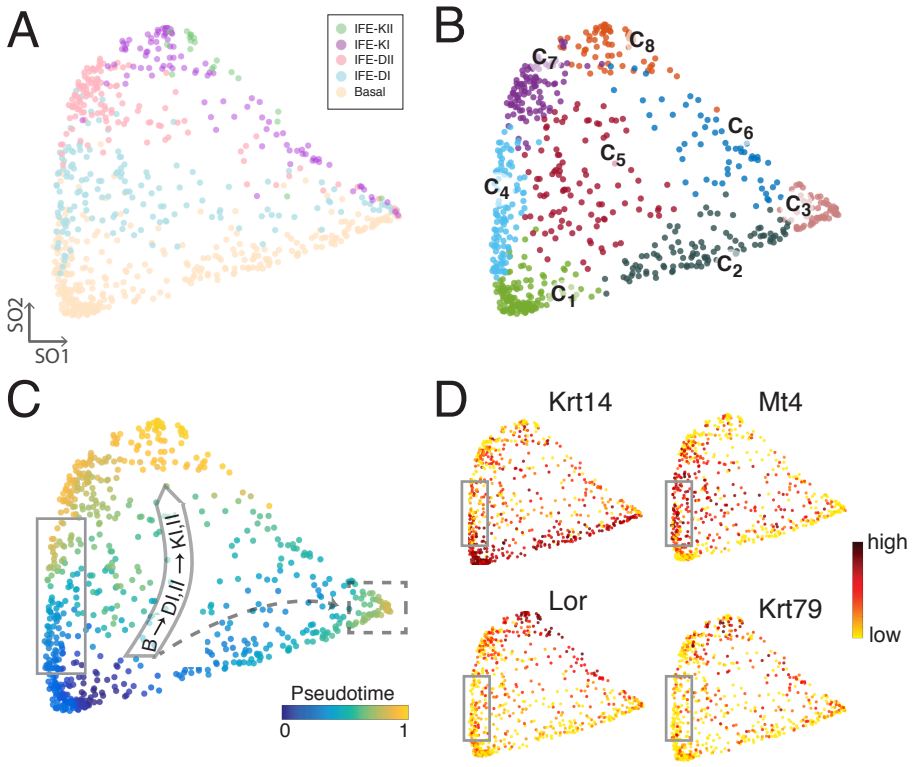


Figure 6: **SoptSC identifies subpopulation structure within the IFE.** (A). Low dimensional projection using SoptSC of 720 single cells from the interfollicular epidermis (IFE); data and cluster labels from [59]. (B). Subpopulations of the IFE identified by SoptSC. (C) Pseudotemporal ordering of IFE cells by SoptSC. Solid gray box marks cells of cluster C_4; dashed gray box marks cells of cluster C_3. Arrows denote putative transition paths through pseudotime.(D) Gene expression of key epidermal markers; gray box marks cells of cluster C_4.

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

all of the subpopulations meet. SoptSC also captures additional substructure in the IFE, namely within the basal and differentiated cell populations (Fig. 6B). Of particular interest are clusters C_3 and C_4, which we will analyze in greater depth below.

In Fig. 6C we order the cells of the IFE in pseudotime and visualize this developmental trajectory in the same 2D projection by SoptSC. We plot the gene expression of four epidermal markers across the IFE in Fig. 6D. We see broad agreement with known epidermal cell biology whereby basal cells appear earliest and transition through a differentiated cell state to eventually become keratinized (late in pseudotime). However we also see a rather dramatic departure from the expected trajectory for the subpopulation C_3, which appears late in pseudotime even though it contains only a few cells that are (identified as) keratinized, mixed with cells identified as basal and differentiated. This result highlights that the heterogeneity within IFE subpopulations is greater than previously known, spanning the whole compartment: putative basal cells appear both at the beginning and at the end of pseudotime. Whereas Joost et al. hinted at this by saying that "all basal cellsbefore reaching this pointare to some extent plastic" [59], our results go even further in their prediction of the extent of heterogeneity within the IFE.

A priori, one could suggest at least three possible hypotheses to explain the composition of cluster C_3: (*i*) the basal cells of C_3 differentiate late; (*ii*) a subpopulation of differentiated/keratinized cells retains basal cell markers; or that (*iii*) a subpopulation of differentiated cells is able to dedifferentiate back to a basal state. Dedifferentiation would imply transition through mid stages (DI,DII) that are marked by the gene Mt4 (Fig. 6D), however we see very low expression of Mt4 for C_3, thus providing putative evidence against hypothesis (*iii*). In addition we know of no examples from the literature of dedifferentiation occurring in the IFE. Higher expression of stem cell marker Krt14 than keratinized cell marker Lor (Fig. 6D) leads us to suggest that hypothesis (*i*) may be more likely than hypothesis (*ii*); to resolve this however, further experiments are needed.

