

SUPPLEMENTARY INFORMATION:**Single-cell topological RNA-Seq analysis reveals
insights into cellular differentiation and development**

Abbas H. Rizvi^{1†}, Pablo G. Camara^{2,3†}, Elena K. Kandror¹, Thomas J. Roberts^{1,3}, Ira Schieren⁴,
Tom Maniatis^{1*}, and Raul Rabadan^{2,3*}

¹Department of Biochemistry and Molecular Biophysics,

²Department of Systems Biology,

³Department of Biomedical Informatics,

⁴Howard Hughes Medical Institute,

Columbia University Medical Center,

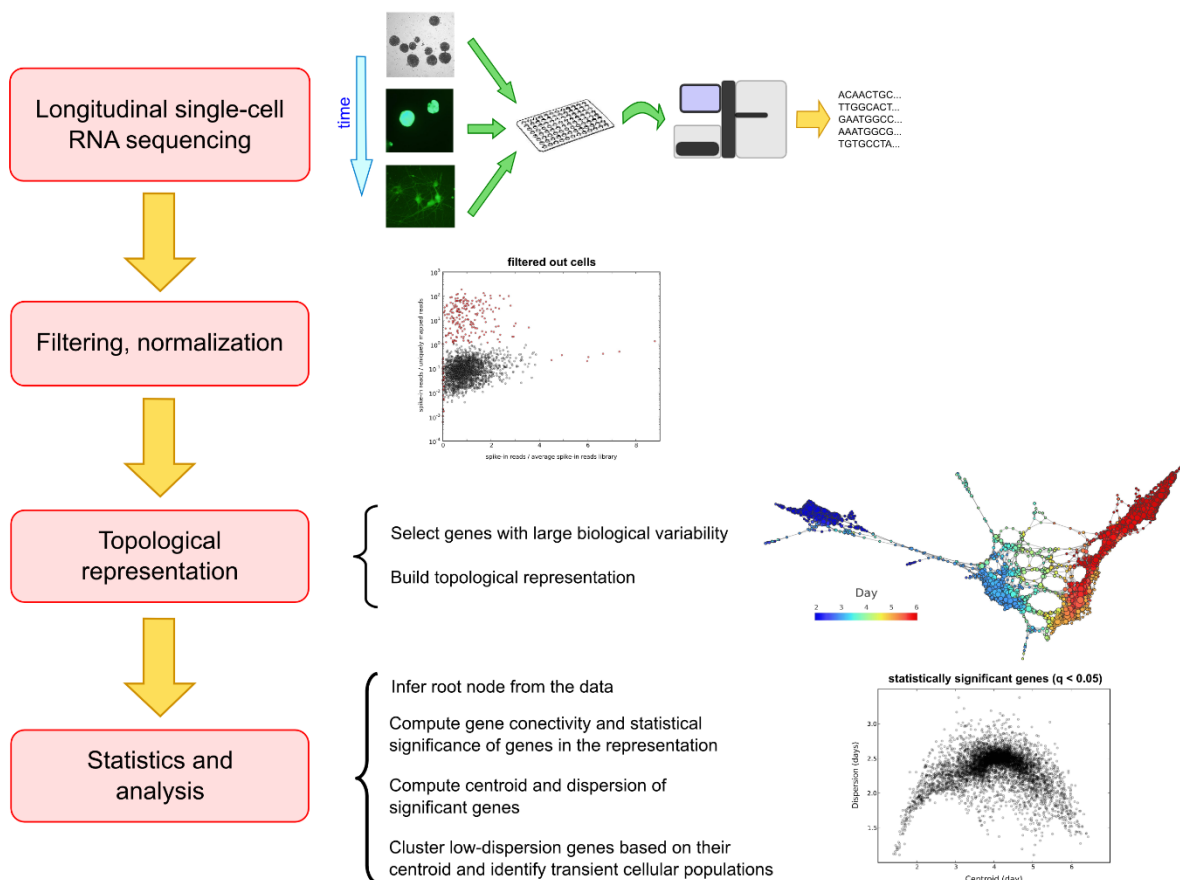
New York, NY 10032.

*Correspondence to: rr2579@cumc.columbia.edu, tm2472@columbia.edu

†These authors contributed equally to this work

Supplementary Figures

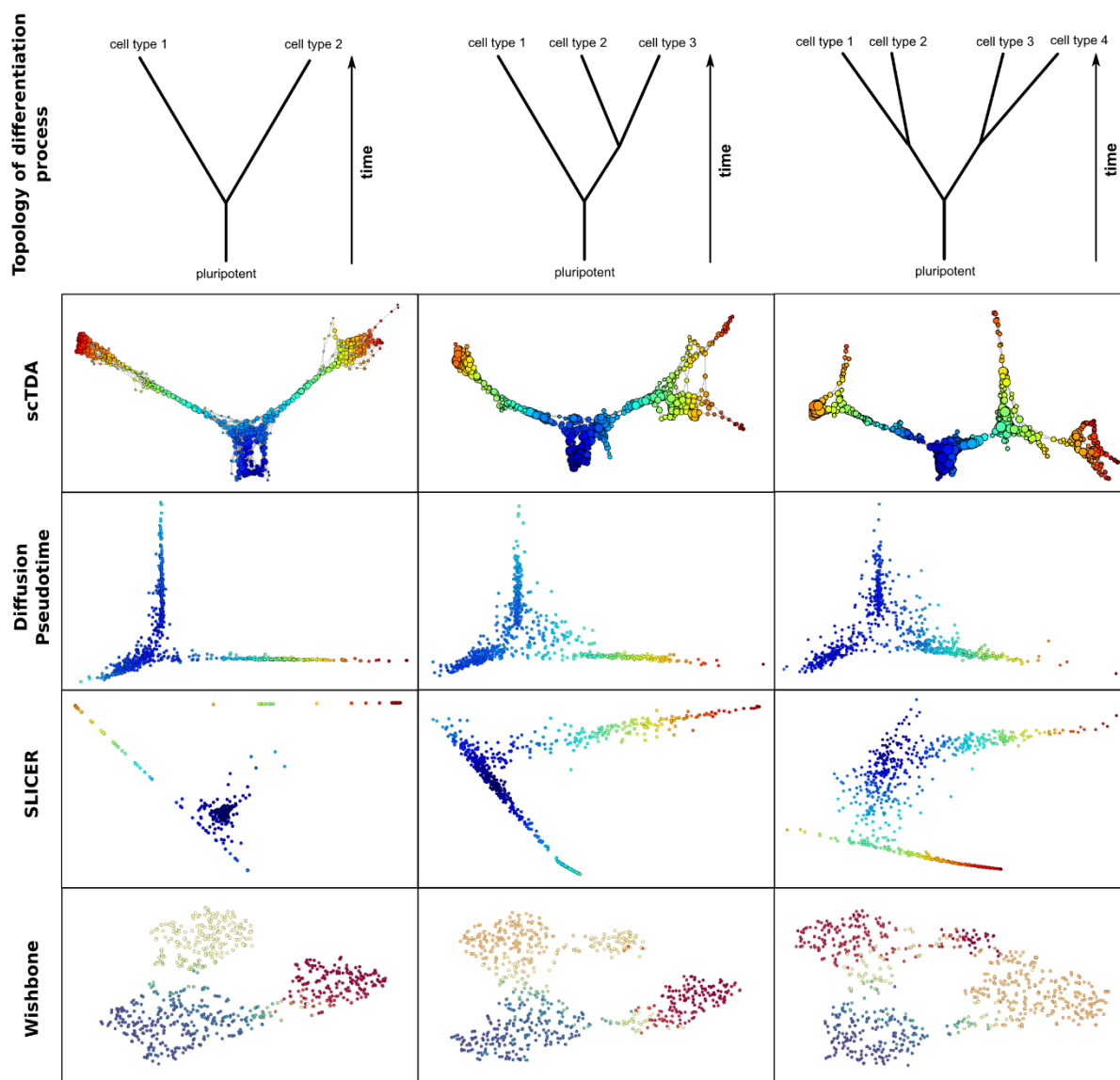
Single Cell Topological Data Analysis (scTDA)



Supplementary Figure 1. Workflow for single cell Topological Data Analysis (scTDA).

Longitudinal single-cell RNA-Seq data is mapped, demultiplexed and pre-processed, which removes cells that do not pass stringent quality controls (QC). A topological representation is built using the Mapper algorithm, and is based on highly-expressed genes that have a large variability. The expression of each transcript is a function that supports the topological representation. A pseudo-time ordering is established and different statistics (gene connectivity,

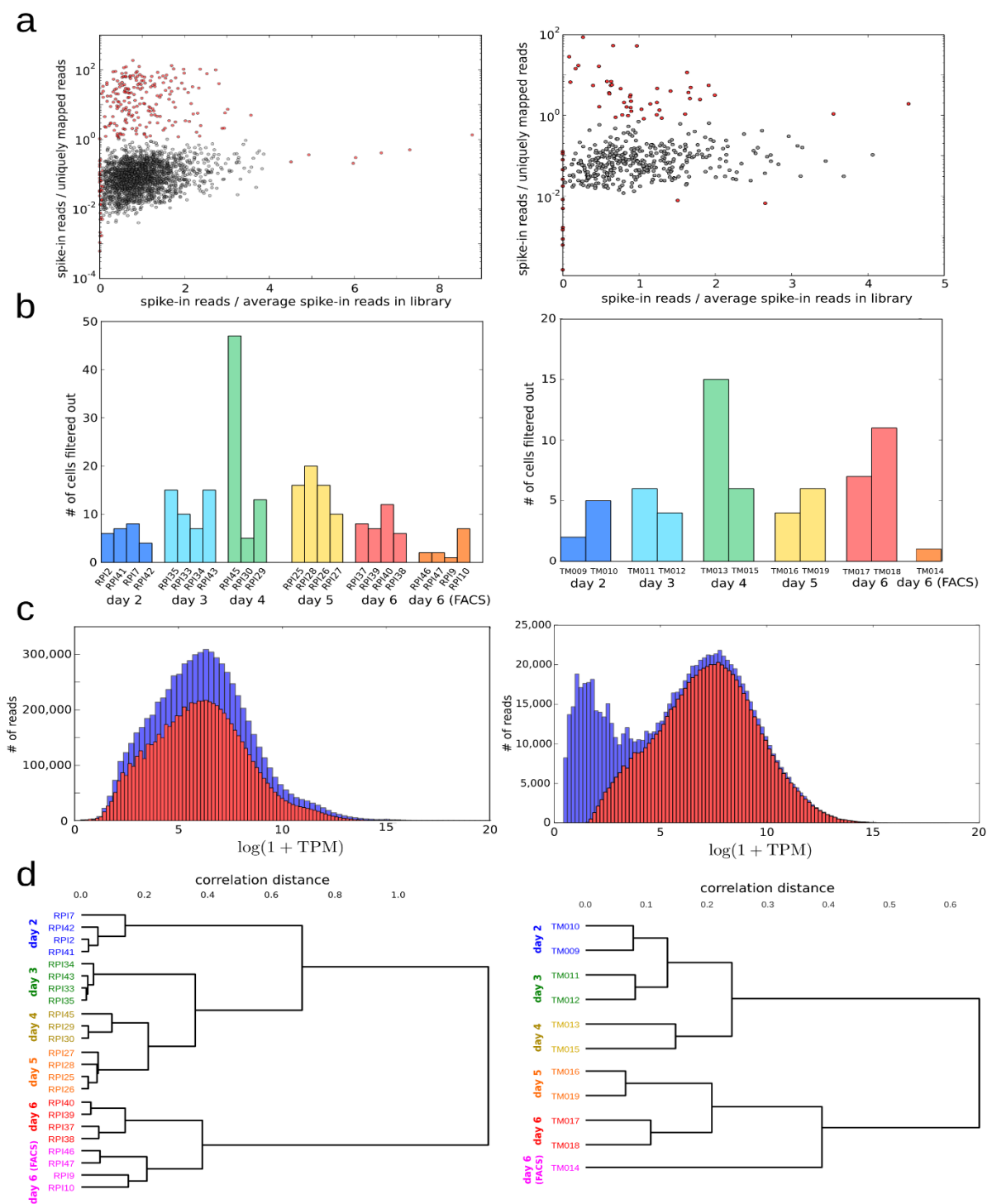
centroid, dispersion) are computed, making it possible to identify expression programs associated with different stages and populations of cells generated throughout the differentiation process.



