

Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development

Abbas H Rizvi^{1,2,6}, Pablo G Camara^{3,4,6}, Elena K Kandror^{1,2}, Thomas J Roberts^{1,2,4}, Ira Schieren^{2,5}, Tom Maniatis^{1,2} & Raul Rabidan^{3,4}

Transcriptional programs control cellular lineage commitment and differentiation during development. Understanding of cell fate has been advanced by studying single-cell RNA-sequencing (RNA-seq) but is limited by the assumptions of current analytic methods regarding the structure of data. We present single-cell topological data analysis (scTDA), an algorithm for topology-based computational analyses to study temporal, unbiased transcriptional regulation. Unlike other methods, scTDA is a nonlinear, model-independent, unsupervised statistical framework that can characterize transient cellular states. We applied scTDA to the analysis of murine embryonic stem cell (mESC) differentiation *in vitro* in response to inducers of motor neuron differentiation. scTDA resolved asynchrony and continuity in cellular identity over time and identified four transient states (pluripotent, precursor, progenitor, and fully differentiated cells) based on changes in stage-dependent combinations of transcription factors, RNA-binding proteins, and long noncoding RNAs (lncRNAs). scTDA can be applied to study asynchronous cellular responses to either developmental cues or environmental perturbations.

The differentiation of motor neurons from neuroepithelial cells in the vertebrate embryonic spinal cord is a well-characterized example of cellular lineage commitment and terminal cellular differentiation¹. Neural precursor cells differentiate in response to spatiotemporally regulated morphogen gradients that are generated in the neural tube by activation of a cascade of specific transcriptional programs¹. Detailed understanding of this process has been hindered by the inability to isolate and purify sufficient quantities of synchronized cellular subpopulations from the developing murine spinal cord. Although *in vitro* approaches have been used to study both the mechanisms of motor neuron differentiation² and motor neuron disease^{3,4}, these approaches are limited by the fact that embryoid bodies (EBs) are differentially exposed to inductive ligands and the uncharacterized paracrine signaling within EBs. This leads to the generation of heterogeneous populations of differentiated cell types⁵. Motor neuron disease mechanisms are thus currently studied in a heterogeneous background of cell types whose contributions to pathogenesis are unknown. Methods to analyze the transcriptome of individual differentiating motor neurons could provide fundamental insights into the molecular basis of neurogenesis and motor neuron disease mechanisms.

Single-cell RNA-sequencing carried out over time enables the dissection of transcriptional programs during cellular differentiation of individual cells, thereby capturing heterogeneous cellular responses to developmental induction. Several algorithms for the analysis of single-cell RNA-seq data from developmental processes have been

published, including Diffusion Pseudotime⁶, Wishbone⁷, SLICER⁸, Destiny⁹, Monocle¹⁰, and SCUBA¹¹ (**Supplementary Table 1**). All of these methods can be used to order cells according to their expression profiles, and they enable the identification of lineage branching events. However, Destiny⁹ lacks an unsupervised framework for determining the transcriptional events that are statistically associated with each stage of the differentiation process, and the statistical framework of Diffusion Pseudotime, Wishbone, Monocle, and SCUBA is biased, for example by assuming a differentiation process with exactly one branch event^{6,7} or a tree-like structure^{10,11}. Although these methods can reveal the lineage structure when the biological process fits the assumptions, an unsupervised method would be expected to have the advantage of extracting more complex relationships, such as the presence of multiple independent lineages or convergent lineages or the coupling of cell cycle to lineage commitment. Moreover, apart from SCUBA, these methods do not exploit the temporal information available in longitudinal single-cell RNA-seq experiments, and they require the user to explicitly specify the least differentiated state^{6–10}.

We present an unbiased, unsupervised, statistically robust mathematical approach to single-cell RNA-seq data analysis that addresses these limitations. Topological data analysis (TDA) is a mathematical approach used to study the continuous structure of high-dimensional data sets. TDA has been used to study viral reassortment¹², human recombination^{13,14}, cancer¹⁵, and other complex genetic diseases¹⁶. scTDA is applied to study time-dependent gene expression using longitudinal single-cell RNA-seq data. Our scTDA method is a statistical

¹Department of Biochemistry and Molecular Biophysics, Columbia University Medical Center, New York, New York, USA. ²The Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, New York, USA. ³Department of Systems Biology, Columbia University Medical Center, New York, New York, USA.

⁴Department of Biomedical Informatics, Columbia University Medical Center, New York, New York, USA. ⁵Howard Hughes Medical Institute, Columbia University Medical Center, New York, New York, USA. ⁶These authors contributed equally to this work. Correspondence should be addressed to R.R. (rr2579@cumc.columbia.edu) or T.M. (tm2472@columbia.edu).

Received 25 September 2016; accepted 20 March 2017; published online 1 May 2017; doi:10.1038/nbt.3854

framework for the detection of transient cellular populations and their transcriptional repertoires and does not assume a tree-like structure for the expression space or a specific number of branching points. scTDA can be used to assess the significance of topological features of the expression space, such as loops or holes. In addition, it exploits temporal experimental information when available, inferring the least differentiated state from the data.

Here we apply scTDA to analyze the transcriptional programs that regulate developmental decisions as mESCs transition from pluripotency to fully differentiated motor neurons and concomitant cell types.

RESULTS

Overview of scTDA

Single-cell gene expression can be represented as a sparse high-dimensional point cloud, with the number of dimensions equivalent to the number of expressed genes (~10,000). Extracting biological information from such data requires a reduction in the dimensionality of the space. Widely used algorithms, such as multidimensional scaling (MDS), independent component analysis (ICA), and *t*-distributed stochastic neighbor embedding (*t*-SNE), have been adapted to flow cytometry, mass spectrometry^{11,17}, and single-cell RNA-seq measurements^{10,18}. These strategies, however, are not designed to preserve continuous relationships in high dimensions. **Figure 1a** represents a simple example of a one-dimensional continuous manifold (a circle) twisted in three dimensions. Reduction to two dimensions using these algorithms introduces artifacts in the low-dimensional representation, including artifactual intersections (in MDS and ICA) and tears apart the original continuous structure (in *t*-SNE). As cell differentiation is a continuous asynchronous process, we reasoned that longitudinal single-cell data would be best analyzed using a mathematical approach that accounts for continuous structures.

We developed a computational approach to the analysis of longitudinal single-cell RNA-seq database on the TDA algorithm Mapper¹⁹ (**Supplementary Note 1**). Mapper builds upon any given dimensional reduction algorithm, such as MDS, and produces a low-dimensional topological representation of the data that preserves locality. To that end, the projection obtained by the dimensional-reduction algorithm is covered with overlapping bins, and clustering of the data within each bin is performed in the original high-dimensional space (**Fig. 1b**). A network is constructed by assigning a node to each cluster, and clusters that share one or more cells are connected by an edge. The result is a low-dimensional network representation of the data in which nodes represent sets of cells with similar global transcriptional profiles, and edges connect nodes that have at least one cell in common (**Fig. 1b**). An important feature of these networks is that elements that are connected in the low-dimensional representation lie near each other in the original high-dimensional expression space, contrary to what occurs with most of the dimensional-reduction algorithms currently in use, including principal component analysis (PCA), ICA, and MDS (**Fig. 1a**). The scTDA approach can be applied even if there are complex structures in the high-dimensional expression space, including non-tree-like trajectories (**Fig. 1a**). Additionally, since nodes represent clusters of cells, the approach is highly scalable to large data sets.

We adapted topological representations so they could be used to analyze longitudinal single-cell RNA-seq data, introducing the necessary mathematical concepts and statistics (**Supplementary Fig. 1**). To identify expression programs associated with different parts of the topological representation, without predefining any cellular population, we developed the concept of gene connectivity. The connectivity of a gene acquires a significant value when cells that

share similar global transcriptional profiles express that particular gene more than random (**Fig. 1c** and Online Methods). Genes with a significant connectivity are expected to be involved in a particular stage (or stages) of differentiation. To identify the least differentiated state in the topological representation, we introduced a root node, which is the node that maximizes correlation between sampling time and graph distance. Using this root node as a reference, we computed the centroid (**Fig. 1d**) and dispersion (standard deviation) of genes in the topological representation (Online Methods). Genes with low centroids are upregulated in stem-like cells, whereas genes with high centroids are specific to fully differentiated cells.

To identify transient cellular subpopulations arising during differentiation, we clustered low-dispersion genes with significant connectivity in the topological representation based on their centroid (Davies–Bouldin criterion, **Fig. 1e**). Finally, to assess the topological features of the representation, such as loops and holes, we used persistent homology, a framework within TDA for deriving and classifying topological features associated with data (Online Methods, **Supplementary Note 1**). Further details of the scTDA method and its mathematical foundations can be found in the Online Methods and **Supplementary Note 1**. Code implementing the scTDA method is provided as **Supplementary Code**.

scTDA for temporal ordering of differentiating cells

We assessed the capacity of scTDA to order asynchronously differentiating cells using a controlled setting where the truth is known. First, we simulated a noisy, branched, asynchronous cellular differentiation process with two branching points (**Fig. 2a** and Online Methods). 700 cells were sampled at three time points, and the expression levels of 500 genes were simulated. From this data, scTDA correctly reconstructed the topological structure of the differentiation tree and identified the most stem-like state (**Fig. 2b**). In contrast, the algorithms Diffusion Pseudotime⁶, Wishbone⁷, and SLICER⁸ failed to correctly assign some of the branches in our hands, producing artifacts in the inferred pseudotemporal ordering of the cells (**Fig. 2c–e**), even if the most stem-like state was provided by the user of these algorithms.

To quantify the performance of scTDA, we extended the above simulations to processes with one, two, or three branching points (Online Methods). scTDA showed higher correlation between the inferred pseudotime and the actual simulated differentiation time compared to the other three methods, thus reconstructing the underlying differentiation process more accurately from the data (**Fig. 2f**). The improvement was particularly large when dropout events were included in the simulation (**Fig. 2g**), where the graphical representations produced by Diffusion Pseudotime, SLICER, and Wishbone were often unable to correctly infer the structure of the differentiation tree (**Supplementary Fig. 2**).