In order to assess the composition of subpopulation C_4, we study the gene expression of four key epidermal marker genes across the IFE (Fig. 6D, gray box). The basal subpopulations are marked by high expression of Krt14, and the keratinized subpopulations by high expression of Lor, in agreement with Joost et al. [59]. Expression of Mt4 is greatest at mid-pseudotime, around the differentiated cell subpopulations, also in agreement with Joost et al. Investigating the gene expression in subpopulation C_4, we find high Mt4 and

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

relatively high Krt14 expression, but also some expression of Lor and Krt79. This indicates a mixed population of basal and differentiated cells (in agreement with the labels of Fig 6A), but also suggest a contribution from the hair follicle compartment, implicating this subpopulation as the infundibulum subpopulation (INFU-B) identified in [59]. Our data thus recapitulate in a single step the three subpopulations of basal cells that were found during two levels of clustering, and in addition, delineate the heterogeneity present within the major subpopulations of the IFE.

3.4. ***SoptSC resolves the monocytic/granulocytic cell fate decision directly from high-dimensional scRNA-seq hematopoietic data***

Hematopoiesis is the formation of all blood cells, including erythrocytes, leukocytes, and platelets, from a rare stem cell residing in the bone marrow in adult mammals [60]. The rise of single cell sequencing has had a dramatic impact on our understanding of hematopoiesis: progenitor cell populations previously resolved (by cell surface markers) within the developmental trajectory between stem and differentiated cells have had their roles/existence thrown into question, or even putatively discarded, as alternative lineage paths are drawn and we see the role of heterogeneity expand [10, 61, 12]. Given these recent results, the complexity of the hematopoietic hierarchy, and the significant levels of heterogeneity yet to be fully accounted for, we chose to analyse a scRNA-seq dataset describing hematopoiesis in mice [12]. Olsson et al. specifically study myelopoiesis, i.e. the formation of erythrocytes, megakaryocytes, monocytes, and granulocytes. We analyze gene expression in 382 single cells, and select 1567 variable genes as input based on the criteria: coefficient of variation > 3; gene included are expressed in at least 40% of the cells. We then applied SoptSC to analyze the substructure of hematopoietic cell populations, their ordering in pseudotime, and their transition paths during differentiation.

We project the cells into low dimension via SoptSC and label them according to their surface marker expression [12] in order to visualize the developmental trajectory (Fig. 7A). We see that there is a visible trajectory from top left moving downwards and rightwards, but also considerable population overlap leading to a lack of distinguishability between the cell surface marker-labelled subpopulations in this projection. Analysis of the eigenspectra of the graph Laplacian for these data predicts nine subpopulations, the

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

same number of subpopulations as was identified by Olsson et al. via iterative clustering and guide gene selection [12]. The clustering results of SoptSC are shown in Fig. 7B, labelled $C_1 - C_9$; these labels will be discussed in detail below. Due in part to the complex structure of the hematopoietic system, this 2D projection cannot completely resolve all the subpopulations (see also inset for zoom in). We note that we compared SoptSC visualization with tSNE and found that projecting via tSNE was considerably worse at capturing this trajectory of myelopoiesis than SoptSC (see S2 Fig. B and Olsson et al. [12]).

Pseudotemporal ordering of cells contributing to myelopoiesis (Fig. 7C) highlights that there are multiple branch points during differentiation, leading to three distinct subpopulations at the end of pseudotime on the main projection, and a fourth that can be identified in the inset; there may be more branching points that are not resolved on this projection. Inference of the transition path between subpopulations suggests four final subpopulations (Fig. 7D). In light of these results we are able to ascribe functional labels to the subpopulations identified in Fig. 7B. C_1 appears at the top of the tree and at the start of pseudotime, expressing stemness markers (e.g. high Gata2 — Fig. 7E) thus representing multipotent hematopoietic stem/progenitor cells. C_2 and C_4 (Fig. 7B inset) can be identified as erythrocytic and megakaryocytic precursors by the expression of Vwf, Klf1, EpoR (See SI Figs). We note that Fig. 7D infers a transition from erythrocytic to megakaryocytic progenitor cells: this annotation is probably incorrect; these lineages develop concurrently. We have however managed to resolve all the other transitions present in agreement with known biology. Subpopulation C_6 can be identified as a myelocytic population, as described in [12] with high granulocytic markers (Fig. 7A [pink/purple] and 7E [Gfi1]) and high expression of Mmp9 (S2 Fig. C), appearing late in pseudotime.

