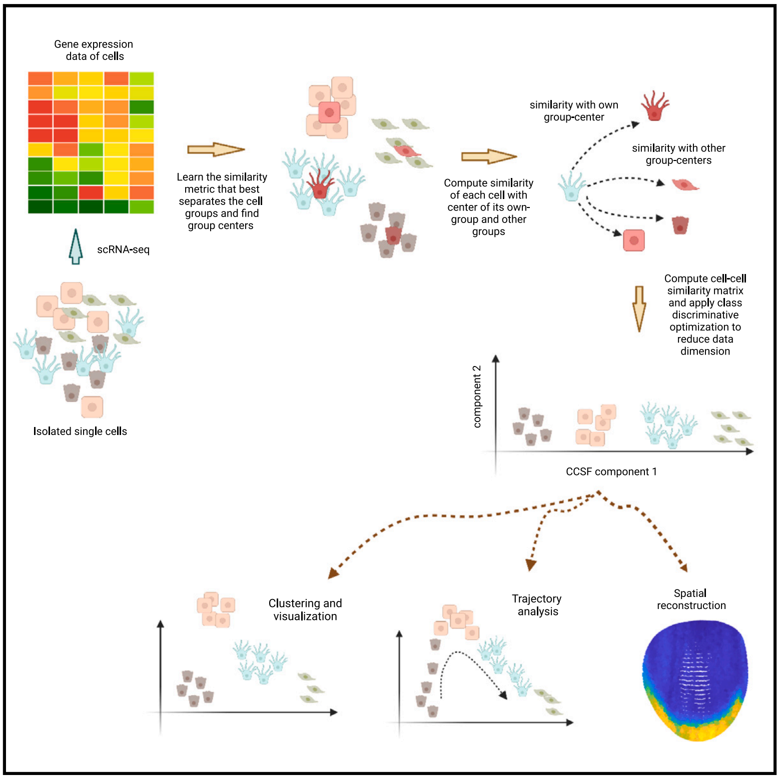
Article

Patterns

Leveraging cell-cell similarity for high-performance spatial and temporal cellular mappings from gene expression data

Graphical abstract


Gene expression data of cells
Learn the similarity metric that best separates the cell groups and find group centers
scRNA-seq
Isolated single cells
Compute similarity of each cell with center of its own-group and other groups
similarity with own group-center
similarity with other group-centers
Compute cell-cell similarity matrix and apply class discriminative optimization to reduce data dimension
component 2
CCSF component 1
Clustering and visualization
Trajectory analysis
Spatial reconstruction

Authors

Md Tauhidul Islam, Lei Xing

Correspondence

lei@stanford.edu

In brief

Scientists use single-cell trajectory mapping and spatial reconstruction to understand how cells form complex tissues and organs. These tools track individual cell paths and map tissue formation. However, their effectiveness depends on data processing tools, which often overlook unique cell characteristics, leading to less accurate results. We have developed a new method that acknowledges cell similarities, drastically improving accuracy and speed. This advancement enhances our exploration of cellular complexity and promises new discoveries in life sciences.

Highlights

- A cell-cell similarity-driven data analysis technique named CCSF
- CCSF allows accurate spatial and temporal cellular maps
- CCSF allows superior clustering and visualization
- CCSF is computationally faster than existing methods

Islam & Xing, 2023, Patterns *4*, 100840
October 13, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.patter.2023.100840

Patterns

CellPress
OPEN ACCESS

Article

Leveraging cell-cell similarity for high-performance spatial and temporal cellular mappings from gene expression data

Md Tauhidul Islam[1] and Lei Xing[1,2,*]

[1]Department of Radiation Oncology, Stanford University, Stanford, CA 94305, USA

[2]Lead contact

*Correspondence: lei@stanford.edu

https://doi.org/10.1016/j.patter.2023.100840

THE BIGGER PICTURE Scientists are keen to understand how cells, the building blocks of life, grow and form complex tissues and organs. They use two methods to do this: single-cell trajectory mapping, which traces the journey of individual cells, and spatial reconstruction, which creates a detailed blueprint of tissue formation. Together, they provide a way to track each cell's path and to understand how they come together to build the grand structure of life. However, the effectiveness of these techniques relies heavily on the analytical tools used to process the vast amounts of data they generate. Existing tools often overlook the biological characteristics of each cell and lead to suboptimal results. We have developed a method that considers the cell features in terms of their similarity with other cells and significantly improves accuracy and speed. This advancement not only enables more efficient and reliable exploration of cellular complexity but also paves the way for new discoveries in life sciences.

1 2 **3** 4 5 **Development/Pre-production:** Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Single-cell trajectory mapping and spatial reconstruction are two important developments in life science and provide a unique means to decode heterogeneous tissue formation, cellular dynamics, and tissue developmental processes. The success of these techniques depends critically on the performance of analytical tools used for high-dimensional (HD) gene expression data processing. Existing methods discern the patterns of the data without explicitly considering the underlying biological characteristics of the system, often leading to suboptimal solutions. Here, we present a cell-cell similarity-driven framework of genomic data analysis for high-fidelity spatial and temporal cellular mappings. The approach exploits the similarity features of the cells to discover discriminative patterns of the data. We show that for a wide variety of datasets, the proposed approach drastically improves the accuracies of spatial and temporal mapping analyses compared with state-of-the-art techniques.

INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) offers an effective means of quantifying the state of individual cells and provides unique opportunities to understand the heterogeneous cell types, cellular dynamics, tissue developmental processes, and the regulatory mechanism controlling the cell functions.[1,2] To comprehend the roles of single cells in multi-cellular functions, decoding the spatial context and temporal evolution of cells from scRNA-seq data is a prerequisite. In reality, separating the cells according to their spatial context is a very challenging task, as single-cell genomics does not encode the spatial locations when the gene expression levels are measured.[3] Cellular trajectory mapping is also challenging because of its high susceptibility to measurement noise.[4]

For trajectory inference, a number of methods, such as monocle,[5] discriminative dimensionality-reduction tree (DDRTree),[6] Slingshot,[7] SCORPIUS,[8] and diffusion pseudo-time (DPT),[9] have been proposed. Commonly used techniques for spatial reconstruction include principal-component analysis (PCA),[10]

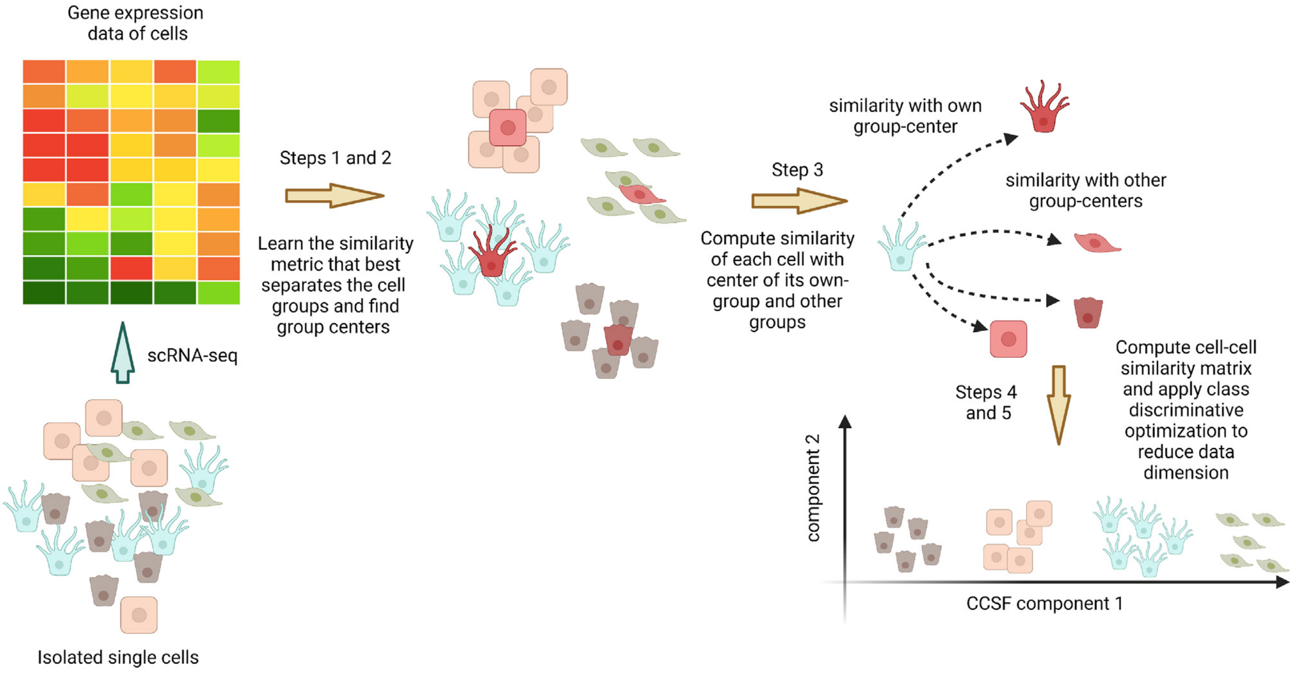

Gene expression data of cells
scRNA-seq
Isolated single cells
Steps 1 and 2
Learn the similarity metric that best separates the cell groups and find group centers
Step 3
Compute similarity of each cell with center of its own-group and other groups
similarity with own group-center
similarity with other group-centers
Steps 4 and 5
Compute cell-cell similarity matrix and apply class discriminative optimization to reduce data dimension
component 2
CCSF component 1

Figure 1. Workflow of the CCSF approach

(1) The cell-cell similarity metric that results in maximum separation of the cell classes in low dimension is learned.
(2) P centroids are located by minimizing the distance between them and all the data points.
(3) The cell-centroid similarity is then computed.
(4) The cell-cell similarity matrix is obtained by computing the gram matrix of the cell-centroid similarity.
(5) A class-discriminative optimization is performed on the similarity matrix to reduce the data dimensionality.