Motor neuron differentiation analysis with scTDA

We applied scTDA to longitudinal single-cell RNA-seq data from *in vitro* motor neuron differentiation experiments. We differentiated mESCs into motor neurons using a well-established protocol². The mESC line expresses enhanced GFP (eGFP) under the control of the early-motor-neuron-specific promoter *Mnx1*, allowing identification of committed motor neurons. Using a modified CEL-seq approach²⁰ (Online Methods and **Supplementary Note 2**), we generated cDNA libraries from individual cells, spanning the transition from pluripotency to postmitotic commitment. To assess reproducibility, we performed two separate differentiations. In the first differentiation (experiment 1) we sequenced 440 cells sampled throughout the entire

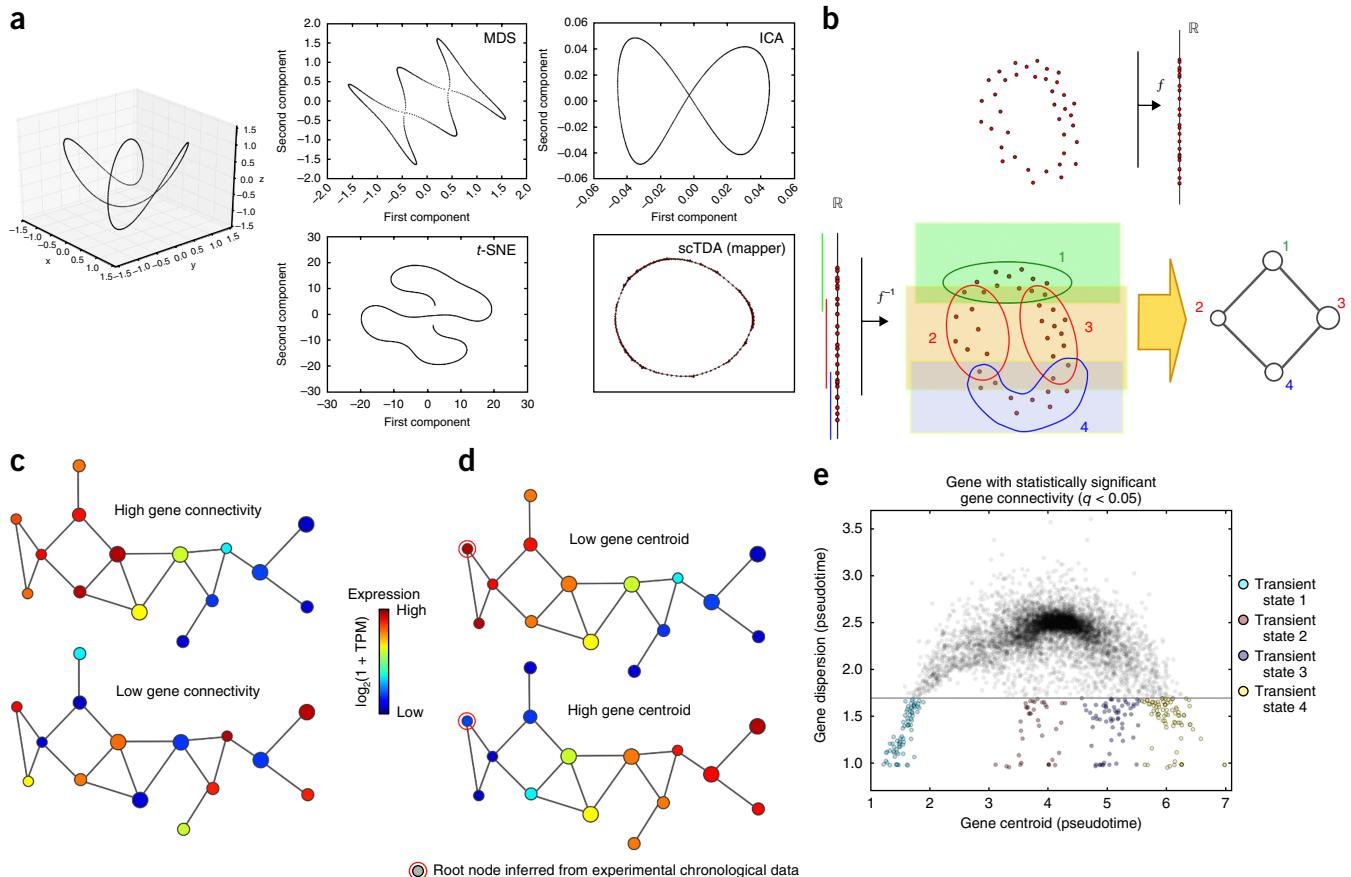


Figure 1 Topological analysis of longitudinal single-cell RNA-seq data. **(a)** Comparison of methods for reducing dataset dimensionality. A toy example is shown, illustrating the artifacts that can emerge when standard dimensional-reduction methods are used to represent differentiation trajectories. A total of 1,000 points are sampled from a twisted circle in 3D space. MDS, ICA, t-SNE, and Mapper were used to represent the above points in two dimensions. Of these methods, only Mapper can capture the continuous circular trajectory of the 3D space without introducing artificial intersections or disrupting the trajectory. **(b)** Mapper algorithm. Top: Mapper builds upon dimensional-reduction function f mapping the high-dimensional single-cell RNA-seq point-cloud data into \mathbb{R}^k (for simplicity we take $k = 1$ in this figure). Bottom: under the inverse function f^{-1} , a covering of \mathbb{R}^k maps into a covering of the single-cell point-cloud data. Clustering is performed independently in each of the induced patches in the high-dimensional space. In the low-dimensional representation, a node is assigned to each cluster of cells. If two clusters intersect, the corresponding nodes are connected by an edge. Topological features in the low-dimensional representation are guaranteed to also be present in the original high-dimensional RNA-seq space. **(c)** Gene connectivity. Gene connectivity allows the identification of genes that are differentially expressed by a cellular subpopulation of the differentiation process, without predefining any cellular subpopulation. Represented is a toy example of two genes with very different gene connectivity on the topological representation. Top: An example of a gene with high gene connectivity in the topological representation. This signifies that there is a set of cells with similar global expression profiles and high expression levels of the gene. Bottom: An example of a gene with low gene connectivity in the topological representation. **(d)** Gene centroid. The centroid of a gene, measured in pseudotime, quantifies where the expression of a gene sits in the topological representation with respect to the root node. The root node represents the least differentiated cellular state and is determined from the experimental sampling times. A toy example of two genes with very different centroid can be used to illustrate the concept. Top: a gene with a low value for the expression centroid mostly associated with pluripotent cells. Bottom: a gene with a high value for the centroid, mostly associated with differentiated cells. **(e)** Transient cellular states. Transient cellular states can be identified in an unsupervised manner by clustering low-dispersion genes with significant gene connectivity according to their centroid in the topological representation. In the figure, an example of distribution of centroids and dispersions for genes with significant gene connectivity is shown. Four clusters of low-dispersion genes are identified, which correspond to four transient cellular states arising throughout the differentiation process.

course of motor neuron differentiation. Cells were concomitantly sampled in bulk at each time point. Using this data, we assessed library saturation and optimal depth of sequencing coverage. In the second differentiation (experiment 2), we sought to increase statistical power and sequenced 2,304 individual cells with the depth of sequencing coverage near the saturation point (~250,000 reads per cell). In both experiments, cells were sampled each day throughout the entire course of differentiation, from day 2 to day 6. In total, we detected 15,609 and 19,009 transcripts, respectively, across experiments 1 and 2. Given the improved statistical power associated with an increased number of

cells, we focused our attention on the results from experiment 2 and assessed the reproducibility of the results using experiment 1.

Cell populations and surface markers of differentiation

We used scTDA to analyze the two longitudinal motor neuron differentiation single-cell RNA-seq data sets. We filtered the data for cells that passed stringent quality-control tests (Online Methods, Supplementary Fig. 3), retaining, respectively, 373 and 1,964 individual cells in experiments 1 and 2. An scTDA analysis recapitulated chronological order based on mRNA-expression levels alone (Fig. 3a,b