Subpopulations $C_7 - C_9$ (Fig. 7D, bottom) define the monocytic/granulocytic cell fate choice, as can be seen from their marker gene expression (Fig. 7E): the mixed progenitor population expresses Gata2 and low levels of Itga2b; the monocytic population expresses Irf8; and the granulocytic population expresses Gfi1. Interestingly, this key monocytic/granulocytic branching point can be clearly visualized in the 2D projection of SoptSC (Fig. 7B), with the monocytic cells located at the bottom of the plot (C_8) and the granulocytic cells on the right hand edge (C_9). Moreover, the subpopulation C_7, closely associates with the population identified by Olsson et al. as 'Multi-Lin': playing a crucial role in the regulation of myelopoiesis. C_7 appears

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

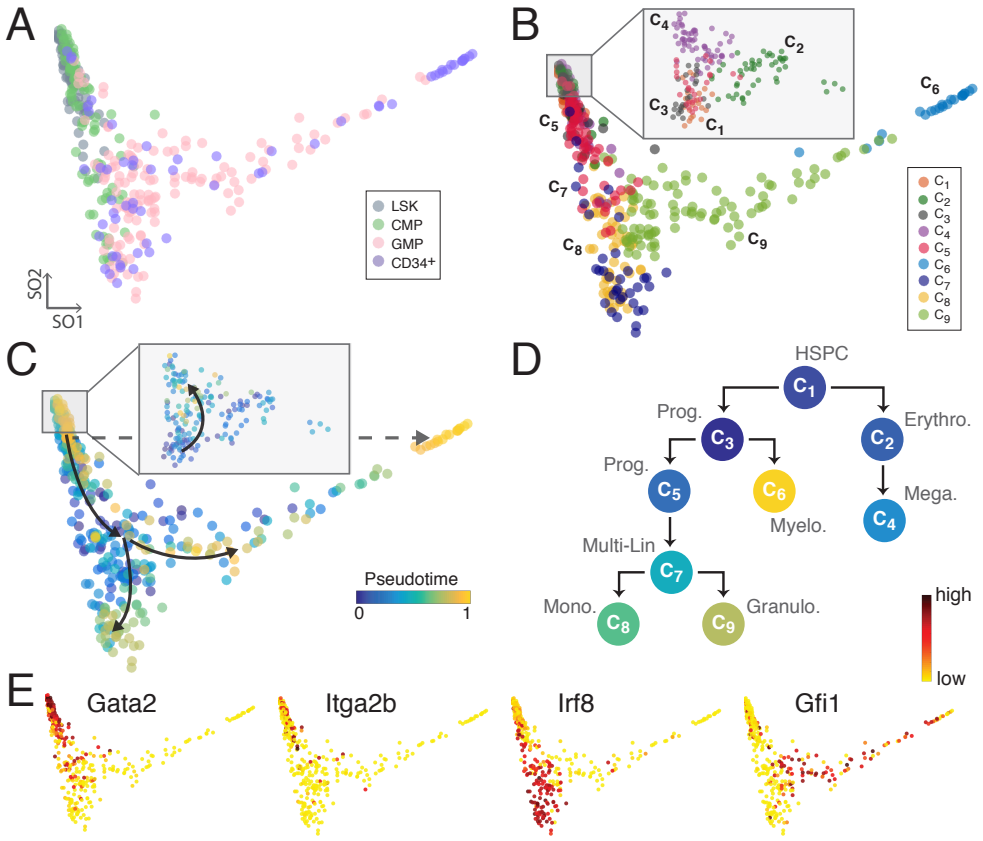


Figure 7: **Sc-RNA-seq subpopulation structure and pseudotemporal ordering during myelopoiesis [12].** (A). Low dimensional projection by SoptSC of 382 single cells from the hematopoietic system. LSK: $Lin^{-}Sca1^{+}c\text{-}Kit^{+}$; CMP: common myeloid progenitor; GMP: granulocyte monocyte progenitor; $CD34^{+}$: LSK $CD34^{+}$ cells. (B). Hematopoietic subpopulation structure identified by SoptSC; inset shows zoom in. (C) Pseudotemporal ordering of hematopoietic cells by SoptSC; inset shows zoom in. Arrows show differentiation paths identified in pseudotime. Dashed arrow indicates that the differentiation path lies on a different manifold than the one SoptSC projects onto here. (D) Lineage hierarchy constructed by SoptSC. Colors correspond to the mean pseudotime value for the subpopulation. Hematopoietic population identities have been curated after construction of the lineage hierarchy. HSPC: hematopoietic stem/progenitor cells; Prog: multipotent progenitor; Multi-Lin: mixed progenitor (see [12]); Mono: monocytic progenitor; Granulo: granulocytic progenitor; Myelo: myelocytic progenitor; Erythro: erythrocytic progenitor; Mega: megakaryocytic progenitor. (E) Gene expression of selected marker genes in single cells.

