Poincaré Maps for Analyzing Complex Hierarchies in Single-Cell Data

Anna Klimovskaia[1*], David Lopez-Paz[1], Léon Bottou[2], and Maximilian Nickel[2*]

[1]Facebook AI, Paris, France
[2]Facebook AI, New York, USA
[*]Corresponding author

July 1, 2019

Abstract

The need to understand cell developmental processes has spawned a plethora of computational methods for discovering hierarchies from scRNAseq data. However, existing techniques are based on Euclidean geometry which is not an optimal choice for modeling complex cell trajectories with multiple branches. To overcome this fundamental representation issue we propose Poincaré maps, a method harnessing the power of hyperbolic geometry into the realm of single-cell data analysis. Often understood as a continuous extension of trees, hyperbolic geometry enables the embedding of complex hierarchical data in as few as two dimensions and well-preserves distances between points in the hierarchy. This enables direct exploratory analysis and the use of our embeddings in a wide variety of downstream data analysis tasks, such as visualization, clustering, lineage detection and pseudotime inference. In contrast to existing methods – which are not able to cover all those important aspects in a single embedding – we show that Poincaré maps produce state-of-the-art two-dimensional representations of cell trajectories on multiple scRNAseq datasets. Specifically, we demonstrate that Poincaré maps allow in a straightforward manner to formulate new hypotheses about biological processes which were not visible with the methods introduced before. Moreover, we show that our embeddings can be used to learn predictive models that estimate gene expressions of unseen cell populations in intermediate developmental stages.

Significance statement

The discovery of hierarchies in biological processes is central to developmental biology. We propose Poincaré maps, a new method based on hyperbolic geometry to infer continuous hierarchies from pairwise similarities. We demonstrate the efficacy of our method on multiple single-cell analysis tasks such as visualization, clustering, lineage identification, and pseudotime inference. Moreover, we show that our metric representation of hierarchies enables the prediction of unknown gene expressions along the discovered lineages.

Introduction

Understanding cellular differentiation, e.g., the transition of immature cells into specialized types, is a central task in modern developmental biology. Recent advances in single-cell technologies, such as single-cell RNA-sequencing and mass cytometry, have led to important insights into these processes based on high-throughput cell measurements [1–4]. Computational methods to accurately discover and represent cell development processes from large datasets and noisy measurements are therefore in great demand. However, this is a challenging task since methods are required to reveal the progression of cells along continuous trajectories with tree-like structure and multiple branches (e.g., as in Waddington's classic epigenetic landscape [5]). Multiple advances have been made towards this goal of discovering and analyzing hierarchical structures from single-cell measurements [6]. In particular, methods that exploit hierarchies for visualization [7;8], clustering [9], and pseudotime inference [10;11] have fueled unprecedented successes in developmental biology. To discover hierarchical relationships in the development of cells, many state-of-the-art methods rely on distances in low-dimensional Euclidean embeddings of cell measurements [7;8;12;13]. However, this approach is limited when modeling complex hierarchies as low-dimensional Euclidean embeddings can cause substantial distortions of distances in these cases. This is not only problematic for visualization but also for clustering and the identification of lineages.

To overcome this issue, we propose Poincaré maps, a novel method to compute embeddings for hierarchy discovery not in Euclidean but in hyperbolic space. This allows us to combine multiple advantages: First, hyperbolic space can be thought of as a

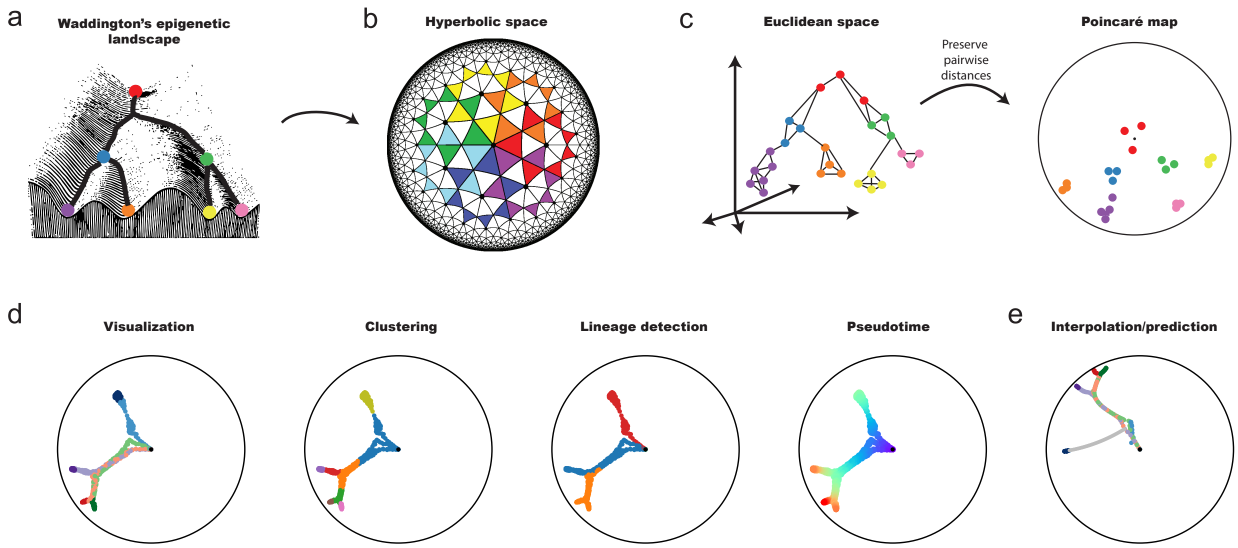


Figure 1 | *Poincaré maps* discover hierarchies and branching processes. (**a**) Our goal is to recover cell developmental processes, depicted here on the Waddington landscape. (**b**) Poincaré disks provide a natural geometry to preserve hierarchical structures and pairwise similarities in two dimensions. Poincaré disks grow as we approach their boundary: all the triangles depicted here are of equal size. (**c**) Poincaré maps first embed the data into RFA similarities, computed from a connected k-nearest neighbor graph. Second, they compute two-dimensional hyperbolic embeddings that preserve these similarities. (**d**) *Poincaré maps* are a tool to perform *standard* single-cell RNA sequencing analysis tasks. (**e**) In addition, Poincaré disks provide meaningful geodesics for hierarchies. Interpolating along hyperbolic geodesics allow the prediction of gene expression of unseen cell populations.

continuous analogue to trees and enables low-distortion embeddings of hierarchical structures in as few as two dimensions [14]. Second, the metric structure of hyperbolic space retains the ability to model continuous trajectories via distances and allows to employ the obtained embeddings for tasks such as clustering, lineage detection, pseudotime inference, and gene expression prediction for unseen cell types. Third, the Riemannian structure of hyperbolic manifolds enables the use of gradient-based optimization methods what is essential to compute embeddings of large-scale measurements. Fourth, by following Nickel et al. [15] and embedding similarities into the two-dimensional Poincaré disk (**Supplementary Note 1**), we can obtain a direct and intuitive visualization of the discovered hierarchies.

Results

Our method, called *Poincaré maps*, is guided by ideas from manifold learning and pseudotemporal ordering [16;17]. Given feature representations of cells such as their gene expressions, we aim to estimate the structure of the underlying tree-like manifold in three main steps (**Fig. 1a–c** and Online Methods): First, we compute a connected k-nearest neighbor graph (kNNG) [18] where each node corresponds to an individual cell and where each edge is weighted proportional to the Euclidean distance between the features of the connected cells. The purpose of this first step is to estimate the local nearest neighbor structure of the underlying manifold for which distances in the feature space are a good approximation. Second, we compute global geodesic distances from the kNN graph by moving along its weighted edges. This step can be computed efficiently using all pairs shortest paths or related measures such as the "Relative Forest Accessibilities" (RFA) index [19]. The purpose of this second step is to estimate the intrinsic geometry of the underlying manifold. Both, step one and two, are commonly used in manifold learning to approximate the structure of an unknown manifold from similarities in the feature space [11;18;20;21]. In the third step, we compute a two-dimensional embedding per cell in the Poincaré disk such that their hyperbolic distances reflect the inferred geodesic distances. The geometry of the Poincaré disk allows us to model continuous hierarchies efficiently: embeddings that are close to the origin of the disk have a relatively small distance to all other points and are thus well-suited to represent the root of a hierarchy or the beginning of a developmental process. On the other hand, embeddings that are close to the boundary of the disk, have a relatively large distance to all other points and are well-suited to represent leaf nodes. In our embedding, we expect therefore that nodes with small distances to many other nodes will be placed close to the origin of the disk. While such cells are likely from an early developmental stage, they do not necessarily belong to the root of the hierarchy (**Supplementary Fig. 1-3**). When a cell belonging to the root stage is known, we perform therefore a global translation on the Poincaré disk, to center this node and ease the visualization of the hierarchy (see Methods).

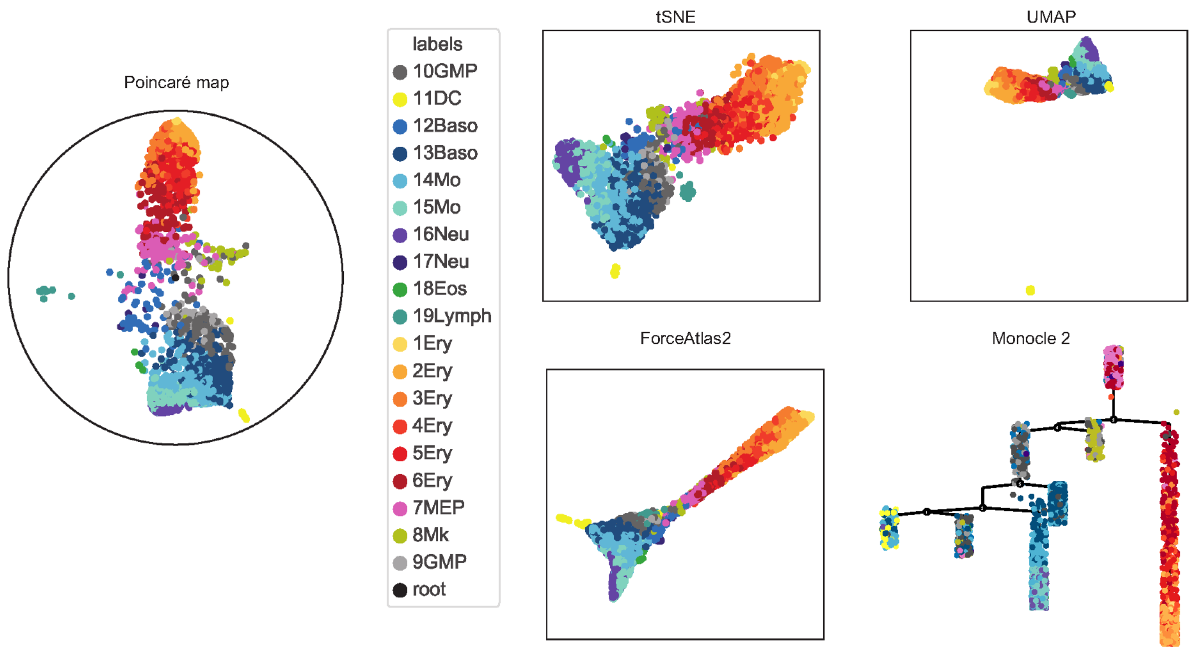


Figure 2 | Comparison of various embeddings for the mouse myeloid progenitors dataset (Paul et al.). Only Poincaré maps and tSNE show that lymphoid cluster is an outlier. However, as tSNE doesn't preserve hierarchy, therefore it is not possible to notice from it and is straightforward to read from Poincaré maps that the original labels of the populations don't correspond to the known biology: it is very surprising that monocytes are progenitors of neutrophils. Wolf et al. with additional analysis of gene expression showed that the labels of the original paper should be re-annotated.