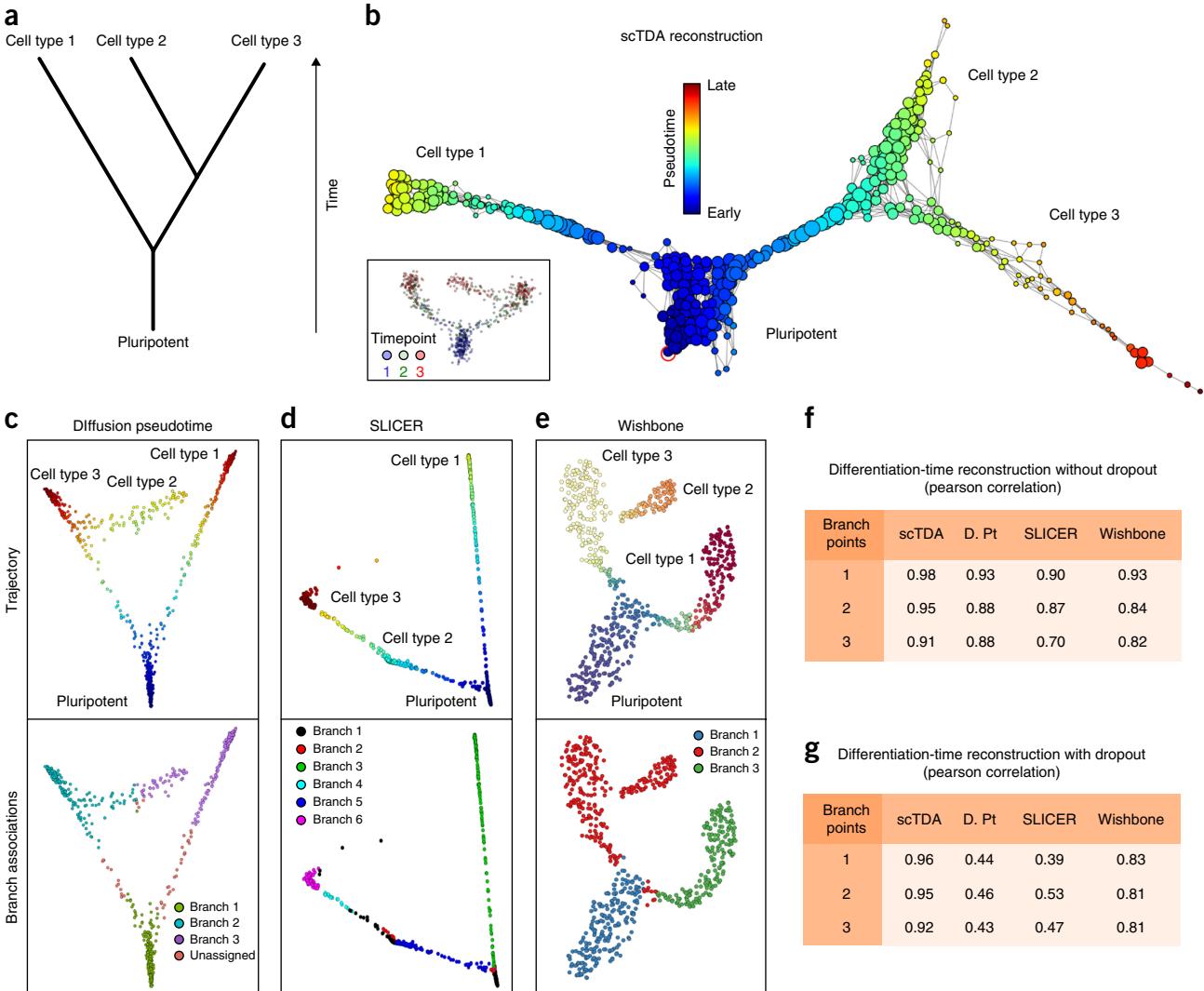


Figure 2 Comparison of algorithms for ordering cellular states. (a) A noisy, branched, asynchronous cellular differentiation process was simulated. The differentiation tree of the process is represented. 700 cells were sampled from this process at three time points. Using these data, we attempted to reconstruct the structure of the differentiation tree with scTDA and compared the outcomes with reconstruction using Diffusion Pseudotime, SLICER and Wishbone, all of which rely on branch assignments for downstream statistical analysis. (b) Reconstructed differentiation trajectory using scTDA. scTDA recovered the structure of the simulated differentiation process and correctly rooted the tree using the experimental information. Nodes correspond to sets of cells sharing similar global transcriptional profiles, with the node sizes proportional to the number of cells in the node. Nodes that are connected by an edge have at least one cell in common. For reference, an inset with the latent MDS representation of the data, colored by sampling day, is also shown. (c) Reconstructed differentiation trajectory using Diffusion Pseudotime. Although the representation of the data using the first two diffusion coefficients reproduces the structure of the differentiation tree, the branch associations are not correctly assigned. (d) Reconstructed differentiation trajectory using SLICER. The representation constructed by SLICER using locally linear embedding was unable to capture the complete structure of the differentiation tree and branch assignments. (e) Reconstructed differentiation trajectory using Wishbone. The t-SNE representation of the data used by Wishbone reproduces the structure of the differentiation tree and correctly identifies the first branching point. However, Wishbone failed to identify the second branching point. (f,g) Pearson's correlation coefficient between the pseudotime, inferred from the scTDA data, Diffusion Pseudotime (D. Pt), SLICER, and Wishbone, and the actual simulated differentiation time. Cellular differentiation processes with one, two, or three branching points were considered, both in the absence (f) and in the presence (g) of dropout events.

and **Supplementary Figs. 4 and 5**) and simultaneously detected detailed transcriptional relationships between individual cells. We observed only a mild dependence of library complexity on the timescale of our differentiation (**Supplementary Fig. 6**, Online Methods), unlike in other experimental settings where differences in cell cycle give rise to substantial variation of library complexity^{21,22}. We did not observe large batch effects (**Supplementary Fig. 7**). Compared with analyses of the same data sets using PCA, MDS, t-SNE, Monocle, Wishbone, Slicer, and Diffusion Pseudotime, we

showed that scTDA had superior ability to identify the continuous chronological structure of the differentiation process (**Fig. 3c** and **Supplementary Figs. 8 and 9**).

Analysis of the data from experiment 2 identified 7,620 genes with significant gene connectivity (q value <0.05 , **Supplementary Fig. 10** and **Supplementary Table 2**), comprising 74% overlap with those identified from experiment 1 (P value $<10^{-100}$, Fisher exact test, **Fig. 3d**). It is likely that the large number of significant genes is due to the transcriptional heterogeneity of the data set, because the samples

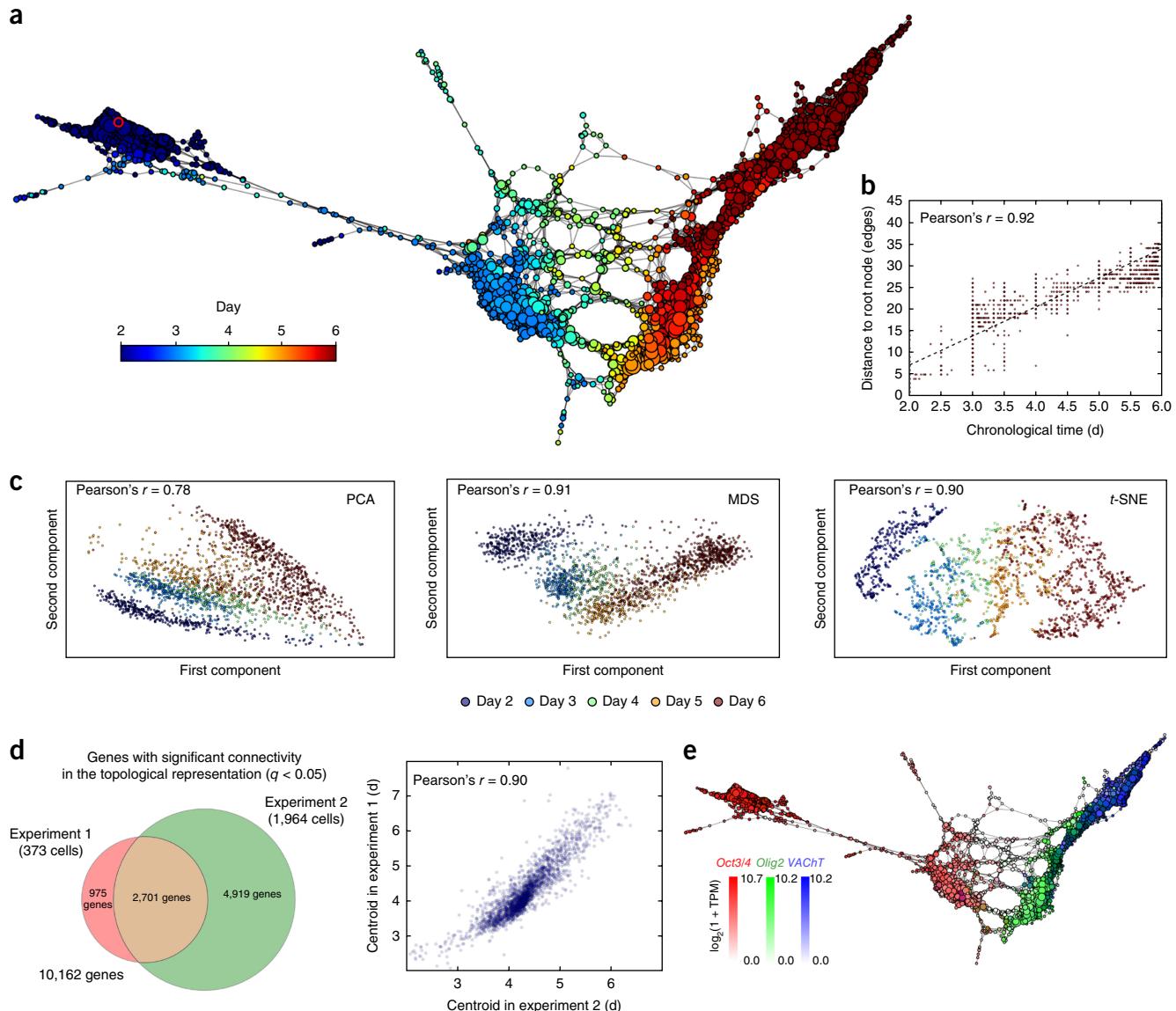


Figure 3 Differentiation of pluripotent mESCs into motor neurons (MN). (a) scTDA recapitulates chronological order based on expression data alone. The topological representation of the expression data of 1,964 single cells, sampled from the differentiation of mESCs into MNs, is labeled by sampling time. The root node, inferred from the experimental chronological information, is indicated with a red circle. (b) The distance of each node to the root node, 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) Comparison with standard dimensional-reduction algorithms. Dimensional reduction of the same expression data of 1,964 single cells, sampled from the differentiation of mESCs into MNs, using PCA, MDS, and t-SNE. The Pearson's correlation coefficient between the sampling time and the two-dimensional Euclidean distance to the root cell (defined as the one that maximizes this correlation) is indicated in each case. (d) Consistency between experiments 1 and 2. Left: Venn diagram of genes with significant gene connectivity ($q < 0.05$) in the topological representations of the two data sets. Both experiments are highly consistent in their calls (Fisher exact test P value $< 10^{-100}$). The number of significant genes is larger in experiment 2, consistent with its higher statistical power (due to the larger number of cells considered). Right: correlation between the centroid (expressed in pseudotime) of the $n = 2,701$ genes with significant ($q < 0.05$) gene connectivity in both topological representations. The distribution of gene centroids is highly consistent across the two experiments. (e) Known markers of pluripotent cells, motor neuron progenitors, and postmitotic neurons are successfully ordered. The topological representation is labeled by mRNA levels of *Oct3/4* (red), *Olig2* (green) and *VACHT* (blue).

of cells represent multiple developmental stages. The centroids of significant genes were consistent between the two data sets (Pearson correlation 0.9, P value $< 10^{-100}$, Fig. 3d). We validated our findings by identifying known markers of pluripotent cells, progenitor cells, and mature motor neurons in the data set over time (Fig. 3e and Supplementary Fig. 4c). We also found that cell-cycle genes were downregulated (as expected) in postmitotic neurons (Supplementary Fig. 11). We verified that the choice of parameters

for building the topological representation (number of genes, bins, and overlap between bins) did not affect the result of the analysis (Supplementary Fig. 12).