Seurat,[11] spatial backmapping[12] and novoSpaRc.[3] Other than novoSpaRc, all of these methods require the use of marker genes or a reference atlas, which adds another layer of complexity and renders the approaches impractical, as reference datasets may not be available in many cases. The achievable accuracy of markerless novoSpaRc in reconstructing cellular spatial map is rather limited.[3] Generally, the temporal and spatial mappings are done by first projecting the high-dimensional (HD) scRNA-seq data onto a low dimension and then optimizing the temporal or spatial positions of the cells. The dimensionality reduction in these algorithms is typically done by using PCA[13] or other similar methods.[14–20] Unfortunately, the embeddings resulting from these methods are often suboptimal because of their inability to explicitly incorporate the underlying biological characteristics of the data.

There are two major hurdles impeding our ability to carry out high-fidelity temporal and spatial mappings: (1) the intrinsic complexity of the scRNA-seq data structure caused by the intertwined relationships among the cells within the data and (2) the humongous size and dimensionality of the data. To overcome these challenges, here we propose a cell-cell similarity-driven framework (CCSF) for genomic data analysis (Figure 1).[21] Our strategy is to transform the gene expression data into a low-dimensional representation according to the cell-cell similarities of the system.[22–27] The proposed CCSF utilizes a data-driven strategy to compute the similarity between expression profiles of cells. In principle, one can compute the cell-cell similarity matrix directly from the gene expression data. However, this may not provide us any information about the cell groups in the data. Moreover, the calculation of the similarity matrix can be computationally prohibitive because of the curse of data size and dimensionality.[19] CCSF first sorts the M cells into P groups (where P is determined automatically using an unsupervised method) according to the similarity of the cells as measured by the learned similarity metric from the data. For efficient computation, we use the centroid expression of each group to represent the gene expressions of the cells in the group. The similarity values of the cells to these centroids form an $M \times P$ matrix. The Gram matrix[28] of the cell-centroid similarity matrix yields the cell-cell similarity matrix while avoiding potentially excessive computational burdens. The symmetric Gram (i.e., kernel) matrix makes it possible for us to extract the dominant components of the cell-cell similarity and use the information to discover the underlying discriminative patterns of the data via mathematical operations like singular value decomposition (SVD).

CCSF contributes to genomic studies and biomedical data science in multiple aspects. It introduces an effective framework for leveraging the characteristics of cell-cell similarity for temporal and spatial mappings. CCSF enables us to effectively describe a large HD genomic dataset by using only a limited number of CCSF components. Mathematically, one may consider CCSF a way of finding a set of more descriptive principal components of the genomic system that are computed based on cell-cell similarity (in place of variance in PCA). Computationally, the method is about 15 times faster than existing methods such as PCA. Moreover, the CCSF alleviates the

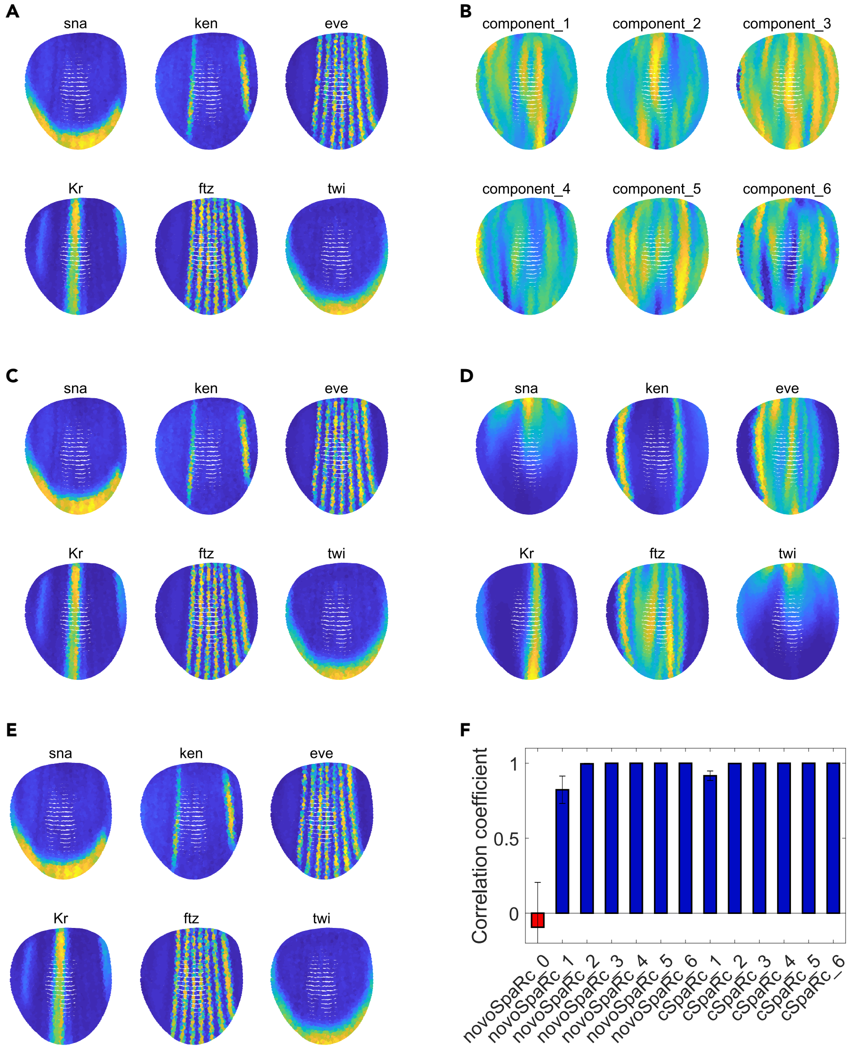


Figure 2. Spatial reconstruction of Drosophila embryos from scRNA-seq data

(A) Ground-truth FISH data for six selected genes.

(B) Projection of the original expression data onto six CCSF components.

(C) Reconstructed spatial maps of the selected genes by CCSF-novoSpaRc using first six CCSF components.

(D) Reconstructed expression data for the selected genes by novoSpaRc without any marker gene.

(E) Reconstructed expression data for the selected genes by novoSpaRc using six marker genes.

(F) Correlation coefficients between the reconstructed spatial maps of the genes by different techniques and the ground-truth FISH data (number of marker genes in the case of novoSpaRc and the component number in the case of CCSF-novoSpaRc [cSpaRc] are shown on the x axis). Error bars represent the standard deviation of the correlation coefficients for different locations and different genes.

bottleneck of the "curse of data dimensionality" often seen in conventional analysis techniques and allows exploration of gene expression data of large size and dimensionality.

RESULTS

In this section, we first demonstrate accurate spatial reconstructions from scRNA-seq data of Drosophila embryo and mouse cerebellum. High-quality trajectory mapping of scRNA-seq data of organoid-derived cells from CCSF components is then presented.

De novo spatial reconstruction of gene expression with CCSF yields accurate results

scRNA-seq datasets of Drosophila embryos, which consist of expression levels of 84 transcription factors quantitatively registered by fluorescence *in situ* hybridization (FISH),[29] were used in this study. We ran CCSF-novoSpaRc with different numbers of CCSF components, along with benchmarking novoSpaRc reconstruction with different marker genes. In our analysis, the first six CCSF components (Figure 2B) were employed to describe the data. It is interesting to observe that the first CCSF component is mainly localized in the middle of the genomic data space, suggesting that this region has the highest differences in the gene expressions. From the variance of these components (Figure S1), we estimated that the first six components explained more than 95% of the spatial variance in the data, indicating the strong ability of CCSF to capture the important information of data with a small number of components. The CCSF-novoSpaRc results matched the ground-truth FISH data very well (Figures 2A and 2C). Significant deviations from the ground-truth spatial maps (obtained from the FISH data) were observed in novoSpaRc reconstruction in the absence of marker genes (Figure 2D). The construction little resembled the ground truth (as also indicated by the poor correlation value of − 0.09). Only when 6 marker genes were included did the spatial maps obtained by novoSpaRc become similar to the ground-truth FISH data (Figure 2E). Remarkably, the CCSF-novoSpaRc approach achieved a perfect correlation with the ground truth when two or more CCSF components were incorporated (Table S1). Even with only a single component, we found that the method could yield a correlation of 0.91, which is much better than that of novoSpaRc with one marker gene (Figure 2F).

The dependence of CCSF-novoSpaRc reconstruction quality on the number of CCSF components was even more pronounced in the second series of experiments with mouse cerebellum data.[3] Here, we use data of mouse cerebellum slices from a recently developed spatial transcriptomics technology named slide-seq.[30] The dataset of sagittal sections contained 46,293 locations, pertaining to one or multiple cells. We first coarse grained the data to 3,900 locations and then performed spatial mapping using different methods as shown in Figure 3.