All-in-one: visualization, clustering, lineage detection and pseudotime inference

In the following, we compare the embedding quality of Poincaré maps to state-of-the-art methods on various single-cell analysis tasks: visualization and lineage detection (Monocle 2[10], PAGA[22], diffusion maps[7], t-SNE[12], UMAP[13], and ForceAtlas2[23]), clustering (Louvain[24], agglomerative, k-means) and pseudotime inference (diffusion pseudotime[11]) (**Fig.1 (d)**).

An important property of Poincaré maps is that it allows to approach all these different tasks in a single embedding by combining the identification of clusters, trajectory, and hierarchy in an unsupervised manner. To the best of our knowledge, this is not possible with existing methods. For instance, t-SNE is a state-of-the-art visualization method that facilitates local similarities to achieve visual separation of the clusters in the data. However, t-SNE does not preserve global similarities between the clusters and therefore there are no guarantees that the global hierarchical structure will be preserved. UMAP computes a low-dimensional representation of data in Euclidean space that preserves the topological structure. However, there are no guarantees that there exists a low-dimensional representation of complex tree topologies in two-dimensional Euclidean space. Diffusion maps specifically tackle the problem of capturing diffusion-like dynamics and continuous branching in the data. However, it only allows to visualize a very simple branching structure in two dimensions. Graph abstractions (PAGA) and Monocle 2 are another class of methods that try to capture and visualize the hierarchical relationships in the data. PAGA produce an "abstracted graph" with nodes corresponding to partitions of the data and the edges representing the relationships of these nodes. PAGA doesn't represent the relationships inside each partition. Monocle 2 forces a tree-like topology on the data using "reversed graph embedding" in a low-dimensional Euclidean space. However, as in the case of UMAP, such a representation might not exist for complex trees.

To evaluate the performance of Poincaré maps, we perform separate comparisons to the state-of-the-art methods for each task. For this purpose, we employ multiple synthetic datasets generated from known dynamical systems, and three single-cell RNA sequencing datasets[2;3;25], where we compare Poincaré maps with the canonical hematopoetic cell lineage tree[26], and various state-of-the-art embeddings (**Supplementary Note 2**).

Additionally we compare Poincaré maps to the visualization of k-NN graph with force-directed layout (ForceAtlas2) guided by PAGA. Despite the fact that ForceAtlas2 produces a good visual layout of a tree topology, it doesn't preserve the hierarchical distances. The limitation of this approach we demonstrate in the section below.

An important result from our experiments is that Poincaré maps was the only method that demonstrated the ability to visualize the correct branching structure of developmental processes for *all* datasets (**Supplementary Fig. 1-7**). For example, on the dataset Paul et al.[2] only Poincaré maps and t-SNE identify the lymphoid cluster while this important population would not be visible during exploratory data analysis using UMAP or ForceAtlas2 (**Fig. 2**). Although t-SNE visualizes separate clusters well for Paul et al. dataset it completely disregards the hierarchical structure between clusters (see also example in **Supplementary Fig. 7**). Knowledge of the position of a newly identified cluster in the developmental hierarchy could be further exploited for assigning labels (e.g. "lymphoid population") or, when the population was not known, for designing experiments to test morphological properties. Finally, Poincaré maps places the 16Neu cluster downstream of 15Mo in the hierarchy – in contrast to the canonical hierarchy where neutrophils and monocytes are located at the same level. This result is in line with the analysis of Wolf et al., and indicates the inconsistency is due to a faulty labeling of the clusters.

In addition to these results, we demonstrate in **Supplementary tables 1-2** that Poincaré maps could be directly applied to achieve state-of-the art results on clustering and pseudotime inference.

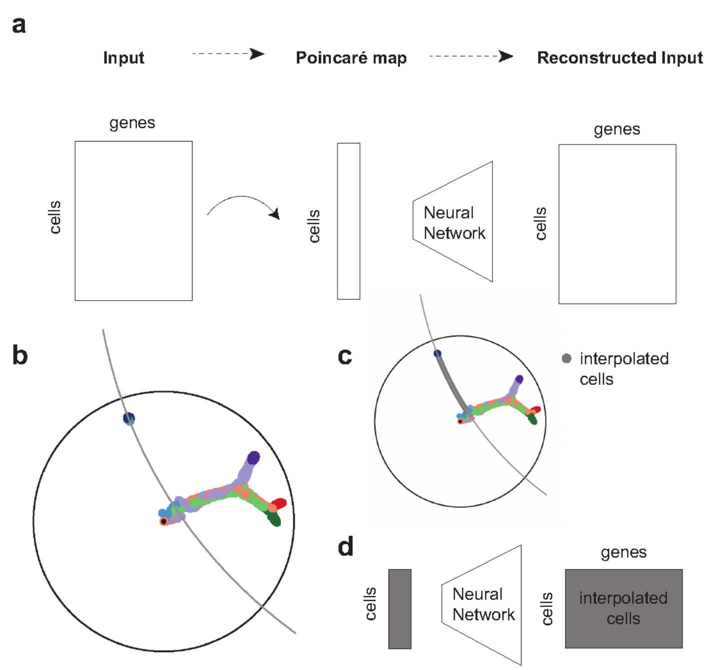


Figure 3 | (a) We obtain a Poincaré map of a dataset (not containing intermediate cell type we want to predict). We train a neural network to perform the mapping from the Poincaré map back to gene expression space. (**b**) Poincaré map visualization of embedded dataset on the synthetic example of myeloid progenitors. Removed neutrophil progenitors were not used to obtain the embedding. Grey line represents a geodesic between a pair of randomly selected points in the populations of interest. (**c**) We sample points uniformly along the geodesics. **d**) We use the pre-trained neural network from (a) to predict gene expression values of interpolated points in (c).

Poincaré maps robustly predict values on unseen intermediate cell types

The ability of two-dimensional Poincaré maps to preserve pairwise similarities unlocks important opportunities to expand the standard biological analysis toolkit. The main reason for this is that geodesics in the Poincaré disk are well-suited to model shortest paths in tree-like structures. We demonstrate this advantage on a prediction task: we remove an intermediate population from a dataset, and predict its gene expression from the computed embeddings (**Fig.1 (e)**). For this purpose, we first estimate a low dimensional embedding of the dataset and then learn a function that maps from these embeddings back into the original gene expression space (**Fig.3 (a)**). If this mapping is performed on a space providing us with meaningful geodesics about the hierarchy (such as the hyperbolic spaces that we propose), it should be able to predict the gene expression values of unseen cells. This could be useful for scenarios where intermediate cell types are not observed.

We demonstrate the performance of interpolations by artificially removing one cell type on several of the datasets described before. In the synthetic example of myeloid progenitors we remove the majority of neutrophil progenitors, in Olsson et al. we remove the HSPC-2 population, and in the Planaria dataset of Plass et al. we remove a part of parenchymal progenitors.

As a first step, we obtain embeddings for each of the datasets (after "shrinking" the dataset by having removed the unseen cell type) using several methods: Poincaré maps, ForceAtlas2 and UMAP. As a second step, we train a neural network to predict gene expression values from the corresponding embeddings, by minimizing the mean squared error between the original gene expression values of "shrinked" dataset and the corresponding predictions (**Fig.3 (b)**). We use the same architecture and training parameters of the neural network for all the embeddings. As a third step, we randomly sample a pair (or multiple pairs) of points that we will consider the end-points of our interpolation. This step relies on some prior knowledge about the developmental hierarchy of the data, yet we consider it to be reasonable for our demonstration purposes, as well as for real case scenarios. Finally, we use the same end-points to construct an uniform interpolation along geodesic in either Poincaré (for Poincaré maps) or Euclidean (for ForceAtlas2 and UMAP) space (**Fig.3 (c)**). We use the previously trained neural network to predict the gene expression for all the unobserved cells that would lie in the chosen interpolation (**Fig.3 (d)**).

In our experiments, we found Poincaré maps to perform this task 1.3-3 times better than other embedding methods on a variety of datasets (Myeloid progenitors, Olsson, Plass). For instance, Poincaré maps recover a faithful two-dimensional hierarchical embedding of the entire Planaria system. By preserving hierarchies and pairwise similarities, these same embeddings are useful for downstream analyses. In particular, Poincaré maps could faithfully reconstruct first 50 principal components of the removed population (**Fig.4 (a) – (c)**).