Interpretation of the distribution of centroids and the dispersion of genes (Supplementary Fig. 13) using scTDA revealed four transcriptionally distinct cellular populations arising during differentiation from mESCs to motor neurons (Fig. 4a). 488 of 7,620 genes were identified in the four expression groups

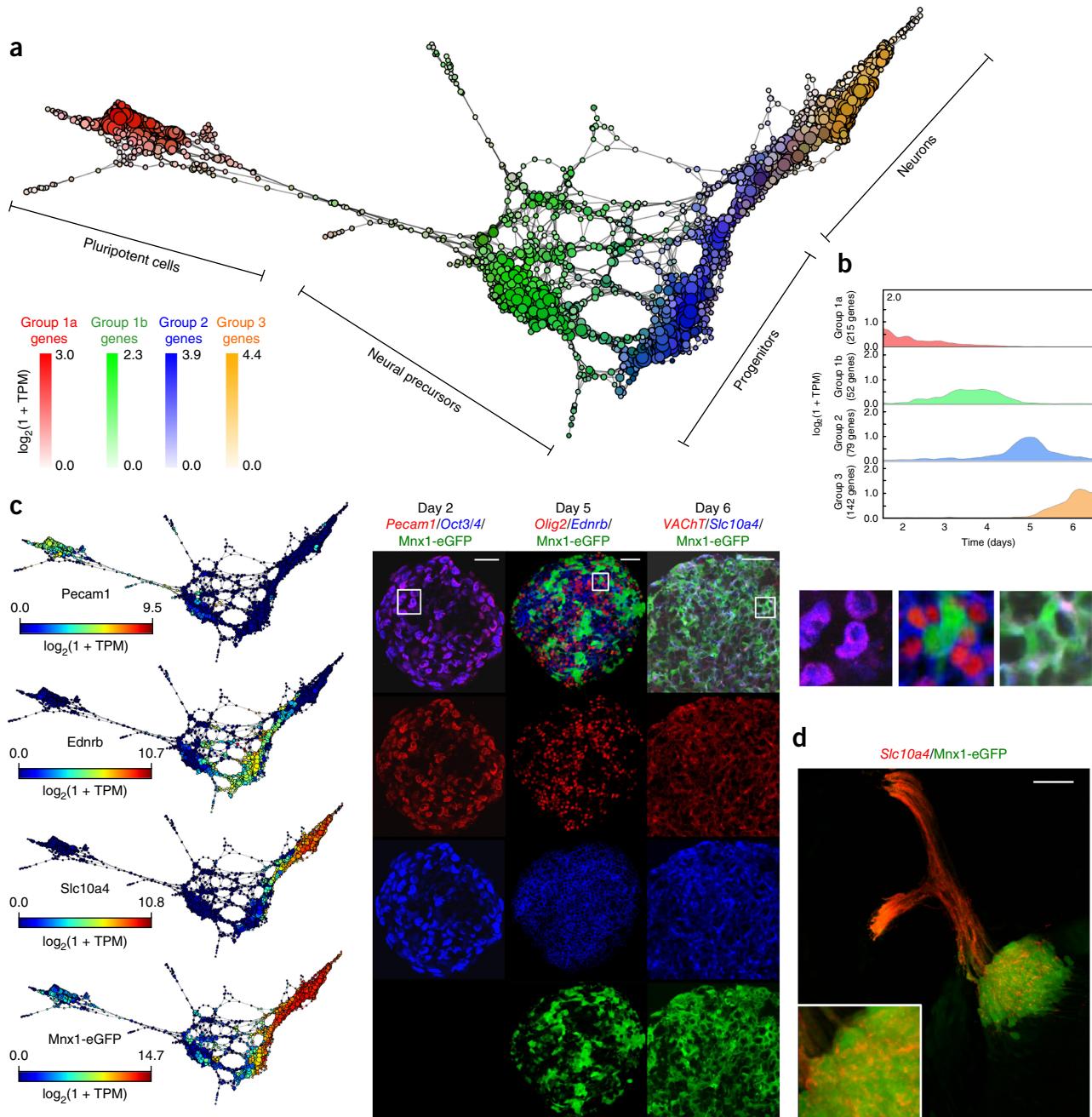


Figure 4 Cellular populations during motor neuron differentiation. (a) scTDA identifies four transient populations in mESC differentiation into MNs. Represented is the topological representation (colored by mRNA levels) of four groups of low-dispersion genes: pluripotent, precursor, progenitor, and postmitotic populations. In total, 488 genes were assigned to one of these four populations based on their expression profiles in the topological representation. TPM, transcripts per million. (b) Reconstructed expression timeline for each of the four groups of low-dispersion genes. (c) Validation by detection of state-specific cell-surface markers identified by scTDA. Left, topological representation (colored by mRNA levels) of surface proteins Pecam1, Ednrb, and Slc10a4; right, immunostaining of cultured EBs. Scale bar, 50 μm . Details of three regions are presented at the far right. For reference, the topological representation colored by mRNA levels of the Mnx1-eGFP reporter is also shown. (d) *In vivo* validation of the motor neuron surface marker *Slc10a4*. Spinal cord section from an E9.5 mouse immunostained for *Slc10a4* (red). The pool of motor neurons is also marked by Mnx1-eGFP expression (green). Scale bar, 50 μm .

(Fig. 4b and Supplementary Table 2). Groups 1a and 1b contain genes that are only expressed in early EBs (*Oct3/4*⁺ cells) and correspond to pluripotent and neural precursor states. Genes in groups 2 and 3 are only expressed in the progenitor (*Olig2*⁺ cells) and postmitotic ensembles (*VACHT*⁺ cells), respectively.

Ontology enrichment analyses using PANTHER²³ revealed developmental genes and genes related to DNA replication in groups 1a, 1b, and 2, whereas group 3 was enriched for genes related to axonogenesis, neuron migration, and regionalization (Supplementary Table 3), consistent with the underlying cellular differentiation process.

The analysis identified 74 transcripts in the four groups that encode proteins with an extracellular domain. These are therefore putative cell-surface-marker candidates that could be used to purify niche populations for further study (Supplementary Table 2). Several of these surface markers have not been reported as markers of neurogenesis. We selected three of these 74 markers that had high expression levels and specific antibodies available for experimental validation.

The three markers were validated using immunohistochemistry *in vitro* in EBs (Fig. 4c and Supplementary Fig. 14) and *in vivo* in the murine embryonic spinal cord (Fig. 4d).

scTDA identifies distinct proliferative states

scTDA analysis of our motor neuron differentiation data revealed numerous loops in the neural precursor population (Fig. 4a) that were

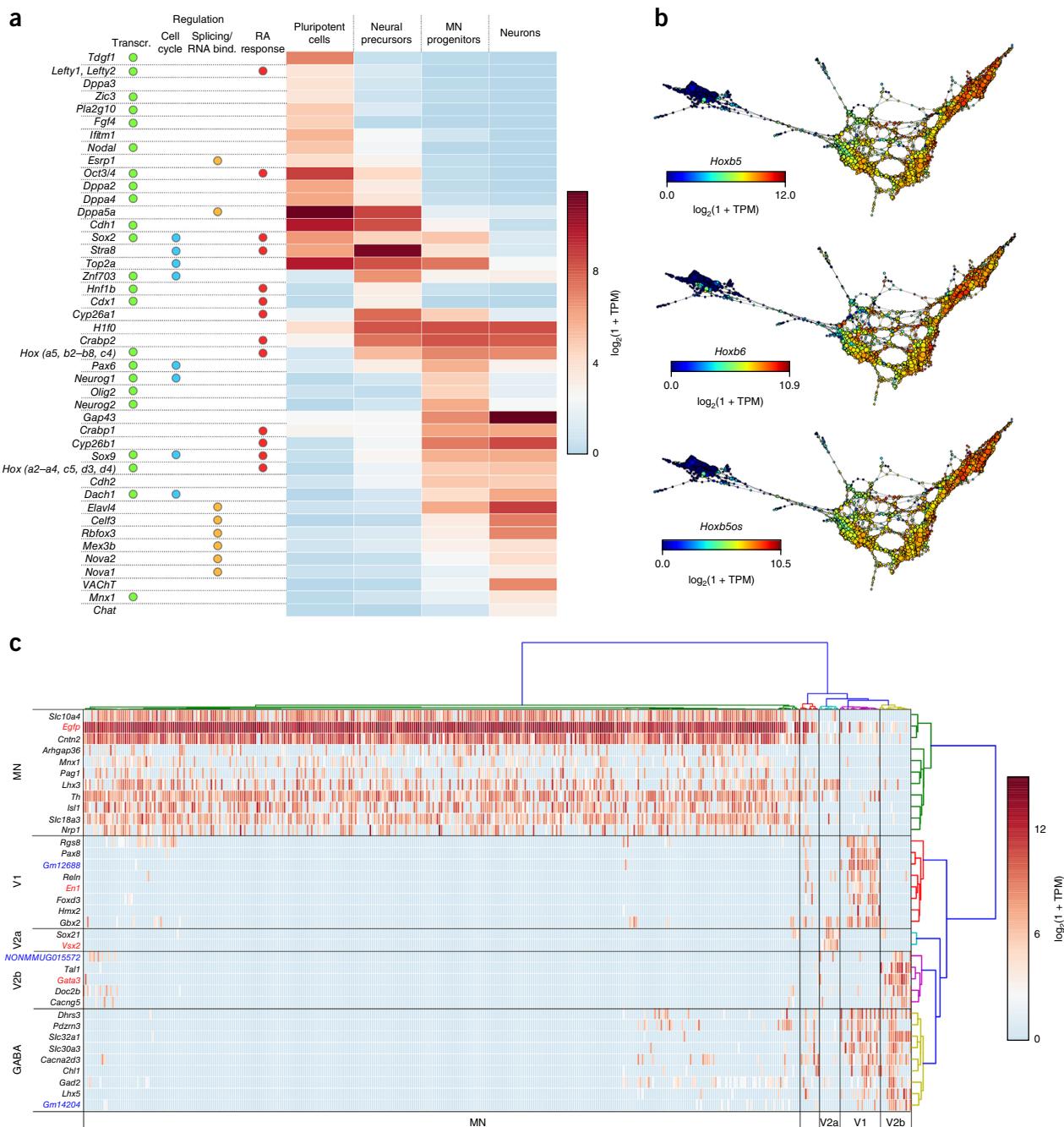


Figure 5 Differential gene expression in neurogenesis. (a) Differentially expressed regulators and downstream genes in the four cell populations are shown. Genes are annotated according to their role in transcriptional, cell cycle, RNA-binding-protein regulation, and RA response. (b) Topological representations (labeled by mRNA levels) of *Hoxb5*, *Hoxb6*, and the antisense lncRNA *Hoxb5os*, showing concordant expression of these transcripts during the generation of motor neurons from mESCs. (c) Postmitotic neuronal populations. Differentially expressed genes between *Vsx2*⁺, *Gata3*⁺, and *En1*⁺ cells that were marked as postmitotic neurons by the scTDA analysis. Hierarchical clustering of cells leads to five groups, four of which correspond to MNs, and V1, V2a, and V2b interneurons. Hierarchical clustering produces four groups of genes uniquely expressed by each of the above cell types and a fifth group associated with GABAergic neurons. lncRNAs are marked in blue.

