Supplemental Information

Cell lineage and communication network inference via optimization for single-cell transcriptomics

Shuxiong Wang[1], Matthew Karikomi[1], Adam L. MacLean[2,*] and Qing Nie[1,3,*]
[1] Department of Mathematics, University of California, Irvine, CA 92697, United States
[2] Department of Biological Sciences, University of Southern California, CA, 90089, USA
[3] Department of Developmental and Cell Biology, University of California, Irvine, CA 92697, United States
[*] Correspondence should be addressed to A.L.M (macleana@usc.edu) or Q.N. (qnie@uci.edu).

Contents

1 Extended Methods for SoptSC

SoptSC has two main components. The first component is a pipeline for discovering and ordering groups of cells and their associated marker genes. The second component evaluates all possible pair-wise interactions among the cells in the data set to provide a normalized prediction of ligand-receptor signaling. In the following sections we describe key elements of the first component in further detail. The overall workflow may be summarized as follows: First, a cell-cell similarity matrix S is constructed based on the gene expression matrix. Second the similarity matrix is used to find: i) cell subpopulations (clusters) via rank-k non-negative matrix factorization (NMF), where k is the number of subpopulations identified by spectral analysis, ii) ranked marker genes for each subpopulation, and iii) pseudotemporal ordering of cells and clusters obtained from a graph whose edges encoded in the similarity matrix.

1.1 Preprocessing of the data

There are two major steps in selecting highly variable genes in SoptSC. First, we remove genes that are expressed in less than $\alpha\%$ of cells or expressed at least $1-\alpha\%$ of cells ($\alpha = 6$ by default). This procedure is the same as SC3 for gene filtering. The motivation here is to filter genes that are not informative for clustering. Second, we perform principal component analysis (PCA) on the filtered single-cell data. Genes with highest loadings in the first k principal components are selected for further analysis. The value of k is chosen as the index at which the largest gap of the principal component variances occurs.

1.2 Cell-to-cell similarity matrix construction

The input to SoptSC is a single-cell gene expression matrix X with m rows of genes and n columns of cells. SoptSC computes the coefficient matrix Z from X by solving the following optimization model [27]:

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

where $||\cdot||_*$ is the nuclear norm, which is the sum of the singular values of a given matrix; $||\cdot||_{2,1}$ is the corresponding $l_{2,1}$ norm, which is the sum of the Euclidean norm of all columns of a given matrix; 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) [13], and a K-nearest neighbors (KNN) algorithm [3] is then applied to the low-dimensional data unlike in the original study in which KNN was directly applied without using dimension reduction. The linear constraint $Z^\top \mathbf{1} = 1$ ensures translational invariance of the data [20].

The optimization model (1) captures relationships between cells by representing each cell as a linear combination of other cells. By restricting coefficients of non-neighboring cells to be zero, the model preserves the local structure in the linear representation. By imposing the low rank constraint, this model can capture the global structure of the original data input, and is then more robust to noise and outliers. Problem (1) can be solved numerically by the Alternating Direction Method of multipliers [27]. Once the optimal solution Z^* in (1) is obtained, we can compute the similarity matrix $S = \max\left\{|Z^*|, |Z^{*\top}|\right\}$. The element $S_{i,j}$ $(= S_{j,i})$ of S thus quantifies the degree of similarity between cell i and cell j.

1.3 Symmetric Non-negative Matrix Factorization (NMF) for Clustering

In order to classify cells into subpopulations based on their similarities, we use symmetric non-negative matrix factorization (NMF), which can be regarded as a graph-based clustering method and is widely used for data clustering [10, 11]. 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:

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

where k is the number of subpopulations of cells and $||\cdot||_F$ is the Frobenius norm (i.e., $||A||_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n |A_{i,j}|^2}$ for any given matrix A). This model is similar to spectral clustering, where the non-negative constraint here is replaced by an orthogonal constraint [10].

It can be shown that the solution of (2) is strictly block-diagonal when the data are sampled from independent subspaces [12]. Thus the similarity matrix S can be approximated by a sum of rank one matrices $H^i H^{i\top}, i = 1, 2, ..., k$, where $H = [H^1, H^2, ..., H^n]$ is solved from the optimization problem (2). Singular value decomposition is used to find H_0, which is an initial low-rank non-negative matrix required as an input for (2) [1]. If $S = [S^1, S^2, ..., S^n]$, then the columns of S can be approximated by the space spanned by the columns of H as:

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

Consequently, 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 of the representation of the corresponding columns of S in the space spanned by the columns of H. We then use H^{T} to classify the N cells into k subpopulations or clusters by assigning the i^{th} cell to the j^{th} subpopulation when the largest element among all components of the i^{th} column of H^{T} lies in the j^{th} position.

1.4 Estimating the Number of Clusters

Determining the number of clusters in a dataset is a fundamental and challenging problem in unsupervised learning. Therefore, many clustering algorithms still require the user to specify the number of clusters. In SoptSC we identify the number of clusters within a dataset by analyzing the graph Laplacian and the consensus matrix [24], an approach similar to that used in a previous study [14]. The major steps required to determine the number of clusters are as follows:

1. Given the inputs S and a range of possible number of clusters $k_i, i = 1, 2, ..., q$, we first partition the cells into k_i subpopulations by solving (2).

2. Find the consensus matrix C [7, 15]. For each $j = 1, 2, ..., q$, 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 C is then defined by $C = \sum_{j=1}^{q} M^j$.

3. Prune the consensus matrix as follows: set a tolerance $\tau \in [0, 0.5]$, and let $C_{i,j} = 0$ if $C_{i,j} \leq \tau q$. This increases the robustness of consensus clustering to biological noise.

4. Compute the graph Laplacian $\mathbb{L}$ and its eigenvalues, given the identity matrix I and a diagonal matrix D such that:

$$\mathbb{L} = I - D^{-1/2} C D^{-1/2},$$

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

5. Find (i) the number of eigenvalues that are close to zero, and (ii) the index at which the largest eigenvalue gap occurs [24].

It has been shown that the number of eigenvalues of $\mathbb{L}$ equal to 0 is equivalent to the number of diagonal blocks of $\mathbb{L}$ [24], so the 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 Figure S1), (ii) can provide a better estimate of the subpopulation structure present in the data. For all analyses performed below, we choose a prior number of clusters ranging from 1 to 25, and we set the tolerance $\tau = 0.3$.

2 Supplementary Tables

2.1 Supplementary Table S1

Dataset	Number of Cells	Number of Genes	Number of Cell-types	Units
Yan [25]	90	20214	7	RPKM
Pollen [19]	249	23730	4, 11	TPM
Deng [2]	268	22431	6, 10	RPKM
Goolam [4]	124	41480	5	CPM
Kolodziejczyk [9]	704	10685	3	CPM
Treutlein [22]	80	23271	5	FPKM
Usoskin [23]	622	25334	4, 8, 11	RPM
Klein [8]	2717	24175	4	UMI
Zeisel [26]	3005	19972	9	UMI

Table S1. Summary of the characteristics of all the single-cell datasets used in clustering performance comparison the paper (for Figure 2).

2.2 Supplementary Table S2

Pathways	Ligand	Receptor	Target Genes
Tgfβ	Tgfβ1	Tgfβr1	Zeb2, Smad2, Wnt4, Wnt11, Bmp7, Sox9, Notch1
	Tgfβ1	Tgfβr2	Zeb2, Smad2, Wnt4, Wnt11, Bmp7, Sox9, Notch1
	Tgfβ2	Tgfβr1	Zeb2, Smad2, Wnt4, Wnt11, Bmp7, Sox9, Notch1
	Tgfβ2	Tgfβr2	Zeb2, Smad2, Wnt4, Wnt11, Bmp7, Sox9, Notch1
Bmp	Bmp1	Bmpr2	Crebbp, Fos, Id1, Jun, Runx1, Smad1, Smad5, Sox4, Cdh1
	Bmp2	Bmpr2	Crebbp, Fos, Id1, Jun, Runx1, Smad1, Smad5, Sox4, Cdh1
	Bmp4	Bmpr2	Crebbp, Fos, Id1, Jun, Runx1, Smad1, Smad5, Sox4, Cdh1
	Bmp7	Bmpr2	SCrebbp, Fos, Id1, Jun, Runx1, Smad1, Smad5, Sox4, Cdh1
Wnt	Wnt3	Fzd1	Ctnnb1, Lgr5, Runx2, Apc, Mmp7, Dkk1, Ccnd1
	Wnt4	Fzd1	Ctnnb1, Lgr5, Runx2, Apc, Mmp7, Dkk1, Ccnd1
	Wnt5a	Fzd1	Ctnnb1, Lgr5, Runx2, Apc, Mmp7, Dkk1, Ccnd1
	Wnt6	Fzd1	Ctnnb1, Lgr5, Runx2, Apc, Mmp7, Dkk1, Ccnd1
	Wnt10a	Fzd1	Ctnnb1, Lgr5, Runx2, Apc, Mmp7, Dkk1, Ccnd1

Table S2. Signaling pathways used for generating cell-to-cell signaling networks and cluster-to-cluster signaling networks from Joost data [6].

2.3 Supplementary Table S3

Pathways	Ligand	Receptor	Target Genes
Tgfβ	Tgfβ1	Tgfβr1	Zeb2, Smad2, Wnt4, Wnt11, Bmp7, Sox9, Notch1
	Tgfβ1	Tgfβr2	Zeb2, Smad2, Wnt4, Wnt11, Bmp7, Sox9, Notch1
	Tgfβ2	Tgfβr1	Zeb2, Smad2, Wnt4, Wnt11, Bmp7, Sox9, Notch1
	Tgfβ2	Tgfβr2	Zeb2, Smad2, Wnt4, Wnt11, Bmp7, Sox9, Notch1
Bmp	Bmp1	Bmpr2	Crebbp, Fos, Id1, Jun, Runx1, Smad1, Smad5, Sox4, Cdh1
	Bmp2	Bmpr2	Crebbp, Fos, Id1, Jun, Runx1, Smad1, Smad5, Sox4, Cdh1
	Bmp4	Bmpr2	Crebbp, Fos, Id1, Jun, Runx1, Smad1, Smad5, Sox4, Cdh1
	Bmp7	Bmpr2	SCrebbp, Fos, Id1, Jun, Runx1, Smad1, Smad5, Sox4, Cdh1
Wnt	Wnt3	Fzd1	Ctnnb1, Lgr5, Runx2, Apc, Mmp7, Dkk1, Ccnd1
	Wnt4	Fzd1	Ctnnb1, Lgr5, Runx2, Apc, Mmp7, Dkk1, Ccnd1
	Wnt5a	Fzd1	Ctnnb1, Lgr5, Runx2, Apc, Mmp7, Dkk1, Ccnd1
	Wnt6	Fzd1	Ctnnb1, Lgr5, Runx2, Apc, Mmp7, Dkk1, Ccnd1
	Wnt10a	Fzd1	Ctnnb1, Lgr5, Runx2, Apc, Mmp7, Dkk1, Ccnd1

Table S3. Signaling pathways used for generating cell-to-cell signaling networks and cluster-to-cluster signaling networks from Olsson data [17].