Supplementary Figure 2. Comparison of various algorithms for the analysis of single cell expression data from developmental processes, using simulated data.

Noisy, branched, asynchronous differentiation processes were simulated, including the effect of drop-out events. 700 cells were sampled at three time points. The reconstruction of the

differentiation process produced by scTDA, Differentiation Pseudotime, Slicer, and Wishbone is shown, colored by the inferred pseudo-time, for processes with one, two, or three branching points. scTDA reproduces the topology of the differentiation process and its temporal structure in complex situations with more than one branching point.



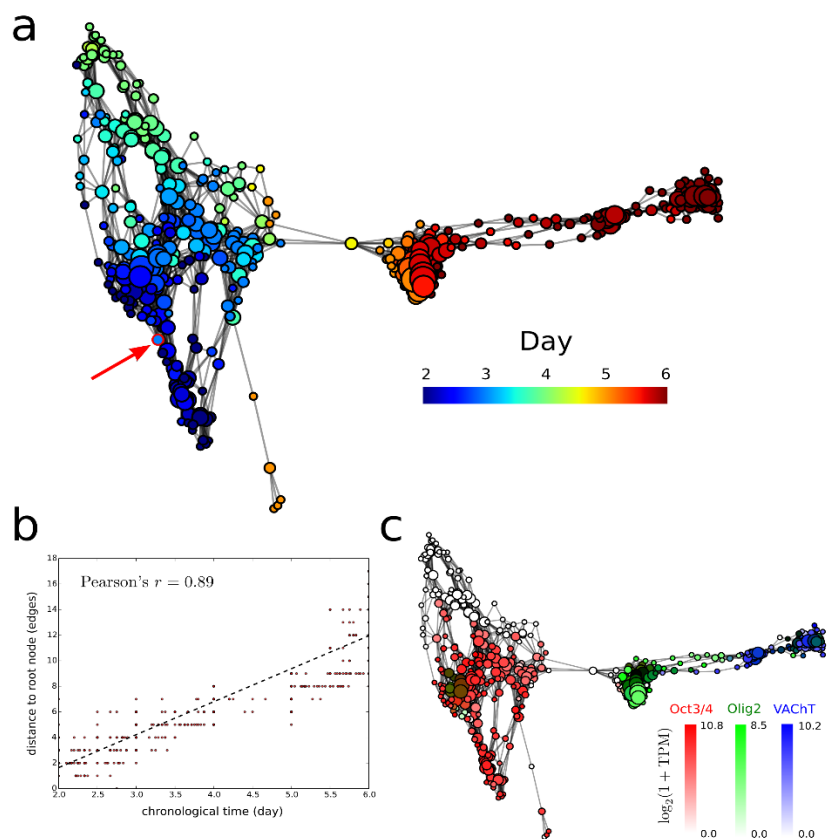
experiments 2 (left) and 1 (right).

a. The fraction of reads from each cell, represented by ERCCs (the ratio between ERCC reads and uniquely mapped reads), was plotted against the relative level of ERCCs per cell (the ratio between ERCC reads and the average number of ERCC reads in the library). This was done for each of the 2,208 cells sequenced in experiment 2 (one library demonstrating a large batch effect was removed from the analysis), and the 440 cells in experiment 1. Cells that are filtered out are represented in red. The ordinate is represented in logarithmic scale.

b. The distribution of filtered-out cells across each of the libraries.

c. The distribution of expression values for all genes in the complete set of cells before (blue) and after (red) filtering. In the experiment 1 (right), the expression of genes supported by less than 5 reads in a cell was set to zero to remove noise near the mRNA capture threshold.

d. Hierarchical clustering of the expression centroids of each library, based on correlation distance. Libraries cluster according to day, with substantial overlap between libraries from the same day.

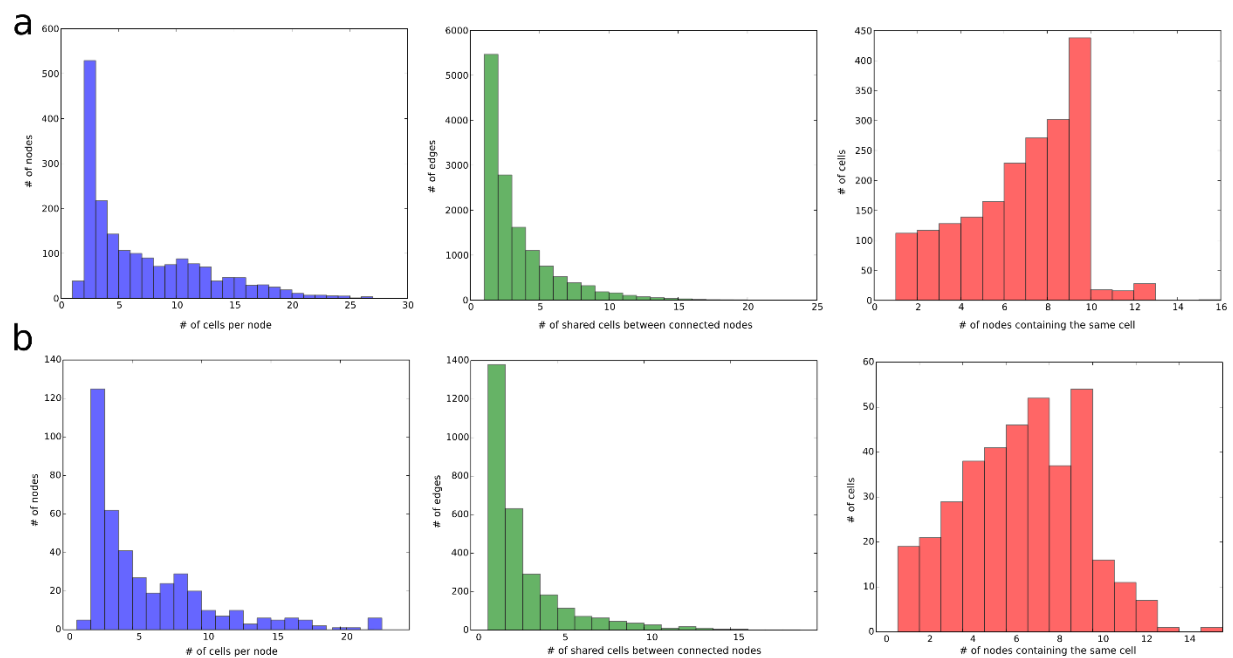


Supplementary Figure 4. Topological representation of longitudinal single-cell RNA-seq data from mESC differentiation into motor neurons in experiment 1.

a. The topological representation labeled by sampling time.

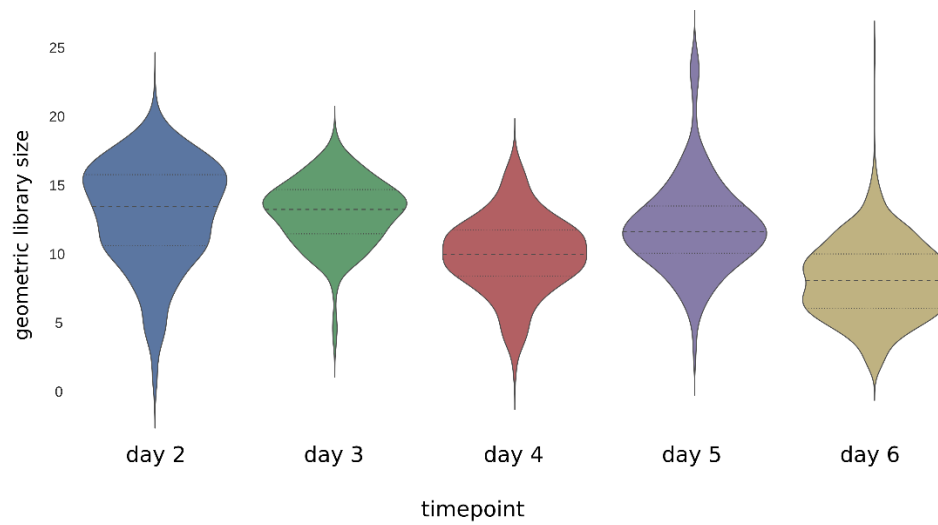
b. The distance of each node to the root node (marked with a red arrow in **a**), represented as a function of sampling time. The chronological time of a node is defined as the mean of the sampling times of the cells in the node.

c. The topological representation labeled by mRNA levels of known markers of pluripotent cells (*Oct3/4*, red), motor neuron progenitors (*Olig2*, green) and post-mitotic neurons (*VACHT*, blue).



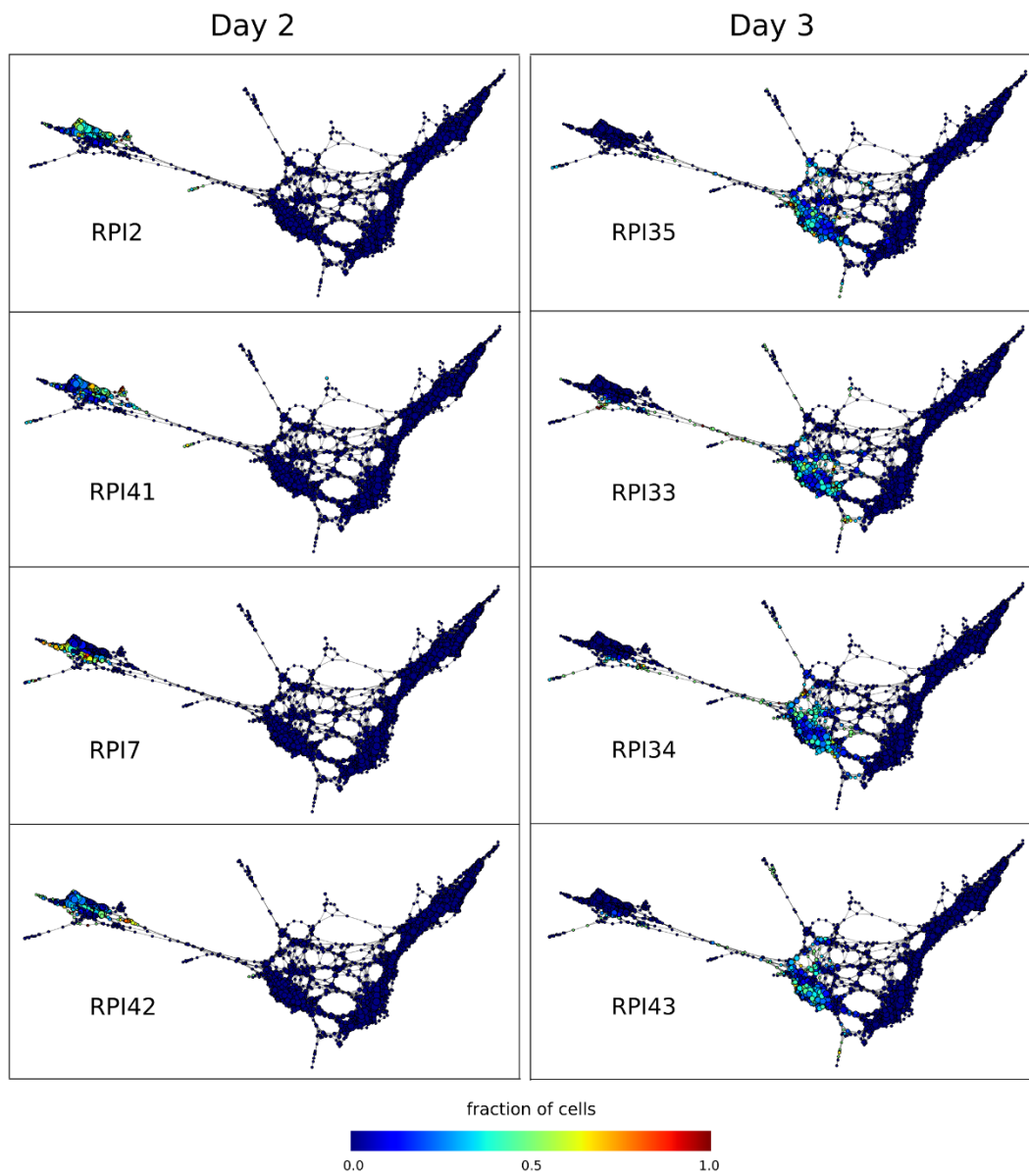
Supplementary Figure 5. Statistics of the topological representation in experiments 2 (a) and 1 (b).