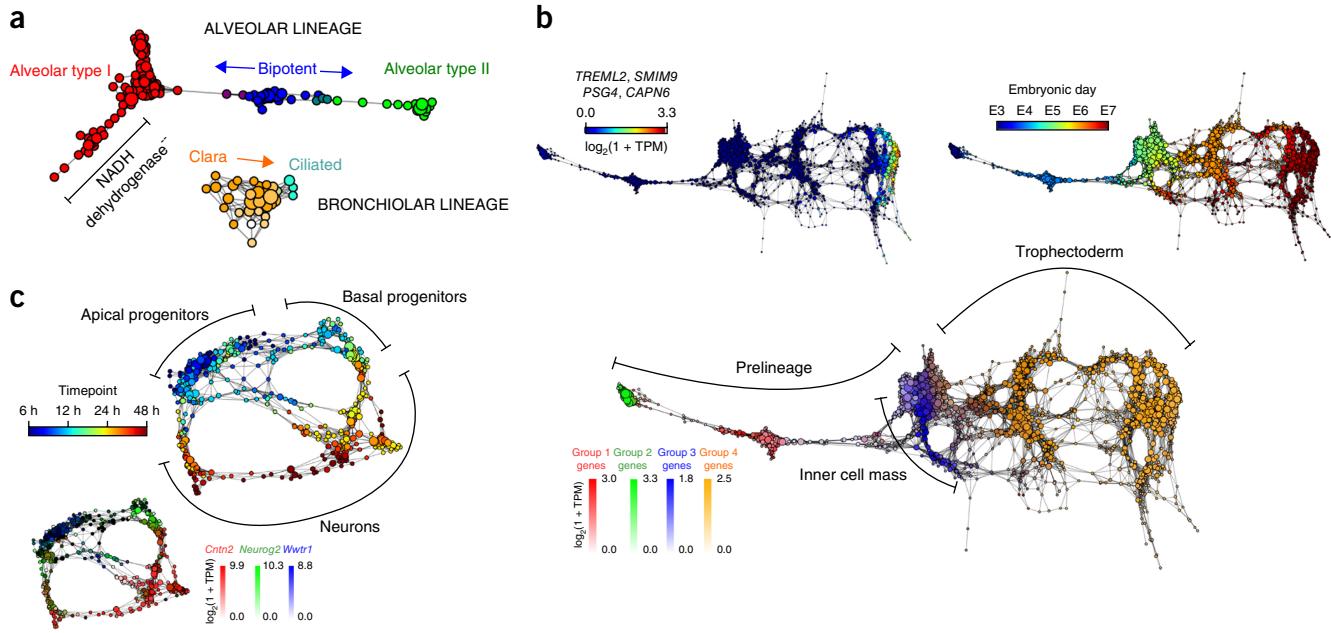


Figure 6 scTDA analysis of mouse and human developmental data sets. **(a)** Topological representation of 80 embryonic (E18.5) mouse lung epithelial cells³⁵ labeled according to cell type. scTDA resolves the alveolar and bronchiolar lineages that were originally identified using PCA, and identifies a putative set of cells with low NADH dehydrogenase expression that were not identified in the original analysis. **(b)** Topological representation of 1,529 individual cells from 88 human pre-implantation embryos³⁶. Top left and bottom: topological representation (labeled by expression levels of genes) associated with cellular populations identified during differentiation. scTDA resolved, without supervision, the segregation of the trophectoderm and inner-cell mass from prelineage cells (bottom) as well as a polar trophectoderm (top left) that were originally identified using a supervised analysis based on PCA and *k*-medoid clustering. Top right: topological representation labeled by embryonic day. **(c)** Topological representation of 272 newborn neurons from the mouse neocortex³⁷ labeled by sampling time after mitosis (top) and expression levels of *Cntr2*, *Neurog2*, and *Wwtr1* (bottom). scTDA recapitulated the converging developmental relations between apical and basal progenitors and neurons that were originally identified using hierarchical clustering and additionally found lineage convergence.

not present in the mESCs or the motor neurons. These loops were also observed in the scTDA analysis of experiment 1 (**Supplementary Fig. 4**). Neural precursor cells proliferate rapidly upon induction with retinoic acid (RA). This is in keeping with biological expectations, given that the cell cycle will give rise to periodic structures in the expression space. To evaluate whether the loops in the topological representation of neural precursors are caused by differences in the cell cycle stage, we built a topological representation using only cell-cycle genes (**Supplementary Fig. 15a**, Online Methods). The cell-cycle topological representation contained substantially larger loops in the same region (**Supplementary Fig. 15a**), separating *Stra8*-upregulated neural precursors and progenitors into proliferative and nonproliferative populations according to the expression of *Mki67* and other proliferation markers (**Supplementary Figs. 15b,c**). We used persistent homology to statistically assess the significance of these loops in the topological representation, asking whether loops of similar size could arise from noise effects (Online Methods). This analysis showed a strong statistical significance for some of the larger loops (q value < 0.05 , **Supplementary Fig. 15d**), consistent with a biological, rather than technical, origin for these features.

Characterization of developmental transitions

We next used scTDA to characterize the developmental transitions that occur during the differentiation of motor neurons. These results are available through an online database (Online Methods). We used scTDA to identify transcriptional programs that are associated with the transitions between the four transient cellular populations we identified (**Fig. 5a**). These findings were consistent across both experiments.

First, the transition from a pluripotent cell to a neural precursor population is characterized by the transcriptional dynamics of pathways involved in RA signaling and downstream effector proteins^{24,25} (**Fig. 5a**). Our analysis resolved, with unprecedented resolution, the timing of the transcriptional events, identifying upregulation of *Stra8* and downregulation of *Fgf4* as some of the earliest events that mark the transition (**Supplementary Fig. 16**). Subsequently, there is transcriptional activation of a subset of the homeobox gene family, including *Hoxa1* and *Hoxb2–Hoxb8*, which continue to be expressed during later stages of the differentiation process, and *Hoxb1*, which is transiently expressed along with caudalizing transcription factors (**Fig. 5a,b** and **Supplementary Fig. 17**).

A second wave of RA-inducible gene activation was identified during the formation of neural progenitors. This is accompanied by transcriptional up- and downregulation of *Stra8* and activation of *Crabp1* and a second set of homeobox genes: *Hoxa2*, *Hoxa3*, *Hoxc5*, *Hoxd3* and *Hoxd4* (**Fig. 5a**). This pattern of Hox-gene activation suggests that the linear chromosomal arrangement of the Hox-gene clusters does not necessitate temporal co-linearity in anterior Hox-gene expression, a phenomenon that has been reported in the developing spinal cord²⁶. Both waves of homeobox gene expression were accompanied by the upregulation of several lncRNAs derived from the antisense RNA strand (**Fig. 5b**, **Supplementary Fig. 17**, and **Supplementary Table 4**), consistent with previously identified lncRNA-based regulation of homeobox gene clusters^{27–30}.

The transition to mature neurons was marked by exit from the cell cycle and by postmitotic differentiation. In keeping with expectations from previous data³¹, scTDA identified *Neurog1* and *Neurog2* (modulators of neuronal specification, cyclin regulation, and cell cycle

exit) as well as *Ascl1* and *Sox9*, four genes that are among the known mediators of neuronal commitment, concomitant with a marked repression of topoisomerase 2A in the neuron population (**Fig. 5a**). We observed variability in eGFP expression levels and fluorescence intensities among differentiated neurons, which might indicate heterogeneity in later stages of differentiation. We performed differential expression analysis of *En1⁺*, *Gata3⁺*, *Vsx2⁺*, and *Egfp⁺* cells within this cellular population (Online Methods) and classified differentiated neurons into motor neurons ($n = 343$), and V1 ($n = 19$, *En1⁺*), V2a ($n = 10$, *Vsx2⁺*), and V2b interneurons ($n = 15$, *Gata3⁺*) (**Fig. 5c**), confirming the presence of cellular heterogeneity.

Interestingly, we observed the presence of several lncRNAs among the differentially expressed genes. V1 interneurons specifically express *Gm12688*, an intergenic lncRNA located near *Foxd3* and transcribed from the opposite strand. We validated this finding using qPCR (**Supplementary Fig. 18**). Similarly, we identified a lncRNA on chromosome 15 (NONCODEv4 accession number NONMMUG015572) that was expressed only in V2b interneurons. *Gm14204*, an intergenic lncRNA located near *Slc32a1* and transcribed from the opposite strand, was only expressed in V1 and V2b GABAergic interneurons. These results may suggest a role for lncRNAs in neural diversification.

Among the genes identified by scTDA as associated with developmental transitions were 20 genes encoding RNA-binding proteins (**Supplementary Table 2**). These include known developmental-state-dependent pre-mRNA splicing factors, as well as stage-specific but uncharacterized RNA-binding proteins, which may guide cellular differentiation and postmitotic commitment. In the context of the progenitor-to-neuron transition, our analysis identified *Nova1*, *Nova2*, *Rbfox3*, *Srrm4*, and the Elav-like transcripts *Elavl4* and *Celf3* (**Fig. 5a**), consistent with previous data^{32,33}. scTDA further revealed upregulation of *Mex3b* in the progenitor and postmitotic cell populations and constitutive expression of *Ptbp1* and *Ptbp2*. Previously published studies have documented *Srrm4*-directed inclusion of neural-specific exon 10 in *Ptbp1* (ref. 34). Our results may indicate a transcriptional switch in splicing-factor regulation that culminates in neural-specific splicing.

scTDA of heterogeneous cellular responses

Our analyses of mESC differentiation using scTDA showed that it can chronologically order asynchronous populations of single cells while simultaneously preserving high-dimensional relationships among their transcriptional programs. We next tested whether scTDA could be applied to other data sets to provide insights not obtained with other analytical approaches. We analyzed three different *in vivo* cellular-differentiation data sets^{35–37} (Online Methods). In each case, scTDA produced an accurate reconstruction of developmental trajectories that confirmed and extended the published observations.

First, we examined 80 single cells from differentiating distal-lung epithelium of mouse embryos³⁵ (**Supplementary Table 5**). All cells were sampled from the same time point (embryonic day (E) 18.5). As shown in **Figure 6a**, scTDA recapitulated the proposed relationships between differentiating cells in both the bronchiolar and alveolar lineages. Moreover, our analysis uncovered a set of alveolar type I cells in which genes associated with mitochondrial respiration are downregulated (**Supplementary Fig. 19**), which might indicate cellular stress or quiescence.