2.4 Supplementary Table S4

Pathways	**Ligand**	**Receptor**	**Target Genes**
Tgfβ	Tgfβ1	Tgfβr1	Zeb2, Smad2, Wnt4, Wnt11, Bmp7, Sox9, Notch1
	Tgfβ1	Tgfβr2	Zeb2, Smad2, Wnt4, Wnt11, Bmp7, Sox9, Notch1
	Tgfβ2	Tgfβr1	Zeb2, Smad2, Wnt4, Wnt11, Bmp7, Sox9, Notch1
	Tgfβ2	Tgfβr2	Zeb2, Smad2, Wnt4, Wnt11, Bmp7, Sox9, Notch1
Bmp	Bmp1	Bmpr2	Crebbp, Fos, Id1, Jun, Runx1, Smad1, Smad5, Cdh1
	Bmp2	Bmpr2	Crebbp, Fos, Id1, Jun, Runx1, Smad1, Smad5, Cdh1
	Bmp4	Bmpr2	Crebbp, Fos, Id1, Jun, Runx1, Smad1, Smad5, Cdh1
	Bmp7	Bmpr2	SCrebbp, Fos, Id1, Jun, Runx1, Smad1, Smad5, Cdh1
Wnt	Wnt3	Fzd10	Ctnnb1, Lgr5, Runx2, Apc, Mmp7, Dkk1, Ccnd1
	Wnt4	Fzd10	Ctnnb1, Lgr5, Runx2, Apc, Mmp7, Dkk1, Ccnd1
	Wnt5a	Fzd10	Ctnnb1, Lgr5, Runx2, Apc, Mmp7, Dkk1, Ccnd1
	Wnt8a	Fzd10	Ctnnb1, Lgr5, Runx2, Apc, Mmp7, Dkk1, Ccnd1
	Wnt8b	Fzd10	Ctnnb1, Lgr5, Runx2, Apc, Mmp7, Dkk1, Ccnd1

Table S4. Signaling pathways used for generating cell-to-cell signaling networks and cluster-to-cluster signaling networks for mouse hematopoietic stem cell differentiation [16].

3 Supplementary Figures

3.1 Supplementary Figure S1

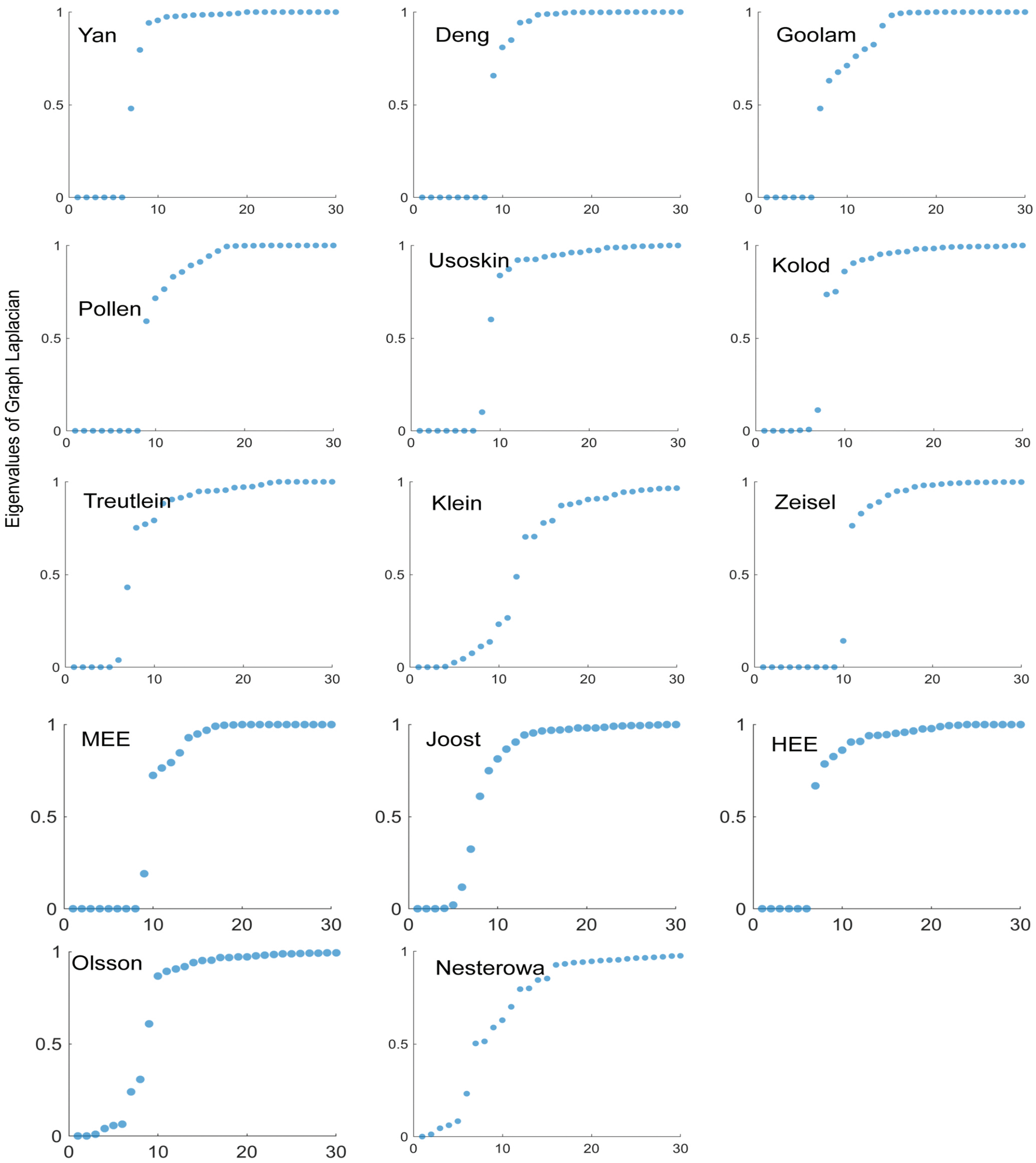


Figure S1. Spectra of the graph Laplacian of the truncated consensus matrix predict the number of clusters. The first 30 sorted eigenvalues of the graph Laplacian of the constructed consensus matrix for each data set is plotted.

3.2 Supplementary Figure S2

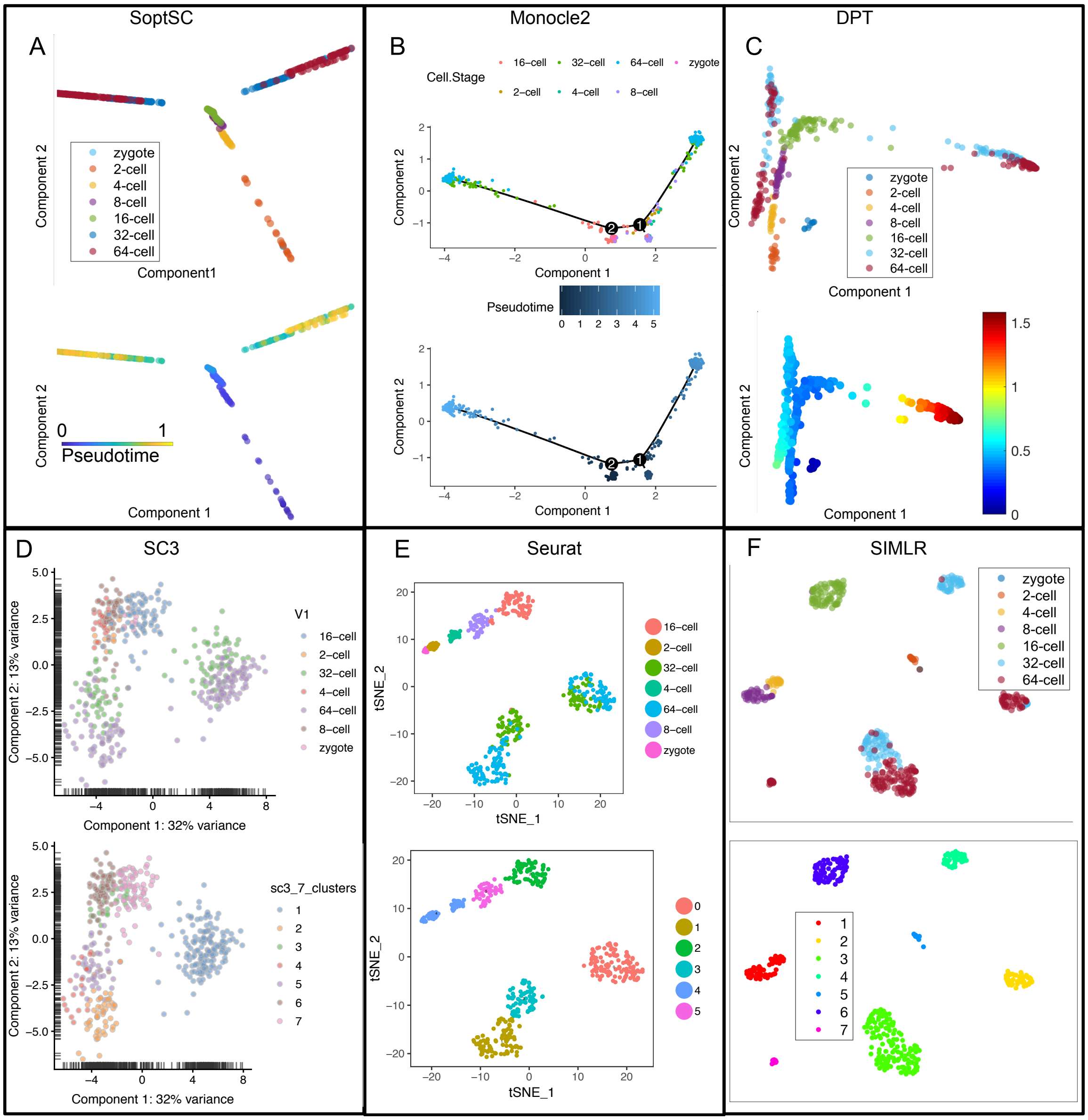


Figure S2. Pseudotime inference by SoptSC, Monocle2 and DPT; as well as clusters identified by SC3, Seurat, and SIMLR for mouse early embryonic data [5]. (A,B,C) Visualization of 2-dimensional trajectory of cells by (SoptSC, Monocle2, DPT) with true experimental time labels and pseudotime inferred by (SoptSC, Monocle2, DPT). (D,E,F) Visualization of low-dimensional projection of cells by (SC3,Seurat,SIMLR) with cell-stage labels and cluster labels identified by (SC3,Seurat,SIMLR).

3.3 Supplementary Figure S3

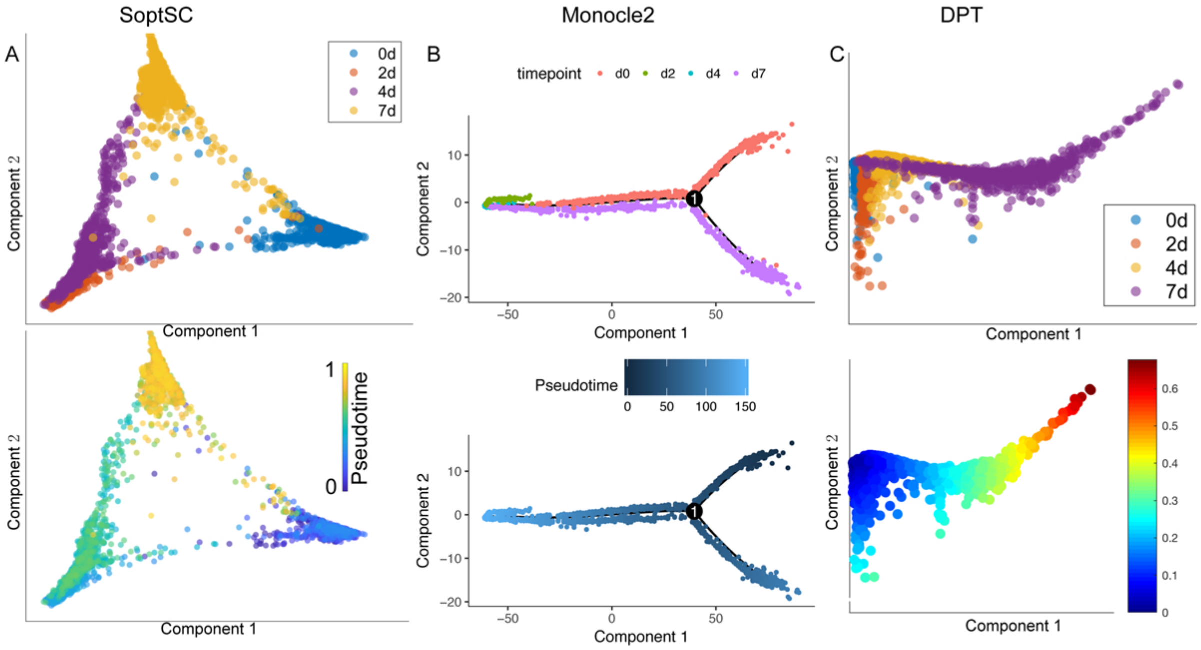


Figure S3. Pseudotime inference by SoptSC, Monocle2 and DPT for ESC data. [8] (A) Visualization of 2-dimensional projection of cells by SoptSC with true experimental time labels and pseudotime inferred by SoptSC. (B) Visualization of low-dimensional trajectory identified by Monocle2 with true time labels and pseudotime inferred by Monocle2. (C) Visualization of low-dimensional projection of cells by DPT with true time labels and pseudotime inferred by DPT.