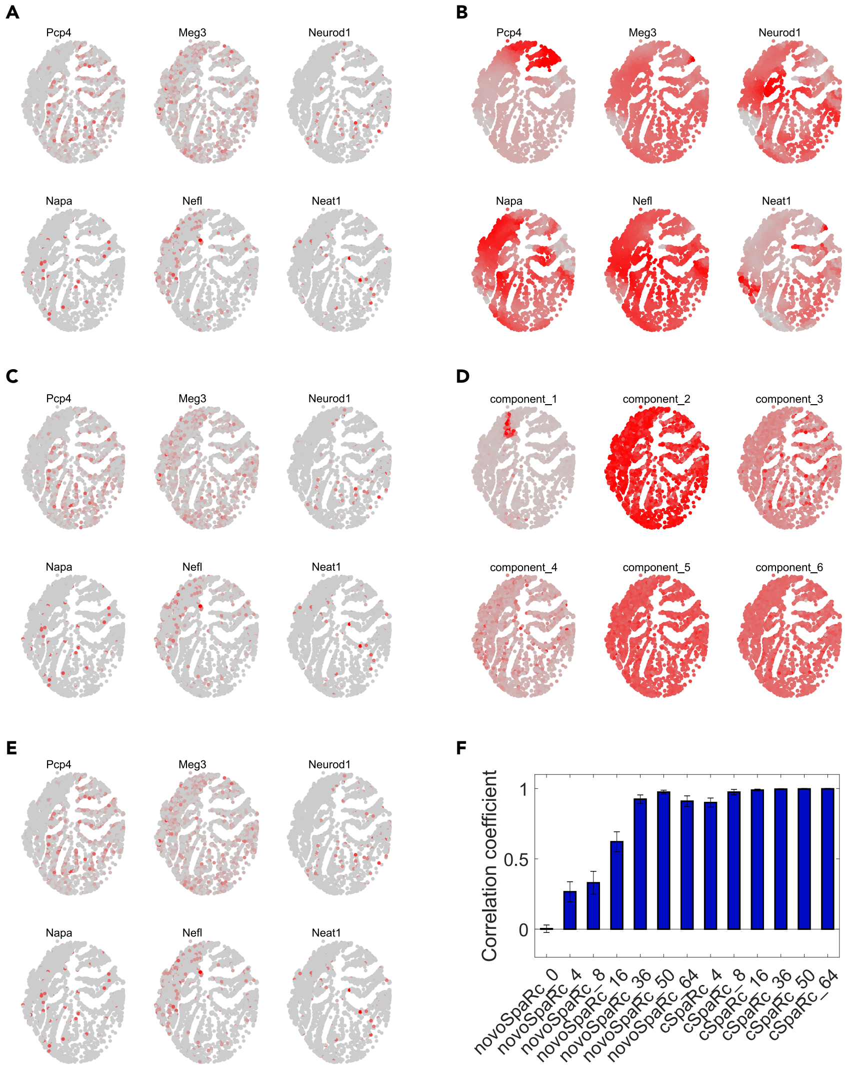


Figure 3. Spatial reconstruction of mouse cerebellum tissue

(A and B) Slide-seq data (A) and novoSpaRc (B) reconstruction without any marker genes.

(C and D) novoSpaRc reconstruction with 64 marker genes (C) and projection of the data onto the first six CCSF components (D).

(E) Reconstructed spatial maps using 64 CCSF components as markers. Overall, novoSpaRc without any marker gene provides spatial reconstruction of almost no similarity to the ground truth, whereas cSpaRc provides perfect reconstruction without any marker gene information (see the correlation values in Table S2).

(F) Correlation coefficients between the reconstructed spatial maps of all the genes by different techniques and ground-truth slide-seq data. Error bars represent the standard deviation of the correlation coefficients for different locations and different genes.

Again, the novoSpaceRc strongly relies on the availability of a large number of markers. The correlation coefficient with the ground truth was found to be nearly zero in the absence of marker genes (Figure 3F). With the use of 4 or more marker genes, the correlation coefficient increased monotonically from 0.26 to the highest level of 0.97 for 50 marker genes (Table S2). Our approach, on the other hand, succeeded without any marker genes and yielded a correlation coefficient of 0.90 with only 4 CCSF components. Remarkably, the correlation coefficients reached 0.99 and 1 for 16 and 32 CCSF components, respectively (Table S2).

In summary, although CCSF-novoSpaRc is a *de novo* reconstruction process, it always provides spatial reconstruction with very high accuracy, which is not achievable by novoSpaRc alone with or without marker genes. Thus, CCSF-novoSpaRc provides a new paradigm of spatial reconstruction where no marker gene is required to achieve accurate reconstruction. CCSF is expected to be a preferred tool to understand the heterogeneous cell types and decode their spatial functionality.

CCSF yields highly accurate cellular trajectories from scRNA-seq data

For cellular trajectory analysis, we used a large dataset obtained by profiling a total of 166,242 organoid-derived cells in two different configurations.[31] In the first configuration, scRNA-seq profiling of 78,379 cells from 9 organoids from two stem cell lines, PGP1 and HUES66, was performed after 3 months of growth. In the second configuration, scRNA-seq profiling of 87,863 cells from 11 organoids was performed. These organoids were grown from PGP1, GM08330, and 11a stem cell lines for 6 months. By comparing the signatures of differentially expressed genes with existing datasets of endogenous cell types, all cells from 20 organoids were classified into sixteen distinct cell types. We produced 2D representations of the scRNA-seq data for organoids 1–3 by using existing methods (Figure 4A) and CCSF (Figure 4B). In the t-distributed stochastic neighbor embedding (t-SNE),[32] uniform manifold approximation and projection (UMAP),[33] and PHATE[34] visualizations, the cells at different stages were separated to a certain extent (Figure 4A), but the stage-to-stage transitions were not clear at all. CCSF-PHATE overcame these limitations and clearly showed the cell transitions with 32 CCSF components (Figure 4B, left). In the CCSF-PHATE visualization, for example, it is seen that the brown cells representing radial glia (RGs) could branch into the following six different paths (shown with black arrows): (1) immature projection neurons (PNs) (dark blue); (2) immature callosal PNs (CPNs) (blue) to mature CPNs (bright blue); (3) immature corticofugal PNs (CFuPNs) (light blue) to mature CFuPNs (cyan blue); (4) induced pluripotent stem cells (iPSCs) (cyan), immature interneurons (cyan green), and ventral precursors (light green); (5) astroglia; and 6) Cajal-Retzius (dark red) and cycling cells (light red). These branches were biologically verified and discussed in detail by

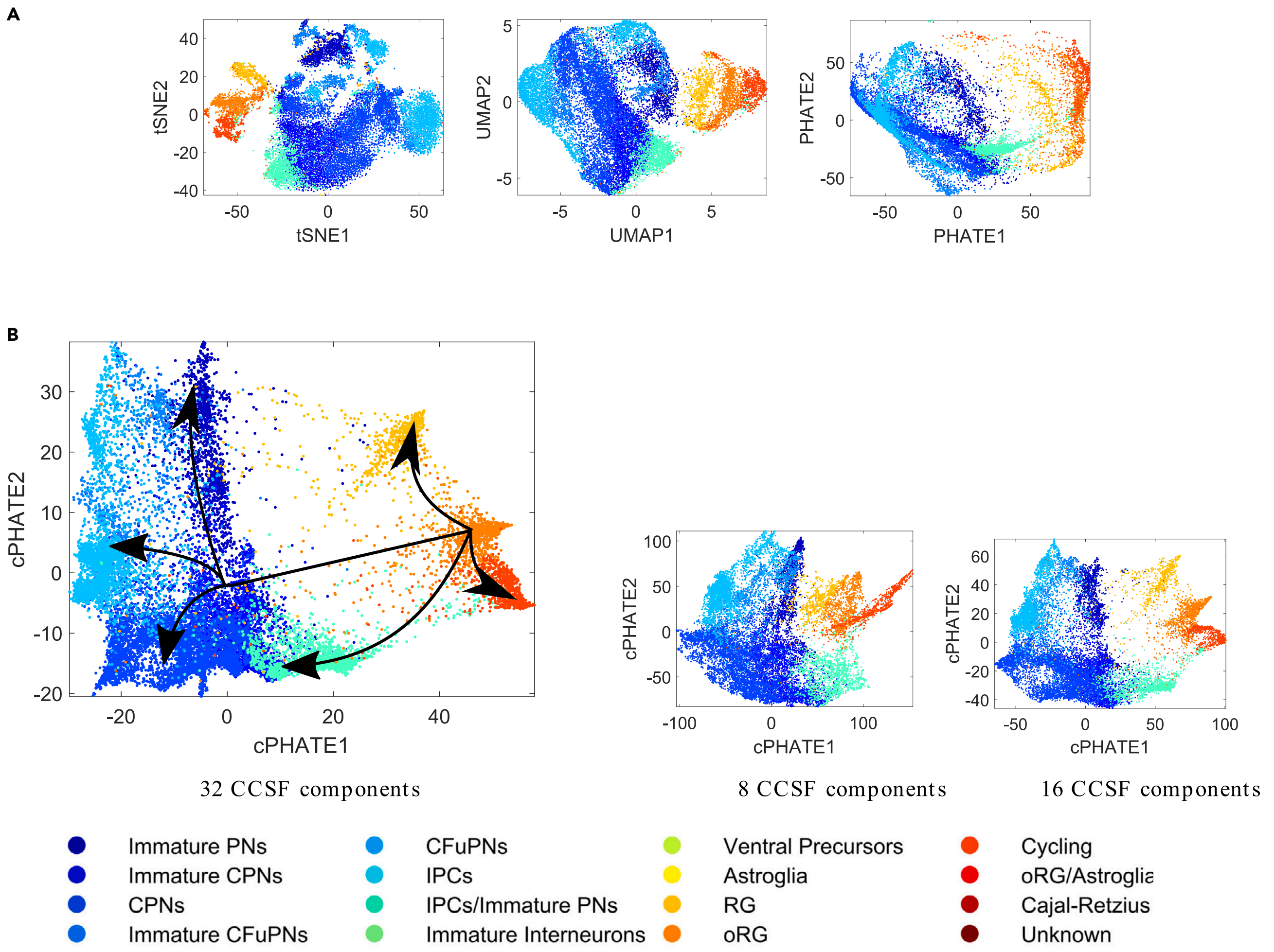


Figure 4. Brain organoids cultured for 3 and 6 months generate the cellular diversity of the human cerebral cortex

The visualization of cells from organoids 1–3 by (A) t-SNE, UMAP, and PHATE (from left to right) and (B) CCSF-PHATE with 32, 8, and 16 CCSF components (from left to right). The black arrows indicate the cells branching into different paths. The components of CCSF-PHATE embedding are denoted by "cPHATE1" and "cPHATE2."

Greig et al.[35] It should be noted that these branches reported in the original study[31] by the pseudo-time and trajectory analysis were not as obvious as those in the CCSF-PHATE visualization (Figure 4B, left).