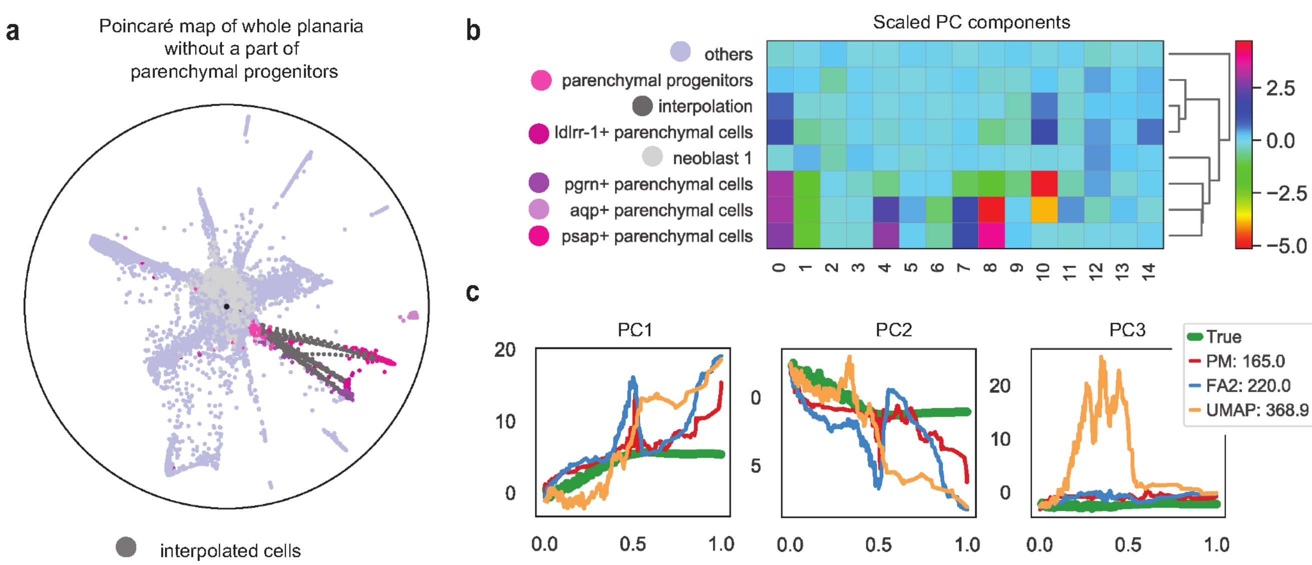


Figure 4 | (a) *Poincaré map* of Plass et al. dataset after removing the parenchymal progenitors cluster. Grey points depict interpolated values along the geodesic in the Poincaré disk. (**b**) Comparison of first 15 normalized features for interpolated points and original values. Interpolated values predict well the gene expression of the removed cluster (parenchymal progenitors). (**c**) Comparison of first 3 features (largest PCA components) in terms of pseudotime for various embeddings used for interpolation.

Poincaré maps generate new hypothesis about early blood development in mice

As a deeper case study of Poincaré maps, we analyze the dataset of early blood development in mice, previously studied by Moignard et al.[1]. This dataset contains measurements of cells captured *in vivo* with qRT-PCR at different development stages: primitive streak (PS), neural plate (NP), head fold (HF), four somite GFP (Runx1) negative (4SG-) and four somite GFP positive (4SG+) (**Fig. 5 (a)**). The stages correspond to different physical times of the experiment between embryonic day 7 and day 8.25. We compare our results obtained with Poincaré maps to Moignard's diffusion maps study[1], and to Haghverdi's reconstruction of diffusion pseudotime[11]. Poincaré maps provide a qualitatively different visualization of the developmental process, where we are able to visualize the whole spectrum of the heterogeneity arising from the onset of the process. Neither PCA, nor diffusion maps are able to provide a visualization of this process. While Moignard's and Haghverdi's analyses suspected an asynchrony in the developmental process, neither their application of PCA or diffusion maps were able to reveal this. In particular, previous studies suggest that the split into endothelial and erythroid sub-populations happens in the head fold. Our analysis using Poincaré maps indicates that the sub-population fate of the cells is already predefined at primitive strike. Additionally, Poincaré maps reveal a separate cluster consisting of a mixture of cells at different developmental stages (**Supplementary Fig. 12**). This cluster is referred to as "mesodermal" cells by Moignard et al., while by Haghverdi et al. considers it as the root of the developmental process. However, as we demonstrate in **Supplementary Fig. 12 – 13**, assigning this cluster as the root of the hierarchy would lead to a contradiction with the physical direction of time. By virtue of the Poincaré visualization, we reassigned the root of the developmental process to the furthest PS cell not belonging to the "mesodermal" cluster. Given our reassigned root, we separate the dataset into five potential lineages (see **Methods**), to find the asynchrony in the developmental process in terms of marker expressions (**Fig. 5 (b)**). Analysis of the composition of cells belonging to each lineage (**Fig. 5 (c)**) indicates that erythroid cells belong only to lineage 0 and this lineage contains no endothelial cells. **Fig. 5 (d)** shows a substantially improved agreement of Poincaré pseudotime (with the newly reassigned root) with the experimental time (stages) compared to the pseudotime ordering proposed by Haghverdi et al. The analysis of gene expressions of main endothelial and hemogenic markers agrees with the known pattern of gene activation for endothelial and erythroid branches (**Supplementary Fig. 14 – 15**). **Fig. 5 (e)** also demonstrates that the main hemogenic genes for the erythroid population are already expressed at the PS stage (details in **Supplementary Note 3**) and that the differences in gene expression aparent at all the stages between the lineages. Our analysis using Poincaré maps suggests therefore that the fate of erythroid and endothelial cells could already be defined at primitive streak.

Discussion

The rapid onset of popularity and accessibility of single-cell RNA sequencing technologies facilitated the development of new computational approaches to analyze these data. While many computational methods exist, their results often don't agree between each other. The choice of the right computational approach at a very early stage of exploratory data analysis, will dictate the generated hypotheses about the underlying biology. Here we demonstrated, that Poincaré maps without strong assumptions on the data and just by leveraging advantages of hyperbolic geometry for hierarchical structures, reveal complex cell developmental processes that would remain undiscovered by other methods. While any hypothesis generated via computational analysis should be validated in the lab before being converted into strong statements, a properly chosen computational approach will facilitate the selection of appropriate experiments and right conclusions.

For this purpose, Poincaré maps aids the discovery of complex hierarchies from single-cell data by embedding large-scale cell measurements in a two-dimensional Poincaré disk. The resulting embeddings are easy to interpret during exploratory analysis and provide a faithful representation of similarities in hierarchies due to the underlying geometry. This property make Poincaré maps to stand out among other embeddings as it allows to simultaneously handle visualization, clustering, lineage detection, and pseudotime inference. In our experiments, we showed that our embeddings can not only be used for reading out hierarchical relations between cell types and for identifying the presence the outliers, but also to predict gene expression values of unseen intermediate populations. Moreover, Poincaré maps are able to capture average *dynamics* of the unseen population and not only an average expression values.

With Poincaré maps, we also hope to bring interest about hyperbolic embeddings in general to the biology community. Due to their advantageous properties for modeling hierarchical data, they could provide substantial benefits for a wide variety of problems such as studying transcriptional heterogeneity and lineage development in cancer from single-cell RNA and DNA sequencing data, reconstructing the developmental hierarchy of blood development, and reconstructing embryogenesis branching trajectories.

Methods

In the following we discuss the main stages of our method, i.e., estimating proximities that are informative about hierarchical structure and embedding these proximities into the Poincare disk.

Let $\mathcal{X} = \{x_i\}_{i=1}^n$ be a high-dimensional dataset of n samples $\boldsymbol{x}_i \in \mathbb{R}^p$ (e.g., individual cells) with p features (e.g., gene expression measurements).

Local Connectivity We first estimate local connectivity structures as typically done in manifold learning.[18;20;21] In particular, let $\mathcal{N}(i,k)$ denote the k nearest neighbors of $\boldsymbol{x}_i$ in $\mathcal{X} \setminus \boldsymbol{x}_i$ according to the Euclidean distance. We then create a symmetric k-nearest-neighbor graph $G = (V, E, w)$, where the set of vertices $V = \{v\}_{i=1}^n$ represents the samples in $\mathcal{X}$ and the set of edges $E = \{v_i \sim v_j : i \in \mathcal{N}(j,k) \wedge j \in \mathcal{N}(i,k)\}$ represent the nearest neighbor relations. Furthermore, each nearest neighbor relation is