3.4 Supplementary Figure S4

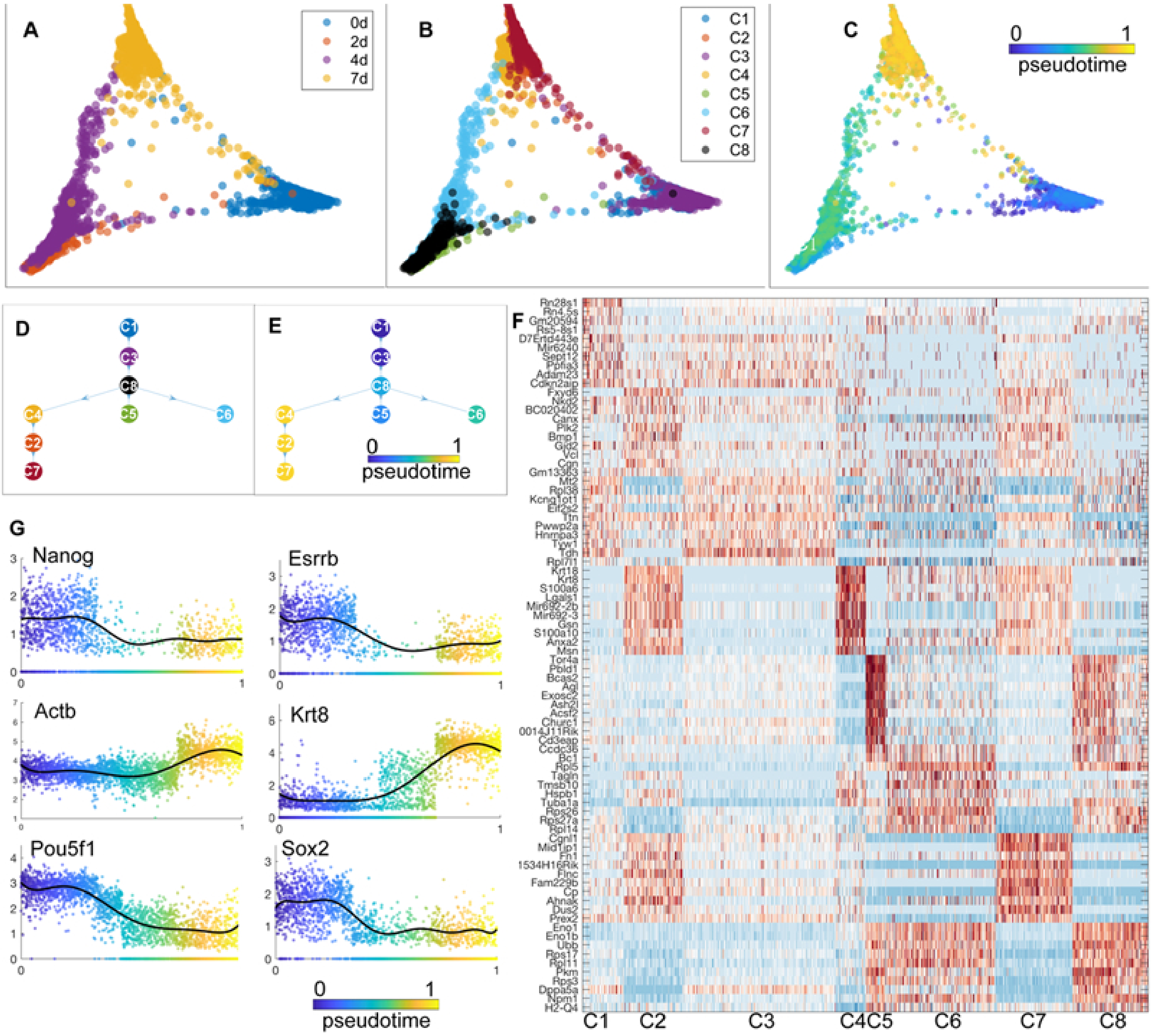


Figure S4. Analysis of the developmental trajectories of mouse ESCs [8]. (A) Visualization of cells labeled by true experimental time. (B) Cell subpopulations inferred by SoptSC. (C) Pseudotime for all cells. (D) Lineage inferred by SoptSC. (E) Pseudotime for cell states where the pseudotime of each state is calculated by the average of the temporal ordering of cells within the state. (F) Clustered gene-cell heatmap of genes from top 10 markers for each cluster identified by SoptSC. (G) Expression of selected marker genes along pseudotime.

3.5 Supplementary Figure S5

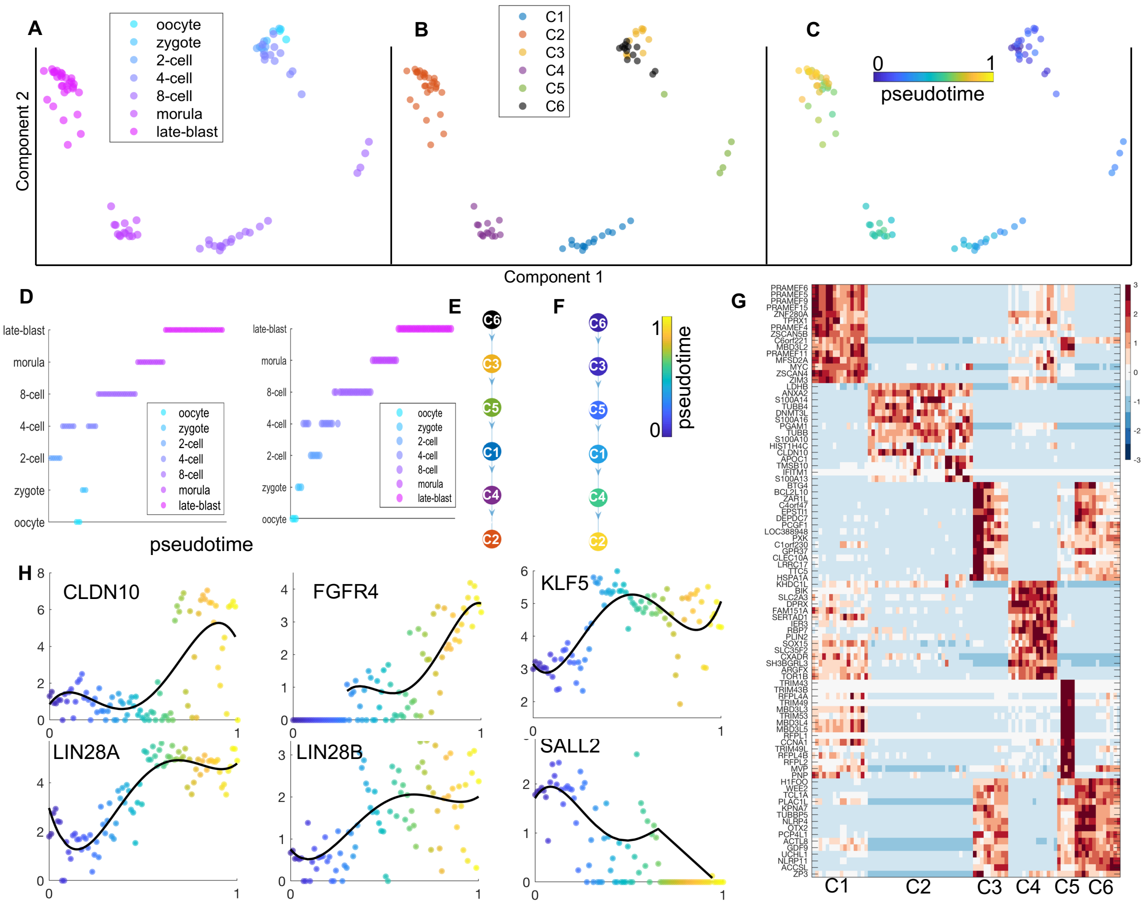


Figure S5. SoptSC identifies cell subpopulations, markers, lineage, and pseudotime in human early embryonic data. [25] (A) Visualization of cell in two-dimensional space labeled by true experimental time information. (B) Unsupervised subpopulations identified by SoptSC. (C) Pseudotime inferred by SoptSC. (D) Pseudotemporal cell ordering with initial state inferred unsupervised (**left**) or set by the user (**right**) is compared with known experimental stages. The Kendall rank correlation measured with experimental stages against i) the temporal ordering of cells inferred by SoptSC unsupervised is 0.84, and against ii) the temporal ordering with initial state set is 0.86. Due to the variability between oocyte and 4-cell stages and the very small number of cells, SoptSC was not able to infer the earliest developmental cell orderings unsupervised; if the initial state is provided, SoptSC predicts the trajectory with high consistency. (E) Lineage inferred by SoptSC, indicating a linear trajectory. (F) Pseudotime along lineage, where the pseudotime of each subpopulation is calculated by the average of the temporal ordering of cells within the subpopulation. (G) Clustered gene-cell heatmap of genes from top 15 markers for each cluster identified by SoptSC. (H) Expression of selected marker genes along pseudotime.