Left: The distribution of the number of cells associated to each node of the topological representation. Middle: The distribution of the number of cells that are shared between nodes connected by an edge. Right: The distribution of the number of nodes in which an individual cell appears.



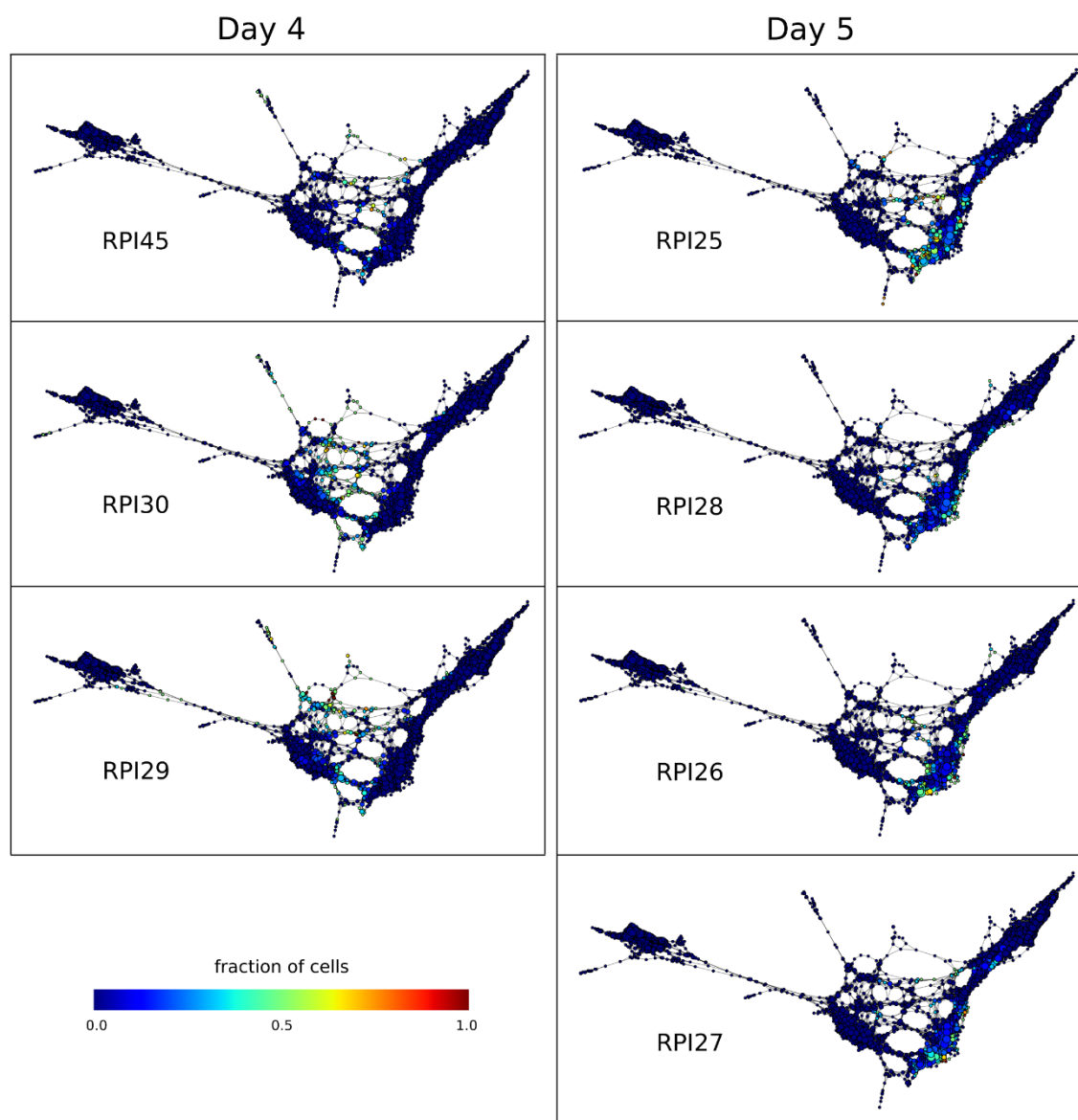
Supplementary Figure 6. Dependence of the library complexity on the sampling time in the motor neuron differentiation experiment 2.

The distribution of the geometric library size is plotted as a function of the sampling time, showing a mild dependence between these two quantities.



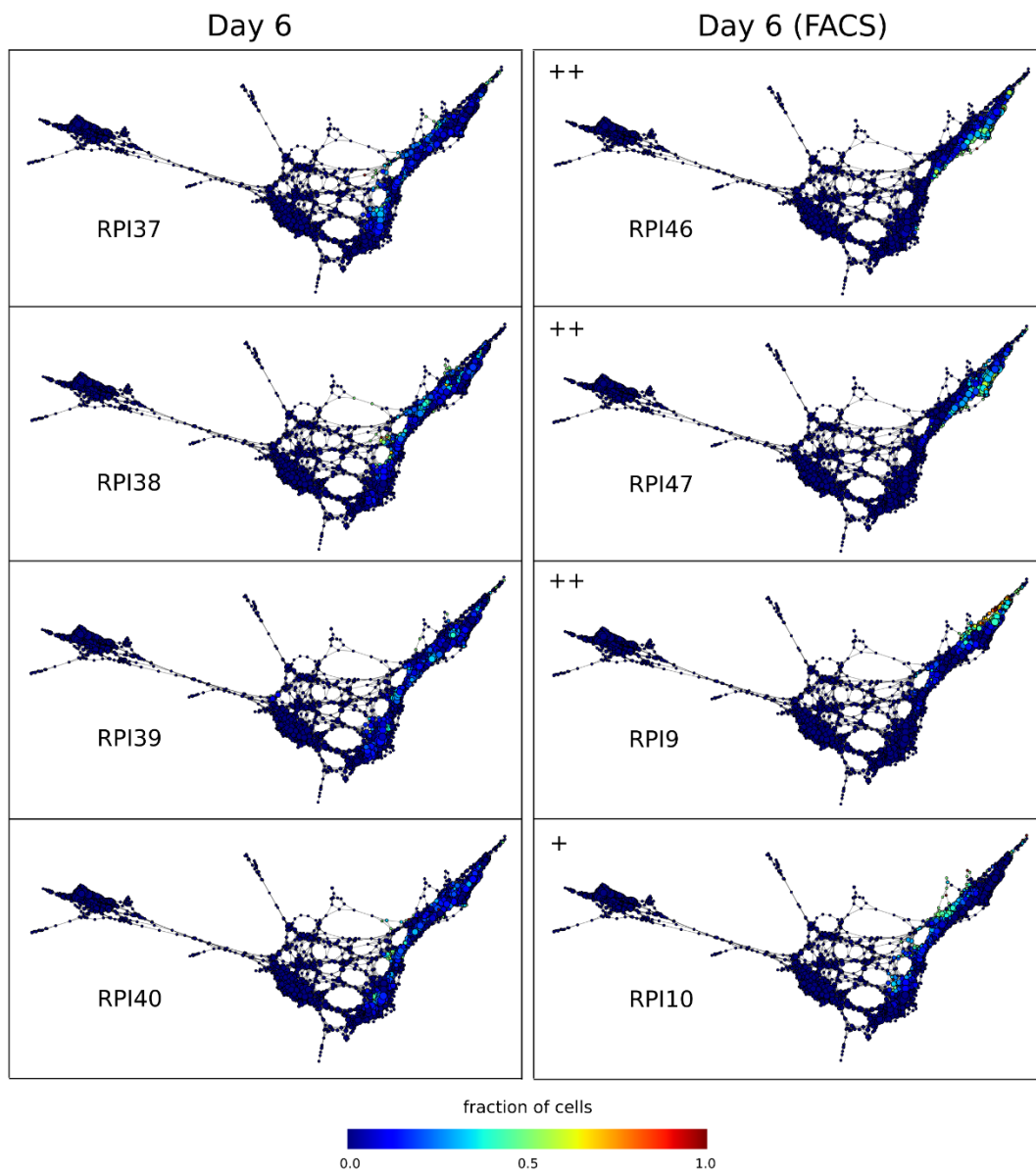
Supplementary Figure 7 (1/3). Distribution of different sequencing libraries across the topological representation in experiment 2.

Libraries from the same day have a substantial overlap in the topological representation, reflecting the absence of large batch effects.



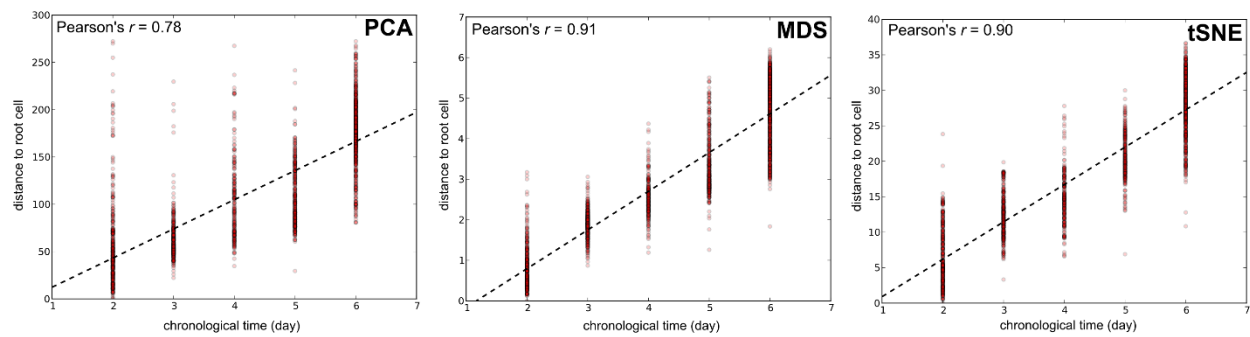
Supplementary Figure 7 (cont., 2/3). Distribution of different sequencing libraries across the topological representation in experiment 2.

Libraries from the same day have a substantial overlap in the topological representation, reflecting the absence of large batch effects.