The CCSF-PHATE study of the data with 8 and 16 CCSF components (Figure 4B, middle, right) revealed two interesting observations. First, the trend of cellular transition can be better visualized in CCSF-PHATE with 8 CCSF components. All seven main types of cells were clustered, and RG cells were located around the middle regions. Other clusters representing the six higher-order cell types were around the RG cell cluster and indicated the probable transitions from RGs into these higher-order cells. Fine details were not apparent in this case. However, because of the smaller distances, the probable stage-to-stage transitions were more clear here than that in the 16 and 32 CCSF component CCSF-PHATE visualizations. Second, with the increase of CCSF components, increasingly fine details were added, and branches became more evident (Figure 4B, left, right). Thus, CCSF-PHATE visualizations with different numbers of components provided a complete representation of cell transitions and branches. Similarly, we studied the transitions and branches of cells from seventeen other organoids and found that CCSF-PHATE yielded superior results over the existing methods (Figures 5A–5F). Each row in Figure 5 shows visualizations from t-SNE, UMAP, PHATE, and CCSF-PHATE with 32 CCSF components (from left to right).

We applied SlingShot[7] to infer the trajectory from CCSF-PHATE and PHATE results and show the results in Figure 6 (1st column). We see that CCSF-PHATE outperforms PHATE by at least 20% in terms of trajectory inference accuracy in most cases. The clustering accuracies and mean inter-class distances for the studies above are shown in Figures 6A–6G. The CCSF-PHATE approach achieved the highest accuracy (2nd column). To demonstrate the performance of CCSF-PHATE in finding the correct trajectories, the geodesic distances among the classes on the trajectories were computed, and the distance matrices are presented in terms of heatmaps. As can be seen from the 4th and 5th columns of Figure 6, the distance increases in both CCSF-PHATE and PHATE visualizations as the number of classes increases, but this tendency is more evident in

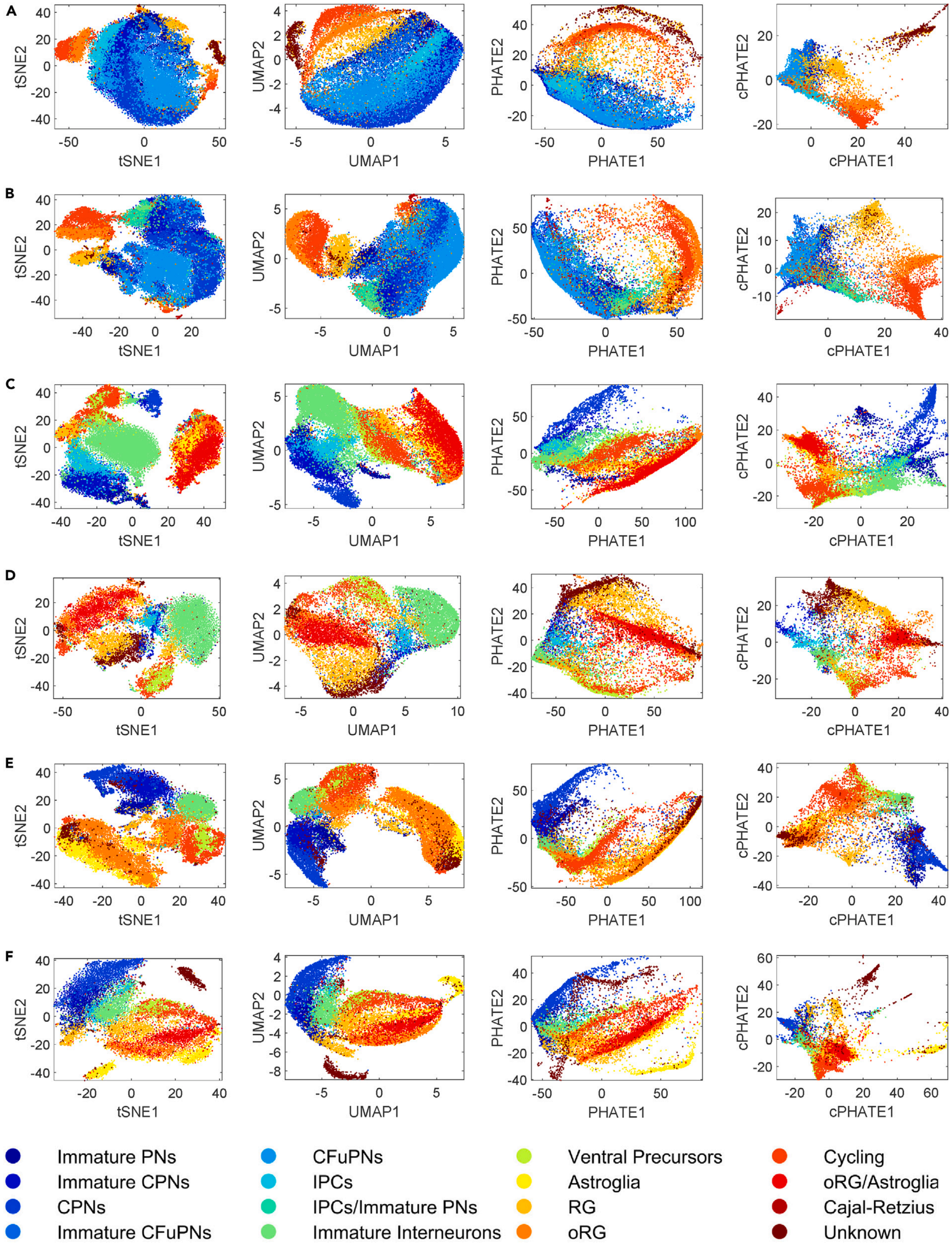


Figure 5. Trajectory mapping of brain organoid scRNA-seq data

The cells from organoids 4–6 are shown in (A), the cells from organoids 7–9 are shown in (B), the cells from organoids 10–12 are shown in (C), the cells from organoids 13–15 are shown in (D), the cells from organoids 16–18 are shown in (E), and the cells from organoids 19 and 20 are shown in (F). The cells are visualized

(legend continued on next page)

CCSF-PHATE (5th column) than in PHATE (4th column). PCA, on the other hand, could not preserve the distance among the classes on the trajectory (3rd column). We also compared the performance of CCSF-PHATE with PCA-PHATE when different number of highly variable genes are used in Figure S12. Again, CCSF-PHATE outperforms PCA-PHATE by a substantial margin. Following Moon et al.,[34] we also have simulated 1,300 simulation datasets using the Splatter simulator and compared the performance of CCSF-PHATE and PHATE in terms of the DeMAP index.[34] Again, as shown in Note S8, CCSF-PHATE outperforms PHATE by >10% in most cases.

In summary, CCSF improves the data quality significantly and finds clusters more accurately than the available methods. Thus, the number and shape of lineages found by CCSF are more accurate than traditional methods, which has been found both visually and quantitatively in our analyses. CCSF would thus be a useful tool to understand the cellular dynamics and tissue developmental processes in many single-cell applications.

Finally, two additional trajectory visualizations of zebrafish embryogenesis and embryoid bodies by CCSF-PHATE are shown in Figure S2 (Note S1) and Figure S3 (Note S2), respectively (see also Note S8 for simulation studies and Figures S10, S16, and S17). It is seen that, compared with PHATE alone, CCSF-PHATE also improved trajectory mapping in these studies.

CCSF can also be used for high-performance clustering analysis of gene expression data. We have added CCSF-based cluster analysis results in Figures S4 and S5 and Tables S3, S8, S9, and Figures S13–S15 for both simulated and experimental datasets (Notes S3, S5, and S10–S12). We first compare the performance of CCSF-UMAP (CCSF components as input to UMAP) with PCA-UMAP for better visualization of the clusters for experimental datasets (Tables S4 and S5 and Notes S10–S12). We then use simulation datasets simulated by the Splatter[36] simulator to quantitatively show that CCSF improves the clustering performance. The simulation parameters were selected by following Moon et al.[34] From Note S8, it is seen that CCSF improves the clustering results by >10% in terms of the adjusted Rand (AR) index[37] when compared to PCA. We also have added the benchmark results of our technique against the Leiden method[38] in terms of two cluster indices in Figure S11. It is seen from Figure S11 that our approach provides an improvement of clustering accuracy of at least 11% compared to this technique.

The computational speed of CCSF was benchmarked for different numbers of data points (Figure S6; Tables S6 and S7; Note S6). We found that the CCSF was computationally much more efficient than existing data analysis techniques. On a personal computer with a Core I9 processor and 64 RAM, for instance, CCSF was found to be 16.7 times faster than PCA (which is known to be the fastest dimensionality-reduction technique to date) for 10,000 cells with 20,000 genes. For the same dataset, CCSF was 52, 29, and 20 times faster than PHATE, t-SNE, and UMAP, respectively. The unprecedented enhancement in computational efficiency is attributed to the fact that CCSF uses a fast clustering operation to reduce the data dimensionality into an intermediate number (P in Figure 1) before projecting to the actual desired number (Q in Figure 1).