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

spread across large portions of the SoptSC projection, located both around monocytic cells and granulocytic cells near the bottom of the plot, and interspersed with less differentiated cells situated on the projection above these. This substantial heterogeneity present within the Multi-Lin subpopulation corroborates the findings of Olsson et al. Comparison of clusters C_8 and C_9 in pseudotime (Fig. 7C) also reveals an intriguing prediction: that upon differentiation of the granulocyte-monocyte progenitor, granulocytes appear earlier, and develop more slowly (i.e. span more of pseudotime) than their monocytic counterparts.

4. Discussion

Here we have presented SoptSC: similarity matrix optimization for single-cell analysis, a new method for the identification of subpopulations, and the reconstruction of pseudotime and cellular transition paths from single cell gene expression data. SoptSC is based on cell-to-cell similarity scores, which are obtained by introducing a low-rank optimization model in which the relationships among cells are represented by a structured similarity matrix [31]. This method preserves the intrinsic geometric structure of the manifold under study by allowing coefficients to be nonzero only in a local neighborhood of each data point. This low-rank constraint enables the model to better capture the global structure of a dataset, and improves its robustness to noise and outliers. These methods lead to a particular strength of SoptSC — exemplified in the previous two applications — namely its ability to extract pertinent information from data directly from a high dimensional space with a large number of clusters; previous analyses of these data [59, 12] first projected the data into a lower dimension or selected out particular clusters for study. Other recent approaches have also focussed on analysis of the geometrical properties of high dimensional data [62].

We applied SoptSC to four published datasets on three biological systems: embryonic development, epidermal homeostasis, and hematopoiesis. SoptSC showed very good agreement with previously determined subpopulation structure and developmental trajectories, in particular recapitulating with high accuracy the branching events occurring during early mouse embryo development [49], and development of early human embryo from oocyte to blastocyst [50]. In addition, scRNA-seq data analysis via SoptSC generated a number of unintuitive predictions: we found evidence for a label-promiscuous population of cells within the interfollicular epidermis, marked

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

by a combination of both basal stem and differentiated cell markers, and an overal higher than previously reported [59] degree of variability between the epidermal cell populations studied. Analysis of hematopoiesis (specifically myelopoiesis) in mice [12] led us to identify differences in the developmental trajectories of granulocytic and monocytic progenitor cells, with the former appearing earlier and developing more slowly than the latter.

Clustering cells and pseudotime reconstruction are performed in SoptSC using rank-k non-negative matrix factorization (NMF), where $k > 1$ is the number of clusters in the data, or for pseudotime, $k = 1$. As we have shown through detailed analyses, high levels of noise lead to significant challenges for these tasks. We thus proposed a modified algorithm for pseudotemporal ordering in the presence of noise that proceeds by projecting the similarity matrix into a lower dimension defined by a set of eigenvectors, before performing rank-1 NMF. We found that this yields more reliable predictions, however it produces an additional parameter (the number of eigenvectors used to project) that needs to be set by the user. An improvement to SoptSC would be to fix the number of eigenvectors according to some criteria, however defining this generally, rather than in a data-dependent manner, remains challenging. Another potential extension to SoptSC is to relax the constraint on Z that forces non-neighboring cell coefficients to be zero, this could be achieved using sparse regularization, i.e. by adding a (L_1 regularization) penalty term to the objective function such that the neighbors of each cell can be inferred directly (by the non-zero coefficients) from the solution to the optimization problem.

SoptSC provides a prediction of the number of clusters present in a dataset; this is performed by constructing the graph Laplacian of the consensus similarity matrix, and calculating the number of zeros and the largest gap in its eigenvalue spectrum. The advantage of this method lies in the step that inputs (simultaneously) the structures of several similarity matrices to the construction of a consensus matrix; we find that this helps to increase robustness of the method to noise in the data, relative to, for example, approaches based on ensemble methods or iterative consensus clustering [39, 38]. However challenges remain; it is not always possible to obtain a good prediction for the number of clusters in a dataset. This ambiguity is in part inherent to single cell analysis: depending on the level of focus, the number of relevant subpopulations may change, and this can be confounded by mixed discrete and continuous cell state transitions [9].