3.6 Supplementary Figure S6

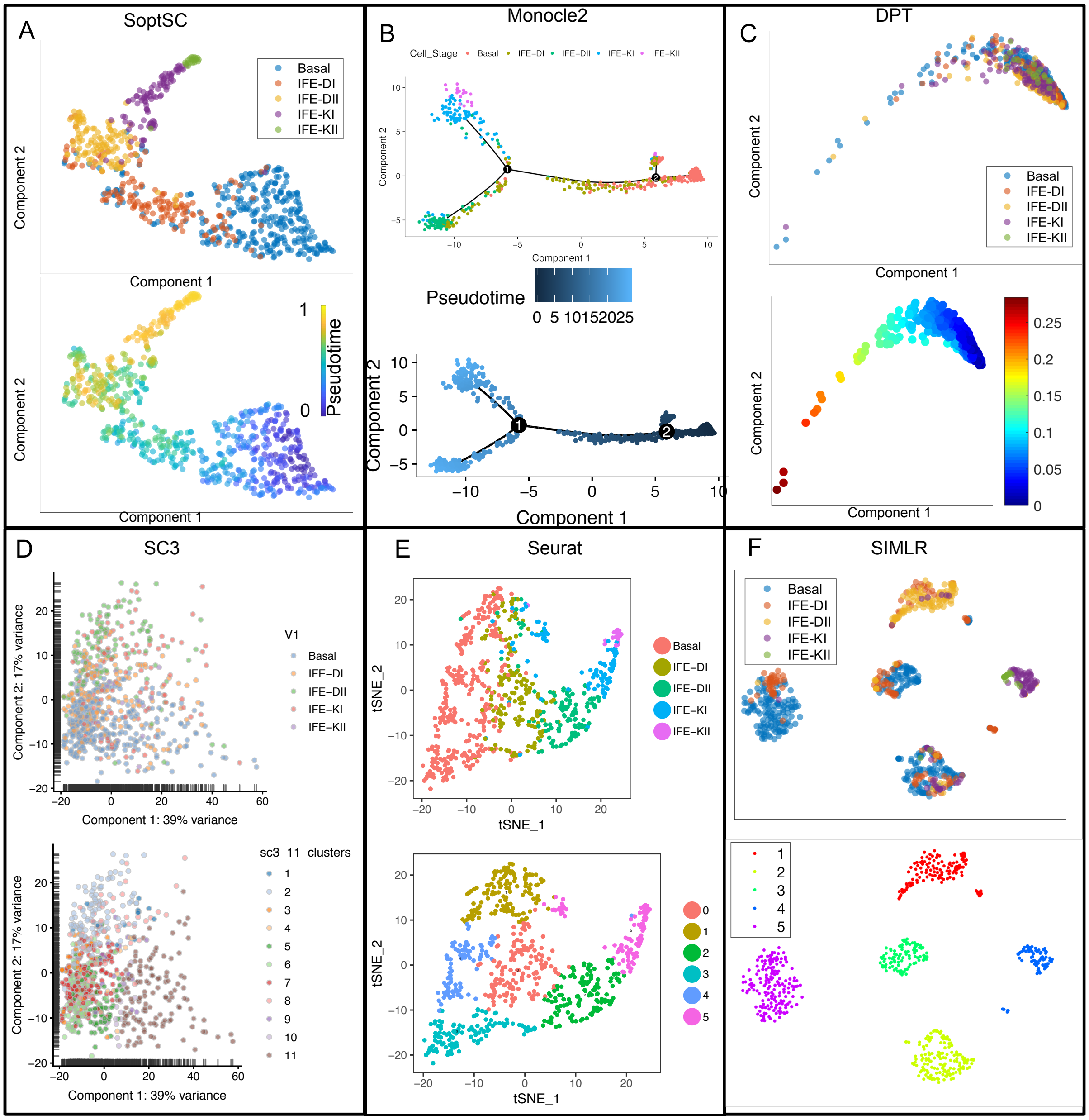


Figure S6. Pseudotime inference by SoptSC, Monocle2 and DPT; and clusters identified by SC3, Seurat, and SIMLR for IFE data [6]. (A, B, C) Visualization of two-dimensional projection of cells by (SoptSC, Monocle2, DPT) with true experimental time labels and pseudotime inferred by (SoptSC, Monocle2, DPT). (D, E, F) Visualization of low-dimensional projection of cells by (SC3,Seurat,SIMLR) with cell-stage labels and cluster labels identified by (SC3,Seurat,SIMLR).

3.7 Supplementary Figure S7

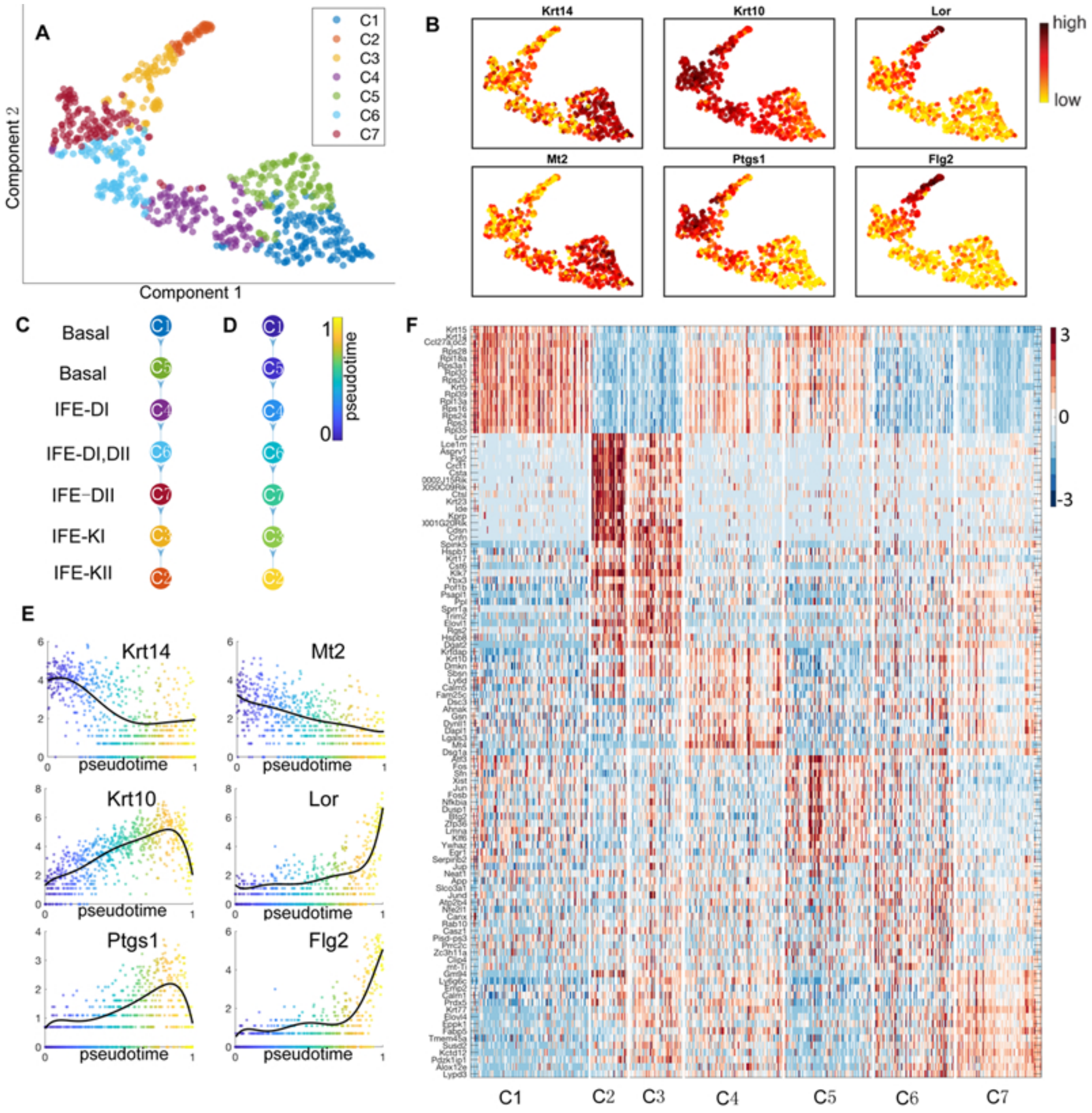


Figure S7. Analysis of epidermal differentiation in the IFE. [**6**] (A) Cell subpopulations inferred by SoptSC and the low dimensional visualization. (B) Selected marker genes and their expression in each population. (C) Lineage tree identified by SoptSC. (D) Pseudotime for cell states where the pseudotime of each subpopulation is calculated by the average of the temporal ordering of cells within the subpopulation. (E) Expression of selected marker genes along pseudotime. (F) Clustered gene-cell heatmap of genes from top 15 markers for each cluster identified by SoptSC.

3.8 Supplementary Figure S8

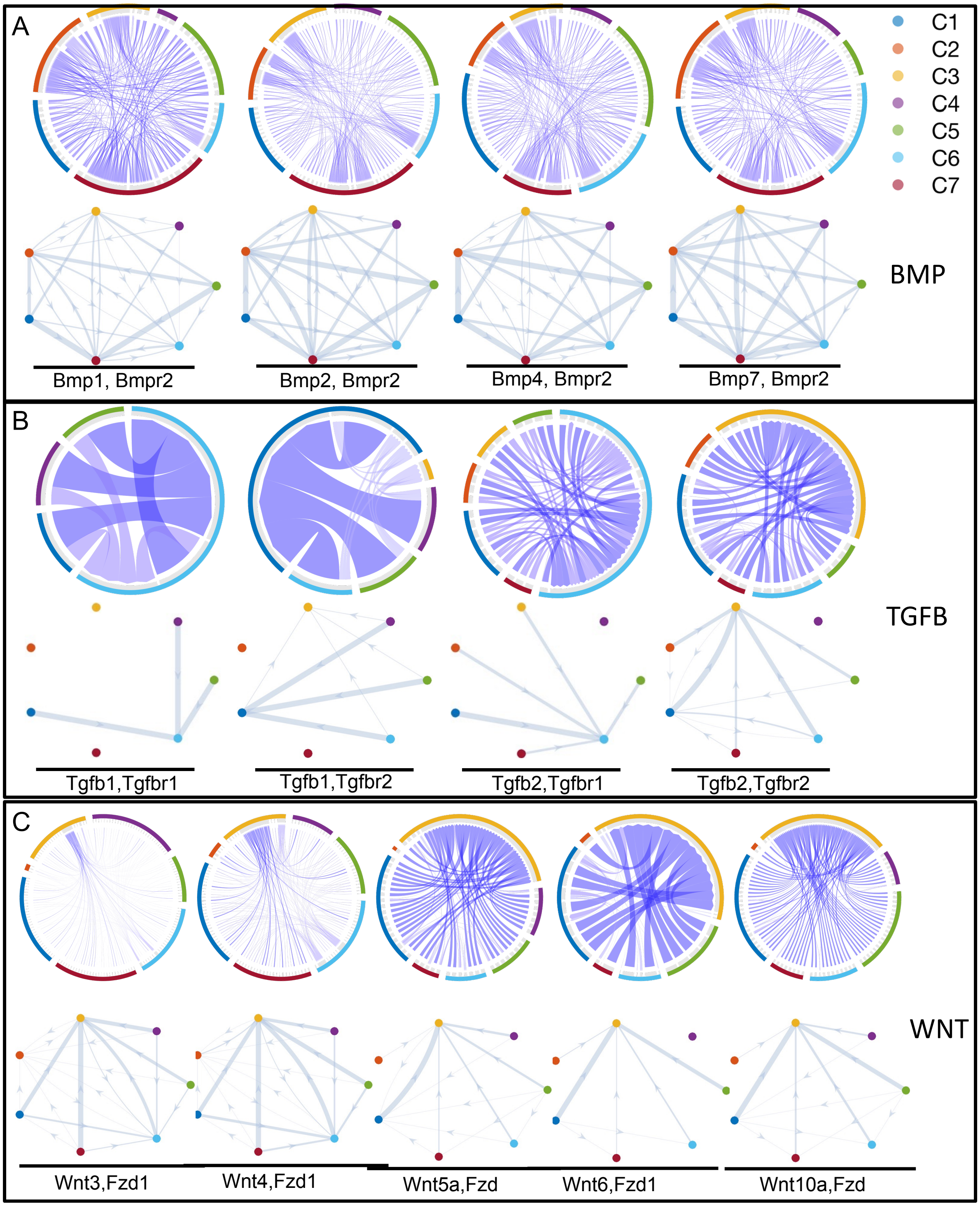
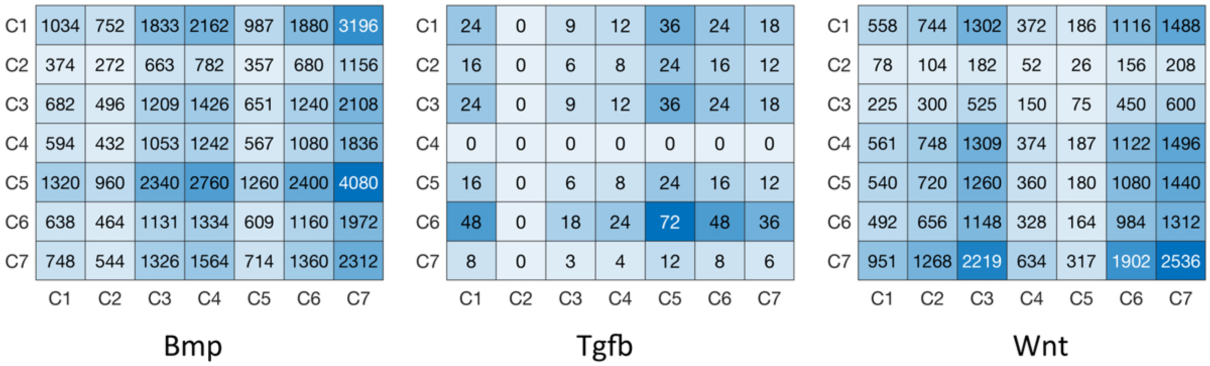


Figure S8. Single-cell signaling networks predicted for the BMP, TGFβ, and Wnt pathways from Joost et al. data [6]. (A) Cell-cell signaling network and cluster-cluster signaling network for individual ligand-receptor pair from BMP pathway. Top row: single-cell signaling networks for ligand-receptor pairs, with edge weights corresponding to the probability of a signal passed between cells. Bottom row: cluster-to-cluster signaling interactions with edge weights corresponding to the probability of a signal passed between clusters. Colors correspond to the cluster labels. (B) Cell-cell signaling network and cluster-cluster signaling network for individual ligand-receptor pair from TGFβ pathway. (C) Cell-cell signaling network and cluster-cluster signaling network for individual ligand-receptor pair from Wnt pathway. Target gene list is summarized in **Table S2**.