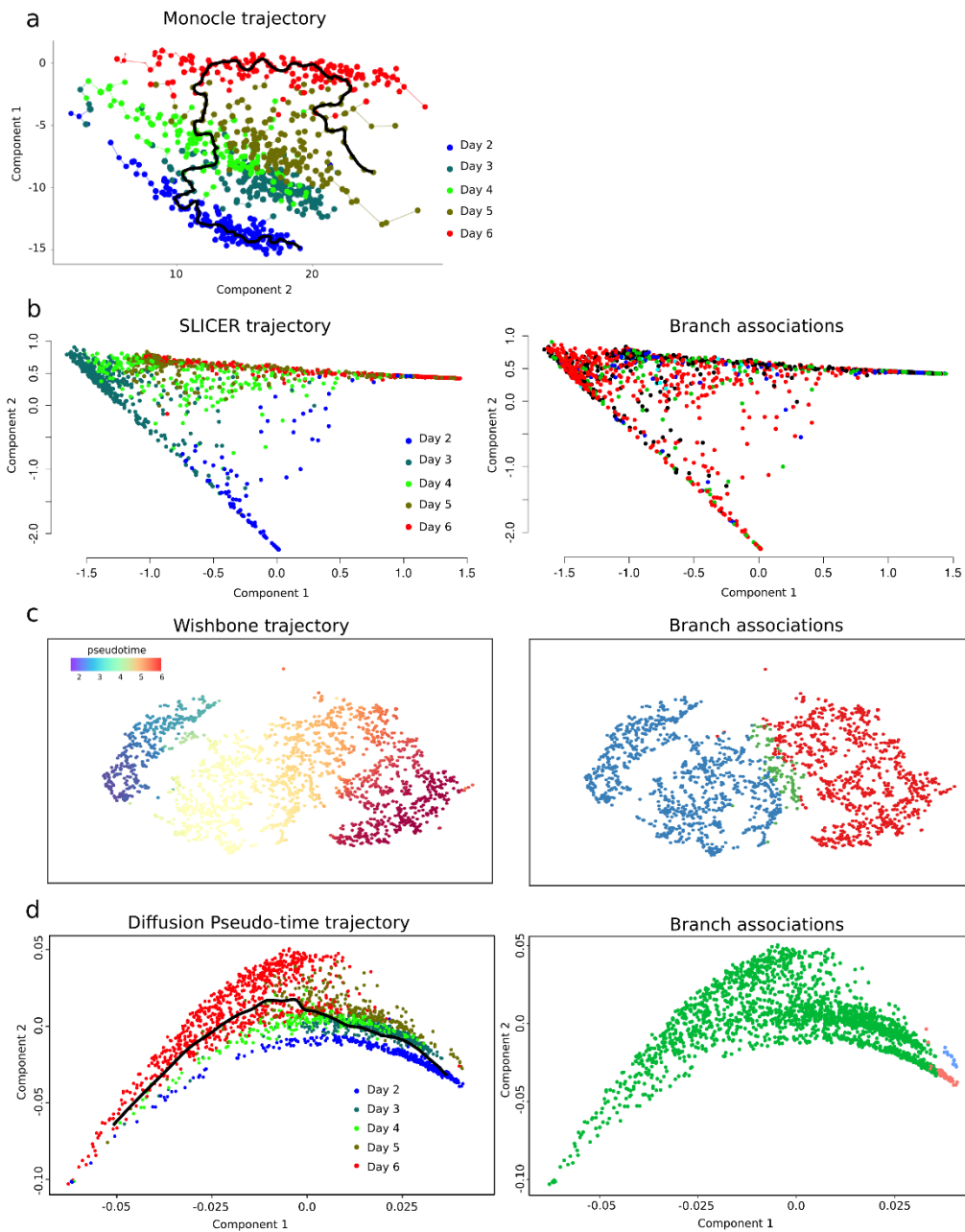
Supplementary Figure 7 (cont., 3/3). Distribution of different sequencing libraries across the topological representation in experiment 2.

Libraries from the same day have a substantial overlap in the topological representation, reflecting the absence of large batch effects.



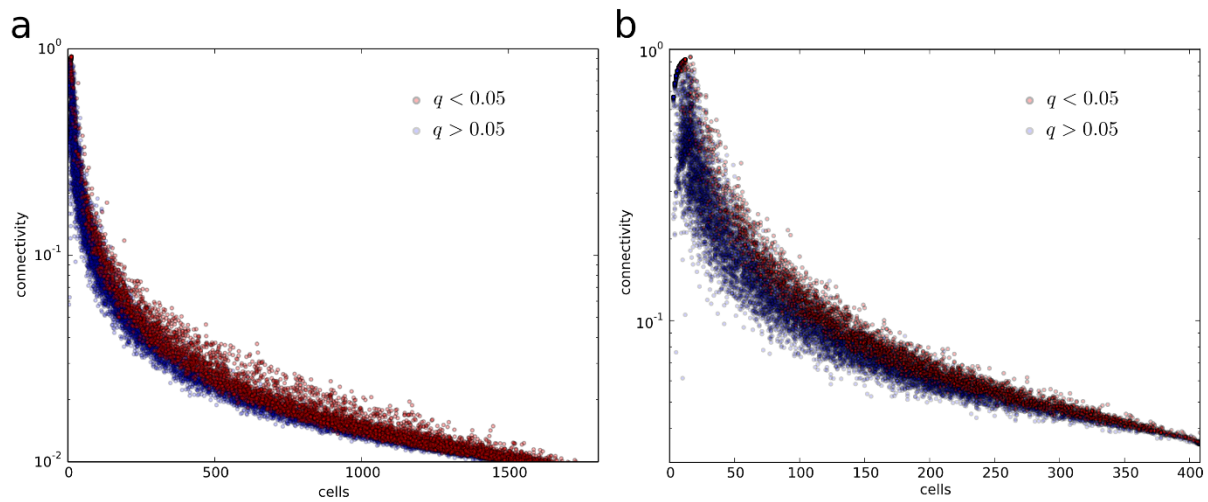
Supplementary Figure 8. Correlation between distance to root node and chronological time for PCA, MDS, and tSNE representations of the single cell expression data from mESC differentiation into motor neurons.

The two-dimensional Euclidean distance to the root cell (defined as the one that maximizes the correlation with the chronological sampling time) is shown as a function of the chronological sampling time.



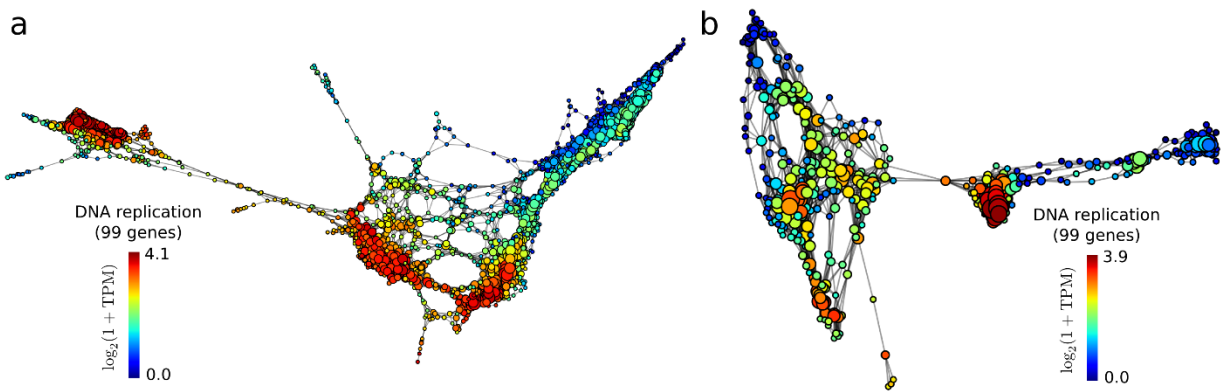
Supplementary Figure 9. Comparison to other algorithms for the analysis of longitudinal single-cell RNA-seq experiments, using data from mESC differentiation into motor neurons.

- a.** Output produced by Monocle. Monocle was run on 834 cells from experiment 2 to build a two dimensional minimum spanning tree representation of the data. Cells are colored by sampling day.
- b.** Output produced by SLICER. SLICER was run on the 1,964 cells from experiment 2 to build a two dimensional representation of the data, where cells are assigned a branch in the differentiation tree inferred from the data. Left: Cells are colored by sampling day. Right: Cells are colored according to the branch assignments made by SLICER.
- c.** Output produced by Wishbone. Wishbone was run on the 1,964 cells from experiment 2 to build a two dimensional t-SNE representation of the data where cells are assigned a pseudo-time and a branch in the differentiation structure inferred from the data. Left: Cells are colored by pseudo-time. Right: Cells are colored according to the branch assignments made by Wishbone.
- d.** Output produced by Diffusion Pseudotime. Diffusion Pseudotime was run on the 1,964 cells from experiment 2 to build a two dimensional representation of the data based on the first two diffusion components, where cells are assigned a pseudo-time and a branch in the differentiation structure inferred from the data. Left: Cells are colored by pseudo-time. Right: Cells are colored according to the branch assignments made by Diffusion Pseudotime.



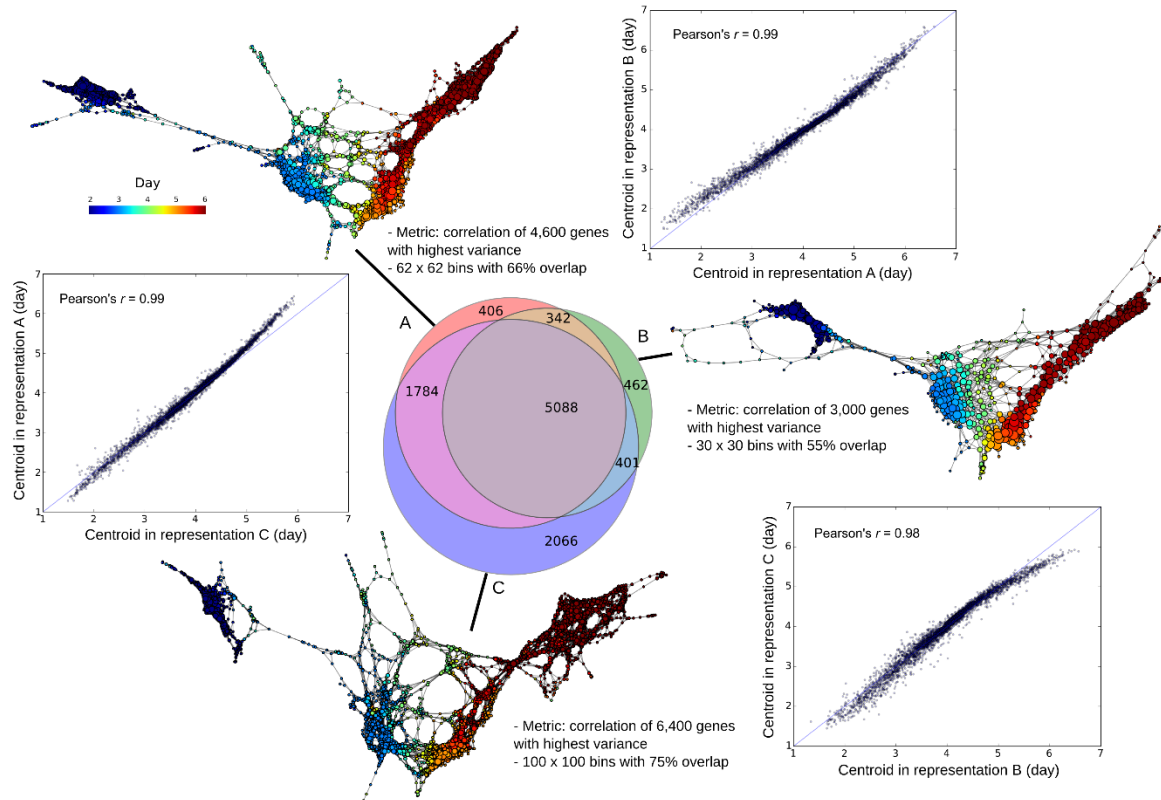
Supplementary Figure 10. Distribution of the gene connectivity score against the number of cells expressing the gene in the motor neuron differentiation experiments 2 (left) and 1 (right).

Statistical significance was evaluated by means of a permutation test (Online Methods). Genes with a significant gene connectivity ($q < 0.05$) after controlling for multiple hypothesis testing are shown in red.



Supplementary Figure 11. Expression levels of genes associated to DNA replication in the motor neuron differentiation experiments 2 (a) and 1 (b).