Single cell data analysis comes with a particular set of promises and

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

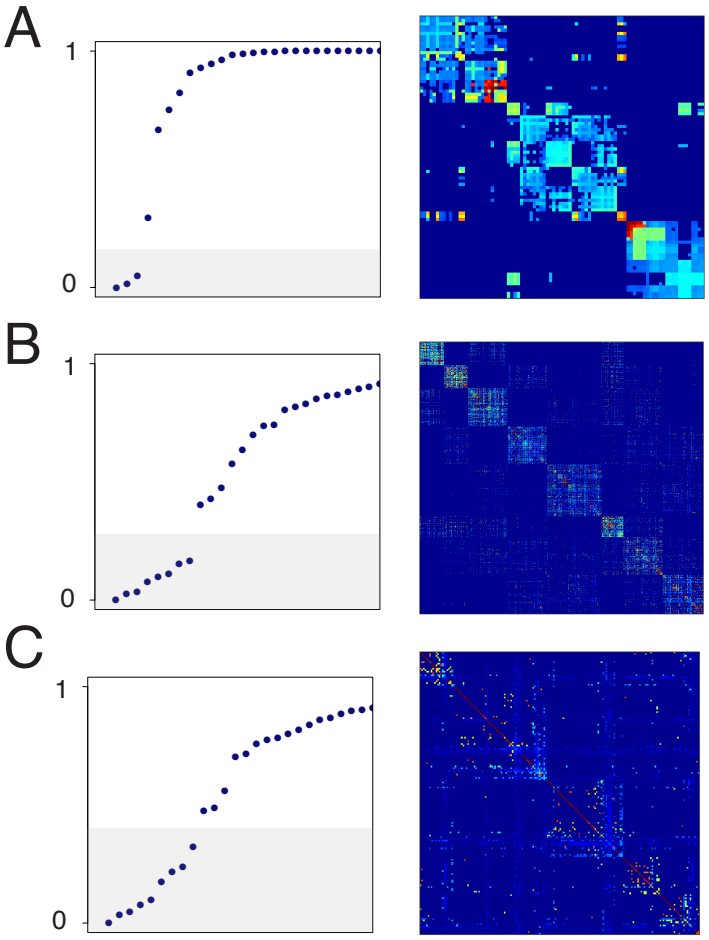
pitfalls. The key strength of scRNA-seq lies in its ability to measure many signals simultaneously and provide global quantification of the transcriptional state of a cell. On the other hand, technical challenges (due to amplification, alignment, dropout, etc [63]) — as well as challenges inherent to the biological system — make appropriately accounting for the heterogeneity in these data is a difficult problem. By presenting SoptSC, an optimization-based pipeline for clustering, pseudotemporal ordering, and cell lineage path reconstruction, we offer new methods for the analysis of scRNA-seq datasets that can stand alone or be integrated into existing workflows. We hope that as such, SoptSC will help to generate insight into emergent biological phenomena in complex tissues.

5. Availability

SoptSC is implemented in MATLAB under a GNU license (GPLv3). The code is available on GitHub: `https://github.com/WangShuxiong/SoptSC`.

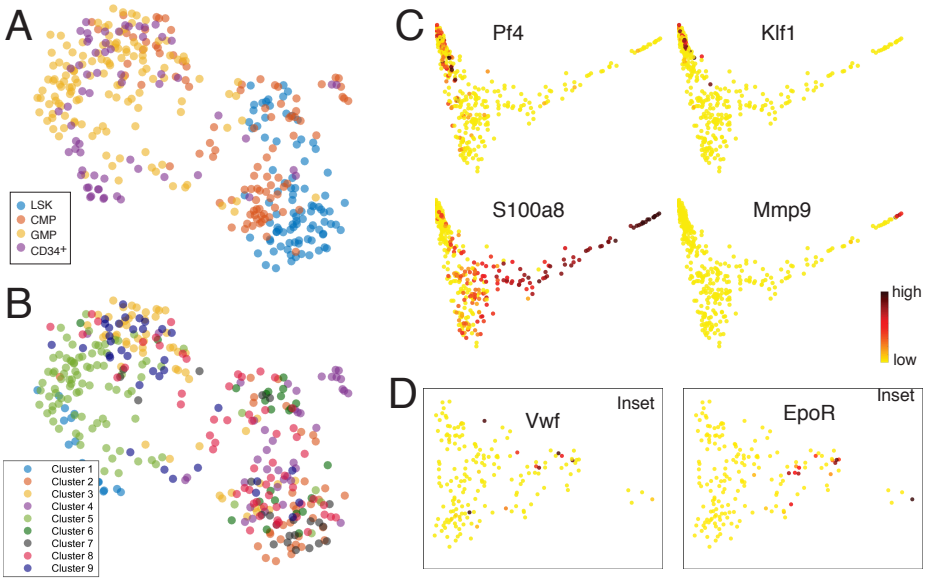
bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