Second, we applied scTDA to 1,529 cells from human preimplantation embryos³⁶ (**Supplementary Table 6**). In this case, scTDA correctly ordered cells according to embryonic developmental time and identified the segregation of several lineages, including the inner-cell mass, the early trophectoderm, and a polar trophectoderm³⁶ (**Fig. 6b**).

In contrast with the published analysis, the identification and characterization of these cellular populations was completely unsupervised, highlighting the value of scTDA for characterizing cell differentiation processes, *a priori*.

Finally, we used scTDA to analyze 272 differentiating neurons in the mouse neocortex³⁷ (**Supplementary Table 7**). scTDA identified a continuum of cellular states with a bifurcation between apical and basal progenitors and neurons (**Fig. 6c**). The topological representation more accurately reflects a basal-to-apical progenitor migration and a split in potency, with apical progenitors sharing transcriptional profiles that more closely reflect their neuronal counterparts. scTDA faithfully represented nonlinear and converging cellular lineages. The convergence of distinct lineages cannot be identified by algorithms based on tree-like structures.

DISCUSSION

In order to preserve the continuity of cellular differentiation, account for asynchronous development, and enable statistical interpretation of patterns of gene activation, we used a method that exploits topological data analysis. Our method, named scTDA, enables unsupervised analyses of single-cell RNA-seq data sets over time.

We applied scTDA to study the *in vitro* differentiation of mESCs into neurons^{1,2} and showed that it accurately recapitulated cellular differentiation and provided the statistics needed to identify and characterize the transcriptional programs that accompany lineage commitment. We comprehensively characterized the dynamic appearance of mRNAs encoding signaling proteins, transcription factors, RNA splicing factors, and lncRNAs. These transcripts were dynamically regulated during the transition from pluripotent cells to neural precursors, progenitors, motor neurons, and interneurons, thus providing a valuable resource for studies of stem-cell differentiation and neurogenesis, and more generally, for any cellular differentiation process amenable to single-cell transcriptomic analysis. Inclusion of topological data enabled scTDA to identify additional processes coupled to differentiation, such as the cell cycle, which cannot be captured by tree-like structures. Furthermore, scTDA identified extensive transcriptional coregulation of thousands of coding and noncoding genes in precursor, progenitor, and neuronal populations. Some of these, such as the surface markers *Pecam1*, *Ednrb*, and *Slc10a4*, respectively, for the pluripotent, neural progenitor, and mature motor neuron populations, as well as the unique expression of the lncRNA *Gm12688* by V1 interneurons, were validated.

In principle, scTDA can be applied to any biological system responding to inductive cues or environmental perturbations (**Fig. 6**). For example, scTDA could be used to study cellular differentiation processes such as hematopoiesis, the evolution of cancer cells, neurodegeneration, or developmental disorders, all of which arise from extracellular signals and mutations that culminate in heterogeneous transcriptional responses and cellular behavior.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the [online version of the paper](#).

Note: Any Supplementary Information and Source Data files are available in the [online version of the paper](#).

ACKNOWLEDGMENTS

We thank T. Jessell, N. Francis, and H. Phatnani for critical reading of the manuscript. A.H.R. and T.M. thank the New York Genome Center and D. Goldstein for sequencing support, S. Morton for providing Engrailed antibody,

and P. Sims for experimental discussions. P.G.C. and R.R. thank A.J. Levine, G. Carlsson, F. Abate, I. Filip, S. Zairis, U. Rubin, and P. van Nieuwenhuizen for useful comments and discussions, O.T. Elliott for technical support with the online database, and Ayasdi Inc. for technical support. The work of P.G.C. and R.R. is supported by the NIH grants U54-CA193313-01 and R01GM117591. The work of A.H.R., E.K.K., T.J.R. and T.M. is supported by ALS Therapy Alliance grant ATA-2013-F-056 and NIH grant NS088992.

AUTHOR CONTRIBUTIONS

P.G.C. and R.R. developed the topology-based computational approach (scTDA) and applied it to single cell RNA sequencing data. A.H.R., E.K.K., and T.M. designed all experiments. A.H.R., E.K.K., and T.J.R. conducted experiments. I.S. conducted all flow cytometry. A.H.R., P.G.C., E.K.K., T.M., and R.R. analyzed the data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

15. Nicolau, M., Levine, A.J. & Carlsson, G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Natl. Acad. Sci. USA* **108**, 7265–7270 (2011).
16. Li, L. *et al.* Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Transl. Med.* **7**, 311ra174 (2015).
17. Bendall, S.C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).
18. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
19. Singh, G., Mémoli, F. & Carlsson, G.E. in SPBG 91–100 (Citeseer, 2007).
20. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
21. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
22. McDavid, A., Finak, G. & Gottardo, R. The contribution of cell cycle to heterogeneity in single-cell RNA-seq data. *Nat. Biotechnol.* **34**, 591–593 (2016).
23. Mi, H., Muruganujan, A., Casagrande, J.T. & Thomas, P.D. Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc.* **8**, 1551–1566 (2013).
24. Balmer, J.E. & Blomhoff, R. Gene expression regulation by retinoic acid. *J. Lipid Res.* **43**, 1773–1808 (2002).
25. Rhinn, M. & Dollé, P. Retinoic acid signalling during development. *Development* **139**, 843–858 (2012).
26. Gaunt, S.J. & Strachan, L. Temporal colinearity in expression of anterior Hox genes in developing chick embryos. *Dev. Dyn.* **207**, 270–280 (1996).
27. Zhang, X., Weissman, S.M. & Newburger, P.E. Long intergenic non-coding RNA HOTAIRM1 regulates cell cycle progression during myeloid maturation in NB4 human promyelocytic leukemia cells. *RNA Biol.* **11**, 777–787 (2014).
28. Lin, M. *et al.* RNA-Seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders. *PLoS One* **6**, e23356 (2011).
29. Mallo, M. & Alonso, C.R. The regulation of Hox gene expression during animal development. *Development* **140**, 3951–3963 (2013).
30. Dinger, M.E. *et al.* Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res.* **18**, 1433–1445 (2008).
31. Sommer, L., Ma, Q. & Anderson, D.J. neurogenins, a novel family of atonal-related bHLH transcription factors, are putative mammalian neuronal determination genes that reveal progenitor cell heterogeneity in the developing CNS and PNS. *Mol. Cell. Neurosci.* **8**, 221–241 (1996).
32. Darnell, R.B. RNA protein interaction in neurons. *Annu. Rev. Neurosci.* **36**, 243–270 (2013).
33. Quesnel-Vallières, M., Irimia, M., Cordes, S.P. & Blencowe, B.J. Essential roles for the splicing regulator nSR100/SRRM4 during nervous system development. *Genes Dev.* **29**, 746–759 (2015).
34. Calarco, J.A. *et al.* Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. *Cell* **138**, 898–910 (2009).
35. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
36. Petropoulos, S. *et al.* Single-cell RNA-Seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **165**, 1012–1026 (2016).
37. Telley, L. *et al.* Sequential transcriptional waves direct the differentiation of newborn neurons in the mouse neocortex. *Science* **351**, 1443–1446 (2016).

ONLINE METHODS

Cell culture and single-cell isolation. Murine embryonic stem cell (mESC)-based differentiations were performed using the method of Wichterle *et al.*². All cells were tested for mycoplasma contamination via PCR examination. In brief, stem cell colonies were expanded on an adherent substrate, after which they were dissociated and monodispersed in a serum-free suspension (day 0). Individual stem cells aggregated into embryoid bodies (EBs), which maintained exclusive expression of pluripotency markers until they were induced down the neuron lineage by the addition of RA and Smoothened Agonist (SAG) (day 2). Metabolized culture medium was replenished on day 5, coincident with the appearance of early eGFP-positive cells within the EBs. EBs were dissociated into single cells using the Worthington Biochemical Papain Dissociation System (LK003178). Single-cell deposition was accomplished using a Beckman Coulter MoFlo Astrios EQ cell sorter into 96 well plates. Cells were then snap frozen and subsequently lysed.

Single-cell library generation. To obtain single-cell expression profiles, modified CEL-Seq²⁰ was carried out as described in the **Supplementary Note 2**. Briefly, we reverse transcribed the mRNA in each cell lysate with barcoded primers (**Supplementary Table 8**), pooled the single-cell samples, synthesized second-strand cDNA, and then linearly amplified by *in vitro* transcription with T7 RNA polymerase. The amplified RNA (aRNA) was chemically fragmented, and T4 RNA ligase was used to ligate Illumina 3'-RNA adapters. The aRNA was run on an Agilent Bioanalyzer to assess proper fragmentation and then reverse transcribed to generate cDNA and subjected to PCR enrichment. The subsequent multiplexed samples were paired-end sequenced (2 × 125 bps) on an Illumina HiSeq 2500. Bulk RNA samples were purified from 2 × 10⁶ cells in 1 mL of Trizol and standard Qiaquick RNA-extraction protocols with an RIN 9.8 or higher. Stranded RNA-seq libraries were generated at the New York Genome Center using a TruSeq Stranded Total RNA Library Prep Kit. Stranded cDNA libraries were paired-end sequenced (2 × 125 bps) on a HiSeq 2500, operating in high output mode, yielding 30 M reads per indexed library.

Immunofluorescence. EBs were washed three times in ice cold PBS and fixed for two hours at room temperature in 4% PFA. They were then washed with ice cold PBS and either embedded into OCT and stored at -80 °C for sectioning or stored in PBS at 4 °C for whole EB staining. 20 µm sections were cut using a Leica CM 1950 cryotome and mounted onto a glass slide. Whole and sectioned EBs were incubated in blocking solution (0.1% Tween-20, 5% donkey serum, 1% BSA) for two hours at room temperature followed by primary antibodies diluted in blocking solution overnight.