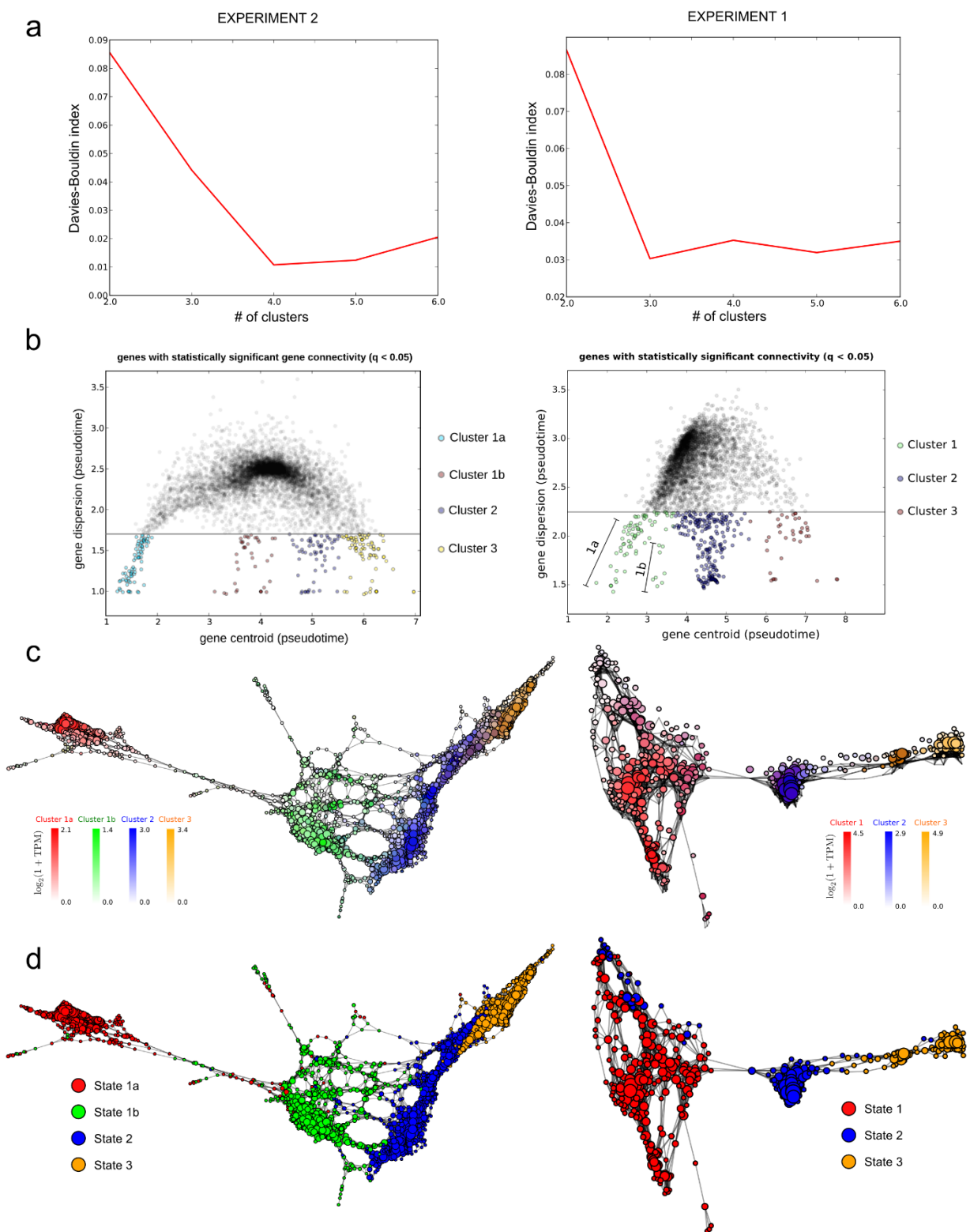
Topological representation labeled according to the expression of 99 genes related to DNA replication. Differentiated neurons exhibit very low levels of DNA replication, consistent with post-mitotic arrest.



Supplementary Figure 12. Stability of topological representations against different parameter choices.

Topological representations of the single cell RNA-seq data of the motor neuron differentiation experiment 2 using different choices for the number of patches and their overlap considered in the Mapper algorithm, as well as for the number of genes that are used to compute the distance matrix. The Venn diagram displays the number of genes with significant gene connectivity ($q < 0.05$) in three topological representations, showing a large degree of consistency across different representations. For each pair of representations, the correlation between the centroids (expressed in pseudo-time) of the $n = 5,088$ genes with significant ($q < 0.05$) connectivity in all representations

is also shown. The distribution of gene centroids is highly consistent across different choices of parameters, with Pearson's correlations in the range 0.98 – 0.99.



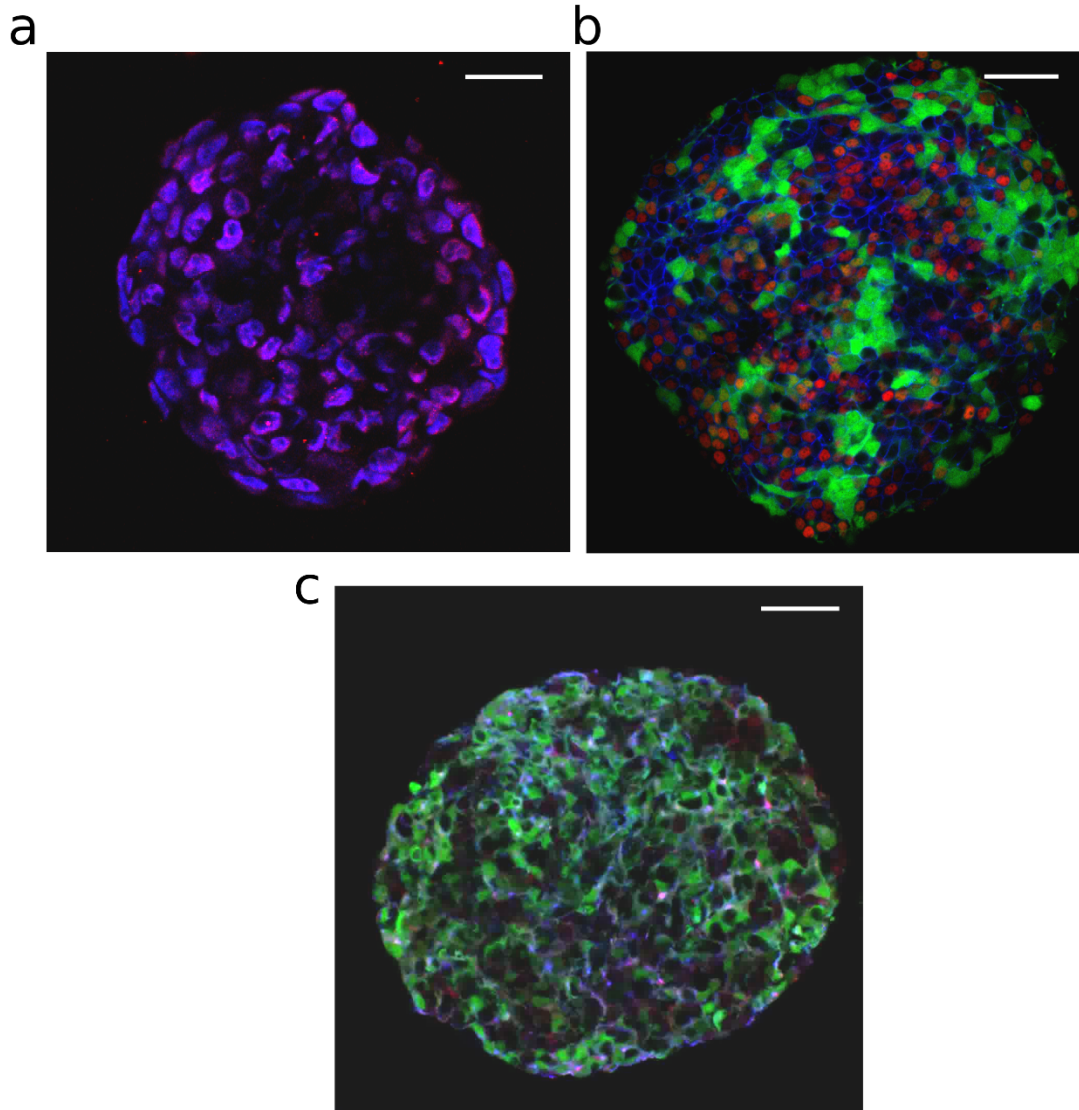
Supplementary Figure 13. Cellular populations arising throughout the differentiation of mESCs into motor neurons in experiments 2 (left) and 1 (right).

a. Davies-Bouldin index for k-means clustering of the centroid of low-dispersion genes in the topological representation of experiment 2 ($k_i < 1.7$) and experiment 1 ($k_i < 2.25$). The minimum is achieved for four and three clusters, respectively, in experiments 2 and 1.

b. Distribution of the centroid and dispersion of significant genes ($q < 0.05$) in the topological representation of experiment 2. Four principal gene groups (1a, 1b, 2 and 3) naturally arise from clustering the centroids of low-dispersion genes, corresponding to four transient cellular populations arising throughout the differentiation.

c. Topological representation of the two experiments labeled by mRNA levels of genes in each of the gene clusters obtained by k-means.

d. Topological representation of the two experiments labeled by the state assigned to each node based on the expression levels of each of the gene clusters.



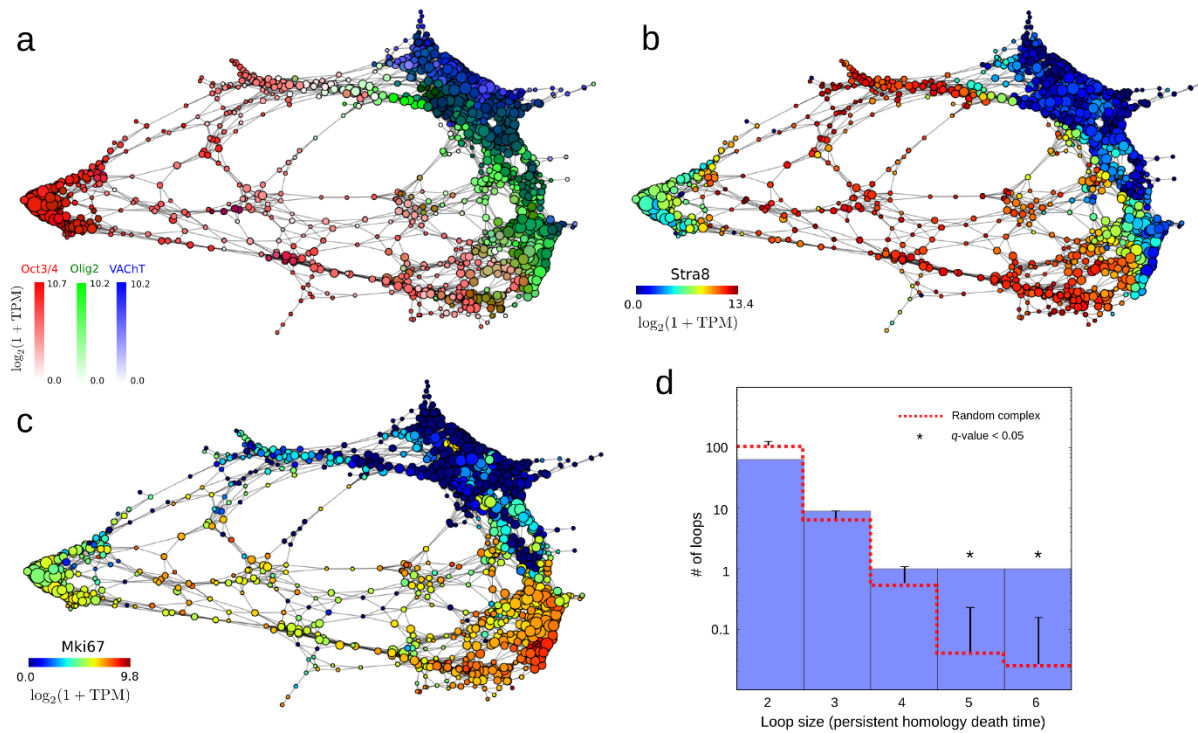
Supplementary Figure 14. Immunostaining of murine EBs against stage-specific surface markers identified by scTDA.

a. Immunostaining against the pluripotent surface marker *Pecam1*. Immunostaining of day 2 EBs against *Pecam1* (red), *Oct3/4* (blue), and *MNx1::eGFP* (green), showing overlap between *Pecam1*⁺ cells and *Oct3/4*⁺ cells.

b. Immunostaining against the neural progenitor surface marker *Ednrb*. Immunostaining of day 5 EBs against *Ednrb* (red), *Olig2* (blue), and *Mnx1::eGFP* (green), showing partial overlap between *Ednrb*⁺ cells and *Olig2*⁺ cells, and mutual exclusivity between *Ednrb*⁺ cells and *Mnx1::eGFP*⁺ cells.

c. Immunostaining against the post-mitotic neuron surface marker *Slc10a4*. Immunostaining of day 6 EBs against *VACHT* (red), *Slc10a4* (blue), and *Mnx1::eGFP* (green), showing overlap between *VACHT*⁺ cells and *Slc10a4*⁺ cells, and partial overlap with *Mnx1::eGFP*⁺ cells.