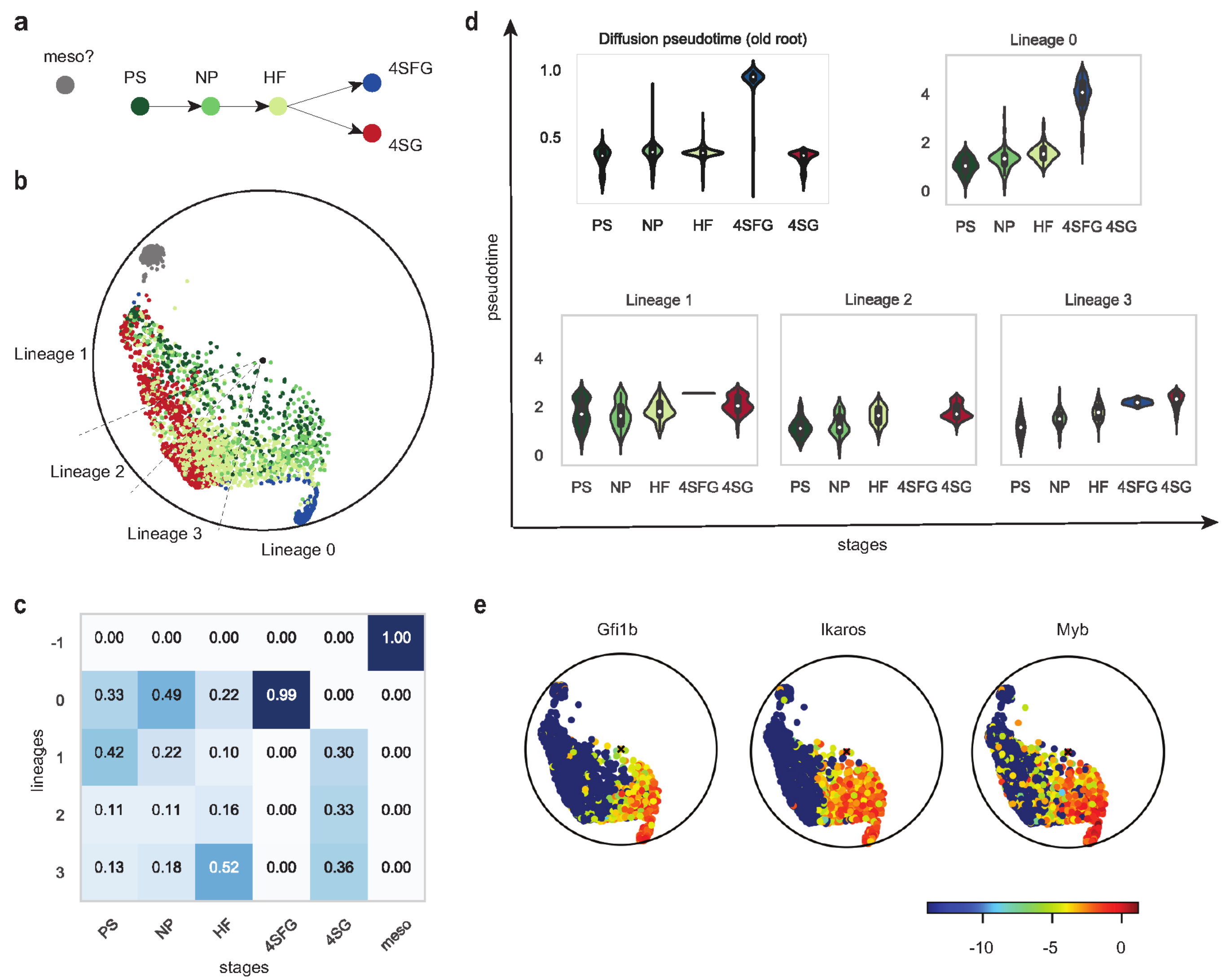


Figure 5 | (**a**) Developmental hierarchy proposed by Moignard et al. and Haghverdi et al. (**b**) Rotated Poincaré map with respect to reassigned root. Grey cluster represent a cluster of potential outliers or “mesodermal” cells as suggested by Moignard et al. Lineage slices were obtained with Poincaré maps (see Methods). (**c**) Composition of detected lineages in terms of the presence of cells from different developmental stages. (**d**) New ordering of cells proposed by Poincaré maps here has a much better agreement with developmental stages than ordering originally proposed by Haghverdi et al.: we see a very clear correlation of Poincaré pseudotime with actual developmental time. (**e**) Gene expression of main hemogenic genes. Hemogenic genes of erythroid lineage are already expressed at the PS and NP stages.

weighted using the Gaussian kernel over distances

$$w(i,j) = \begin{cases} \exp\left(-\frac{\|\boldsymbol{x_i} - \boldsymbol{x_j}\|_2^2}{2\sigma^2}\right) & \text{if } i \sim j \in E, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

where σ is a hyperparameter that controls the kernel width. By enforcing connectivity of G, we preserve finite distances between all measurements.

Global Proximites To estimate the underlying manifold structure from distances on the kNN graph G, we can employ all-pairs shortest paths or related methods such as the *Relative Forest Accessibility* (RFA) index, which is defined as follows: Let $L = D - A$

denote the graph Laplacian of the graph G, where $A_{ij} = w(i,j)$ is the corresponding adjacency matrix and $D_{ii} = \sum_j w(i,j)$ is the degree matrix. The RFA matrix P is then given as[19]

$$P = (I + L)^{-1}. \tag{2}$$

P is a doubly stochastic matrix where each entry p_{ij} corresponds to the probability that a spanning forest of G includes a tree rooted at i which also includes j (i.e., where j is accessible from i)[19;27] Compared to shortest-paths, the RFA index has the advantage to increase the similarity between nodes that belong to many shortest paths. This can provide an important signal to discover hierarchical structures as nodes that participate in many shortest paths are likely close to the root of the hierarchy. In all experiments, we use the RFA index to estimate global proximities.

Hyperbolic Embedding Given P, we aim at finding an embedding $\boldsymbol{y_i}$ of each $\boldsymbol{x}_i$ that highlights the hierarchical relationships between the samples. For this purpose, we embed P into two-dimensional hyperbolic space.

The Poincaré disk is the Riemannian manifold $\mathcal{P} = (\mathcal{B}, d_p)$, where $\mathcal{B} = \{\boldsymbol{y} \in \mathbb{R}^2 : \|\boldsymbol{y}\| < 1\}$ is the open n-dimensional unit ball. The distance function on $\mathcal{P}$ is then defined as

$$d_p(\boldsymbol{y}_i, \boldsymbol{y}_j) = \operatorname{acosh}\left(1 + 2\frac{\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2}{(1 - \|\boldsymbol{y}_i\|^2)(1 - \|\boldsymbol{y}_j\|^2)}\right) \tag{3}$$

It can be seen from Equation (3), that the Euclidean distance within $\mathcal{B}$ is amplified smoothly with respect to the norm of $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$. This property of the distance is key for learning continuous embeddings of hierarchies. For instance, by placing the root node of a tree at the origin of $\mathcal{B}$, it would have relatively small distance to all other nodes, as its norm is zero. On the other hand, leaf nodes can be placed close to the boundary of the ball, as the distance between points grows quickly with a norm close to one.

To compute the embedding we use an approach similar to t-SNE[12] and approximate the RFA probabilities in P via distances in the embedding space. In particular, we define the similarity q_{ij} between the embeddings v_i and v_j as:

$$q_{ij} = \frac{\exp(-d_p(\boldsymbol{y}_i, \boldsymbol{y}_j)/\gamma)}{\sum_k \exp(-d_p(\boldsymbol{y}_i, \boldsymbol{y}_k)/\gamma)}, \tag{4}$$

where $\boldsymbol{y}_i, \boldsymbol{y}_j \in \mathcal{P}$. A natural measure for the quality of the embedding is then the symmetric Kullback-Leibler divergence between both probability distributions:

$$\mathcal{L}(P; \mathcal{Y}) = \sum_i \mathrm{KL}(P_i || Q_i) + \mathrm{KL}(Q_i || P_i) \tag{5}$$

Details on the optimization To compute the embeddings, we minimize Equation (5) via *Riemannian Stochastic Gradient Descent* (RSGD).[28] In particular, we update the embedding of $\boldsymbol{y}_i$ in epoch t using

$$\boldsymbol{y}_i^{t+1} \leftarrow \mathfrak{R}_{\boldsymbol{y}_i^t}(-\eta \operatorname{grad}(\mathcal{L}, \boldsymbol{y}_i^t)), \tag{6}$$

where $\operatorname{grad}(\mathcal{L}, \boldsymbol{y}_i^t)$ denotes the Riemannian gradient of Equation (5) with respect to $\boldsymbol{y}_i^t$, $\mathfrak{R}_{\boldsymbol{y}_i^t}$ denotes a retraction (or the exponential map) from the tangent space of $\boldsymbol{y}_i^t$ onto $\mathcal{P}$, and $\eta > 0$ denotes the learning rate. The optimization can be performed directly in the Poincaré ball or, alternatively, in the Lorentz model of hyperbolic space which provides improved the numerical properties and efficient computation of the exponential map.[29]

Translation in $\mathcal{P}$ Equation (5) favors embeddings where nodes with short distances to all other nodes are placed close to the origin of the disk. While such nodes correspond often to nodes that are close to the root of the underlying tree, it is not guaranteed that the root is the closest embedding to the origin. However, when the root node is known, we can perform an isometric transformation of the entire embedding that places this node at the origin and preserves all distances between the points. In particular, to translate the disk such that the origin of the Poincaré disk is translated to $\boldsymbol{v}$, $\boldsymbol{x}$ is translated to

$$\tau(\boldsymbol{x}, \boldsymbol{v}) = \frac{(1 + 2\langle \boldsymbol{v}, \boldsymbol{x}\rangle + \|\boldsymbol{x}\|^2)\boldsymbol{v} + (1 - \|\boldsymbol{v}\|^2)\boldsymbol{x}}{1 + 2\langle \boldsymbol{v}, \boldsymbol{x}\rangle + \|\boldsymbol{v}\|^2\|\boldsymbol{x}\|^2} \tag{7}$$

Since the spatial resolution is amplified close to the origin of the disk, provides also a method to "zoom" into different parts of the embedding by moving the area of interest to the origin.

Clustering Hyperbolic space is a metric space and thus allows us to compute distances between any pair of points. This makes Poincaré maps straightforwardly applicable to clustering techniques that rely only on pairwise (dis)similarity measurements such as spectral clustering, agglomerative clustering and kmedoids.

Lineages As a naive approach for lineage detection we suggest to use agglomerative clustering by the angle between a pair of points in the Poincaré disk after the rotation with respect to the root node.

Poincaré pseudotime "Pseudotime" is typically reffered as "a measure of how much progress an individual cell has made through a process such as cell differentiation"[17]. As Poincaré pseudotime we propose to use the distance from the root node in the Poincaré ball.

Interpolation with Poincaré maps Given two classes, interpolation predicts the gene expression values for an intermediate population. For a given dataset, we obtain its Poincaré map and train a neural network to map elements in the Poincaré disk back to the gene expression space. Then, we sample pairs of points from the two classes, as well as points along the Poincaré geodesic between them. We use the trained neural network to predict the gene expression values for the interpolated points. Since temporal dynamics are very important for developmental processes, we compare the reconstruction using dynamic time warping between the diffusion pseudotime series for the removed population and the prediction provided by the different embeddings.

Choice of Hyperparameters In the following, we discuss the function of different hyperparameters in Poincaré maps and propose typical value ranges. The number of nearest neighbors k reflects the average connectivity of the clusters and is typically set to $k \in [15, 30]$. The Gaussian kernel width σ is responsible for the weights for the k-NN graph in the original space and is typically set to $\sigma \in [1.0, 2.0]$. The softmax temperature γ controls the scattering of embeddings and is typically set to $\gamma \in [1.0, 2.0]$.

Code availability The code to reproduce our analyses is available at `https://github.com/klanita/Poincare-maps/tree/master/release`.

Data availability Several public datasets were used in this study: three synthetic datasets generated with Scanpy, Olsson et al. (synapse ID syn4975060), Paul et al. (accession code GSE72857), Moignard et al. (accession code GSE61470), and Plass et al. (accession code GSE103633, preprocessed data available at `https://shiny.mdc-berlin.de/psca/`).

Acknowledgments

We would like to thank Ioana Sandu and Will Macnair for valuable discussions.

Author information

Contributions

L.B., M.N. and A.K. conceived the idea. A.K., M.N. and D. L.-P. designed and implemented the computational tools. A.K. performed the analysis and biological interpretation of results. L.B. contributed to the design of the study. M.N. supervised the study. A.K., M.N. and D.L.-P. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Corresponding author

Correspondence to Anna Klimovskaia or Maximilian Nickel.

References

[1] Moignard, V. *et al.* Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature biotechnology* **33**, 269 (2015).

[2] Paul, F. *et al.* Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).