DISCUSSION

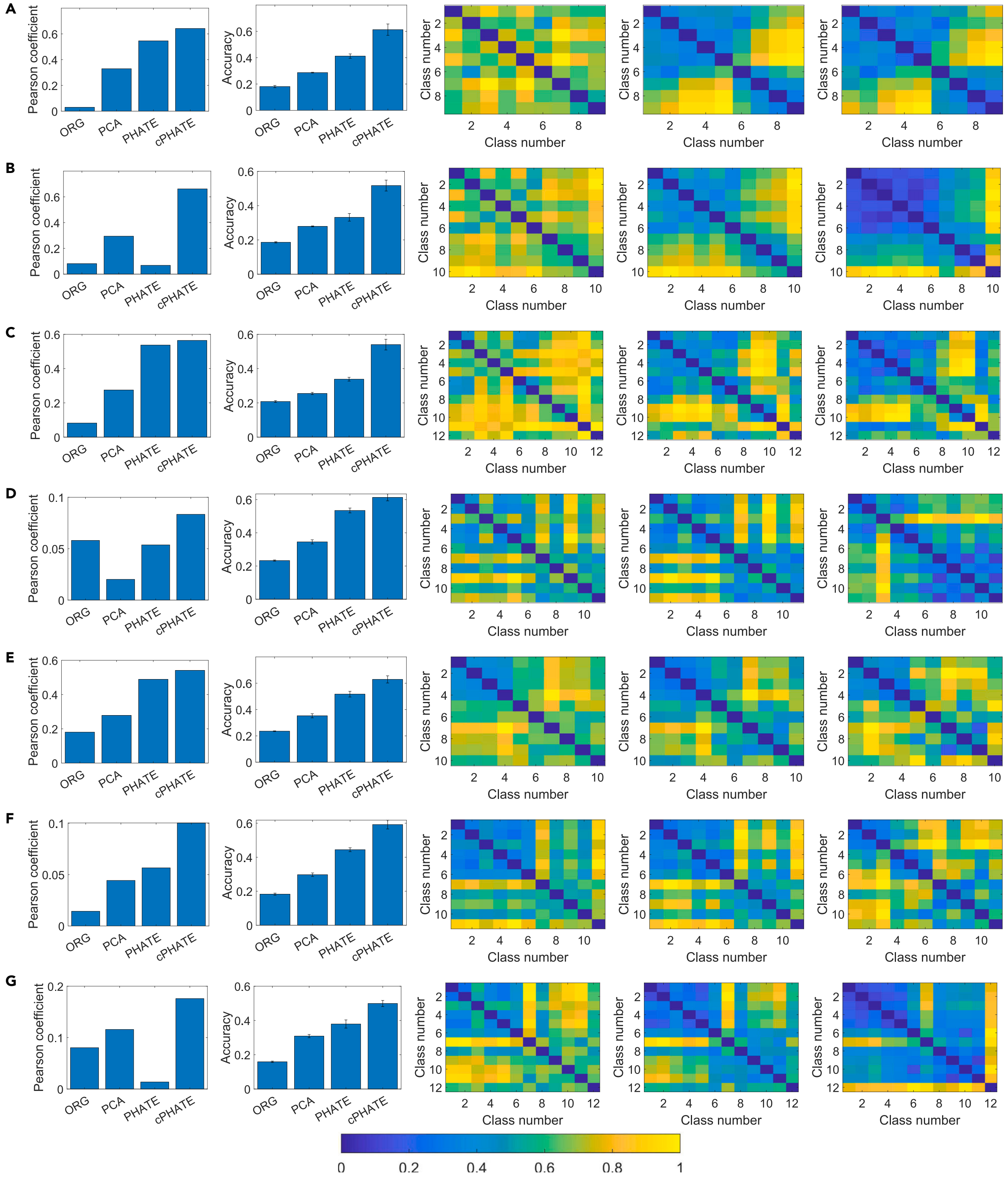
In this work, CCSF is proposed to meet the ever-increasing demand for accurate spatial and temporal mapping of the genomic data. Superior performance of the technique has been demonstrated by using a variety of datasets. The proposed CCSF approach offers several unique advantages over the existing techniques, including (1) high-fidelity *de novo* cellular spatial reconstruction; (2) highly accurate cell trajectory identification; (3) significant enhancement in computational speed; and (4) capability of analyzing large-sized genomic data without being cursed by data size and dimensionality as commonly seen in conventional approaches.[13–20] In reality, accomplishing any of the above features would represent a significant advance in biomedical data science; however, all of them are intrinsic features of the CCSF approach.

The emerging spatial transcriptomics technology promises to provide insights into spatial context of cells and how multicellular functions are orchestrated by individual cells.[39,40] These methods are generally divided into two main types: (1) experimental approaches such as FISH,[41] FISSEQ,[42] seqFISH+,[43] and TIVA[44] and (2) computational approaches to estimate the spatial locations of cells from scRNA-seq data. The former approaches, which sequence cellular RNA *in situ*, require highly specialized experimental tools and do not yet offer widespread applicability or molecular sensitivity.[45] The latter approach is also challenging, as tissues must be dissociated into single cells before scRNA-seq can be performed. During this process, the original spatial context and relationships between cells are lost. Although new computational approaches for spatial reconstruction have been proposed in recent years, they often require an extensive reference database, which may not be available. Thus, computational techniques capable of accurate spatial reconstruction without any reference database (marker genes) are urgently needed. The proposed CCSF promises to fill the gap and lend a valuable tool for genomic data analysis. It is remarkable that perfect spatial reconstructions are attainable by CCSF, which has not even been possible by using marker genes or reference atlases.

In CCSF, the centers of the cell groups are randomly initialized using the k-means++ algorithm.[46] Different initializations can be employed as the cell-to-cell similarity remains unchanged (Figures S7 and S8; Note S7). For quantitative evaluation, we assessed the results of CCSF-PHATE analysis of zebrafish data for 1,000 different CCSF initializations and observed little change (<1%) in the resultant DEMaP (Figure S7). In the case of Drosophila embryo data, we found no change in the spatial reconstruction results for all 1,000 different initializations of CCSF.

In CCSF, a linear class-discriminative optimization is applied on the cell-cell similarity matrix to maximize the separation of

with t-SNE, UMAP, PHATE, and CCSF-PHATE (with 32 CCSF components), and the results for an organoid group are presented in the corresponding row. The components of CCSF-PHATE embedding are denoted by "cPHATE1" and "cPHATE2." In CCSF-PHATE visualization, the transition and branching of RG cells into different higher-order cells are clearly visible. Although t-SNE, UMAP, and PHATE are able to cluster the cells into different types, they completely fail to show the transitions and branches.



(legend on next page)

cell groups, primarily because the cell groups become linearly separable after the computation of similarities (see Figure S9 and Note S9). A nonlinear technique such as quadratic discriminant analysis (QDA)[47] and generalized discriminant analysis (GDA)[48] would not be a good choice for linearly separable data, as proven in the literature.[47,49] For illustration, we have tried to replace the class-discriminative optimization by these techniques in CCSF and reported the performance comparison in Note S4. From the results, it is seen that the CCSF with the linear class-discriminative optimization performs the best. We also have included extensive ablation studies with more than 1,300 simulated datasets and showed the effectiveness of other steps of CCSF (Note S4). CCSF is very robust to dropout effect. We have added simulation results with different levels of dropout in Note S8 for simulation datasets. It is seen that our method is highly robust and provides accurate results for a large range of dropouts in the data.

Most of the traditional and state-of-the-art techniques[7,8,50–56] perform the analysis of developmental and differentiation trajectories by first finding the data clusters. In these analyses, the idea is to first find the clusters, and then a minimum spanning tree is constructed on the clusters to determine the number and rough shape of lineages. Then, simultaneous principal curves or orthogonal projections are used to obtain smooth representations of each lineage. The existing methods perform this in an *ad hoc* step-by-step manner using different classical techniques in different steps. As an example, in Slingshot,[7] at first, the dimension of the data is reduced using UMAP/t-SNE, and a number of cluster centers are computed from the dimension-reduced data. Next, the minimum spanning tree is computed from the cluster centers, and principal graphs are used to infer the trajectory. In CCSF, we developed a theoretical framework for the dimensionality-reduction and clustering steps. In CCSF, (1) we first find the factors that best discriminate between the clusters, (2) the separation between clusters is rigidly encoded through the block diagonal matrix omega, and (3) we maximize the ratio of inter- and intra-cluster distances in low dimensions. In our study, we used PHATE for the trajectory analysis from the CCSF results. However, the minimum spanning tree can be computed from the clusters from CCSF, and then principal graphs can be used to infer the trajectory. We show one such example result in Figure S10. Thus, CCSF uses the same workflow as the existing techniques for continuous data analysis. However, unlike other methods, which are *ad hoc*, CCSF is based on a solid theoretical foundation and provides a more robust and accurate data analysis.

There are a number of data dimensionality-reduction techniques in the literature based on similarity metric learning, including SIMLR[19] and FEM.[57] However, the objective, focus, approach, and algorithm here are quite different. Both SIMLR and FEM are focused on reducing the data dimensionality to a low value (2 or 3) for visualization and clustering purposes. Because of the limited number of low-dimensional components, the information provided by these models is insufficient for reliable trajectory analysis and spatial reconstruction. On the other hand, the focus of CCSF is to create an optimum representation of the data with an appropriate number of components (such as 32, 64, and 96) to meet the specific requirements of a downstream task (such as trajectory analysis and spatial reconstruction) (Figures 2, 3, 4, 5, and 6). Finally, gene expression data represent a very special type of HD data, as their dimensionality and size ($M\times$ dimensionality, where M is the number of cells) may curse any existing dimensionality-reduction algorithm, including FEM and SIMLR methods. CCSF provides an effective means to overcome this bottleneck.

Spatial reconstruction is a difficult optimization problem and depends on the geometry of space and quality of the single-cell data. In our results, we found that CCSF leads to perfect spatial reconstruction in two representative problems. However, this may not always be the case when the space geometry is complex (such as tissue cross-section) and the acquired data contain inaccurate measurements of gene expression count. CCSF does have some limitations; as CCSF is a method of dimensionality reduction guided by the contrastive nature of cell classes, if there is no contrast between the data points (i.e., the dataset contains data points from only one class), CCSF works similarly to kernel PCA. In such cases, computation of the distance from cluster centers in the first step of CCSF works as a data linearization process, and the linear discriminant analysis step works similarly to simple PCA. In these cases, data linearization and computational speed are the only benefits of using CCSF over PCA.

In conclusion, we have presented a computational framework for spatial reconstruction and trajectory mapping of genomic data. The technique offers a number of unique features and greatly enhances our ability to analyze large genomic datasets with high speed and accuracy. We have shown that our *de novo* calculation process can provide spatial maps and cell trajectory with remarkable accuracies. The CCSF components can also be used as input to many other data visualization and analysis techniques such UMAP and t-SNE for efficient data exploration (see Note S3). The CCSF technique thus lays a technical foundation for the analysis of genomic data of increasing scale and complexity. Given its generality and effectiveness, the strategy should also be useful to analyze and solve big data problems from many other disciplines.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Lei Xing (lei@stanford.edu).