The scale bar in the immunostaining images denotes a length of 50 μm .



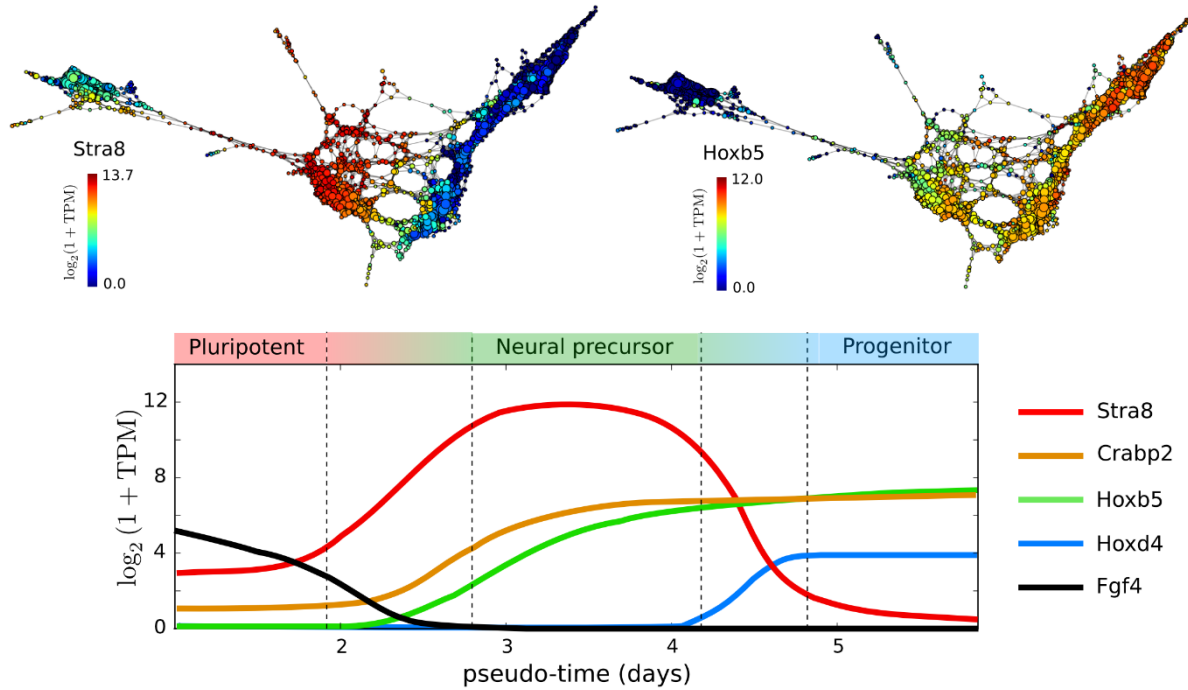
Supplementary Figure 15. Topological representation of single cell expression data from mESC differentiation into motor neurons experiment 2, based on cell cycle genes only.

a. The topological representation is labeled by mRNA levels of known markers of pluripotent cells (*Oct3/4*, red), motor neuron progenitors (*Olig2*, green) and post-mitotic neurons (*VachT*, blue).

b. The region of the topological representation corresponding to neural precursors has numerous loops. The topological representation is labeled by mRNA levels of *Stra8*, a marker of neural precursors.

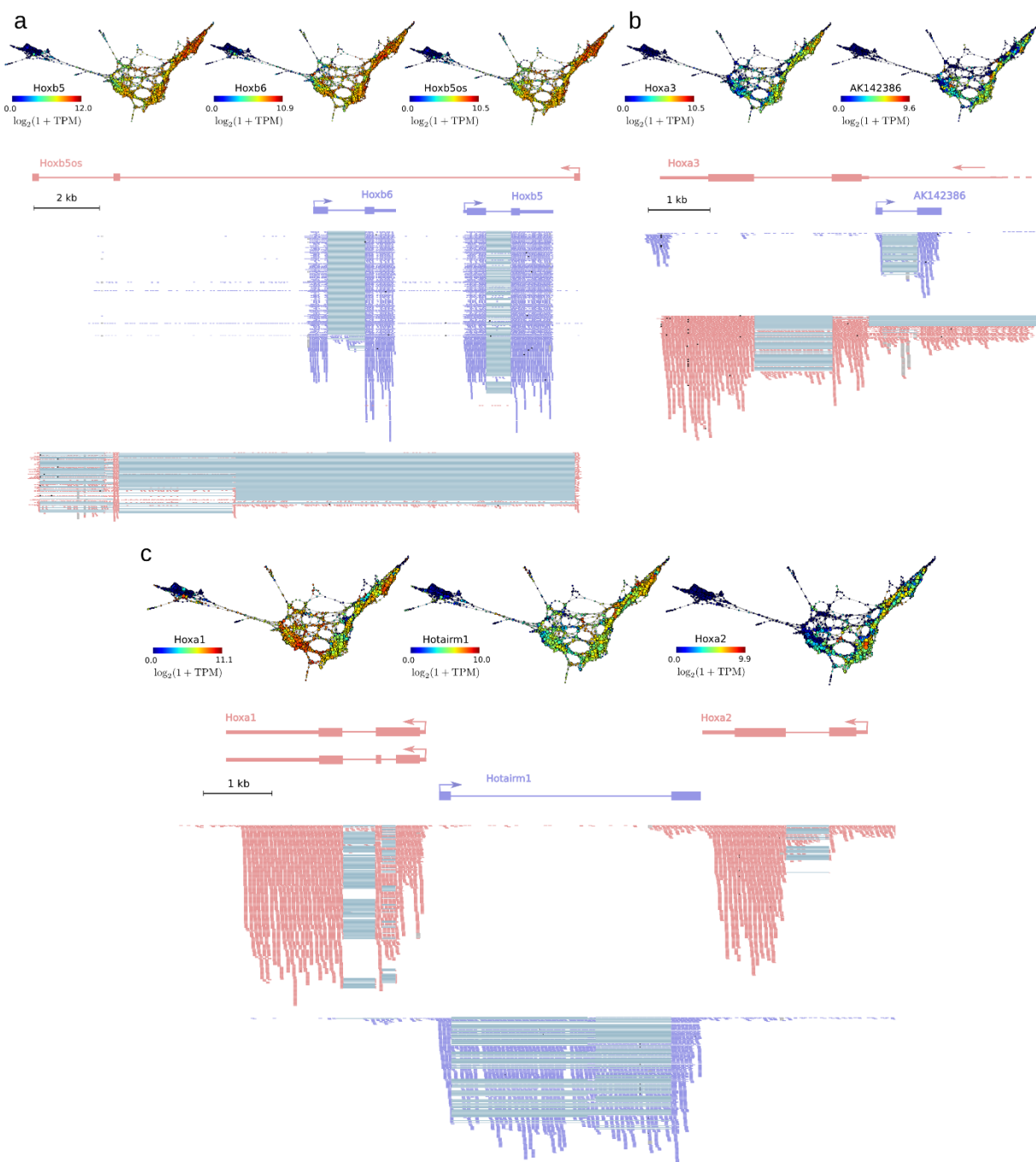
c. Neural precursors separate into proliferative and non-proliferative populations. The topological representation is labeled by *Mki67*, a known marker of cellular proliferation.

d. Distribution of the number of loops of a given size (as measured by their first persistent homology death time) across the representation (blue), compared to that of random complexes built by permuting the genes independently in each cell. Some of the larger loops in the topological representation are statistically significant, consistent with a biological origin for these loops.



Supplementary Figure 16. Transcriptional events occurring during the transition between pluripotent and neural precursor, and between neural precursor and neural progenitor populations of cells.

Reconstructed expression timeline for some of the transcriptional changes identified by scTDA in the transitions between pluripotent and neural precursor, and between neural precursor and neural progenitor populations, in the motor neuron differentiation experiment 2. scTDA identified upregulation of *Stra8* and downregulation of *Fgf4* as one of the early events in the transition between a pluripotent and a neural precursor state. The topological representation labeled according to the mRNA expression levels of *Stra8* and *Hoxb5* is also shown.

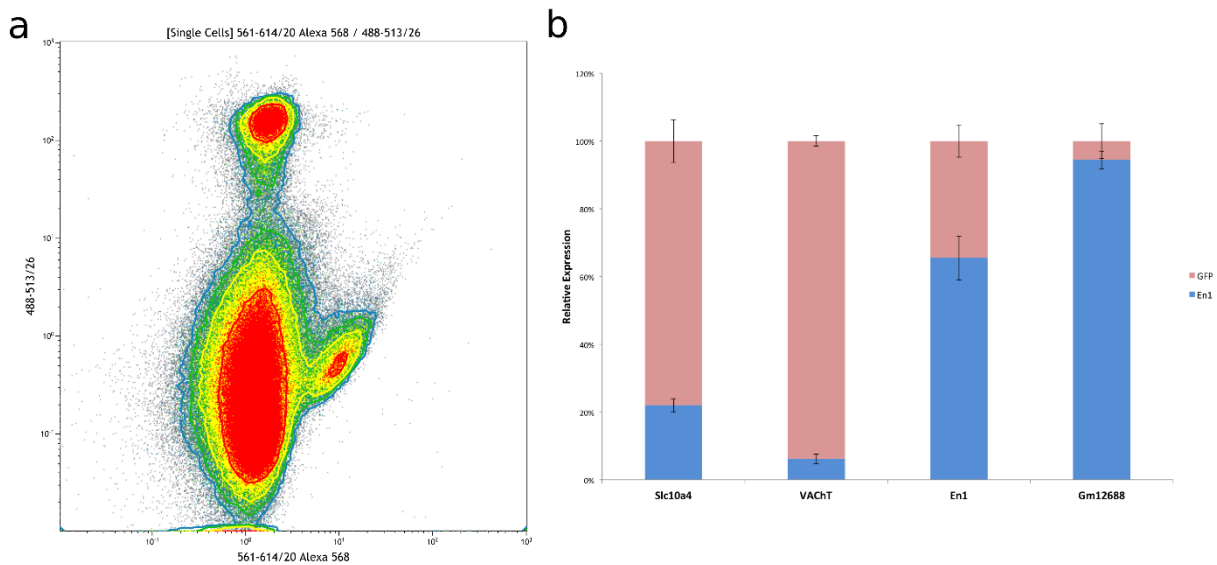


Supplementary Figure 17. Expression of antisense lncRNAs derived from homeobox gene clusters during the differentiation of mESCs into motor neurons.

a. Topological representations of experiment 2 labeled by mRNA levels of *Hoxb5*, *Hoxb6*, and the antisense lncRNA *Hoxb5os*, showing concordant expression of these genes during the generation of motor neurons from mESCs. Stranded bulk RNA-seq reads from day 5 of the differentiation process are also depicted.

b. Topological representations of experiment 2 labeled by mRNA levels of *Hoxa3* and the antisense lncRNA *AK142386*, and stranded bulk RNA-seq reads.

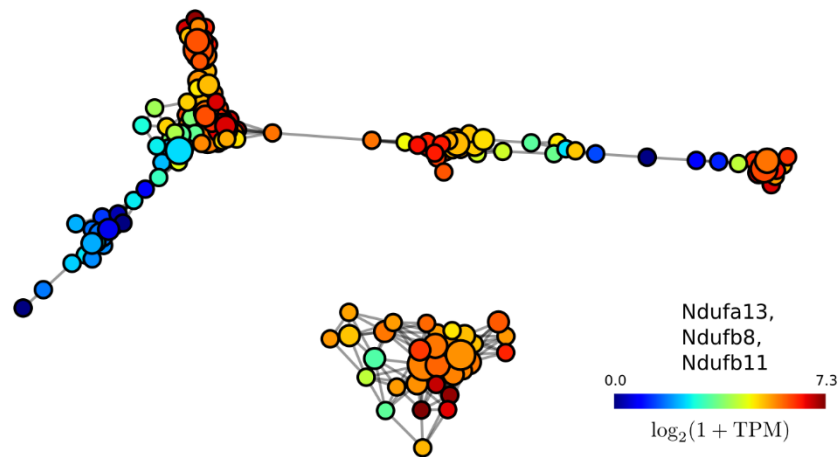
c. Topological representations of experiment 2 labeled by mRNA levels of *Hoxa1*, *Hoxa2*, and the antisense lncRNA *Hotairm1*, and stranded bulk RNA-seq reads.