6. Supporting information



S1 Fig.. **Eigenspectra and similarity matrices for the identification of subpopulations.** The first 25 eigenvalues of the graph Laplacian of the consensus matrix (left) and the similarity matrix constructed via SoptSC for (A) the human embryonic dataset studied [50]; (B) the interfollicular epidermal dataset studied [59]; and (C) the hematopoiesis dataset studied [12].

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



S2 Fig.. **Additional analysis of subpopulations and marker genes from Olsson et al. [12].** (A) tSNE projection of cells labelled by their cell surface marker expression: LSK: Lin$^-$Sca1$^+$c-Kit$^+$; CMP: common myeloid progenitor; GMP: granulocyte monocyte progenitor; CD34$^+$: LSK CD34$^+$ cells. (B) tSNE projection of cells labelled by their clusters as identified by SOptSC. (C) Gene expression of selected genes projected into 2D via SoptSC . (D) Gene expression in the inset projections as specified in Fig. 7.

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

[1] V. Svensson, R. Vento-Tormo, S. A. Teichmann, Moore's Law in Single Cell Transcriptomics, arXiv (2017) arXiv:1704.01379.

[2] F. Tang, C. Barbacioru, E. Nordman, B. Li, N. Xu, V. I. Bashkirov, K. Lao, M. A. Surani, RNA-Seq analysis to capture the transcriptome landscape of a single cell, Nature Protocols 5 (2010) 516–535.

[3] D. Ramsk ld, S. Luo, Y.-C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, G. P. Schroth, R. Sandberg, Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells, Nature biotechnology 30 (2012) 777–782.

[4] S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. L nnerberg, S. Linnarsson, Quantitative single-cell RNA-seq with unique molecular identifiers, Nature Methods 11 (2013) 163–166.

[5] D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, I. Amit, Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types, Science 343 (2014) 776–779.

[6] C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, et al., Molecular portraits of human breast tumours, Nature 406 (2000) 747–752.

[7] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by rna-seq, Nature methods 5 (2008) 621–628.

[8] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, M. Snyder, The transcriptional landscape of the yeast genome defined by rna sequencing, Science 320 (2008) 1344–1349.

[9] N. Moris, C. Pina, A. Martinez Arias, Transition states and cell fate decisions in epigenetic landscapes., Nature Reviews Genetics 17 (2016) 693–703.

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[10] F. Paul, Y. Arkin, A. Giladi, D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, D. Winter, D. Lara-Astiaso, M. Gury, A. Weiner, E. David, N. Cohen, F. K. B. Lauridsen, S. Haas, A. Schlitzer, A. Mildner, F. Ginhoux, S. Jung, A. Trumpp, B. T. Porse, A. Tanay, I. Amit, Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors, Cell 163 (2015) 1663–1677.

[11] D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, A. van Oudenaarden, Single-cell messenger RNA sequencing reveals rare intestinal cell types, Nature 525 (2015) 251–255.

[12] A. Olsson, M. Venkatasubramanian, V. K. Chaudhri, B. J. Aronow, N. Salomonis, H. Singh, H. L. Grimes, Single-cell analysis of mixed-lineage states leading to a binary cell fate choice, Nature 537 (2016) 698–702.

[13] B. Treutlein, D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantalas, F. H. Espinoza, T. J. Desai, M. A. Krasnow, S. R. Quake, Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq, Nature 509 (2014) 371–375.

[14] V. Moignard, S. Woodhouse, L. Haghverdi, A. J. Lilly, Y. Tanaka, A. C. Wilkinson, F. Buettner, I. C. Macaulay, W. Jawaid, E. Diamanti, S.-I. Nishikawa, N. Piterman, V. Kouskoff, F. J. Theis, J. Fisher, B. Gottgens, Decoding the regulatory network of early blood development from single-cell gene expression measurements, Nature biotechnology 33 (2015) 269–276.