3.9 Supplementary Figure S9



	C1	C2	C3	C4	C5	C6	C7
C1	1034	752	1833	2162	987	1880	3196
C2	374	272	663	782	357	680	1156
C3	682	496	1209	1426	651	1240	2108
C4	594	432	1053	1242	567	1080	1836
C5	1320	960	2340	2760	1260	2400	4080
C6	638	464	1131	1334	609	1160	1972
C7	748	544	1326	1564	714	1360	2312

Bmp

	C1	C2	C3	C4	C5	C6	C7
C1	24	0	9	12	36	24	18
C2	16	0	6	8	24	16	12
C3	24	0	9	12	36	24	18
C4	0	0	0	0	0	0	0
C5	16	0	6	8	24	16	12
C6	48	0	18	24	72	48	36
C7	8	0	3	4	12	8	6

Tgfb

	C1	C2	C3	C4	C5	C6	C7
C1	558	744	1302	372	186	1116	1488
C2	78	104	182	52	26	156	208
C3	225	300	525	150	75	450	600
C4	561	748	1309	374	187	1122	1496
C5	540	720	1260	360	180	1080	1440
C6	492	656	1148	328	164	984	1312
C7	951	1268	2219	634	317	1902	2536

Wnt

Figure S9. The table counts all possible signaling events (for BMP, Tgf-β, and Wnt pathways) between the cells in one cluster C_i and the cells in another cluster C_j, where an event occurs when a cell from cluster C_i expresses ligand and cell from cluster C_j expresses one of its receptors from **Table S2**. The results come from Joost et al. data [6].

3.10 Supplementary Figure S10

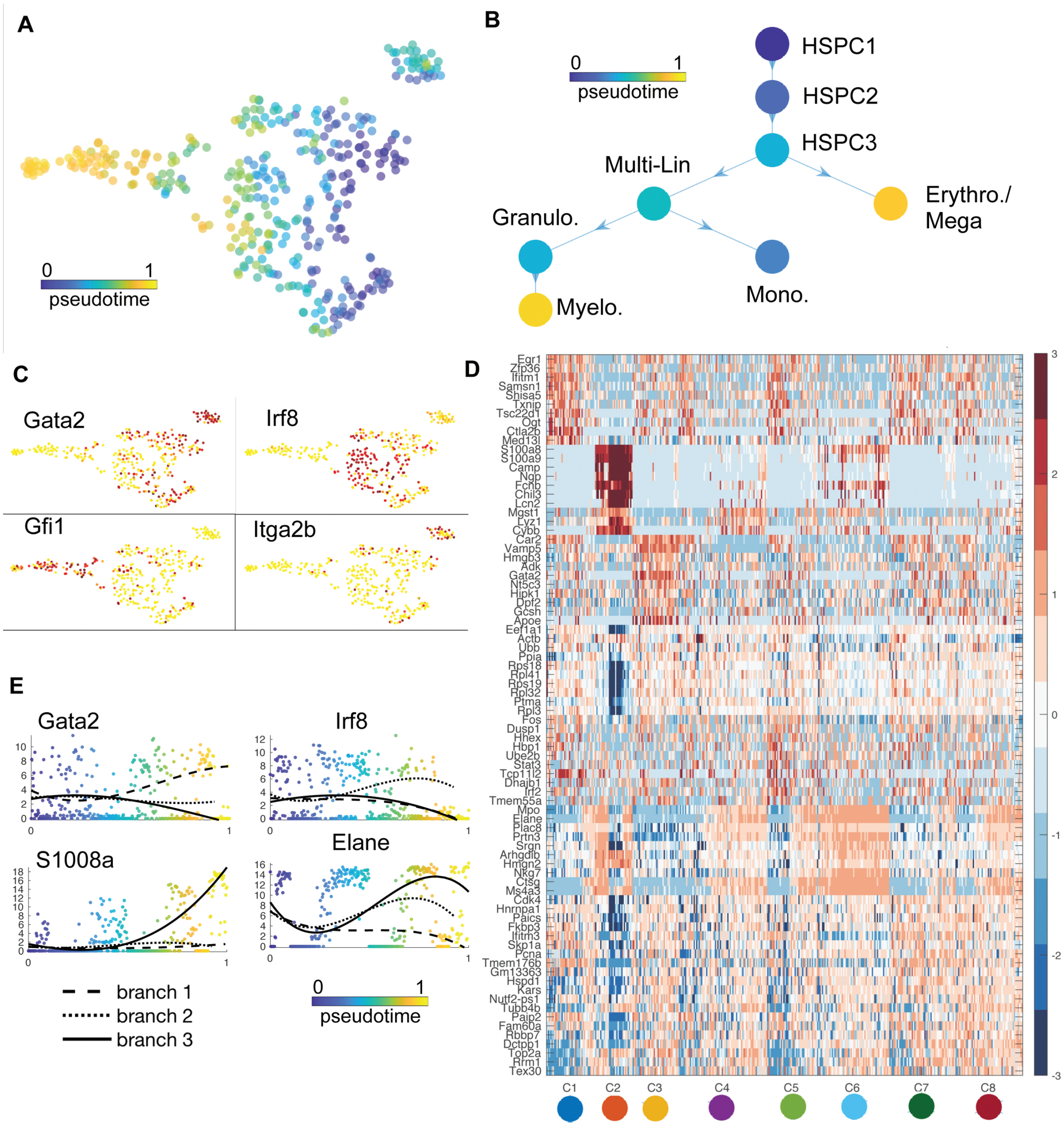


Figure S10. Caption on next page.

Figure S10. Analysis of subpopulations, pseudotime, and lineage paths during myelopoiesis [18]. (A) Pseudotemporal ordering of hematopoietic cells by SoptSC. (B) Lineage hierarchy constructed by SoptSC. Colors correspond to the mean pseudotime value for the subpopulation. Subpopulation identities have been annotated according to marker gene expression. HSPC: hematopoietic stem/progenitor cells; Multi-Lin: mixed progenitor (see [18]); Mono: monocytic progenitor; Granulo: granulocytic progenitor; Myelo: myelocytic progenitor; Erythro: erythrocytic progenitor; Mega: megakaryocytic progenitor. (C) Gene expression of selected markers. (D) Clustered gene-cell heatmap of genes from top 15 markers for each cluster identified by SoptSC; Clusters from C1 to C8 correspond to: HSPC2, Myelo., Erythro./Mega, Mono., HSPC1, Granulo., HSPC3, MultiLin (see legend in **Figure 6B** of the main text). (E) Expression of selected markers along pseudotime where branch 1 corresponds to HSPC1, HSPC2, HSPC3, Erythro./Mega; branch 2 corresponds to HSPC1, HSPC2, HSPC3, Multi-Lin, Mono.; and branch 3 corresponds to HSPC1, HSPC2, HSPC3, Granulo., Myelo.;

3.11 Supplementary Figure S11

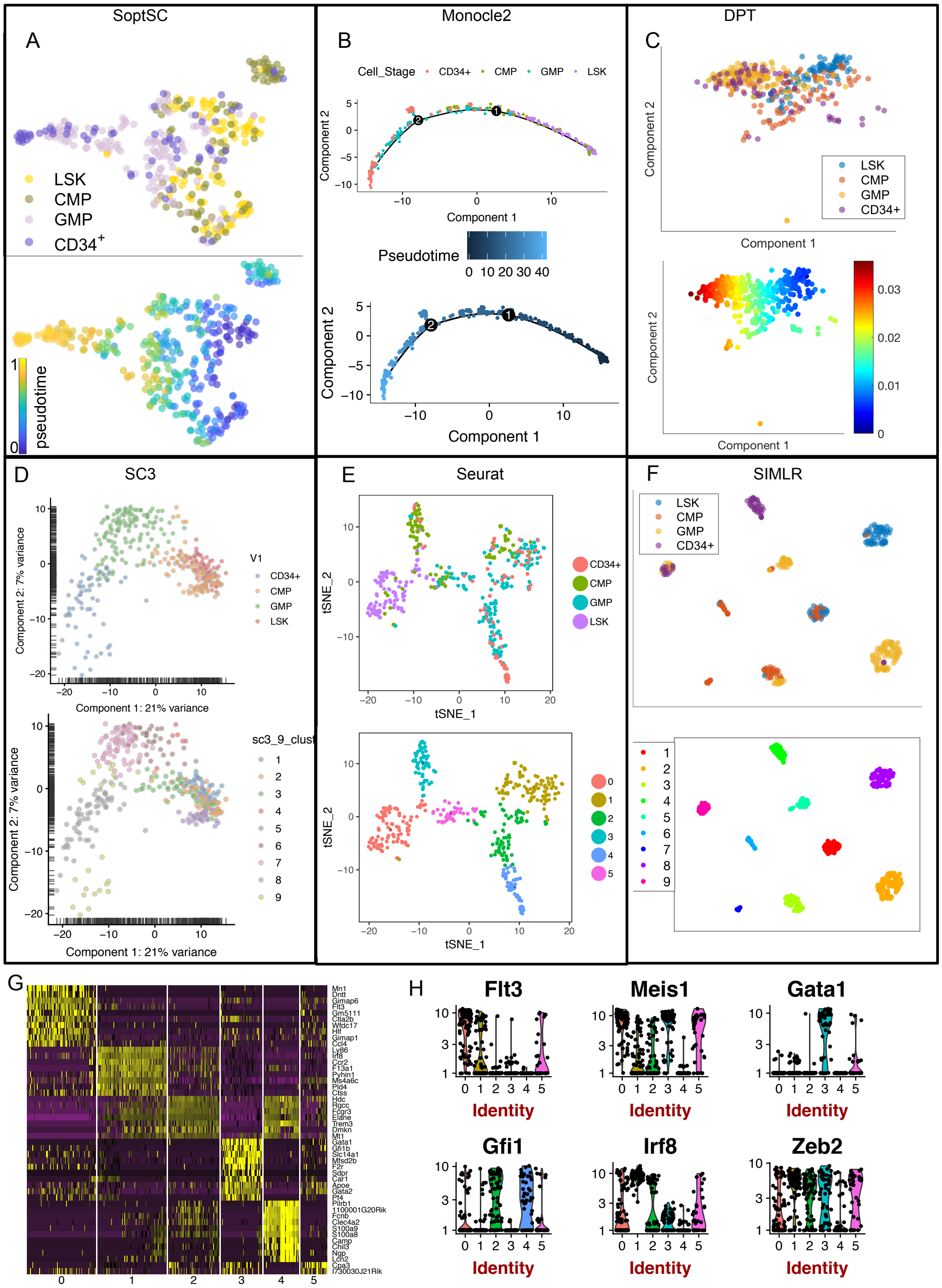


Figure S11. Pseudotime inference by SoptSC, Monocle2 and DPT; and clusters identified by SC3, Seurat, and SIMLR for Olsson et al. data [17]. (A, B, C) Visualization of two-dimensional projection of cells by (SoptSC, Monocle2, DPT) with true labels from the original study and pseudotime inferred by (SoptSC, Monocle2, DPT). (D, E, F) Visualization of low-dimensional projection of cells by (SC3,Seurat,SIMLR) with cell-stage labels and cluster labels identified by (SC3, Seurat, SIMLR). (G) Heatmap of Top 10 marker genes from Seurat. (H) Violin plots of selected markers.