All antibodies used have been validated with specific mouse reactivity for immunohistochemistry or flow cytometry as indicated by the commercial source. Primary antibodies used were: goat-α-Slc9a1 (Santa Cruz sc-16097, 1:100), rabbit-α-Olig2 (Millipore AB9610, 1:100), mouse-α-Olig2: clone 21F1.1 (Millipore MABN50, 1:100), rabbit-α-Ednrb (Novus NLS54, 1:200), guinea pig-αVACHT (gift from N. Schneider's lab, Department of Neurology, Columbia University, 1:400), rabbit-α-Slc10a4 (Novus Biologicals NBP1-81134, 1:100), mouse-α-Pou5f1 (BD Biosciences 611202, 1:100), goat-α-PECAM1 (Santa Cruz sc-1506, 1:100), sheep-α-CD44 (R&D Systems AF6127, 1:250), goat-α-Foxc1 (Santa Cruz sc-21396, 1:50), and goat-α-Sstr2 (Santa Cruz sc-11606, 1:50). Alexa Fluor-conjugated secondary antibodies from Life Technologies were used at a 1:1,000 dilution for two hours at room temperature. CD44, Foxc1, and Sstr2 were respectively conjugated to Cy-5, Alexa-405, and Cy-3 fluorophores using the DyLight system from Abcam (ab201798, ab188287, ab188288). Coverslips were mounted with Vectashield and EBs were imaged on an Olympus Fluoview FV1000 microscope using Olympus Fluoview v4.1.

Processing of single-cell RNA-seq data. Paired-end 125-bp reads were demultiplexed, trimmed, and mapped to the UCSC mouse reference (mm10) using Tophat³⁸. Gene expression was quantified using transcript read counts as derived from HTSeq³⁹. Read counts were normalized as

$$e_i = \log_2 \left(1 + \frac{10^6 \cdot r_i}{r} \right)$$

where r_i denotes the unambiguous read count for transcript i , and r denotes the total number of reads that are mapped to transcripts in the cell. A strategy based on spike-in read counts, as described by Stegle *et al.*⁴⁰, was implemented to filter out cells with a low content of mapped RNA and/or low sequencing depth. Specifically, the ratio ϵ_1 between average spike-in reads in the cell and average spike-in reads in the library (for spike-ins with an average of more than five reads in the library) was used to discard cells with low sequencing coverage. Similarly, the ratio ϵ_2 between the total number of spike-in reads in the cell and the total number of mapped reads was used to discard cells with a low number of mapped reads, relative to the sequencing coverage. Those cells showed very low expression of housekeeping genes, possibly representing cells under stressed conditions and/or with large amounts of degraded RNA. Based on the distribution of ϵ_1 and ϵ_2 in each of the two experiments (**Supplementary Fig. 2a**), cells with $\epsilon_1 > 5 \times 10^{-4}$ and $0.7 > \epsilon_2 > 0.01$ in experiment 1 and cells with $4.0 > \epsilon_1 > 0.05$ and $1.0 > \epsilon_2 > 0.01$ in experiment 2, were kept for subsequent analysis. The distribution of filtered-out cells across different libraries was uniform in both experiments (**Supplementary Fig. 2b**). To reduce the noise observed near detection threshold in experiment 1, read counts with $r_i < 5$ were set to zero. Additionally, one of the libraries (RPI36) was discarded from subsequent analysis because it presented a large-batch effect. To assess the dependence of the library complexity on the differentiation time (**Supplementary Fig. 3**), we computed at each time point the distribution of the geometric library size, defined as the sum of log-expression values over all genes in a cell²².

Topological representation. The algorithm Mapper¹⁹ (**Supplementary Note 1**) was used to build topological representations of the RNA-seq data through the implementation provided by Ayasdi Inc. Several open-source implementations of Mapper are available (<https://github.com/MLWave/kepler-mapper>, <http://danifold.net/mapper/>, <https://github.com/RabadanLab/sakmapper>, <https://github.com/paultpearson/TDAmapper>). In brief, the processed RNA-seq data was endowed with a dissimilarity matrix by taking pairwise correlation distance (1 - Pearson correlation). To minimize the effect of dropout events present in single-cell data, we only considered the 5,000 genes (for experiment 1) and the 4,600 genes (for experiment 2) with highest variance across each data set. These are highly expressed transcripts for which the probability of not being captured by the RNA amplification (dropout events) is small⁴¹⁻⁴⁴. The space was reduced to \mathbb{R}^2 using MDS, as displayed in **Figure 3c**. A covering of \mathbb{R}^2 consisting of 26 × 26 and 62 × 62 rectangular patches was considered for experiments 1 and 2, respectively. The size of the patches was chosen such that the number of cells in each row or column of patches was the same, avoiding sampling-density biases. The overlap between patches was 66% (on average). Single-linkage clustering was performed in each of the pre-images of the patches using the algorithm described in Singh *et al.*¹⁹. A network was constructed in which each vertex corresponds to a cluster, and edges correspond to nonvanishing intersections between clusters. We checked for the absence of batch effects in the topological representation by verifying that batches from the same time point had a substantial overlap in the representation (**Supplementary Figs. 3d and 7**) and the stability against different choices for the threshold for the number of genes used to compute the distance matrix and for the covering of \mathbb{R}^2 (**Supplementary Fig. 12**).

We verified that our analysis was reproducible using a freely available open-source implementation of Mapper. To that end, we repeated the analysis of experiment 1 using a topological representation built with SakMapper (<https://github.com/RabadanLab/sakmapper>) (**Supplementary Fig. 20**). Comparison with the results of the scTDA analysis presented in the main text shows a clear consistency between the two analyses. In particular, 78% of the genes with significant gene connectivity (q value < 0.05) in the representation generated with SakMapper also displayed significant gene connectivity in the Ayasdi representation (Fisher exact test P value < 10⁻¹⁰⁰). The centroid and dispersion of those genes were also consistent between the two topological representations, with Pearson's correlation coefficient $r = 0.98$ (P value < 10⁻¹⁰⁰) for the centroid (**Supplementary Fig. 20b**) and $r = 0.94$ (P value < 10⁻¹⁰⁰) for the dispersion (**Supplementary Fig. 20c**) of significant genes.

Gene connectivity, centroid, and dispersion within the topological representation. A notion of gene connectivity in the topological representation was introduced, defined as

$$s_i = \frac{N}{N-1} \sum_{\alpha, \beta \in \Gamma} \frac{e_{i,\alpha} A_{\alpha\beta} e_{i,\beta}}{\left(\sum_{\gamma \in \Gamma} e_{i,\gamma} \right)^2}$$

where $e_{i,\alpha}$ represents the average expression of gene i in node α of the topological representation, normalized as described in the paragraph “Processing of single cell RNA-seq data”. Γ denotes the set of nodes of the topological representation, $A_{\alpha\beta}$ is its adjacency matrix, and N is the total number of nodes. With this normalization, s_i takes values between 0 and 1.

The gene connectivity score depends on the distribution of expression values of the specific gene (**Supplementary Fig. 10**), and therefore genes cannot be ranked accordingly to their gene-connectivity score in a meaningful way. To assess the magnitude of the connectivity score relative to genes with the same expression profile and rank genes accordingly, we introduced a nonparametric statistical test. We tested for the null hypothesis of a randomly expressed gene with the same distribution of expression values having a higher gene-connectivity score. To that end, a null distribution was built for each gene i using a permutation test. Cell labels were randomly permuted 5,000 times for each gene, computing s_i after each permutation. A P value was estimated by counting the fraction of permutations that led to a larger value of s_i than the original one. Gene connectivity and its statistical significance were computed for each gene expressed in at least three cells. The resulting P values were adjusted for multiple testing by using the Benjamini–Hochberg procedure for controlling the false discovery rate.

To establish a pseudotemporal ordering within the topological representation, the notion of root node was introduced, defined as the node that maximizes the function

$$r(\alpha) = \text{corr}(d_\alpha, t)$$

where $\text{corr}(x,y)$ denotes the Pearson’s correlation coefficient between x and y , d_α is the graph distance function to node α that assigns to each node of the topological representation a value corresponding to the number of edges that are crossed in the shortest path from that node to node α , and t is the chronological sampling-time function that assigns to each node of the topological representation the average sampling time (expressed in days) of the cells contained in the node.

Least-squares linear regression was performed to determine the best fit for the coefficients a_0 and a_1 in the relation

$$d_{\text{root}} \equiv a_1 t + a_0$$

where d_{root} is the graph-distance function to the root node, determined in the previous paragraph. These coefficients were used to define the centroid and dispersion of each gene in the topological representation, expressed in days and given respectively by

$$c_i = \frac{1}{a_1} \left(\frac{\sum_{\alpha \in \Gamma} d_{\text{root}}(\alpha) e_{i,\alpha}}{\sum_{\beta \in \Gamma} e_{i,\beta}} - a_0 \right)$$

and

$$k_i = \frac{1}{a_1} \left(\sqrt{\frac{\sum_{\alpha \in \Gamma} (d_{\text{root}}(\alpha) - c_i a_1 + a_0)^2 e_{i,\alpha}}{\sum_{\beta \in \Gamma} e_{i,\beta}}} - a_0 \right)$$

Such normalization, using coefficients a_0 and a_1 to express the centroid and dispersion in units of pseudotime (days), allows comparison of the connectivity and dispersion of a gene across different topological representations or studies.

Significance of topological features. We computed the first persistent homology group^{45,46} using the graph distance of the topological representation. Given the pairwise distances of a set of points sampled from a space, persistent homology enables the quantification of topological features (connected components, loops, cavities, etc., preserved under continuous deformations of the space) compatible with the data at each scale. The first homology group, in particular, classifies loops of the space (**Supplementary Note 1**). We used

persistent homology death times as a proxy of the size of the loops and evaluated their statistical significance using a permutation test. To that end, we randomly permuted the labels of the genes 500 times for each cell independently. For each permutation we built a topological representation using the same parameters as in the original representation and computed the first persistent homology group. A P value for each of the loops was estimated from the distribution of the number of loops as a function of their death time. The resulting P values were adjusted for multiple testing by using Benjamini–Hochberg procedure for controlling the false-discovery rate.

Comparison with other methods for analyzing longitudinal single-cell RNA-seq data. We dimensionally reduced the processed single-cell RNA-seq data of experiment 2 using MDS, ICA, and t-SNE. For MDS and ICA, we used the same set of highly expressed, highly variant genes that we used for building the topological representation. For t-SNE, we used the top ten principal components of the gene space. In each representation we determined the cell that maximized the Pearson correlation coefficient between the 2D Euclidean distance to the cell and chronological sampling time, corresponding to the least differentiated cellular state.

Additionally, we compared scTDA to the single-cell software Monocle¹⁰, based on ICA and minimum-spanning trees, Diffusion Pseudotime⁶, and Wishbone⁷, based on diffusion coefficients, and SLICER⁸, based on locally linear embedding. We followed all recommendations in the documentation of these algorithms. In our tests, Monocle failed in running over the complete data set from experiment 2, consisting of 2,304 cells, and only a partial set of 834 cells sampled from all time points was analyzed.