[15] L.-F. Chu, N. Leng, J. Zhang, Z. Hou, D. Mamott, D. T. Vereide, J. Choi, C. Kendziorski, R. Stewart, J. A. Thomson, Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm, Genome Biology 17 (2016) 173.

[16] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data, Nature biotechnology 33 (2015) 495–502.

[17] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green,

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

M. Hemberg, SC3: consensus clustering of single-cell RNA-seq data., Nature Methods 9 (2017) 2579.

[18] M. Ringnér, What is principal component analysis?, Nature Biotechnology 26 (2008) 303–304.

[19] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (2008) 2579–2605.

[20] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, 2nd edition, 2009.

[21] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, J. L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells, Nature biotechnology 32 (2014) 381–386.

[22] G.-C. Yuan, L. Cai, M. Elowitz, T. Enver, G. Fan, G. Guo, R. Irizarry, P. Kharchenko, J. Kim, S. Orkin, et al., Challenges and emerging directions in single-cell analysis, Genome biology 18 (2017) 84.

[23] O. B. Poirion, X. Zhu, T. Ching, L. Garmire, Single-cell transcriptomics bioinformatics and computational challenges, Frontiers in Genetics 7 (2016).

[24] J. D. Welch, A. J. Hartemink, J. F. Prins, Slicer: inferring branched, nonlinear cellular trajectories from single cell rna-seq data, Genome biology 17 (2016) 106.

[25] L. Haghverdi, M. Buettner, F. A. Wolf, F. Buettner, F. J. Theis, Diffusion pseudotime robustly reconstructs lineage branching, Nature Methods 13 (2016) 845–848.

[26] J. Nocedal, S. Wright, Numerical optimization, Springer Science & Business Media, 2006.

[27] W. Sun, Y.-X. Yuan, Optimization theory and methods: nonlinear programming, volume 1, Springer Science & Business Media, 2006.

[28] H. Sayyaadi, M. H. Ahmadi, S. Dehghani, Optimal design of a solar-driven heat engine based on thermal and ecological criteria, Journal of Energy Engineering 141 (2014) 04014012.

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[29] F. Brauer, C. Castillo-Chavez, Mathematical models in population biology and epidemiology, volume 40, Springer Science & Business Media, 2011.

[30] J. Monk, J. Nogales, B. O. Palsson, Optimizing genome-scale network reconstructions, Nature biotechnology 32 (2014) 447–452.

[31] L. Zhuang, J. Wang, Z. Lin, A. Y. Yang, Y. Ma, N. Yu, Locality-preserving low-rank representation for graph construction from nonlinear manifolds, Neurocomputing 175 (2016) 715–722.

[32] J. H. Friedman, J. L. Bentley, R. A. Finkel, An algorithm for finding best matches in logarithmic expected time, ACM Transactions on Mathematical Software (TOMS) 3 (1977) 209–226.

[33] S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, science 290 (2000) 2323–2326.

[34] D. Kuang, S. Yun, H. Park, Symnmf: nonnegative low-rank approximation of a similarity matrix for graph clustering, Journal of Global Optimization 62 (2015) 545–574.

[35] D. Kuang, C. Ding, H. Park, Symmetric nonnegative matrix factorization for graph clustering, in: Proceedings of the 2012 SIAM international conference on data mining, SIAM, pp. 106–117.

[36] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, Y. Ma, Robust recovery of subspace structures by low-rank representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (2013) 171–184.

[37] C. Boutsidis, E. Gallopoulos, Svd based initialization: A head start for nonnegative matrix factorization, Pattern Recognition 41 (2008) 1350–1362.

[38] U. Von Luxburg, A tutorial on spectral clustering, Statistics and computing 17 (2007) 395–416.

[39] C. Meyer, S. Race, K. Valakuzhy, Determining the number of clusters via iterative consensus clustering, in: Proceedings of the 2013 SIAM International Conference on Data Mining, SIAM, pp. 94–102.

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[40] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, et al., Sc3-consensus clustering of single-cell rna-seq data, bioRxiv (2016) 036558.

[41] M. Nascimento, F. de Toledo, A. Carvalho, Consensus clustering using spectral theory, Advances in Neuro-Information Processing (2009) 461–468.

[42] I. Jolliffe, Principal component analysis, Wiley Online Library, 2002.

[43] B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, S. Batzoglou, Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning, Nature Methods 14 (2017) 414–416.

[44] D. Arthur, S. Vassilvitskii, 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, pp. 1027–1035.

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

[46] N. X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance, Journal of Machine Learning Research 11 (2010) 2837–2854.