3.12 Supplementary Figure S12

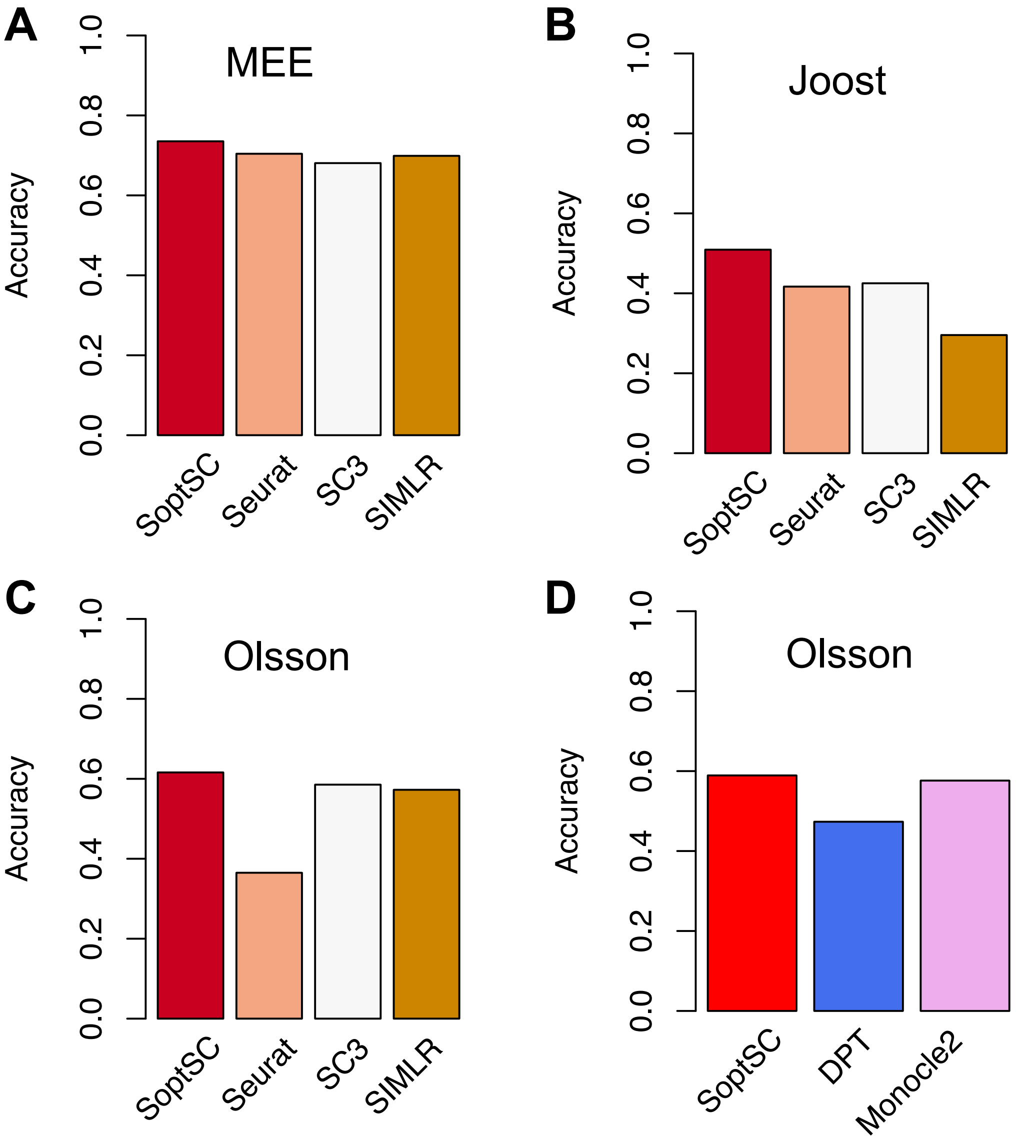


Figure S12. Comparison of performance of clustering for methods SoptSC, Seurat, SC3, SIMLR, and pseudotime inference for SoptSC, Monocle2, DPT (A) Accuracy of clusters identified by SoptSC, Seurat, SC3 and SIMLR for mouse early embryonic data [5]. (B) Accuracy of clusters identified by SoptSC, Seurat, SC3 and SIMLR for IFE data [6]. (C) Accuracy of clusters identified by SoptSC, Seurat, SC3 and SIMLR for Olsson et al. data data [17]. (D) Accuracy of pseudotime inferred by SoptSC, Monocle2 and DPT for Olsson et al. data data [17].

3.13 Supplementary Figure S13

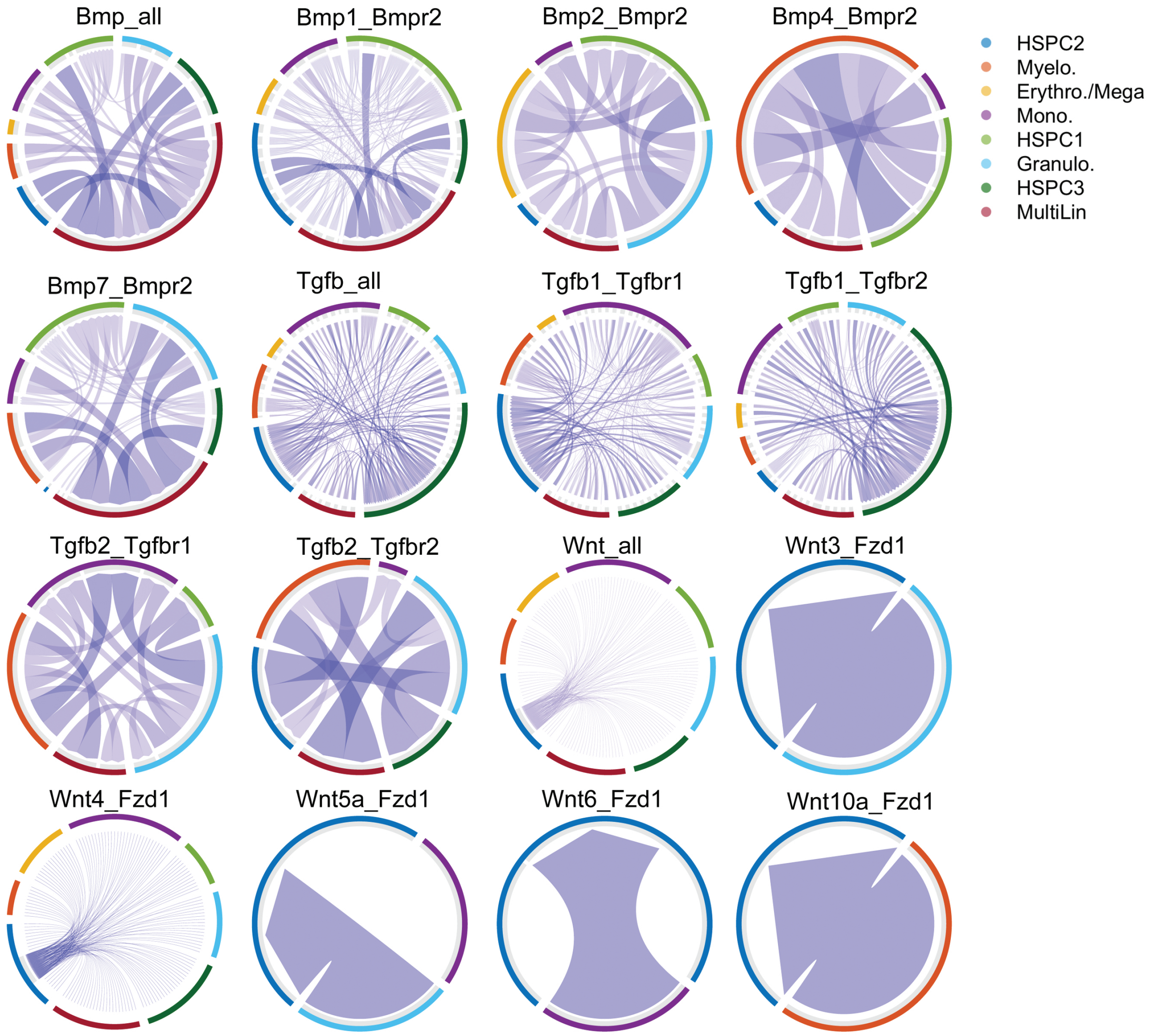


Figure S13. Cell-to-cell signaling networks predicted for pathways in data from Olsson et al. [18] Single-cell signaling networks for ligand-receptor pairs, with edge weights corresponding to the probability of a signal passed between cells. Members of the pathways analyzed for Bmp, Tgf-β and Wnt are summarized in **Table S3**.

3.14 Supplementary Figure S14

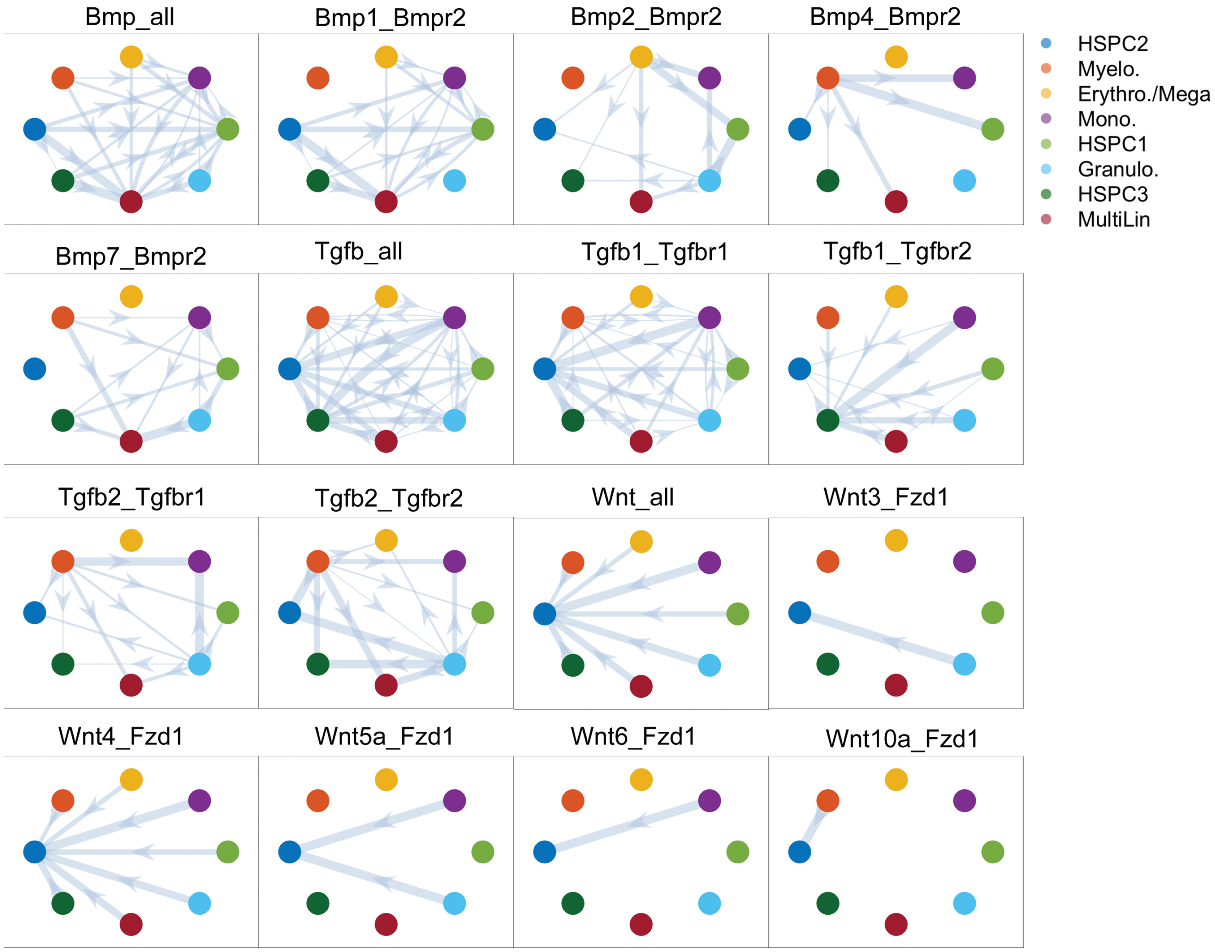
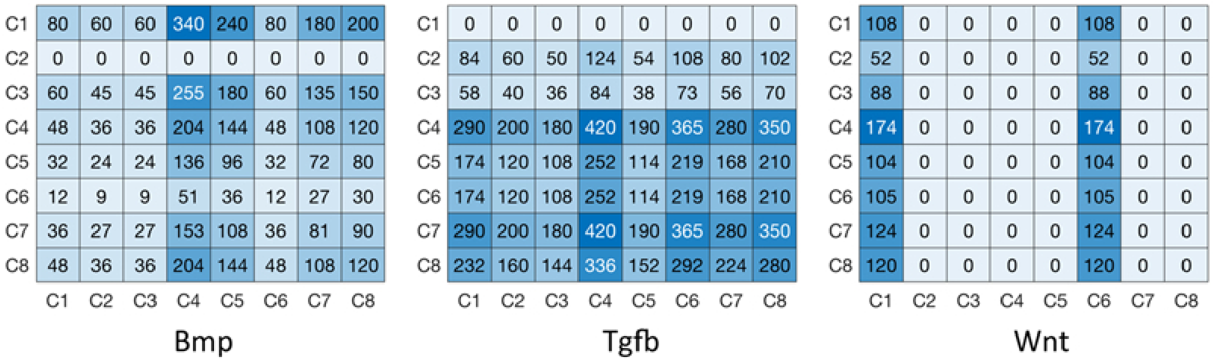