Figure 6. The Pearson coefficient between ground truth and estimated trajectory, clustering accuracy, and inter-class distances in the embeddings from PCA, PHATE (with PCA preprocessing), and CCSF-PHATE

Pearson coefficient (1st column), clustering accuracy (2nd column), and mean inter-class geodesic distances (3rd–5th columns) are for the data shown in Figures 4 and 5 for organoids 1–3 (A), organoids 4–6 (B), organoids 7–9 (C), organoids 10–12 (D), organoids 13–15 (E), organoids 16–18 (F), and organoids 19 and 20 (G). Distance heatmaps from PCA, PHATE, and CCSF-PHATE (cPHATE) results are shown in the 3rd–5th columns, respectively. The increment of the geodesic distance across the trajectory is more clear in the cPHATE heatmaps than in the PHATE heatmaps, which is also reflected in the Pearson coefficient values. Error bars represent the standard deviation of the indices for 1,000 different initializations of k-means clustering.

Table 1. Resource availability

Resource	Source
Original codes	https://doi.org/10.24433/CO.4232425.v2[72]
	https://github.com/xinglab-ai/ccsf
The Cancer Genome Atlas (TCGA) dataset	https://portal.gdc.cancer.gov/
BDTNP dataset	http://bdtnp.lbl.gov:8080/Fly-Net/
Cerebellum slide-seq dataset	https://portals.broadinstitute.org/single_cell/study/slide-seq-study
Brain organoid dataset	https://portals.broadinstitute.org/single_cell/study/reproducible-brain-organoids

Materials availability

All data are available in the main text or the supplemental information.

Data and code availability

This article analyzes existing, publicly available data. These sources of the datasets are listed in the resource availability table (Table 1). All original codes have been deposited at Code Ocean and GitHub and are publicly available as of the date of publication. DOIs are listed in Table 1. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Methods details

Computation of cell-centroid similarity

We use an unsupervised approach to learn the similarity metric for CCSF. For this purpose, we sequentially examine the distance of all the data points to two centroids for a number of commonly used similarity metrics (square Euclidean, Chebyshev, correlation, and cityblock) with k-means algorithm.[58] For each chosen metric, this led to a distance matrix. The similarity metric resulting in the best two cluster quality indices (Silhouette, Davies-Bouldin, Calinski-Harabasz)[57] computed from the distance matrix is chosen as the optimal metric for subsequent calculations of CCSF. Here, an implicit assumption is used that the best metric that describes the similarity among the cells would result in the maximum separation of cell classes. The way of optimizing the metric here is a well-established procedure.[19,32–34,59] As examples, in t-SNE and UMAP,[32,33] the cluster separation is optimized in two dimensions to obtain the best data visualization.

Next, we compute the similarity between the cells and the P numbers of centers of cell groups. Let us assume a data matrix X of size $M \times N$ (M cells and N genes), which can be treated as M ($1 \times N$) expression vectors $x_1, x_2, \ldots.., x_M$. The P numbers of group centers in CCSF[60] are computed by optimizing the following function:

$$J_{\mathrm{P}} = \sum_{p=1}^{P} \sum_{i \in C_p} d(\mathbf{x}_i, \mathbf{c}_p), \qquad \text{(Equation 1)}$$

where $\mathbf{c}_p = \sum_{i \in C_p} \mathbf{x}_i / n_p$ is the centroid of group C_p and n_p is the number of cells in C_p. The distance ($d()$) between a cell expression vector and the group centroid in the above equation is the learned similarity metric from the first step. The optimum matrix $D_d \in \mathbb{R}^{M \times P} = d(X, \boldsymbol{c}_{opt})$ is obtained from the first optimization in CCSF, which contains the distance (dissimilarity) of the cells from the centroids. Rows of D_d are scaled from 0 to 1 and then subtracted from 1 to obtain the cell-centroid similarity matrix (D). Here, $\boldsymbol{c}_{opt}$ denotes the collection of P centroids obtained by minimizing Equation 1.

The objective function (Equation 1) is minimized using the following iterative method.

(1) Choose P initial group centroids from the input data by k-means++ algorithm.[61].
(2) Compute cell-to-centroid distances for all the cells to each centroid.
(3) Assign each cell to the group with the closest centroid based on distance.
(4) Compute the average of the expression values of cells in each group to obtain P new centroids.
(5) Repeat steps 2 through 4 until group assignments do not change or until the maximum number of iterations is reached.

We note that the only major calculation for obtaining the cell-centroid similarity in CCSF is the distance computation between the expression values of the cells and centroids. The complexity of the distance calculation is $\mathcal{O}(N)$, i.e., the computational complexity increases linearly with the number of genes. Moreover, the data dimensionality reduces to a small number P after this step. Thus, this step and the remaining CCSF's computations are not cursed by the data dimensionality. On the other hand, for PCA, for example, the computational complexity of eigen decomposition is $\mathcal{O}(N^3)$, which increases nonlinearly with the number of genes.

Computation of cell-cell similarity

The computed cell-centroid similarity matrix D from the last step consists of M data points of P dimension. Let us denote the centered matrix computed from it by $\overline{D} = [\overline{D}^{(1)}, \cdots, \overline{D}^{(P)}]$. Here, the centered matrix is computed by subtracting the mean of each column from the data of that column. The cell-cell similarity matrix can now be obtained by computing the Gram matrix of $\overline{D}$ as $S_t = \overline{D}\overline{D}^T$. Computationally, the Gram matrix computes all possible dot products (similarity) of the gene expression vector in the cell-centroid similarity matrix, which in turn computes the cell-to-cell similarity (denoting similarity) matrix. Here, we used the fact that if two cells are similar to a centroid, they are also similar themselves. Let us now assume that $\overline{d}_k^{(i)} = d_k^{(i)} - \mu$ (μ is the column-wise mean vector of D of size $1 \times P$) denotes the k-th centered cell-centroid similarity data point of the i-th class ($i = 1, \ldots, P$) and that $\overline{D}^{(i)} = [\overline{d}_1^{(i)}, \cdots, \overline{d}_{M_i}^{(i)}]$ denotes the centered data matrix of i-th class. Here, M_i is the number of data points in the i-th class. We can express the between-class cell similarity matrix as

$$\begin{aligned} S_b &= \sum_{i=1}^{P} M_i \left(\mu^{(i)} - \mu\right)\left(\mu^{(i)} - \mu\right)^T \\ &= \sum_{i=1}^{P} M_i \left(\frac{1}{M_i} \sum_{j=1}^{M_i} \left(\mathrm{d}_j^{(i)} - \mu\right)\right)\left(\frac{1}{M_i} \sum_{j=1}^{M_i} \left(d_j^{(i)} - \mu\right)\right)^T \\ &= \sum_{i=1}^{P} \frac{1}{M_i} \left(\sum_{j=1}^{M_i} \overline{d}_j^{(i)} \sum_{j=1}^{M_i} \left(\overline{d}_j^{(i)}\right)^T\right) \\ &= \sum_{i=1}^{P} \overline{D}^{(i)} \Omega^{(i)} \left(\overline{D}^{(i)}\right)^T, \end{aligned}$$

where $\mu^{(i)}$ is the mean vector of i-th class and $\Omega^{(i)}$ is a $M_i \times M_i$ matrix with all the elements equal to $1/M_i$.

Let us denote a $M \times M$ matrix Ω as

$$\Omega = \begin{bmatrix} \Omega^{(1)} & 0 & \cdots & 0 \\ 0 & \Omega^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega^{(P)} \end{bmatrix}.$$

We can write

$$S_b = \sum_{i=1}^{P} \overline{D}^{(i)} \Omega^{(i)} \left(\overline{D}^{(i)}\right)^T = \overline{D} \Omega \overline{D}^T.$$

We can now express the within-class cell-cell similarity matrix as[62]

$$S_w = S_t - S_b = \overline{D}(I - \Omega)\overline{D}^T = \overline{D} L \overline{D}^T,$$

where I is an identity matrix and $L = I - \Omega$ is called the "graph Laplacian."[63]

Class-discriminative optimization

The goal of class-discriminative optimization is to find a projection matrix $W \in \mathbb{R}^{P \times Q}$ such that in the projected datasets, the points with the same cell type remain close to each other and the points of different classes remain distant. To maximize the separation between the cell classes in the data, we optimize the cost function

$$W^* = \underset{W}{\operatorname{argmax}}\frac{\operatorname{tr}\left(W^T S_b W\right)}{\operatorname{tr}\left(W^T S_w W\right)}, \qquad \text{(Equation 2)}$$

which is equivalent to optimizing

$$W^* = \underset{W}{\operatorname{argmax}}\frac{\operatorname{tr}\left(W^T S_b W\right)}{\operatorname{tr}\left(W^T S_t W\right)}. \qquad \text{(Equation 3)}$$

Here, $\operatorname{tr}(F)$ denotes the trace of the matrix F and W^* denotes the optimum projection matrix. The above equation can also be written as

$$W^* = \underset{W}{\operatorname{argmax}}\operatorname{tr}\left(\left(W^T S_t W\right)^{-1}\left(W^T S_b W\right)\right). \qquad \text{(Equation 4)}$$

The optimization problem of Equation 4 can be solved by performing the following generalized eigenvalue decomposition[64,65]:

$$S_b W = \kappa S_t W, \qquad \text{(Equation 5)}$$

where κ is a diagonal matrix containing the eigen values and each column of A contains one eigen vector. For invertible S_t, Equation 5 can be further written as

$$S_t^{-1} S_b W = \kappa W, \qquad \text{(Equation 6)}$$

which is a standard eigen decomposition of $S_t^{-1} S_b$ but is computationally prohibitive for large datasets.[65] We solve this problem efficiently using two SVDs as detailed below.

The SVD of $\overline{D}$ can be written as

$$\overline{D} = U \Sigma V^T, \qquad \text{(Equation 7)}$$

where $\Sigma = \operatorname{diag}(\sigma_1, \cdots, \sigma_r)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ are the singular values of $\overline{D}$; $U = [u_1, \cdots, u_r] \in \mathbb{R}^{N \times r}$ and u_is are the left singular vectors; and $V = [v_1, \cdots, v_r] \in \mathbb{R}^{M \times r}$ and v_is are the right singular vectors. Here, $rank(\overline{D}) = r$.