[3] Olsson, A. *et al.* Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature* **537**, 698 (2016).

[4] Nestorowa, S. *et al.* A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**, e20–e31 (2016).

[5] Ferrell Jr, J. E. Bistability, bifurcations, and waddington's epigenetic landscape. *Current biology* **22**, R458–R466 (2012).

[6] Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331 (2017).

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

[8] Wolf, F. A. *et al.* Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *bioRxiv* 208819 (2018).

[9] Levine, J. H. *et al.* Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).

[10] Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nature methods* **14**, 979 (2017).

[11] Haghverdi, L., Buettner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods* **13**, 845 (2016).

[12] Maaten, L. v. d. & Hinton, G. Visualizing data using t-sne. *Journal of machine learning research* **9**, 2579–2605 (2008).

[13] McInnes, L. & Healy, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[14] Gromov, M. *Metric structures for Riemannian and non-Riemannian spaces* (Springer Science & Business Media, 2007).

[15] Nickel, M. & Kiela, D. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, 6341–6350 (2017).

[16] Magwene, P. M., Lizardi, P. & Kim, J. Reconstructing the temporal ordering of biological samples using microarray data. *Bioinformatics* **19**, 842–850 (2003).

[17] Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32**, 381 (2014).

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

[19] Chebotarev, P. & Shamis, E. The matrix-forest theorem and measuring relations in small social groups. *https://arxiv.org/abs/math/0602070* (2006).

[20] Tenenbaum, J. B., De Silva, V. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *science* **290**, 2319–2323 (2000).

[21] Belkin, M. & Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation* **15**, 1373–1396 (2003).

[22] Wolf, F. A. *et al.* Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology* **20**, 59 (2019).

[23] Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one* **9**, e98679 (2014).

[24] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**, P10008 (2008).

[25] Plass, M. *et al.* Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, eaaq1723 (2018).

[26] Murphy, K. & Weaver, C. *Janeway's immunobiology* (Garland Science, 2016).

[27] Chebotarev, P. Spanning forests and the golden ratio. *Discrete Applied Mathematics* **156**, 813–821 (2008).

[28] Bonnabel, S. Stochastic gradient descent on Riemannian manifolds. *IEEE Trans. Automat. Contr.* **58**, 2217–2229 (2013).

[29] Nickel, M. & Kiela, D. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. *arXiv preprint arXiv:1806.03417* (2018).

Supplementary Notes for: Poincaré Maps for Analyzing Complex Hierarchies in Single-Cell Data

June 28, 2019

Supplementary Note 1: Poincaré maps for learning hierarchical representations

Hyperbolic space is a Riemannian manifold whose structure is well-suited to represent hierarchical and tree-like relationships. For our work, this combines two important advantages: First, the metric structure of hyperbolic space allows us to capture continuous hierarchical relationships and interpolate between points. Second – and in contrast to other metric spaces – hierarchies can already be represented in two-dimensional hyperbolic space with small distortion [1, 2, 3, 4].

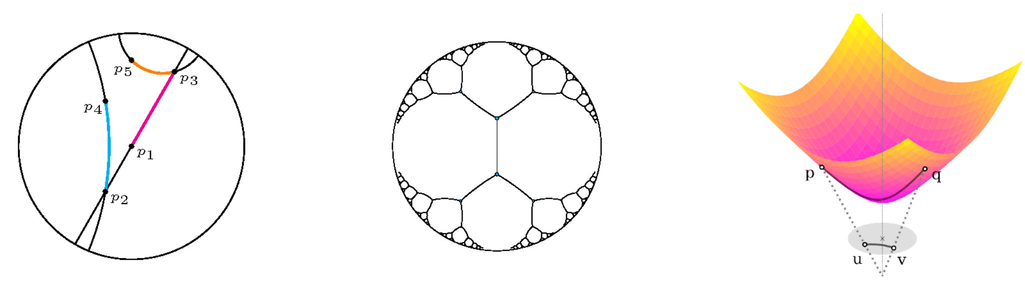
Poincaré disk model

There exist multiple, equivalent models of hyperbolic space, such as the Beltrami-Klein, the Lorentz, and the Poincaré half-plane model. In this work, we base our approach on the Poincaré disk model, as it is best suited for visual analysis. The Poincare disk is defined as follows: $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^2 \mid \|\boldsymbol{x}\| < 1\}$ be the *open* unit disk, where $\|\cdot\|$ denotes the Euclidean norm. The Poincaré disk corresponds then to the Riemannian manifold $(\mathcal{P}, g_{\boldsymbol{x}})$, i.e., the open unit disk equipped with the Riemannian metric tensor

$$g_{\boldsymbol{x}} = \left(\frac{2}{1 - \|\boldsymbol{x}\|^2}\right)^2 g^E,$$

where $\boldsymbol{x} \in \mathcal{P}$ and g^E denotes the Euclidean metric tensor. Furthermore, the distance between points $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{P}$ is given as

$$d(\boldsymbol{u}, \boldsymbol{v}) = \operatorname{acosh}\left(1 + 2\frac{\|\boldsymbol{u} - \boldsymbol{v}\|^2}{(1 - \|\boldsymbol{u}\|^2)(1 - \|\boldsymbol{v}\|^2)}\right). \tag{1}$$



(a) Geodesics in the Poincaré disk **(b)** Tree Embedding **(c)** Lorentz model

Supplementary Figure 1. a) Geodesics in the Poincaré disk model of hyperbolic space. Due to the negative curvature of the space, geodesics between points are arc that are perpendicular to the boundary of the disk. For curved arcs, midpoints are closer to the origin of the disk (p1) than the associated points, e.g. (p3, p5). c) Points (p,q) lie on the surface of the upper sheet of a two-sheeted hyperboloid. Mapping of points on the hyperboloid (p, q) onto the Poincaré disk.

The boundary of the disk is denoted by $\partial\mathcal{B}$ and is not part of the manifold, but represents infinitely distant points. Geodesics in $\mathcal{P}$ are then arcs that are orthogonal to $\partial\mathcal{B}$ (as well as all diameters). See Figure 1a for an illustration.

It can be seen from Equation (1) that the Euclidean distance of two points in the Poincaré disk is amplified with respect to their distance to the origin of the disk. This locality property of the Poincaré distance is key for continuous embeddings of hierarchies. For instance, by placing the root node of a tree at the origin of $\mathcal{B}^d$ it would have a relatively small distance to all other nodes as its Euclidean norm is zero. On the other hand, leaf nodes can be placed close to the boundary of the Poincaré disk as the distance grows fast between points with a norm close to one. Furthermore, Equation (1) is symmetric and the hierarchical organization of the space is solely determined by the distance of points to the origin. Due to this property, Equation (1) is applicable in an unsupervised setting where the hierarchical order of objects is not specified in advance. Importantly, this allows us to learn embeddings that simultaneously capture the hierarchy of objects (through their norm) as well a their similarity (through their distance).

The Riemannian manifold structure of hyperbolic space enables the use Riemannian Stochastic Gradient Descent (RSGD) [5] to compute the embeddings. In RSGD, parameter updates are performed via

$$\boldsymbol{y}_{t+1} = \mathfrak{R}_{\boldsymbol{y}_t}(-\eta\,\mathrm{grad}(\mathcal{L}, \boldsymbol{y}_t))$$

where $\mathfrak{R}_{\boldsymbol{y}}$ denotes a retraction from the tangent space at $\boldsymbol{y}$ onto the manifold, $\mathrm{grad}(\mathcal{L}, \boldsymbol{y}_t)$ denotes the Riemannian gradient of the scalar function $\mathcal{L}$, and $\eta > 0$ denotes the learning rate. The embeddings can be learned directly in the Poincaré disk $\mathcal{P}$ or, alternatively, in the Lorentz model of hyperbolic space $\mathcal{H}$ which has advantageous properties for stochastic optimization. We refer to [2] and [4] for the detailed optimization procedure on both hyperbolic manifolds. When optimization is performed in the Lorentz model, we can map the learned embeddings into the Poincaré disk via the diffeomorphism $p : \mathcal{H} \to \mathcal{P}$

$$p(x_0, x_1, \ldots, x_n) = \frac{(x_1, \ldots, x_n)}{x_0 + 1}$$

which preserves all geometric properties including isometry (see also Supplementary Figure 1).

Supplementary Note 2: Benchmarks on datasets with known hierarchy

Visualization

We compare Poincaré maps to several methods frequently used for visualization: tSNE [6], UMAP [7], diffusion maps [8], graph abstractions (PAGA) [9], ForceAtlas2 [10] and Monocle 2 [11]. For all competing methods, we used the default parameters provided by the authors. If no parameters were provided, we performed a parameter search to achieve the best performance for each method.

While methods such as diffusion maps, PAGA and Monocle 2 can be used by a knowledgeable user to infer the correct structure form data with several post-processing iterations, here we would like to demonstrate how Poincaré maps extracts meaningful insights from data without further post-processing. The ability to recover hidden hierarchies automatically and in one shot makes Poincaré maps an attractive tool for the analysis of branching processes and complex hierarchical structures.

Synthetic datasets

To demonstrate the performance of Poincaré maps we used several synthetic datasets available as Jupyter notebooks with Scanpy [12]: a simple toggle switch, myeloid progenitors and myeloid progenitors with Gaussian blobs. These datasets were previously used to demonstrate the performance of diffusion maps and graph abstractions, and constitute great examples of manifolds with a hierarchical structure of increasing levels of complexity. All models consist of Boolean equations, which were translated into ordinary differential equations and simulated with Scanpy as stochastic differential equations with Gaussian noise [13].

A simple toggle switch model [14, 15] is a process with two branches, which are defined by the expression of two markers. **Supplementary Figure 2** demonstrates that all competing methods produce rather correct results for this simple problem. However, Poincaré maps gives a more clear separation of the intermediate states of terminal fates (inter1 vs inter2). In this example, only tSNE, diffusion maps, and Poincaré maps produce embeddings with meaningful pairwise distances.

A synthetic dataset for myeloid differentiation [16] represents cell differentiation progresses of a common myeloid progenitor state towards one of four different branches: erythrocyte, neutrophil, monocyte and megakaryocyte. **Supplementary Figure 3** shows the provided embeddings for all methods. Poincaré maps produce an embedding which is visually similar to the other methods, but has neither discontinuities, nor overlaps in the trajectories, since it preserves all the pairwise distances. Given the known root, the rotation of the Poincaré map (by means of translation) allows to easily read out the hierarchy. Diffusion maps produce embeddings consistent with one main branch, but more Euclidean dimensions would be necessary to separate the rest. Monocle 2 produce a tree layout consistent with the hierarchical structure of the data, but is not able to reconstruct the temporal connection (trajectory) of the cell differentiation process.