Figure S14. Cluster-to-cluster signaling networks predicted for pathways in data from Olsson et al. [18] Bmp_all, Tgfβ_all and Wnt_all represent the signaling network between clusters using all provided ligand-receptor pairs for the specific pathways.

3.15 Supplementary Figure S15



	C1	C2	C3	C4	C5	C6	C7	C8
C1	80	60	60	340	240	80	180	200
C2	0	0	0	0	0	0	0	0
C3	60	45	45	255	180	60	135	150
C4	48	36	36	204	144	48	108	120
C5	32	24	24	136	96	32	72	80
C6	12	9	9	51	36	12	27	30
C7	36	27	27	153	108	36	81	90
C8	48	36	36	204	144	48	108	120

Bmp

	C1	C2	C3	C4	C5	C6	C7	C8
C1	0	0	0	0	0	0	0	0
C2	84	60	50	124	54	108	80	102
C3	58	40	36	84	38	73	56	70
C4	290	200	180	420	190	365	280	350
C5	174	120	108	252	114	219	168	210
C6	174	120	108	252	114	219	168	210
C7	290	200	180	420	190	365	280	350
C8	232	160	144	336	152	292	224	280

Tgfb

	C1	C2	C3	C4	C5	C6	C7	C8
C1	108	0	0	0	0	108	0	0
C2	52	0	0	0	0	52	0	0
C3	88	0	0	0	0	88	0	0
C4	174	0	0	0	0	174	0	0
C5	104	0	0	0	0	104	0	0
C6	105	0	0	0	0	105	0	0
C7	124	0	0	0	0	124	0	0
C8	120	0	0	0	0	120	0	0

Wnt

Figure S15. The table counts all possible signaling events (for BMP, Tgf-β, and Wnt pathways) between the cells in one cluster C_i and the cells in another cluster C_j, where an event occurs when a cell from cluster C_i expresses ligand and cell from cluster C_j expresses one of its receptors from **Table S3**. The resuls come from Olsson et al. data [18]. {C1,C2,C3,C4,C5,C6,C7,C8} corresponds to {HSPC2, Myelo., Erythro./Mega, Mono., HSPC1, Granulo., HSPC3, MultiLin}.

3.16 Supplementary Figure S16

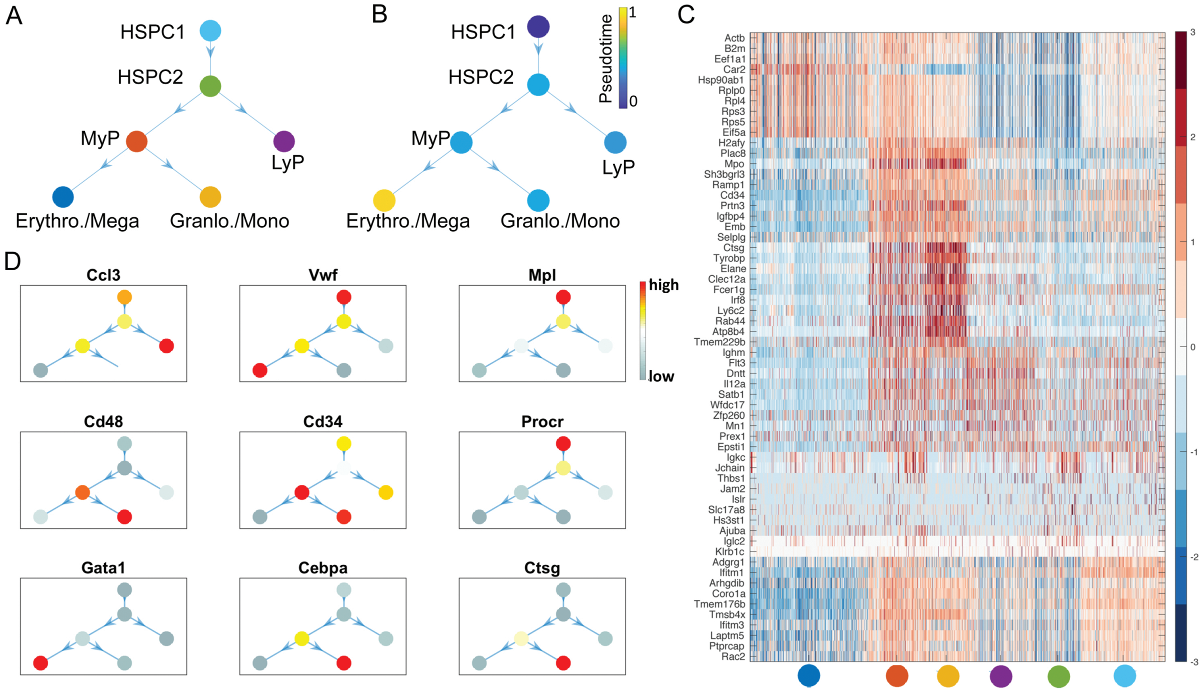


Figure S16. Analysis of pseudotime, and lineage paths for mouse hematopoietic stem cell differentiation [16]. (A) Lineage inferred by SoptSC. (B) Lineage hierarchy constructed by SoptSC. Colors correspond to the mean pseudotime value for the subpopulation. (C) Clustered gene-cell heatmap of genes from top 10 markers for each cluster identified by SoptSC; (D) Gene expression of selected markers. Colors represent the mean expression for each gene within each cluster.

3.17 Supplementary Figure S17

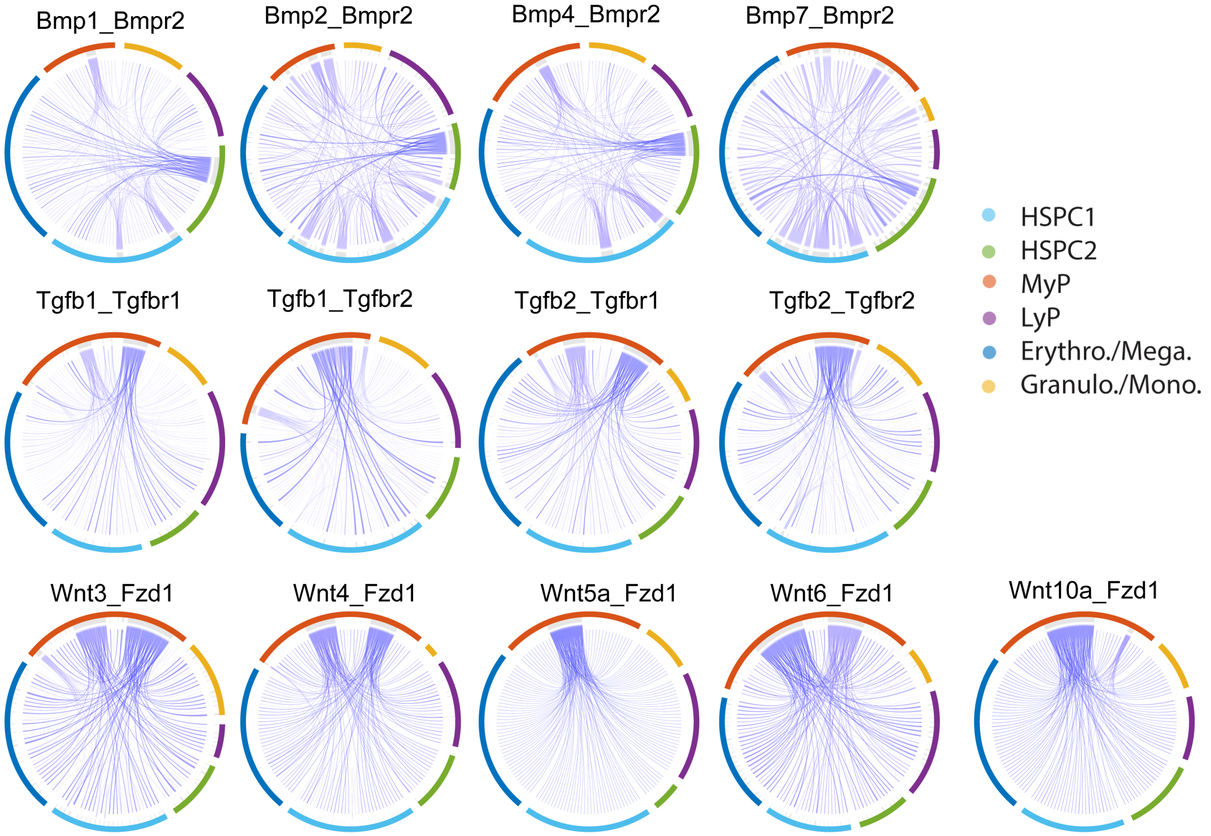


Figure S17. Cell-to-cell signaling networks predicted for pathways in data for mouse hematopoietic stem cell differentiation [16]. Single-cell signaling networks for ligand-receptor pairs, with edge weights corresponding to the probability of a signal passed between cells. Members of the pathways analyzed for Bmp, Tgf-β and Wnt are summarized in **Table S4**.