Supplementary Figure 18. Sorting of V1 interneurons validates the exclusive expression of lncRNA *Gm12688*.

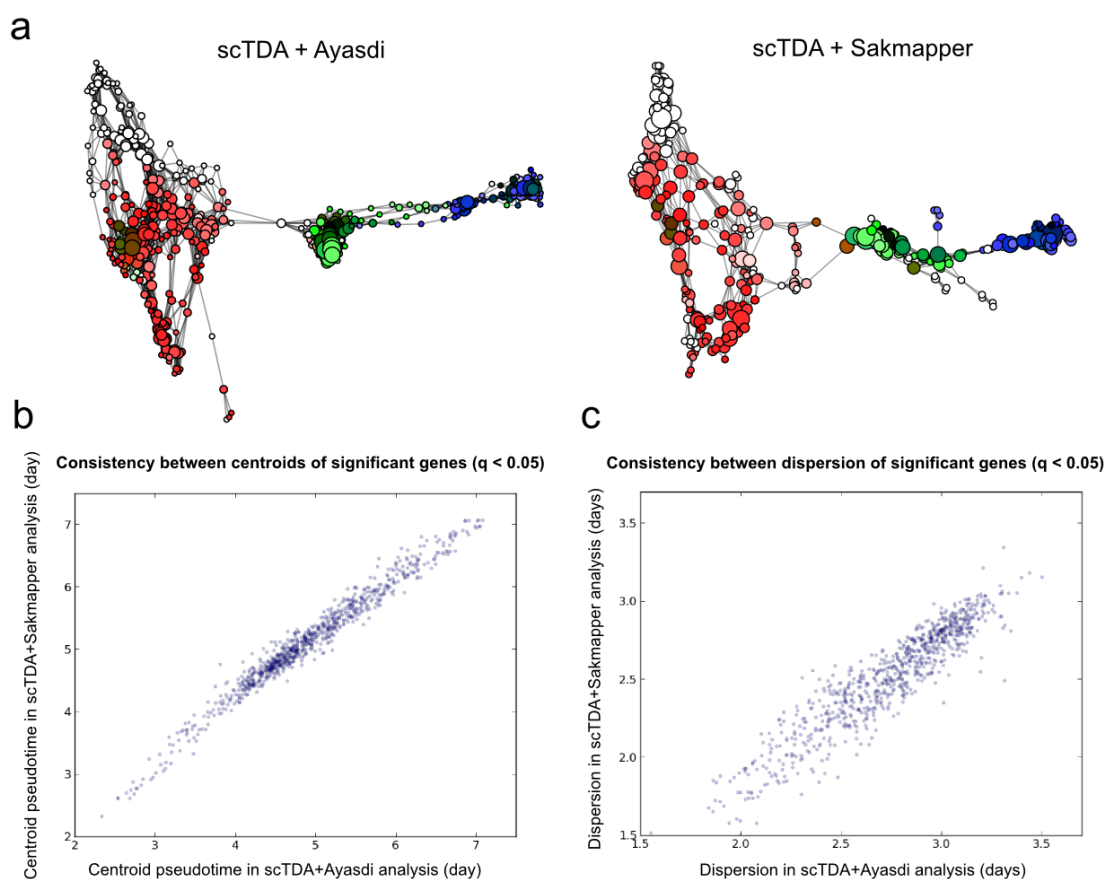
a. Dissociation of day 6 EBs and immunostaining against Engrailed (*En1*) enabled purification of V1 interneurons through flow cytometry. Alexa 568 expressing cells are *En1* positive. *Mnx1::eGFP* expressing motor neurons are low in Alexa 568 signal, but high in eGFP fluorescence.

b. RT-PCR of eGFP expressing motor neurons and *En1*⁺ cells demonstrates enrichment of expected markers, and a high enrichment of *Gm12688* in V1 interneurons.



Supplementary Figure 19. Differential expression of genes across the two populations of alveolar type I cells in the developing distal lung epithelium.

The topological representation is labeled by the mRNA levels of genes coding for proteins from the mitochondrial respiratory chain NADH dehydrogenase that are differentially expressed between the two populations of alveolar type I cells identified by scTDA.



Supplementary Figure 20. Comparison of scTDA analyses using Ayasdi and Sakmapper topological representations of the single-cell expression data from the motor neuron differentiation experiment 1.

(a) The topological representations obtained using Ayasdi (left) and Sakmapper (right) are labeled by mRNA levels of known markers of pluripotent cells (*Oct3/4*, red), motor neuron progenitors (*Olig2*, green), and post-mitotic neurons (*VACHT*, blue).

(b, c) Consistency between the centroid (b) and dispersion (c) of genes that have a significant ($q < 0.05$) gene connectivity in both topological representations.

Supplementary Tables

	Diffusion pseudotime	Wishbone	SLICER	Destiny	Monocle	SCUBA	scTDA
Unbiased	No	No	Yes	Yes	No	No	Yes
Statistics	Yes*	Yes*	Yes*	No	Yes*	Yes*	Yes
Exploits longitudinal information	No	No	No	No	No	Yes	Yes
Reference	6	7	8	9	10	11	-

* Only for branching events

Supplementary Table 1. Comparison of various features among existing algorithms for single-cell RNA-seq data analysis.

Supplementary Table 2 (Provided as a separate spreadsheet). Characterization of the expression profile in the topological representation of the two motor neuron differentiation experiments for all RefSeq genes. For each gene and experiment, the number of nodes with non-zero expression, the mean, minimum and maximum expression values, the value of the gene connectivity, the statistical significance before (p -value) and after (q -value) adjusting for the false discovery rate (Benjamini-Hochberg), the centroid and dispersion (expressed in days), and the gene group assignment are presented. Several gene ontology annotations are also shown.

Supplementary Table 3 (Provided as a separate spreadsheet). Gene ontology enrichment analysis for each gene group in the two motor neuron differentiation experiments. Statistical significant ($q < 0.05$, Bonferroni) biological process gene ontologies are presented for each gene group in the topological representation.

Supplementary Table 4 (Provided as a separate spreadsheet). Characterization of the expression profile in the topological representation of the two motor neuron differentiation experiments for significant lncRNAs. For each antisense or intergenic NONCODEv4 lncRNA with significant ($q < 0.05$) gene connectivity in both experiments gene and experiment, the number of nodes with non-zero expression, the mean, minimum and maximum expression values, the value of the gene connectivity, the statistical significance before (p -value) and after (q -value) adjusting for the false discovery rate (Benjamini-Hochberg), the centroid and dispersion (expressed in days), the gene group assignment, and the number of reads in bulk stranded RNA-seq data from days 2 to 6 of the differentiation are presented. Alternate RefSeq name is shown, when available. For antisense lncRNAs, the name of the coding genes in the opposite strand is presented. Only lncRNAs supported by at least 50 reads in one day of the bulk stranded RNA-seq data are considered.

Supplementary Table 5 (Provided as a separate spreadsheet). Characterization of the expression profile in the topological representation of 80 embryonic (E18.5) mouse lung

epithelial cells. For each gene, the number of nodes with non-zero expression, the mean, minimum and maximum expression values, the value of the gene connectivity, and the statistical significance before (p -value) and after (q -value) adjusting for the false discovery rate (Benjamini-Hochberg), are presented.

Supplementary Table 6 (Provided as a separate spreadsheet). Characterization of the expression profile in the topological representation of 1,529 individual cells from 88 human preimplantation embryos. For each gene, the number of nodes with non-zero expression, the mean, minimum and maximum expression values, the value of the gene connectivity, the statistical significance before (p -value) and after (q -value) adjusting for the false discovery rate (Benjamini-Hochberg), the centroid and dispersion (expressed in days), and the gene group assignment are presented.

Supplementary Table 7 (Provided as a separate spreadsheet). Characterization of the expression profile in the topological representation of 272 newborn neurons from the mouse neocortex. For each gene, the number of nodes with non-zero expression, the mean, minimum and maximum expression values, the value of the gene connectivity, the statistical significance before (p -value) and after (q -value) adjusting for the false discovery rate (Benjamini-Hochberg), the centroid and dispersion (expressed in days) are presented.

Supplementary Table 8 (Provided as a separate spreadsheet). Barcoded reverse transcription primers utilized in motor neuron differentiation experiment 2.

Supplementary Note 1

A mathematical primer.

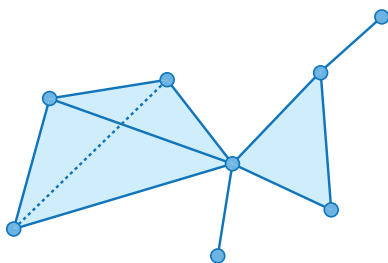
In this note some of the fundamental mathematical objects used to study single cell expression data are further explained. scTDA builds upon recent developments in topological data analysis, or TDA. The overarching aim of TDA is to infer global properties of spaces from samples of points. Most of the constructions in TDA are based on generating sets of simplicial complexes (generalizations of networks) and exploit the global structure of these complexes. In particular, we are interested in a space that is a simplified, lower dimensional version of the original space. In 1904¹ Poincaré called this concept the skeleton of the space, or in more modern terms, a Reeb graph or space^{2,3}. A Reeb graph is a graph (a one dimensional object) that summarizes some, but not necessarily all, topological features of a space, like number of connected components or loops. Mapper is the algorithm developed by Singh, Memoli, and Carlsson⁴ that constructs simplicial complexes from finite metric spaces (points with distances) as approximations to Reeb spaces.

We are interested in studying dynamic biological processes (like development or evolution) from samples of points, reconstructing and inferring properties of the underlying or some derived spaces (as Reeb spaces), showing how these properties relate to time, and studying functions on these spaces by performing statistical analysis. Although TDA is able to represent, summarize and quantify properties of spaces from finite sampling, TDA is not adapted to study dynamical processes or to perform statistical analysis on functions on associated objects (e.g. simplicial complexes). The primary intellectual contributions of this paper are the following: to extend the construction of Reeb spaces in finite metric spaces with temporal sampling, to study statistics on

functions defined on the data, and to apply these methods in the context of single cell transcriptomic data.

Starting from single cell transcriptomic data, the final reconstructed object is an annotated simplicial complex, with a single vertex labeling the initial state and statistics associated with functions on the data. The marked single point in the construction allows the association of a real-valued function to the rest of the complex representing the imputed pseudotime. The statistics on functions on the simplicial complex measure how expression profiles are localized.

Simplicial complexes and topological data analysis.



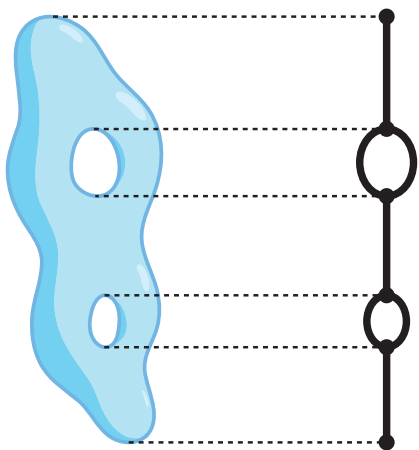
The final object we calculate is an annotated simplicial complex together with statistics associated to functions defined on the complex. The basic “atomic” objects that compose a simplicial complex are simplices. An n -dimensional simplex spanned by points $[v_0, v_1, \dots, v_k]$ is the convex hull of $k+1$ affine-independent points in \mathbb{R}^n . A zero-dimensional simplex is just a point, a one-dimensional simplex a line segment, a two-dimensional simplex a triangle, a three dimensional a tetrahedron, etc. Each n -simplex is a collection of $n+1$ points, $n(n+1)/2$ lines joining the points, etc. Simplicial complexes are collections of simplices of different dimensions (see figure). Indeed, a simplicial complex can be defined as a set of simplices with the property that a non-empty intersection between two simplices in the set is a face in each of the two simplices forming the intersection.