The third dataset shows the stability of Poincaré maps with respect to the existence of clusters not related to the main cell development process. To this end we use the synthetic dataset of myeloid differentiation with two Gaussian blobs, added as proposed by Wolf et al. [9] (**Supplementary Figure 4**). None of the benchmark methods except ForceAtlas2 is able to capture the hierarchy.

Mouse myelopoesis dataset (single-cell RNA seq)

To demonstrate the performance of Poincaré maps on single-cell RNA seq data, we used the mouse myelopoesis dataset (wild type only) from Olsson et al. [17]. The data was downloaded and preprocessed according to the pipeline from Qiu et al. [11]. The processed dataset contained 532 features for 382 cells. Nine cell types were annotated corresponding to the original study: HSCP-1 (haematopoietic stem cell progenitor), HSCP-2, megakaryocytic, erythrocytic, Multi-Lin* (multi-lineage primed), MDP (monocyte-dendritic cell precursor), monocytic, granulocytic and myelocyte (myelocytes and metamyelocytes). In order to obtain best results for Monocle 2, we used the analysis pipeline provided by the authors (`https://github.com/cole-trapnell-lab/monocle2-rge-paper`). As the reference hierarchy, we used canonical hematopoetic cell lineage tree [18] (**Supplementary Figure 5 (a)**).

Poincaré maps, after rotation (**Supplementary Figure 5 (b)**), reveal the known hierarchy and suggest that part of HSPC-2 cluster actually corresponds to the megakaryocyte/erythrocyte progenitor (MEP), and that the cluster named Multi-Lin corresponds to the granylocyte/monocyte progenitor (GMP). Also according to Poincaré maps, the cluster annotated as myelocyte does not belong to the hierarchy, or constitutes a mature state of granulocytes. However, the validation of these hypotheses requires a detailed differential expression analysis.

Supplementary Figure 5 (c) shows how widely used methods such as tSNE distort the pairwise distances, therefore making more difficult to draw conclusions about hierarchies. Similarly, two dimensions of diffusion maps are not enough to represent the branching. UMAP and ForceAtlas2 results overall agree with the Poincaré maps, but don't allow to reason about the subtle hierarchical relations between HSCP-1/2 clusters and MDP. Monocle 2 captures the global branching, but fails to depict more fine-grained relations: between erythrocytytes and megakaryocytes or granulocytes and myelocytes.

Mouse myeloid progenitors dataset (MARS-Seq)

As an example of a dataset with multiple intermediate populations, we use a dataset provided by Paul et al. [19]. Myeloid progenitor cells were separated by sorting the c-Kit+ Sca1 lineage from mouse bone marrow and sequenced with MARS-seq. We followed the data preprocessing procedure recipe_zheng17 (Scanpy-recipe [20]), which selects the 1000 most highly-variable genes for 2730 cells. In the original study, the authors identify 19 clusters. We use these labels and canonical hematopoetic cell lineage tree (**Supplementary Figure 6 (a)**) to compare the performance of all methods. We run all methods except Monocle 2 on the 20 top principal components of the preprocessed data. For Monocle 2, we used the Jupyter notebook provided by the authors (the lymphoid cluster was separated as described in the original study).

Supplementary Figure 6 (b) shows the embeddings provided by Poincaré maps. For this dataset, the root is supposed to be at CMP cluster, which is not observed. We chose the root as the medoid (with respect to Poincaré distances) of the MEP and GMP clusters combined.

Supplementary Figure 6 (c) shows the hierarchy that could be read out from the Poincaré map. We would like to point out that Poincaré maps clearly separate lymphoid cells and dendritic cells as outliers, which agrees with the canonical tree as they are part of lymphoid lineage. None of the other methods **(Supplementary Figure 6 (d)**) were able to capture this fact. Poincaré maps also suggest that some of the clusters (13-15) could be relabeled to better reflect the canonical hierarchy. After the removal of the lymphoid cluster, Monocle 2 captures the main lineage branching between the MEP and GMP lineages, but it does not separate dendritic cells, and destroys the eosonphils cluster. Wolf et al. demonstrated that Monocle 2 results without the removal of the lymphoid cluster only worsen.

Finally, Poincaré maps places the 16Neu cluster downstream of 15Mo in the hierarchy. However, the canonical hierarchy shows neutrophils and monocytes at the same level. As noted by Wolf et al., we suppose that this inconsistency is due to a faulty labeling of the clusters.

Planaria dataset (Drop-seq)

To demonstrate scalability of Poincaré maps to large datasets, we analyzed the entire Planaria dataset of Plass et al. [21]. The dataset comprises 11 individual experiments capturing a total of 21,612 cells with droplet-based single-cell transcriptomics (Drop-seq). To obtain the Poincaré maps we used the pre-processed data provided by the authors: https://nbviewer.jupyter.org/github/rajewsky-lab/planarian_lineages/blob/master/paga/planaria.ipynb. The preprocessed dataset comes in the form of 50 principal components, which were used by the authors to apply tSNE, PAGA and ForceAtlas2. (**Supplementary Figure 7**) illustrates that Poincaré maps agree with tSNE and ForceAtlas2 embeddings, significantly outperforms PCA and UMAP, and agrees with the PAGA hierarchy annotation (Figure 4 in Plass et al.).

Clustering

Poincaré maps provide embeddings useful beyond visualization. Since Poincaré maps preserve pairwise similarities, their embeddings are suitable for downstream tasks, such as clustering. We compared several clustering approached using Poincaré maps and benchmark embeddings. We additionally provided Louvain clustering and clustering in the original gene expression space. Since the datasets comprise several continuous trajectories and there is no true separation for progenitor populations of different branches, we used the Adjusted Rand Index (ARI) and Fowlkes-Mallows scores (FMS) to measure cluster quality.

Adjusted Rand Index. The Adjusted Rand Index (ARI) is a function that measures the similarity between two cluster assignments. ARI is bonded bewteen $[-1, 1]$, where negative values correspond to independent labelings, similar clusterings have a positive ARI, and 1.0 is the perfect match score. Lets denote C a ground truth class assignment and K the clustering. Adjusted Rand Index is defined through raw Rand Index (RI):

$$RI = \frac{a+b}{C_2^{n_{samples}}}, \tag{2}$$

where a is the number of pairs of elements that are in the same set in C and in the same set in K, b is the number of pairs of elements that are in different sets in C and in different sets in K, and $C_2^{n_{samples}}$ is the total number of possible pairs in the dataset (without ordering). ARI is after adjusting for random labelings:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}, \tag{3}$$

where a is the number of pairs of elements belonging to the *same* cluster in predicted and true labels, b is the number of pairs of elements belonging to *different* clusters in predicted and true labels, and $C_2^{n_{samples}}$ is the number of all possible combinations of pairs of elements in the dataset.

Fowlkes-Mallows scores. The Fowlkes-Mallows score FMI is defined as the geometric mean of the pairwise precision and recall:

$$FMI = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}}, \tag{4}$$

where TP is the number of pairs of points that belong to the same cluster in both the true labels and the predicted labels (true positives), FP is the number of pairs of points that belong to the same clusters in the true labels but not in the predicted labels (false positives), and FN is the number of pairs of points that belong in the same clusters in the predicted labels but not in the true labels (false negatives). The FMI ranges from 0 to 1. A high value indicates a good similarity between two clusterings.

More details on these metrics can be found at: `https://scikit-learn.org/stable/modules/clustering.html#clustering-evaluation`

Supplementary Table 1 shows the clustering results on synthetic datasets. Poincaré maps achieve very similar score to louvain clustering and significantly outperform clustering approaches using other embedding methods except tSNE embedding, which combined with spectral clustering allows to achieve the best scores. However, as we demonstrated before, tSNE is not preserving the hierarchy and therefore would be less useful for other downstream tasks.

Pseudotime

We demonstrated Poincaré pseudotime performance by comparison with real time and diffusion pseudotime on synthetic datasets. **Supplementary Table 2** demonstrates that Poincaré pseudotime as well as diffusion pseudotime achieve high correlation scores with the actual time on all synthetic datasets. This is unsurprising, since these two measures are related in their nature. The performance of both pseudotime approaches is probably bounded by the construction of kNNG.

Interpolation

We demonstrate the advantage of Poincaré maps in one interpolation task. Our goal is to predict values of unseen intermediate cell types. This could be useful for scenarios where intermediate cell type are not observed. We demonstrate the performance of interpolations on several of the datasets described before by artificially removing one cell type. In the synthetic example of myeloid progenitors we remove the majority of neutrophil progenitors, in Olsson et al. we remove the HSPC-2 population, and in the Planaria dataset of Plass et al. we remove a part of parenchymal progenitors.

As a first step, we obtain embeddings for each of the datasets (after "shrinking" the dataset by having removed the unseen cell type) using several methods: Poincaré maps, ForceAtlas2 and UMAP. As as second step, we train a neural network to predict gene expression values from the corresponding embeddings, by minimizing the mean squared error between the original gene expression values of "shrinked" dataset and the corresponding predictions **Supplementary Figure 8 (a), (b), 10 (a), 11 (a)**. We use the same architecture and training parameters of the neural network for all the embeddings. As a third step, we randomly sample a pair (or multiple pairs) of points that we will consider the end-points of our interpolation (**Supplementary Figure 8 (c)**). This step relies on some prior knowledge about the developmental hierarchy of the data, yet we consider it to be reasonable for our demonstration purposes, as well as for real case scenarios. Finally, we use the same end-points to construct an uniform interpolation along geodesic in either Poincaré (for Poincaré maps) or Euclidean (for ForceAtlas2) space (**Supplementary Figure 8 (d), (e)**). We use the previously trained neural network to predict the gene expression for all the unobserved cells that would lie in the chosen interpolation (**Supplementary Figure 8 (f), 10 (b), 11 (b)**.

Since temporal dynamics are very important for developmental processes, we compare the reconstruction using dynamic time warping between the diffusion pseudotime series for the removed population and the prediction provided by the different embeddings. **Supplementary Table 3** demonstrates that Poincaré maps provide twice better prediction performance in datasets with a complex hierarchy, such as myeloid progenitors and Olsson datasets. In the case of Planaria, the hierarchy is rather shallow, so the advantage of Poincaré maps is less pronounced.