Simulated data. Noisy branched-asynchronous cellular differentiation processes were simulated, from which 700 cells were sampled at three time points. To that end, we used the following strategy:

- (1) We simulated a noisy, branched tree-like structure with three parameters (t , u , and v) and performed a nonlinear transformation into 4D embedding space (spanned by the variables $g^{(1)}$, $g^{(2)}$, $g^{(3)}$, and $g^{(4)}$). This space provided the structure for four different groups of correlated genes.
- (2) We randomly sampled 700 points from this space, corresponding to 700 cells, and assigned a sampling day based on a multinomial distribution with probabilities given by a logistic function of t , to simulate asynchrony.
- (3) We simulated the expression of 300 genes driven by the variables $g^{(1)}$, $g^{(2)}$, $g^{(3)}$, and $g^{(4)}$ in addition to 200 genes with noncorrelated expression sampled from normal distributions.
- (4) We simulated the effect of dropout events using the standard logistic dependence of the dropout probability as a function of expression.

In what follows, we provide a detailed description of each of these steps.

First, we simulated four groups of correlated genes. These were defined by the equations

$$\begin{aligned} g_i^{(1)} &= \frac{200 u_i}{u_i^2 + t_i^2 + 0.2} + N_i^{(1)} \\ g_i^{(2)} &= \frac{200 t_i}{u_i^2 + t_i^2 + 0.6} + N_i^{(2)} \\ g_i^{(3)} &= \frac{100(u_i^2 + t_i^2) - 20}{u_i^2 + t_i^2 + 0.2} + N_i^{(3)} \\ g_i^{(4)} &= 100\sqrt{v_i} + N_i^{(4)} \end{aligned}$$

where $N_i^{(k)}$ are normally distributed random variables with mean $\mu = 150$ and s.d. $\sigma = 8$. The index $i = 1, \dots, 700$ runs across the sampled cells, and u_i , v_i , and t_i are randomly sampled from

$$\begin{aligned} u_i &= 0, \quad v_i = 0, \quad 0 \leq t_i < 0.2, \quad \text{or} \\ u_i &= 0.2 - t_i, \quad v_i = 0, \quad 0.2 \leq t_i < 0.7, \quad \text{or} \\ u_i &= t_i - 0.2, \quad v_i = 0, \quad 0.2 \leq t_i < 0.7, \quad \text{or} \\ u_i &= 0.6 - t_i, \quad v_i = 0, \quad 0.4 \leq t_i < 0.7, \quad \text{or} \\ u_i &= -0.2, \quad v_i = t_i - 0.4, \quad 0.4 \leq t_i < 0.7, \quad \text{or} \end{aligned}$$

where only the first $2 + r$ equations are considered in simulated differentiation processes with r lineage-branching points. The variable t_i represents the differentiation pseudotime of cell i at the time of sampling. To simulate asynchrony, each sampled cell was assigned a sampling day according to

$$\begin{aligned} 1 \geq p_i > \tau(t_i, 0.23) &\rightarrow \text{day 1} \\ \tau(t_i, 0.23) \geq p_i > \tau(t_i, 0.47) &\rightarrow \text{day 2} \\ \tau(t_i, 0.47) \geq p_i \geq 0 &\rightarrow \text{day 3} \end{aligned}$$

where $p_i \in [0,1]$ is a random number uniformly distributed, and τ is the logistic function

$$\tau(a, b) = \frac{1}{1 + e^{15(b-a)}}$$

We simulated 75 genes in each of the four groups of genes, with expression values given by

$$m_{l,i}^{(k)} = g_i^{(k)} N_l'^{(k)} N_l''^{(k)}, \quad k=1, \dots, 4, \quad l=1, \dots, 75, \quad i=1, \dots, 700$$

where $N_l'^{(k)}$ and $N_l''^{(k)}$ are normally distributed random variables with mean $\mu = 1$ and s.d. $\sigma = 0.2$. In addition, we simulated 200 extra genes with noncorrelated expression

$$m_{l,i}^{(0)} = N_l^{(0)} N_l''^{(0)}, \quad l=1, \dots, 200, \quad i=1, \dots, 700$$

with $N_l^{(0)}$ and $N_l''^{(0)}$ being normally distributed random variables with mean $\mu = 200$ and 1, and s.d. $\sigma = 50$ and 0.2, respectively. Hence, a total of 500 genes were simulated in each of the 700 cells. To model the effect of dropout events, we randomly set the expression of some of the genes in some of the cells to zero, with probability

$$P = \frac{1}{1 + e^{m-1}}$$

where m is the original expression value of the gene in the cell.

Apart from scTDA, we ran the algorithms Diffusion Pseudotime⁶, Wishbone⁷, and SLICER⁸ in the simulated data sets. We followed all the recommendations in the documentation of these algorithms. We assessed the performance of each algorithm by computing the Pearson's correlation coefficient between the cell pseudotime inferred by each of these algorithms and the simulated variable t .

Gene-ontology annotation. Gene ontologies were obtained from EMBL-EBI QuickGO⁴⁷. Specifically, categories GO:0006355 “Regulation of transcription, DNA-templated”, GO:0008380 “RNA splicing”, GO:0044822 “Poly(A) RNA binding”, GO:0051726 “Regulation of cell cycle”, and GO:0007049 “Cell cycle” were used to annotate genes. Expression of genes associated with DNA replication was analyzed by considering the 99 genes in the category GO:0006260 “DNA replication” expressed in less than 1,400 cells in experiment 2. Genes coding for proteins on the cellular surface were identified by looking in UniProtKB database⁴⁸ for proteins annotated with an extracellular topological domain. Gene-ontology enrichment analysis was performed using PANTHER classification system²³.

Transient cellular populations. Low-dispersion genes ($k_i < 1.7$ days and $k_i < 2.25$ days in experiments 2 and 1, respectively) with significant gene connectivity ($q < 0.05$) in the topological representation were clustered according to their centroid using k -means clustering (Supplementary Fig. 13b,c). The optimal number of clusters according to Davies-Bouldin index was four in experiment 2 (three in experiment 1) (Supplementary Fig. 13a), as it was also evidenced from visual inspection of the centroid distribution for low-dispersion genes. A state $r = 1, \dots, 4$ was assigned to each node of the topological representation based on the average expression of each cluster of low-dispersion genes in the node (Supplementary Fig. 13d). Genes with significant gene connectivity according to the permutation test described in the paragraph “Gene connectivity, centroid, and dispersion within the topological representation” were assigned to each of the four populations based on the number of cells

expressing the gene in each state r . Only genes expressed in at least 80 cells and at most 1,500 cells were considered.

Analysis of long noncoding RNAs. The coordinates of intergenic and anti-sense lncRNAs were downloaded from NONCODEv4 (ref. 49), and read counts were obtained using HTSeq (<http://www-huber.embl.de/users/anders/HTSeq>). The connectivity and statistical significance of each long noncoding gene in the topological representation was computed using scTDA. Only lncRNAs that were significant ($q < 0.05$) in both experiments 1 and 2 and that were supported by at least 50 reads in the longitudinal-stranded RNA-seq data were kept. Curation was performed to remove lncRNAs whose 3' end overlapped the 3' end of another gene, gene assignment therefore being ambiguous, or that corresponded to possible misannotations of the 3' UTR of a nearby gene.

Characterization of interneuron populations. Differential expression analysis between *En1*⁺, *Gata3*⁺, *Vsx2*⁺, and *Egfp*⁺ cells in nodes characterized as postmitotic (Supplementary Fig. 13d, state 3) was performed using the software SCDE⁴² with default parameters.

Analysis of mouse and human developmental data sets. We downloaded the read-count data of the single-cell RNA-seq experiments performed in ref. 35 (mouse distal lung epithelium), ref. 36 (human preimplantation embryos), and ref. 37 (mouse neocortex), normalized it as described in “Processing of single-cell RNA-seq data”, and analyzed it following the same steps described in “Topological representation” and “Gene connectivity, centroid and dispersion within the topological representation”. The topological representations were built using the Ayasdi implementation of Mapper, using correlation distance as metric and 2D MDS as a dimensional-reduction algorithm. We used a covering of \mathbb{R}^2 consisting of 33×33 , 52×52 , and 30×30 rectangular patches with an average overlap of 73%, 70%, and 71%, respectively, for the data from mouse distal lung epithelium, human preimplantation embryos, and mouse neocortex, respectively. In the case of the data from human preimplantation embryos, we also performed an analysis of transient cellular populations, following the same steps described in “Transient cellular populations” and setting a threshold of 2.95 in gene dispersion.

scTDA software. The algorithms described in this work were implemented and documented in an object-oriented python library for topological data analysis of high-throughput longitudinal single-cell RNA-seq data called scTDA. scTDA is provided as **Supplementary Code** and is publicly available at <https://github.com/RabadanLab/scTDA>.

Online database. We established a database for the topological representations and statistics of the two motor neuron differentiation data sets. The database is publicly available at https://rabadan.c2b2.columbia.edu/motor_neurons_tda.

Data availability. We deposited the single-cell RNA-seq data and read counts from the two motor neuron differentiation experiments in the NCBI Gene Expression Omnibus (GEO) database, with accession number **GSE94883**. The single-cell RNA-seq data from the differentiating distal-lung epithelium of mouse embryo³⁵ can be accessed through the GEO database, accession number **GSE52583**. The single-cell RNA-seq data from human pre-implantation embryos³⁶ can be accessed through the ArrayExpress database, with accession number **E-MTAB-3929**. The single-cell RNA-seq data from differentiating neurons in the mouse neocortex³⁷ can be accessed at https://vlp.ias.edu/science2016/DanaInfo=genebrowser.unige.ch+telley_govindan_science2016.tgz. All other data are available from the corresponding author upon reasonable request.

- 38. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
- 39. Anders, S., Pyl, P.T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
- 40. Stegle, O., Teichmann, S.A. & Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16**, 133–145 (2015).
- 41. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).

42. Kharchenko, P.V., Silberstein, L. & Scadden, D.T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
43. Shalek, A.K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363–369 (2014).
44. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).
45. Edelsbrunner, H., Letscher, D. & Zomorodian, A. Topological persistence and simplification. *Discrete Comput. Geom.* **28**, 511–533 (2002).
46. Zomorodian, A. & Carlsson, G. Computing persistent homology. *Discrete Comput. Geom.* **33**, 249–274 (2005).
47. Binns, D. *et al.* QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* **25**, 3045–3046 (2009).
48. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).
49. Zhao, Y. *et al.* NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.* **44**, D203–D208 (2016).