Most of data, including most of biological data, and certainly the data from single cell expression, can be presented as a set of points (e.g. cells) with some notion of similarity or distance (e.g. similarity between expression profiles), i.e. these data constitute a finite metric space (X, ∂) . We can associate different simplicial complexes with the data. For instance, consider a scale ε at which we are going to study our data and define the Vietoris-Rips complex $VR_\varepsilon(X, \partial)$ as

1. vertices are the points of X , and
2. a simplex $[v_0, v_1, \dots, v_k]$ is in the complex when $\partial(v_i, v_j) < \varepsilon$ for all $0 \leq i, j \leq k$.

Topological Data Analysis (or TDA) refers to a set of techniques to characterize the topological properties of data. Typically it involves constructing an auxiliary set of objects, e.g. simplicial complexes.

Morse functions, Reeb graphs and Reeb spaces.



Imagine that we have a topological space M , as in the left side of the figure, and a continuous function $f: M \rightarrow \mathbb{R}$. Let us define an equivalence relation on M between two points $p \sim q$ in M if $f(p)=f(q)$ and p and q in the same connected component in $f^{-1}(p)$. The Reeb graph associated to the space M and the function f is $\text{Reeb}(M, f) = M/\sim$, the quotient of M by the above equivalence relation (see figure). The first mention of this construction was in the 5th Supplement

of the founding paper of topology (*Analysis Situs*) from 1904. Poincaré referred to the Reeb graph of a manifold as the skeleton.

One can generalize the notion of Reeb graph to Reeb spaces by considering functions $f: M \rightarrow \mathbb{R}^n$. Using the above equivalence relation one can define $\text{Reeb}(M, f) = M/\sim$. In the applications to single cell data we will be only using $n=1$ and $n=2$.

Mapper as an approximation to Reeb spaces.

The data that we will be considering in single cell analysis can be understood as points in a high dimensional space with some notion of similarity between them (X, ∂) . We can approximate the “skeleton” or Reeb space of these data by considering an auxiliary function $f: X \rightarrow \mathbb{R}^n$. The Mapper algorithm generates a simplicial complex $\text{Mapper}(X, \partial, C)$ from (X, ∂, f) and $C = \{U_i\}$, a cover on the image of f in \mathbb{R}^n .

Several types of covers are possible, depending on specific applications. The Mapper algorithm proceeds as follows:

- 1.- Cluster the elements of $f^{-1}(U_i)$, the inverse image of each element in the cover. Let us denote each of the clusters by C_μ .
- 2.- Vertices of $\text{Mapper}(X, \partial, C)$ are the clusters C_μ from 1. and a simplex $[C_0, C_1, \dots, C_k]$ is in the simplicial complex $\text{Mapper}(X, \partial, C)$ when the $(k+1)$ -fold intersection of elements in $\{C_0, C_1, \dots, C_k\}$ is not empty.

The results of Mapper can be considered as an approximation of Reeb spaces. Indeed, the work of Elizabeth Munch and Bei Wang have shown the convergence of Mapper and Reeb graphs⁵.

Weinberger theorems.

One natural question is how many points we should sample to get a “good” approximation to the topology of the original space. Without any assumption on connectivity, Weinberger⁶ estimated the number of topological types (potential solution to the topology inference problem, which could be in this context homeomorphisms, diffeomorphisms or homotopy type) from finite sampling in a manifold of dimension n embedded in \mathbb{R}^N . These estimates depend on a scale number, called the condition number (the minimum size at which a tubular neighborhood self-intersects) and the diameter of the manifold (D) measure in units of the scale number (so is dimensionless). For dimension, $n=0$ (clusters) or $n=1$ (graphs) the number grows as D^N , while for $n=2$ the number of types grows like $\exp(D^{N/2})$, and for $N>2$ $\exp(D^N)$. Notice that the problem has a polynomial on size only for inference of low dimensional objects (clusters and graphs), but become exponential on the diameter for higher dimensions. Even to express the right answer, one needs many bits (on the order of \log of the number of possibilities), so one needs at least $O(N \log(D))$ points in dimension 0 and 1, $O(D^{N/2})$ for dimension 2 and $O(D^N)$ for higher dimensions. That implies that in our single cell experiments, where N , the embedding dimension (number of genes used in the reconstruction) is of the same order as the number of cells, we can only aim to capture 0 or 1 dimensional features. Notice that the exponential of number of genes will be always be bigger than the atoms in the universe, then bigger than the number of cells in the universe, so we will never reach the complexity needed for characterizing the topology type in dimension bigger than 1. These obstacles become even more daunting when interested in estimating additional structure (local coordinates, metric structure, etc) as in some manifold learning procedures.

These results that one is more likely to be successful to identify low dimensional features of spaces (like clusters of graphs) than to capture higher dimensional ones. We have two advantages in this direction. First is a technical one. Inference of Reeb graphs or clustering should be easier with finite number of points. Second is a biological presumption. When looking at single cell data we expect data to come in clusters (different cell types) and trajectories (like lineage differentiation, or evolution). These theoretical and practical implications suggest that a potential approach to study single cell data should be through the study of low dimensional objects, as the approximation of Reeb graphs that Mapper generates. However, the Mapper reconstructed complexes lack annotation regarding time and statistics associated to functions on it, that are fundamental for an application to single cell applications.

Studying temporal single cell data, scTDA.

scTDA builds upon Mapper to generate an simplicial complex that captures some of the low dimensional features from the space (like clusters and trajectories) from single cell data. One of the main ideas described in this manuscript is how to use the temporal information to find a single vertex in the inferred complex that represents the closest point to the most “ancestral” state. This ancestral node could be a stem cell state, progenitor, embryonic state, cancer initiating cell, etc. The idea, explained throughout the manuscript, is based on the association between the graph distance from a node and the measured time. The ancestral state is defined as the node that maximizes this correlation. The details are explained in Results and Methods sections of the manuscript.

The second fundamental point in our applications, and in many others as well, is the identification of some features in the space. That requires to study functions from data to real numbers, $g: X \rightarrow \mathbb{R}$ and statistics associated to them. In our case, these functions could be the expression of genes or phenotypic observations in single cells, and we are interested in identifying transcriptional programs associated to states, such as genes expressed on the cell membrane, that could work as markers. A simple idea is to construct functions on the simplicial complex by averaging functions on X over the points representing each simplex (in each cluster). That defines a kind of pushforward of the function on the original data to a function on the simplicial complex $g_{\sim}: \text{Mapper}(X, \partial, f, C) \rightarrow \mathbb{R}$. Statistics are defined by permuting the values of the function on the original data, and comparing the pushforward of the permuted values with the original ones. The details for the single cell analysis are described in the manuscript.

Using scTDA, we refine TDA simplicial representations by providing a marked state, a derived pseudotime, a set of pushforward functions g_{\sim} on the complex and statistics associated to them.

Homology in simplicial complexes and persistent homology.

Now that we have an annotated simplicial complex representing the data and functions on it, we can further use topological tools to characterize the presence of different topological tools on them. In particular, in our applications, we have used simplicial homology and persistent homology. The basic idea of simplicial homology is to associate groups with objects of different dimension. The zero dimensional homology captures the number of connected components in a simplicial complex in the following fashion: we say that two zero simplices a and b are in the same equivalence class

$a \sim b$ if they are the boundary of a chain of one dimensional objects (i.e. if there is a path connecting a and b). The same procedure can be defined to objects of higher dimension: closed loops are related if they form the boundary of two dimensional objects, etc. The rank of the homology groups, or Betti numbers, captures the number of independent components of different dimensions that are in the space. The first Betti number counts the number of loops in a space.

When working with finite metric spaces (X, ∂) one can define a set of related simplicial complexes in the following fashion. At each scale ε one can define a simplicial complex (for instance, the Vietoris-Rips complex $VR_\varepsilon(X, \partial)$). At a bigger scale $\varepsilon' > \varepsilon$ one can define a different complex that will include the scale at complex $VR_\varepsilon(X, \partial) \hookrightarrow VR_{\varepsilon'}(X, \partial)$. This allows to track the different chains of objects at different scales and to identify the relevant scales at which different homology classes are present. A summary of the different scales at which homology classes are present is captured by the notion of a barcode: a multiset of non-empty intervals of the form either $[a_i, b_i) \subset \mathbb{R}$ or $[a_i, \infty)$, representing the scales when the homology class h_i is present. Notice that the zero dimensional homology shows how different clusters are joined as the scale ε increases, and it is formally equivalent to single linkage clustering. The presence of a barcode $[a, b)$ in one dimensional persistent homology, on the other hand, shows the scales $\varepsilon \in [a, b)$ at which there is a loop in the data.

As we show in the manuscript, both simplicial and persistent homology are useful tools to data mine the results from the analysis. We encourage the interested reader to read about the topic in introductory textbooks for simplicial homology⁷ and textbooks and reviews on persistent homology^{8,9}.

References

1. Poincare, H. Cinquième complément à l'analysis situs. *Palermo Rend.* **18**, 45-110 (1904).
2. Adelson-Velskii, G. & Kronrod, A. in Dokl. Akad. Nauk SSSR, Vol. 49 239-241 (1945).
3. Reeb, G. Sur les points singuliers d'une forme de Pfaff complètement intégrable ou d'une fonction numérique. *CR Acad. Sci. Paris* **222**, 2 (1946).
4. Singh, G., Mémoli, F. & Carlsson, G.E. in SPBG 91-100 (Citeseer, 2007).
5. Munch, E. & Wang, B. Convergence between categorical representations of Reeb space and mapper. *arXiv preprint arXiv:1512.04108* (2015).
6. Weinberger, S. The complexity of some topological inference problems. *Foundations of Computational Mathematics* **14**, 1277-1285 (2014).
7. Hatcher, A. Algebraic topology. *Cambridge UP, Cambridge* **606** (2002).
8. Carlsson, G. Topology and data. *Bulletin of the American Mathematical Society* **46**, 255-308 (2009).
9. Edelsbrunner, H. & Harer, J. Computational topology: an introduction. (American Mathematical Soc., 2010).

Supplementary Note 2

Single cell library generation

In one biological replicate, we sorted a small sample size from a differentiation, sequencing 80 cells per differentiation time-point utilizing standard CEL-Seq primers with anchor bases at the 3' end of reverse transcription primers, pooling 40 cells at a time prior to *in vitro* transcription (IVT). To assess library saturation and capture efficiency, two single cell libraries from each differentiation time point (consisting of 40 cells each) from the experiment 1 were paired end sequenced (2x125 bps) on an Illumina HiSeq 2500, operating in high output mode, sequencing

with Illumina v4 chemistry. To increase capture efficiency, enhanced *in vitro* transcription based amplification, and leveraging the library saturation curves from experiment 1, we utilized 96 barcoded CEL-Seq RT primers (**Supplementary Table 4**), forgoing the usage of anchor bases at the 3' terminus. We then conducted a differentiation on a second biological replicate, sampling 384 cells per time-point (inclusive of 96 FACS purified mid-level GFP expressing, and 288 high GFP expressing cells), collected into 96 well plates and implemented CEL-Seq, now pooling 96 cells per IVT reaction. Following IVT, aRNA was fragmented using magnesium (NEBNext Magnesium RNA Fragmentation Module) for 90 seconds and column purified (Zymo Research RNA Clean & Concentrator-5). Purified aRNA was then subjected to treatment with Antarctic Phosphatase and T4 polynucleotide kinase. Ligation of Illumina RA3 adapters was conducted using truncated T4 RNA Ligase 2 for 1 hour at 28 C. Following adapter ligation, adapter ligated aRNA was reverse transcribed using Illumina RTP at 50 C for 1 hour and placed on ice. To avoid amplification based batch effects, the resultant cDNA was PCR amplified with Illumina RPIX primers to no more than 15 cycles. The sequencing libraries were then twice purified using AmpureXP beads, held at a ratio of 1:0.65, yielding size selected libraries with an insert size of ~250 bps. The single cell libraries were then multiplexed to a total representation of 384 cells per lane at equimolar concentrations and mixed with 50% exome libraries generated using an Illumina TruSeq Exome Kit.