3.18 Supplementary Figure S18

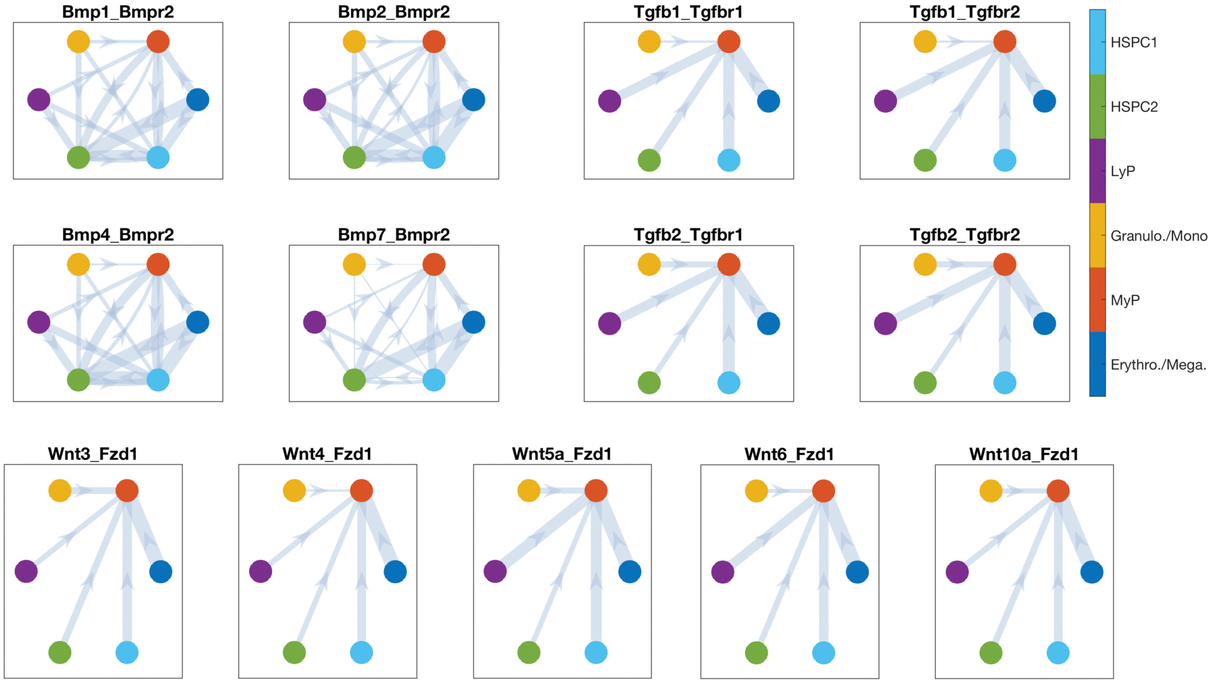
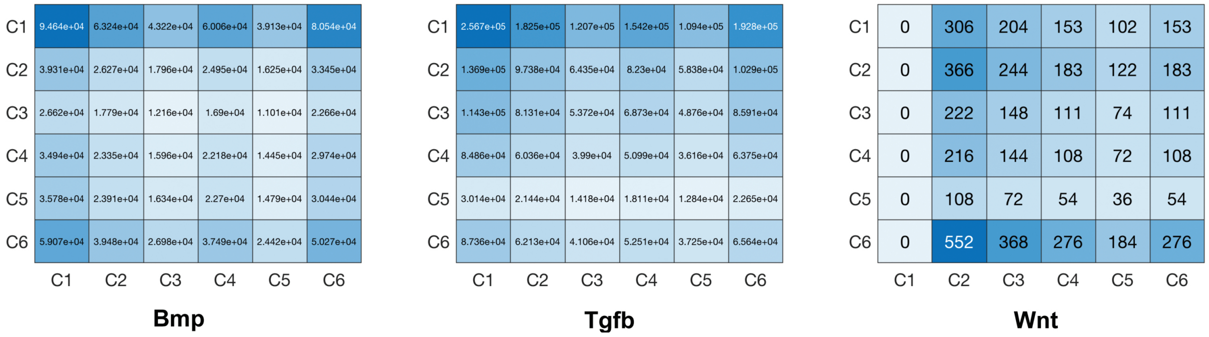


Figure S18. Cluster-to-cluster signaling networks predicted for pathways in data for mouse hematopoietic stem cell differentiation [16]. Bmp_all, Tgfβ_all and Wnt_all represent the signaling network between clusters using all provided ligand-receptor pairs for the specific pathways.

3.19 Supplementary Figure S19



Bmp

	C1	C2	C3	C4	C5	C6
C1	9.464e+04	6.324e+04	4.322e+04	6.006e+04	3.913e+04	8.054e+04
C2	3.931e+04	2.627e+04	1.796e+04	2.495e+04	1.625e+04	3.345e+04
C3	2.662e+04	1.779e+04	1.216e+04	1.69e+04	1.101e+04	2.266e+04
C4	3.494e+04	2.335e+04	1.596e+04	2.218e+04	1.445e+04	2.974e+04
C5	3.578e+04	2.391e+04	1.634e+04	2.27e+04	1.479e+04	3.044e+04
C6	5.907e+04	3.948e+04	2.698e+04	3.749e+04	2.442e+04	5.027e+04

Tgfb

	C1	C2	C3	C4	C5	C6
C1	2.567e+05	1.825e+05	1.207e+05	1.542e+05	1.094e+05	1.928e+05
C2	1.369e+05	9.738e+04	6.435e+04	8.23e+04	5.838e+04	1.029e+05
C3	1.143e+05	8.131e+04	5.372e+04	6.873e+04	4.876e+04	8.591e+04
C4	8.486e+04	6.036e+04	3.99e+04	5.099e+04	3.616e+04	6.375e+04
C5	3.014e+04	2.144e+04	1.418e+04	1.811e+04	1.284e+04	2.265e+04
C6	8.736e+04	6.213e+04	4.106e+04	5.251e+04	3.725e+04	6.564e+04

Wnt

	C1	C2	C3	C4	C5	C6
C1	0	306	204	153	102	153
C2	0	366	244	183	122	183
C3	0	222	148	111	74	111
C4	0	216	144	108	72	108
C5	0	108	72	54	36	54
C6	0	552	368	276	184	276

Figure S19. The table counts all possible signaling events (for BMP, Tgf-β, and Wnt pathways) between the cells in one cluster C_i and the cells in another cluster C_j, where an event occurs when a cell from cluster C_i expresses ligand and cell from cluster C_j expresses one of its receptors from **Table S4**. The results come from single-cell data for mouse hematopoietic stem cell differentiation [16]. {C1,C2,C3,C4,C5,C6} corresponds to {Erythro./Mega., MyP, Granulo./Mono., LyP, HSPC2, , HSPC1}.

4 Extended Details on Data Analysis

4.1 Details of Data Analysis by SoptSC

All the results in this paper is run under MATLAB R2017b on Mac Pro (Late 2013) with 3.5 GHz 6-Core Intel Xeon E5.

For the datasets [2, 4, 8, 9, 19, 22, 23, 25, 26] used in evaluating the clustering performance (**Fig. 2**), we set $\alpha = 0$ for Treutlein[22] and Yan[25] and $\alpha = 1$ for the other datasets. In all cases, the number of selected genes is 2000.

For single-cell qPCR data from mouse early embryo (e.g., [5]), we remove the first two control genes (actb, ahcy) and analyze the remaining 46 with SoptSC.

For scRNA-seq data from human early embryo (ref. [25]), we selected genes that expressed at least 6 cells and at most among the overall 88 cells, which induces 11517 genes to be used in the downstream analysis.

For Joost data set (ref. [6]), we selected 3000 genes for downstream analysis based on our gene filtering technique with parameter α being set as $0.03 \times N$ where N represents the number of cells in the data.

For the Olsson data set (ref. [17]), we selected 2000 genes for downstream analysis based on our gene filtering technique with $\alpha = 0$.

For scRNA-seq data from Nesterowa (ref. [16]), we selected 3000 genes for downstream analysis based on our gene filtering techniques with $\alpha = 0.03 \times N$ where N is the number of cells.

4.2 Details of Data Analysis by SC3

SC3 is run in R under version 3.4.3. For all the datasets used in evaluating the performance clustering in the paper, we used the default setting.

4.3 Details of Data Analysis by Seurat

For all datasets used to evaluate the performance of clustering against different methods (Figure 2), we implement the parameters setting for Seurat as follows. To initialize the Seurat object with the raw (non-normalized data), we keep all genes expressed in at least 3 cells and keep all cells with at least 200 detected genes To select highly variable genes for initial clustering of cells, we performed Principal Component Analysis (PCA) on the scaled data for all genes included in the previous step. We set x.low.cutoff = 0, y.cutoff = 0.8 in FindVariableGenes function. For clustering, we used the function FindClusters using 10 PCs with resolution 0.8. Nonlinear dimensionality reduction method, namely tSNE, was applied to the scaled matrix for visualization of cells in two-dimensional space using first 10 PC components.

For mouse early embryonic data [5], Joost et al. data [6] and Olsson et al. data [18] we used 3 PCs with resolution 0.8.

4.4 Details of Data Analysis by SIMLR

We run SIMLR in MATLAB with default setting for all datasets used in the paper.

4.5 Details of Data Analysis by Monocle2

For two embryonic datasets used in pseudotime comparison, we run Monocle2 for the reduction method set as "DDRTree" with parameters pseudo expr_set as 0 and max_components set as 2. For Joost et al. [6] and Olsson et al. [18] datasets, we selected ordering genes with parameters mean_expression greater than 0.2 and dispersion_empirical larger than 0.5 $*$ dispersion_fit For Shalek et al. [21] data, we selected ordering genes with parameters mean_expression greater than 1 and dispersion_empirical larger than 1.5 $*$ dispersion_fit

4.6 Details of Data Analysis by DPT

We run DPT in MATLAB with default setting for all datasets used in the paper.

References

1. C. Boutsidis and E. Gallopoulos. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008.

2. Q. Deng, D. Ramsköld, B. Reinius, and R. Sandberg. Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196, 2014.

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

4. M. Goolam, A. Scialdone, S. J. Graham, I. C. Macaulay, A. Jedrusik, A. Hupalowska, T. Voet, J. C. Marioni, and M. Zernicka-Goetz. Heterogeneity in oct4 and sox2 targets biases cell fate in 4-cell mouse embryos. *Cell*, 165(1):61–74, 2016.

5. G. Guo, M. Huss, G. Q. Tong, C. Wang, L. L. Sun, N. D. Clarke, and P. Robson. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental cell*, 18(4):675–685, 2010.

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

7. 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*, page 036558, 2016.

8. A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

9. A. A. Kolodziejczyk, J. K. Kim, J. C. Tsang, T. Ilicic, J. Henriksson, K. N. Natarajan, A. C. Tuck, X. Gao, M. Bühler, P. Liu, et al. Single cell rna-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell stem cell*, 17(4):471–485, 2015.

10. D. Kuang, C. Ding, and H. Park. Symmetric nonnegative matrix factorization for graph clustering. In *Proceedings of the 2012 SIAM international conference on data mining*, pages 106–117. SIAM, 2012.

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

12. G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

13. L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

14. C. Meyer, S. Race, and K. Valakuzhy. Determining the number of clusters via iterative consensus clustering. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 94–102. SIAM, 2013.

15. M. Nascimento, F. de Toledo, and A. Carvalho. Consensus clustering using spectral theory. *Advances in Neuro-Information Processing*, pages 461–468, 2009.

16. S. Nestorowa, F. K. Hamey, B. Pijuan Sala, E. Diamanti, M. Shepherd, E. Laurenti, N. K. Wilson, D. G. Kent, and B. Göttgens. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*, 128(8):e20–e31, 2016.

17. A. Olsson, M. Venkatasubramanian, V. K. Chaudhri, B. J. Aronow, N. Salomonis, H. Singh, and H. L. Grimes. Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature*, 537(7622):698–702, Aug. 2016.

18. A. Olsson, M. Venkatasubramanian, V. K. Chaudhri, B. J. Aronow, N. Salomonis, H. Singh, and H. L. Grimes. Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature*, 2016.

19. A. A. Pollen, T. J. Nowakowski, J. Shuga, X. Wang, A. A. Leyrat, J. H. Lui, N. Li, L. Szpankowski, B. Fowler, P. Chen, et al. Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature biotechnology*, 32(10):1053, 2014.

20. S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

21. A. K. Shalek, R. Satija, J. Shuga, J. J. Trombetta, D. Gennert, D. Lu, P. Chen, R. S. Gertner, J. T. Gaublomme, N. Yosef, S. Schwartz, B. Fowler, S. Weaver, J. Wang, X. Wang,

R. Ding, R. Raychowdhury, N. Friedman, N. Hacohen, H. Park, A. P. May, and A. Regev. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, June 2014.

22. B. Treutlein, D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantalas, F. H. Espinoza, T. J. Desai, M. A. Krasnow, and S. R. Quake. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. *Nature*, 509(7500):371, 2014.

23. D. Usoskin, A. Furlan, S. Islam, H. Abdo, P. Lönnerberg, D. Lou, J. Hjerling-Leffler, J. Haeggström, O. Kharchenko, P. V. Kharchenko, et al. Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing. *Nature neuroscience*, 18(1):145, 2015.

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

25. 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(9):1131–1139, 2013.

26. A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.

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