[47] D. A. G. Card, P. B. Hebbar, L. Li, K. W. Trotter, Y. Komatsu, Y. Mishina, T. K. Archer, Oct4/Sox2-regulated miR-302 targets cyclin D1 in human embryonic stem cells., Molecular and Cellular Biology 28 (2008) 6426–6438.

[48] F. von Meyenn, M. Iurlaro, E. Habibi, N. Q. Liu, A. Salehzadeh-Yazdi, F. Santos, E. Petrini, I. Milagre, M. Yu, Z. Xie, L. I. Kroeze, T. B. Nesterova, J. H. Jansen, H. Xie, C. He, W. Reik, H. G. Stunnenberg, Impairment of DNA Methylation Maintenance Is the Main Cause of Global Demethylation in Naive Embryonic Stem Cells, Molecular Cell 62 (2016) 848–861.

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[49] G. Guo, M. Huss, G. Q. Tong, C. Wang, L. L. Sun, N. D. Clarke, P. Robson, Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst, Developmental cell 18 (2010) 675–685.

[50] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, et al., Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells, Nature structural & molecular biology 20 (2013) 1131–1139.

[51] E. Marco, R. L. Karp, G. Guo, P. Robson, A. H. Hart, L. Trippa, G.-C. Yuan, Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape, Proceedings of the National Academy of Sciences 111 (2014) E5643–E5650.

[52] J. Rossant, P. P. Tam, Blastocyst lineage formation, early embryonic asymmetries and axis patterning in the mouse, Development 136 (2009) 701–713.

[53] L. Haghverdi, F. Buettner, F. J. Theis, Diffusion maps for high-dimensional single-cell analysis of differentiation data, Bioinformatics 31 (2015) 2989–2998.

[54] M. Guo, E. L. Bao, M. Wagner, J. A. Whitsett, Y. Xu, Slice: determining cell differentiation and lineage based on single cell entropy, Nucleic Acids Research (2016) gkw1278.

[55] M. Guo, H. Wang, S. S. Potter, J. A. Whitsett, Y. Xu, Sincera: a pipeline for single-cell rna-seq profiling analysis, PLoS Comput Biol 11 (2015) e1004575.

[56] C. Blanpain, E. Fuchs, Epidermal Stem Cells of the Skin, Annual Review of Cell and Developmental Biology 22 (2006) 339–373.

[57] H. Yang, R. C. Adam, Y. Ge, Z. L. Hua, E. Fuchs, Epithelial-Mesenchymal Micro-niches Govern Stem Cell Lineage Choices, Cell 169 (2017) 1–14.

[58] J. W. Oh, J. Kloepper, E. A. Langan, Y. Kim, J. Yeo, M. J. Kim, T.-C. Hsi, C. Rose, G. S. Yoon, S.-J. Lee, J. Seykora, J. C. Kim, Y. K. Sung, M. Kim, R. Paus, M. V. Plikus, A Guide to Studying Human

bioRxiv preprint doi: https://doi.org/10.1101/168922; this version posted July 26, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Hair Follicle Cycling In Vivo, Journal of Investigative Dermatology 136 (2016) 34–44.

[59] S. Joost, A. Zeisel, T. Jacob, X. Sun, G. La Manno, P. Lönnerberg, S. Linnarsson, M. Kasper, Single-Cell Transcriptomics Reveals that Differentiation and Spatial Signatures Shape Epidermal and Hair Follicle Heterogeneity, Cell Systems 3 (2016) 221–237.e9.

[60] S. J. Morrison, D. T. Scadden, The bone marrow niche for haematopoietic stem cells, Nature 505 (2014) 327–334.

[61] F. Notta, S. Zandi, N. Takayama, S. Dobson, O. I. Gan, G. Wilson, K. B. Kaufmann, J. McLeod, E. Laurenti, C. F. Dunant, J. D. McPherson, L. D. Stein, Y. Dror, J. E. Dick, Distinct routes of lineage development reshape the human blood hierarchy across ontogeny, Science 351 (2016) aab2116–aab2116.

[62] Y. Hart, H. Sheftel, J. Hausser, P. Szekely, N. B. Ben-Moshe, Y. Korem, A. Tendler, A. E. Mayo, U. Alon, Inferring biological tasks using Pareto analysis of high-dimensional data, Nature Methods 12 (2015) 233–235.

[63] O. Stegle, S. A. Teichmann, J. C. Marioni, Computational and analytical challenges in single-cell transcriptomics., Nature Reviews Genetics 16 (2015) 133–145.