Technical details. In the myeloid progenitors dataset, to obtain a pair of points between which to draw the interpolation, we first perform a Louvain clustering of the shrinked dataset. Afterwards, we select two clusters from which we sample a corresponding pair: one cluster corresponding to mature neutrophils, and another cluster corresponding to early neutrophil progenitors.

In the Olsson et al. dataset, we use the previously annotated clusters, and randomly sample several pairs of points belonging to "HSPC-1" cluster, and to either the "Multi-lin", "Eryth" and "Meg" clusters. This corresponds to a potential position of "HSPC-2" cluster in the hierarchy.

In the Plass et al. Planaria dataset, we removed parenchymal progenitors. We interpolate between the original cluster of parenchymal progenitors and either the "psap+ parenchymal cells", "pgrn+ parenchymal cells", or "ldlrr-1+ parenchymal cells".

As for the network architecture, we used 5 fully connected layers with ReLU non-linearities. For UMAP and ForceAtlas2, we added batch normalization, as it showed better performance.

Supplementary Note 3: Reconstructing developmental trajectories of asynchronous process: early blood development in mice (qPCR)

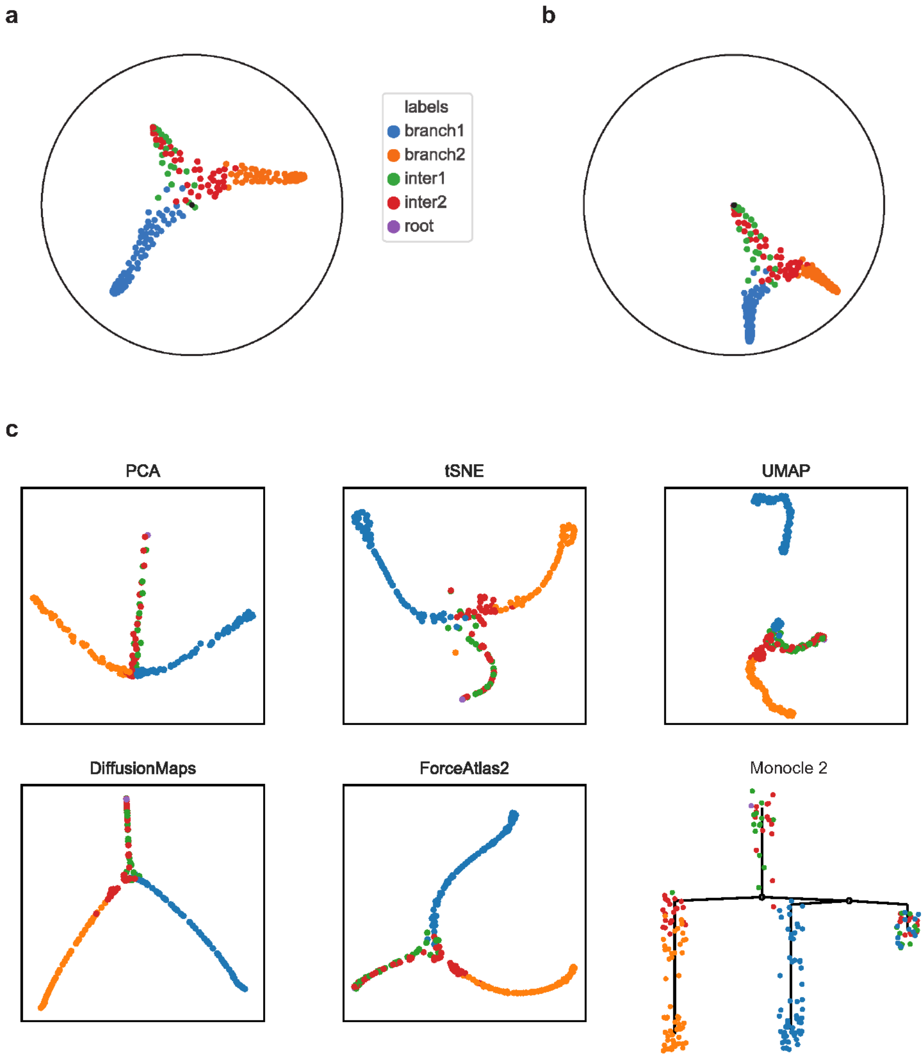
We analyze the single cell qPCR dataset of early blood development in mice [22] using Poincaré maps. We followed the data preprocessing procedure described in Haghverdi et al. [8].

First, we visualized the dataset with a Poincaré map using the labels corresponding to different stages of differentiation [22]: primitive streak (PS), neural plate (NP), head fold (HF), four somite GFP negative (4SG-) and four somite GFP positive (4SG+) (**Supplementary Figure 12 (a)**). We see one cluster standing out. Therefore, we perform spectral clustering with Poincaré distances to break down this cluster for further analysis (**Supplementary Figure 12(b),(c)**). Then, cluster 4 mainly consists of Flk1-Runx1- cells (see **Supplementary Figure 12 (d)**). Moignard et al. [22] refer to this cluster as "mesodermal cells at primitive strike" and suggest that these cells give rise to blood and endothelial cells.

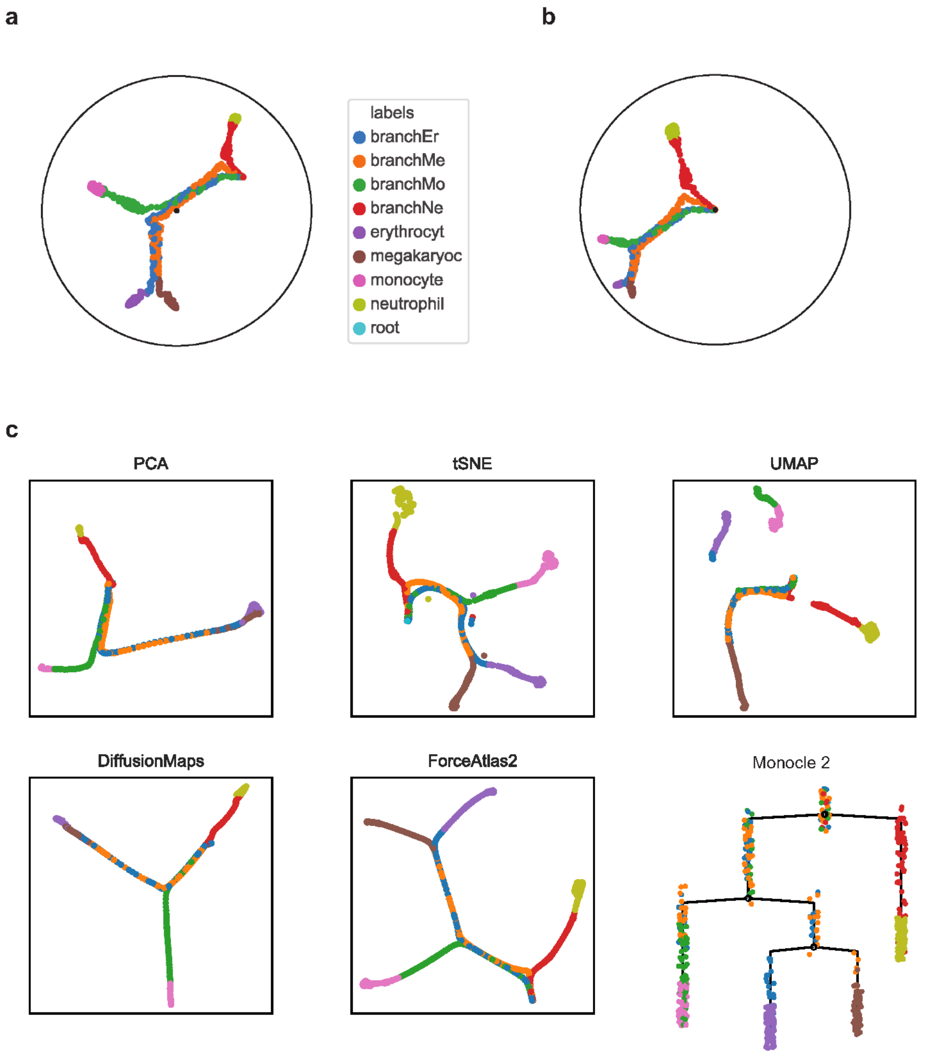
The cell that Haghverdi et al. choose as root of the differentiation for the diffusion pseudotime analysis belongs to the "mesodermal" cluster in our analysis. We visualize (**Supplementary Figure 13**) the diffusion pseudotime and Poincaré pseudotime with the roots (a) suggested by Haghverdi et al., and (b) the most dissimilar point in the PS cluster in terms of Poincaré distance. Undesiredly, the distances from (a) grow orthogonal to the actual developmental stages. It agrees with the conclusion in Haghverdi et al. that such a choice of embedding does not allow to see the asynchronous development. Therefore, cluster 4 may not correspond to cells leading to endothelial and blood cells, but rather to early mesodermal cells, which in their turn lead to some other population (Supplementary Figure 4 in Moignard et al.). We will further refer to the cluster 4 as "mesodermal".

As pointed out by Moignard et al., blood development is a highly asynchronous process, which is hard to capture with PCA or diffusion maps. In **Supplementary Figure 14** we further demonstrate how Poincaré maps could be used to reveal the developmental structure in this process. First, we apply the rotation to the Poincaré map to place the root cell defined above to the center of the disk. Then, we apply our lineage detection procedure and demonstrate that inside of each lineage, the order of the developmental stages is on average preserved. However, if we look at all lineages combined, then the populations from PS, NP, HF stages appear to be a homogeneous mixture. Therefore, the angular information in Poincaré maps adds the additional amount of information crucial to understand asynchronous processes.