Replacing $\overline{D}$ in Equation 3, we obtain

$$W^* = \underset{W}{\operatorname{argmax}}\frac{\operatorname{tr}\left(W^T U \Sigma V^T \Omega V \Sigma U^T W\right)}{\operatorname{tr}\left(W^T U \Sigma V^T V \Sigma U^T W\right)}. \qquad \text{(Equation 8)}$$

If we now modify the variable such that $B = \Sigma U^T W$, we get[66]

$$B^* = \underset{B}{\operatorname{argmax}}\frac{\operatorname{tr}\left(B^T V^T W V B\right)}{\operatorname{tr}\left(B^T B\right)}, \qquad \text{(Equation 9)}$$

and B^* is the matrix containing the eigenvectors of $V^T W V$. From the optimum matrix B^*, we can compute W^* by solving a set of linear equations $\Sigma U^T A = B^*$.[66] We note that for U and B^*, we can obtain infinitely many solutions of A that satisfy the systems of equation. However,

$$W^* = U \Sigma^{-1} B^* \qquad \text{(Equation 10)}$$

is obviously one of the solutions and can be considered as the optimum vector that maximizes the class separation of the data. The proof that obtained components are indeed orthogonal to each other is given in Cai et al.[66] We note that the obtained components can be regarded as uncorrelated discriminant components, which are different from the classical linear discriminant components.[47]

From the above analysis, we conclude that the projection matrix W in Equation 6 can be computed efficiently through the following steps: (1) computing the SVD of $\overline{D}$ to get U, V, and Σ, (2) computing B, the eigenvectors of $V^T \Omega V$, and (3) computing $W = U \Sigma^{-1} B$. In the end, $Y = DW, Y \in \mathbb{R}^{M \times Q}$ contains the Q CCSF components.

Implementation and parameter settings

Both Python and MATLAB 2019a (MathWorks, Natick, MA, USA) implementations of CCSF were performed. UMAP and PHATE implementations by the original authors have been used to produce the respective results. If the number of data points in the analyzed dataset is more than 2,000, then 2,000 randomly selected data points were used for learning the similarity metric in the first step of CCSF. P is the number of data groups in the k-means++-based clustering (first optimization) step of CCSF, which is automatically determined by the Leiden algorithm with a resolution of 1[38] for spatial reconstruction and clustering applications. For trajectory analysis, P was kept fixed at 33. Q was always set to $P - 1$.

Spatial reconstruction

novoSpaRc with default parameters was used to reconstruct all the spatial maps.[3] In the case of marker-based reconstruction, marker genes were chosen randomly following the original work.[3] For CCSF-novoSpaRc, the CCSF components were used as the marker genes.

Computation of clustering accuracy, cluster quality indices, and DEMaP

For calculation of clustering accuracy and cluster quality indices, we at first cluster the data into N_g classes (N_g is the number of classes in ground-truth label) by k-means clustering technique. We then find the best map of cluster labels in comparison to the ground-truth labels. These cluster labels are then used to compute the indices. Accuracy is the number of correctly found class labels divided by the total number of class labels. The silhouette value[67] is a measure of how similar a data point is to its own cluster compared with the other clusters. The silhouette ranges from -1 to +1, where a high value indicates that the data are well clustered. Calinski-Harabasz and Davies-Bouldin indices were computed using the formulations proposed in Caliński and Harabasz[68] and Davies and Bouldin.[69] The AR index is computed following Hubert and Arabie.[37] DEMaP is an index recently proposed in Moon et al.[34] for evaluating the low-dimensional representation from a dimensionality-reduction technique. To compute DEMaP, at first, the geodesic distance[70] among the data points is computed from the HD data, and the Euclidean distance among the data points is computed from the low-dimensional representation. DEMaP is defined as the Spearman correlation coefficient between the geodesic distances and Euclidean distances.

Computation of Pearson correlation coefficient

Let us assume that R_f and R_e are the gene expressions from FISH/slide-seq and are estimated using novoSpaRc/CCSF-novoSpaRc. At first, R_f and R_e were rescaled from 0 to 1. Let us assume that A is the FISH/slide-seq gene expression vector at N different locations from R_f, whereas B is the estimated gene expression vector at the same locations from R_e. Then, the Pearson correlation coefficient between A and B is defined as

$$\rho(A, B) = \frac{1}{N - 1}\sum_{i=1}^{N}\left(\frac{A_i - \mu_A}{\sigma_A}\right)\left(\frac{B_i - \mu_B}{\sigma_B}\right), \qquad \text{(Equation 11)}$$

where μ_A and σ_A are the mean and standard deviation of A, respectively, and μ_B and σ_B are the mean and standard deviation of B.

Ablation study

An ablation study was performed on all the datasets simulated using Splatter[36] for comparing the performance of different configurations of CCSF (see Note S4). In CCSF, the k-means++-based clustering technique was replaced by Fuzzy c-means and partition around medoids (PAM)[71] by keeping class-discriminative optimization. On other condition, k-means++-based clustering was kept fixed, and the class-discriminative optimization was replaced with GDA with Gaussian and polynomial kernels, QDA, MDS, and PCA.

Description of datasets

scRNA-seq data were acquired from 17,774 cells from brain organoids 1–3; 27,646 cells from brain organoids 4–6; 32,959 cells from organoids 7–9; 25,618 cells from organoids 10–12, 15,256 cells from organoids 13–15; 21,213 cells from organoids 16–18; and 14,754 cells from organoids 19 and 20. The numbers of cells in the analyzed BDTNP and cerebellum datasets were 3,039 and 7,704, respectively. The number of genes in the cerebellum dataset was 19,782.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2023.100840.

ACKNOWLEDGMENTS

This work was partially supported by NIH (1R01 CA223667 and R01CA227713) and a Faculty Research Award from Google, Inc.

AUTHOR CONTRIBUTIONS

L.X. conceived the experiment(s) and M.T.I. conducted the experiment(s) and analyzed the results. Both authors reviewed the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 14, 2022
Revised: June 29, 2023
Accepted: August 10, 2023
Published: September 7, 2023

REFERENCES

1. Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat. Rev. Genet. *14*, 618–630. https://doi.org/10.1038/nrg3542.
2. Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015). The Technology and Biology of Single-Cell RNA Sequencing. Mol. Cell *58*, 610–620. https://doi.org/10.1016/j.molcel.2015.04.005.
3. Nitzan, M., Karaiskos, N., Friedman, N., and Rajewsky, N. (2019). Gene expression cartography. Nature *576*, 132–137. https://doi.org/10.1038/s41586-019-1773-3.
4. Liu, Z., Lou, H., Xie, K., Wang, H., Chen, N., Aparicio, O.M., Zhang, M.Q., Jiang, R., and Chen, T. (2017). Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. Nat. Commun. *8*, 22. https://doi.org/10.1038/s41467-017-00039-z.
5. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381–386. https://doi.org/10.1038/nbt.2859.
6. Mao, Q., Wang, L., Goodison, S., and Sun, Y. (2015). Dimensionality Reduction Via Graph Structure Learning. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Association for Computing Machinery). KDD '15, 765–774. https://doi.org/10.1145/2783258.2783309.
7. Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genom. *19*, 477. https://doi.org/10.1186/s12864-018-4772-0.
8. Cannoodt, R., Saelens, W., Sichien, D., Tavernier, S., Janssens, S., Guilliams, M., Lambrecht, B., Preter, K.D., and Saeys, Y. (2016). SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development. Preprint at bioRxiv. https://doi.org/10.1101/079509.
9. Haghverdi, L., Büttner, M., Wolf, F.A., Buettner, F., and Theis, F.J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. Nat. Methods *13*, 845–848. https://doi.org/10.1038/nmeth.3971.
10. Durruthy-Durruthy, R., Gottlieb, A., Hartman, B.H., Waldhaus, J., Laske, R.D., Altman, R., and Heller, S. (2014). Reconstruction of the Mouse Otocyst and Early Neuroblast Lineage at Single-Cell Resolution. Cell *157*, 964–978. https://doi.org/10.1016/j.cell.2014.03.036.
11. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502. https://doi.org/10.1038/nbt.3192.
12. Achim, K., Pettit, J.B., Saraiva, L.R., Gavriouchkina, D., Larsson, T., Arendt, D., and Marioni, J.C. (2015). High-throughput spatial mapping of single-cell RNA-Seq data to tissue of origin. Nat. Biotechnol. *33*, 503–509. https://doi.org/10.1038/nbt.3209.
13. Jolliffe, I.T. (2002). Principal Component Analysis. Springer Series in Statistics, 2 edn (Springer-Verlag).
14. Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature *401*, 788–791. https://doi.org/10.1038/44565.
15. Lawley, D.N., and Maxwell, A.E. (1962). Factor Analysis as a Statistical Method. Journal of the Royal Statistical Society. Series D (The Statistician) *12*, 209–229. https://doi.org/10.2307/2986915. 2986915.
16. Hyvärinen, A., and Oja, E. (2000). Independent component analysis: Algorithms and applications. Neural Network. *13*, 411–430. https://doi.org/10.1016/S0893-6080(00)00026-5.
17. Schölkopf, B., Smola, A., and Müller, K.R. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Comput. *10*, 1299–1319. https://doi.org/10.1162/089976698300017467.
18. Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality Reduction by Learning an Invariant Mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), *2*, pp. 1735–1742. https://doi.org/10.1109/CVPR.2006.100.
19. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat. Methods *14*, 414–416. https://doi.org/10.1038/nmeth.4207.
20. Sohn, K., Lee, H., and Yan, X. (2015). Learning Structured Output Representation using Deep Conditional Generative Models. Advances in Neural Information Processing Systems, *28* (Curran Associates, Inc.).
21. Biorender Created with BioRender.Com. (Science Suite Inc., 2023).
22. Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., 2nd, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189–196. https://doi.org/10.1126/science.aad0501.
23. Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. Cell *144*, 646–674. https://doi.org/10.1016/j.cell.2011.02.013.
24. Vento-Tormo, R., Efremova, M., Botting, R.A., Turco, M.Y., Vento-Tormo, M., Meyer, K.B., Park, J.E., Stephenson, E., Polański, K., Goncalves, A., et al. (2018). Single-cell reconstruction of the early maternal–fetal interface in humans. Nature *563*, 347–353. https://doi.org/10.1038/s41586-018-0698-6.
25. Efremova, M., Vento-Tormo, M., Teichmann, S.A., and Vento-Tormo, R. (2020). Inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nat. Protoc. *15*, 1484–1506. https://doi.org/10.1038/s41596-020-0292-x.
26. Jin, S., Guerrero-Juarez, C.F., Zhang, L., Chang, I., Ramos, R., Kuan, C.H., Myung, P., Plikus, M.V., and Nie, Q. (2021). Inference and analysis of cell-cell communication using CellChat. Nat. Commun. *12*, 1088. https://doi.org/10.1038/s41467-021-21246-9.
27. Noël, F., Massenet-Regad, L., Carmi-Levy, I., Cappuccio, A., Grandclaudon, M., Trichot, C., Kieffer, Y., Mechta-Grigoriou, F., and Soumelis, V. (2021). Dissection of intercellular communication using the transcriptome-based framework ICELLNET. Nat. Commun. *12*, 1089. https://doi.org/10.1038/s41467-021-21244-x.
28. Horn, R.A., and Johnson, C.R. (2012). Matrix Analysis, 2nd edition edn (Cambridge University Press).
29. Berkeley Drosophila Transcription Network Project. (2020).
30. Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F., and Macosko, E.Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science *363*, 1463–1467. https://doi.org/10.1126/science.aaw1219.
31. Velasco, S., Kedaigle, A.J., Simmons, S.K., Nash, A., Rocha, M., Quadrato, G., Paulsen, B., Nguyen, L., Adiconis, X., Regev, A., et al. (2019). Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. Nature *570*, 523–527. https://doi.org/10.1038/s41586-019-1289-x.

32. van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research***9**, 2579–2605.

33. Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. *37*, 38–44. https://doi.org/10.1038/nbt.4314.

34. Moon, K.R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D.B., Chen, W.S., Yim, K., Elzen, A.v.d., Hirn, M.J., Coifman, R.R., et al. (2019). Visualizing structure and transitions in high-dimensional biological data. Nat. Biotechnol. *37*, 1482–1492. https://doi.org/10.1038/s41587-019-0336-3.

35. Greig, L.C., Woodworth, M.B., Galazo, M.J., Padmanabhan, H., and Macklis, J.D. (2013). Molecular logic of neocortical projection neuron specification, development and diversity. Nat. Rev. Neurosci. *14*, 755–769. https://doi.org/10.1038/nrn3586.

36. Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: Simulation of single-cell RNA sequencing data. Genome Biol. *18*, 174. https://doi.org/10.1186/s13059-017-1305-0.

37. Hubert, L., and Arabie, P. (1985). Comparing partitions. J. Classif. *2*, 193–218. https://doi.org/10.1007/BF01908075.

38. Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. Sci. Rep. *9*, 5233. https://doi.org/10.1038/s41598-019-41695-z.

39. Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science *353*, 78–82. https://doi.org/10.1126/science.aaf2403.

40. Marx, V. (2021). Method of the Year: Spatially resolved transcriptomics. Nat. Methods *18*, 9–14. https://doi.org/10.1038/s41592-020-01033-y.

41. Langer-Safer, P.R., Levine, M., and Ward, D.C. (1982). Immunological method for mapping genes on Drosophila polytene chromosomes. Proc. Natl. Acad. Sci. USA *79*, 4381–4385. https://doi.org/10.1073/pnas.79.14.4381.

42. Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Yang, J.L., Ferrante, T.C., Terry, R., Jeanty, S.S.F., Li, C., Amamoto, R., et al. (2014). Highly Multiplexed Subcellular RNA Sequencing in Situ. Science *343*, 1360–1363. https://doi.org/10.1126/science.1250212.

43. Eng, C.-H.L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.C., and Cai, L. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature *568*, 235–239. https://doi.org/10.1038/s41586-019-1049-y.

44. Lovatt, D., Ruble, B.K., Lee, J., Dueck, H., Kim, T.K., Fisher, S., Francis, C., Spaethling, J.M., Wolf, J.A., Grady, M.S., et al. (2014). Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. Nat. Methods *11*, 190–196. https://doi.org/10.1038/nmeth.2804.

45. Burgess, D.J. (2019). Spatial transcriptomics coming of age. Nat. Rev. Genet. *20*, 317. https://doi.org/10.1038/s41576-019-0129-z.

46. Arthur, D., and Vassilvitskii, S.K. (2007). Means++: The Advantages of Careful Seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '07, 1027–1035 (Society for Industrial and Applied Mathematics).

47. Hastie, T., Tibshirani, R., and Friedman, J. (2009). Linear Methods for Classification. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, 101–137, T. Hastie, R. Tibshirani, and J. Friedman, eds. (Springer). https://doi.org/10.1007/978-0-387-84858-74.

48. Mika, S., Rätsch, G., Weston, J., Schölkopf, B., and Müller, K.-R. (1999). Fisher Discriminant Analysis with Kernels.

49. James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). Classification. In *An Introduction to Statistical Learning: With Applications in R*, Springer Texts in Statistics, 127–173, G. James, D. Witten, T. Hastie, and R. Tibshirani, eds. (Springer). https://doi.org/10.1007/978-1-4614-7138-74.

50. Ji, Z., and Ji, H. (2016). Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res. *44*, e117. https://doi.org/10.1093/nar/gkw430.

51. Shin, J., Berg, D.A., Zhu, Y., Shin, J.Y., Song, J., Bonaguidi, M.A., Enikolopov, G., Nauen, D.W., Christian, K.M., Ming, G.l., and Song, H. (2015). Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. Cell Stem Cell *17*, 360–372. https://doi.org/10.1016/j.stem.2015.07.013.

52. Islam, M.T., and Xing, L. (2023). Cartography of Genomic Interactions Enables Deep Analysis of Single-Cell Expression Data. Nat. Commun. *14*, 679. https://doi.org/10.1038/s41467-023-36383-6.

53. Smolander, J., Junttila, S., Venäläinen, M.S., and Elo, L.L. (2022). An ensemble method for fast and accurate linear trajectory inference from single-cell RNA-seq data. Bioinformatics *38*, 1328–1335. https://doi.org/10.1093/bioinformatics/btab831.

54. Campbell, K., Ponting, C.P., and Webber, C. (2015). Laplacian eigenmaps and principal curves for high resolution pseudotemporal ordering of single-cell RNA-seq profiles. Preprint at bioRxiv. https://doi.org/10.1101/027219.

55. Zhang, Y., Tran, D., Nguyen, T., Dascalu, S.M., and Harris, F.C. (2023). A robust and accurate single-cell data trajectory inference method using ensemble pseudotime. BMC Bioinf. *24*, 55. https://doi.org/10.1186/s12859-023-05179-2.

56. Islam, M.T., Wang, J.Y., Ren, H., Li, X., Khuzani, M.B., Sang, S., Yu, L., Shen, L., Zhao, W., and Xing, L. (2022). Leveraging data-driven self-consistency for high-fidelity gene expression recovery. Nat. Commun. *13*, 7142. https://doi.org/10.1038/s41467-022-34595-w.

57. Islam, M.T., and Xing, L. (2021). A data-driven dimensionality-reduction algorithm for the exploration of patterns in biomedical data. Nat. Biomed. Eng. *5*, 624–635. https://doi.org/10.1038/s41551-020-00635-3.

58. Pelleg, D., and Moore, A.W. (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *Proceedings Of the Seventeenth International Conference On Machine Learning*, ICML '00, *727–734* (Morgan Kaufmann Publishers Inc.).

59. Kruskal, J.B., and Wish, M. (1978). Multidimensional Scaling (SAGE).

60. Lloyd, S.M., Jr., and Johnson, D.G. (1982). Least squares quantization in PCM. IEEE Trans. Inf. Theor. *74*, 129–141. https://doi.org/10.1109/TIT.1982.1056489.

61. Arthur, D., and Vassilvitskii, S.K. (2006). Means++: The Advantages of Careful Seeding.

62. Ghojogh, B., Karray, F., and Crowley, M. (2019). Fisher and Kernel Fisher Discriminant Analysis: Tutorial. Preprint at arXiv. https://doi.org/10.48550/arXiv.1906.09436.

63. Chung, F.R.K. (1996). Spectral Graph Theory, uk ed. edition edn (American Mathematical Society).

64. Wang, F., Wang, Q., Nie, F., Li, Z., Yu, W., and Wang, R. (2019). Unsupervised Linear Discriminant Analysis for Jointly Clustering and Subspace Learning. IEEE Trans. Knowl. Data Eng. 1. https://doi.org/10.1109/TKDE.2019.2939524.

65. Ghojogh, B., Karray, F., and Crowley, M. (2019). Eigenvalue and Generalized Eigenvalue Problems: Tutorial. Preprint at arXiv. https://doi.org/10.48550/arXiv.1903.11240.

66. Cai, D., He, X., and Han, J. (2005). An Efficient Algorithm for Large-Scale Discriminant Analysis. IEEE Trans. Knowl. Data Eng. *17*, 1624–1637. https://doi.org/10.1109/TKDE.2007.190669.

67. Rousseeuw, P.J. (1987). A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

68. Caliński, T., and Harabasz, J. (1974). A dendrite method for cluster analysis. Commun. Stat. *3*, 1–27. https://doi.org/10.1080/03610927408827101.

69. Davies, D.L., and Bouldin, D.W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-*1***, 224–227. https://doi.org/10.1109/TPAMI.1979.4766909.

70. Tenenbaum, J.B., de Silva, V., and Langford, J.C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science *290*, 2319–2323. https://doi.org/10.1126/science.290.5500.2319.

71. Kaufman, L., and Rousseeuw, P.J. (2009). Finding Groups in Data: An Introduction to Cluster Analysis (John Wiley & Sons).

72. Islam, M.T., and Xing, L. (2023). Leveraging cell-cell similarity for high-performance spatial and temporal cellular mappings from gene expression data. Code Ocean. https://doi.org/10.24433/CO.4232425.v2.