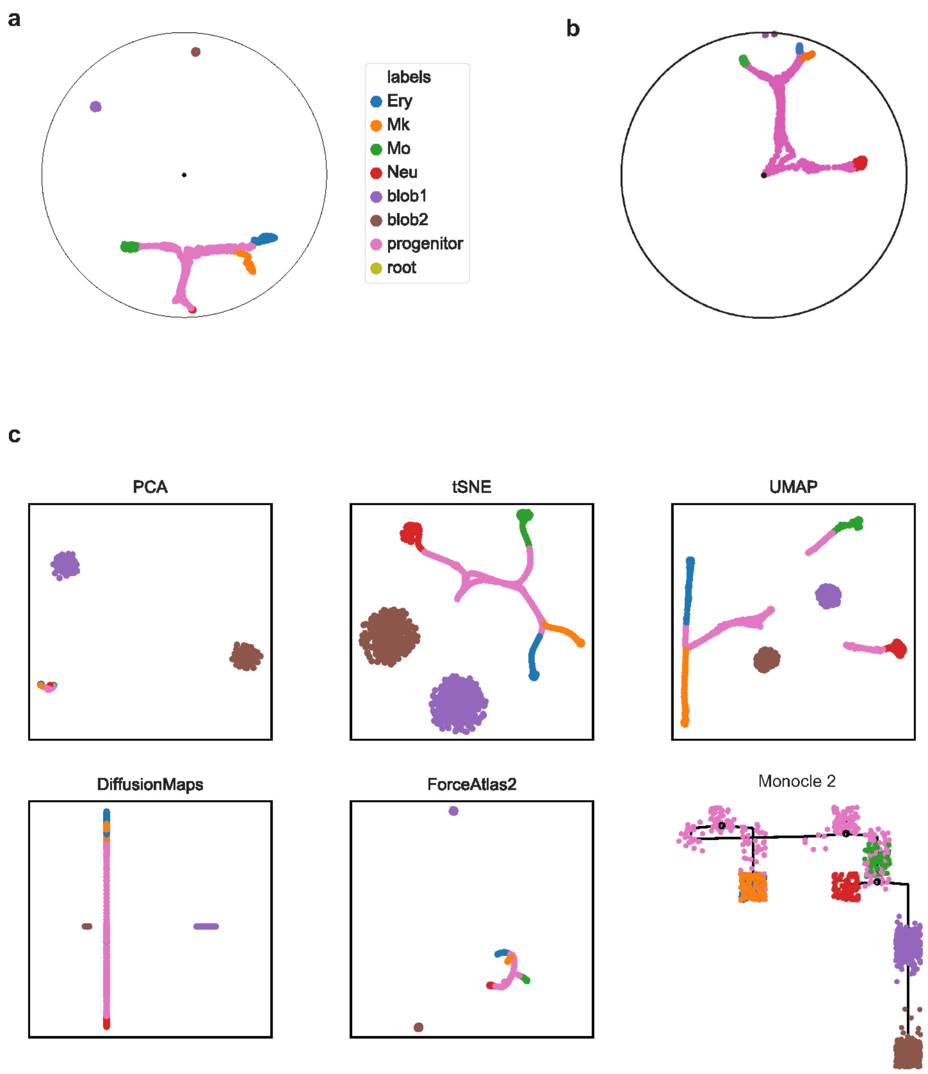
Finally, we analyzed the expression profiles of main endothelial and hematopoetic markers for different lineages (**Supplementary Figure 15**). Poincaré maps suggest that cells make an early decision about which branch to become. In particular, we suggest that cells commit to their future branch as early as in the PS stage.



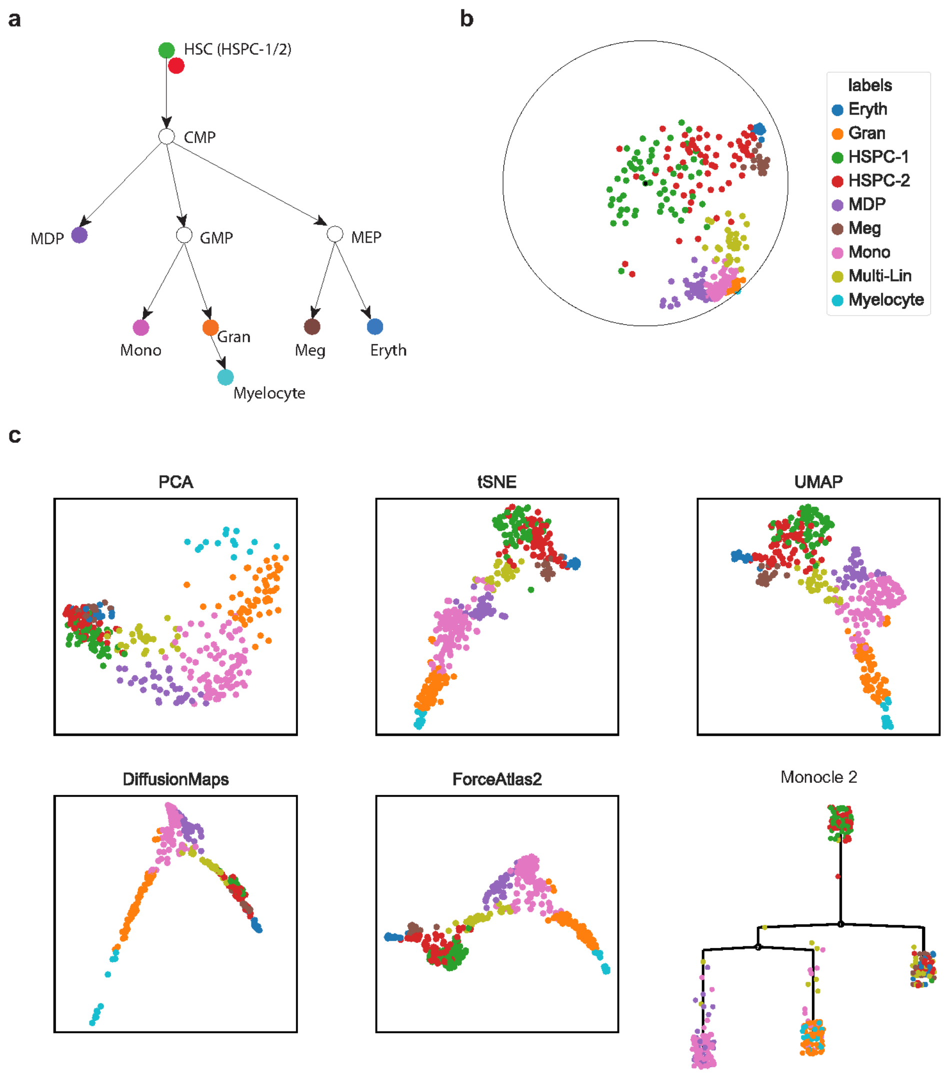
Supplementary Figure 2. Comparison of various embeddings for the simple toggle switch model. There are 2 distinct branches. We additionally labeled intermediate states from the simulations. **(a)** Raw Poincaré map. **(b)** Rotation of the Poincaré map with respect to the known root. **(c)** Benchmark methods.



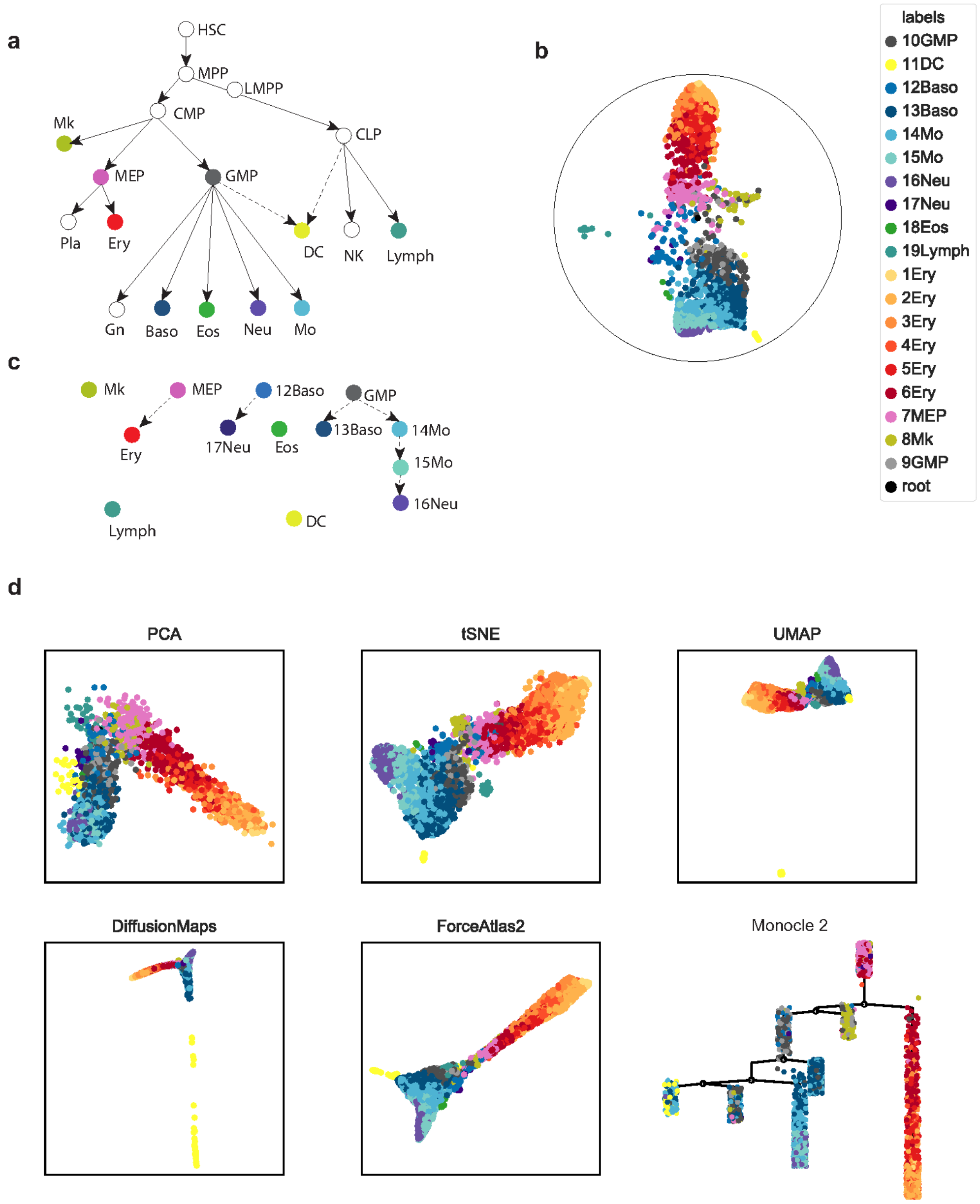
Supplementary Figure 3. Comparison of various embeddings for a synthetic model of myeloid progenitors differentiation. There are 4 distinct branches. We additionally labeled intermediate states from the simulations. **(a)** Raw Poincaré map. **(b)** Rotation of the Poincaré map with respect to the known root. **(c)** Benchmark methods.



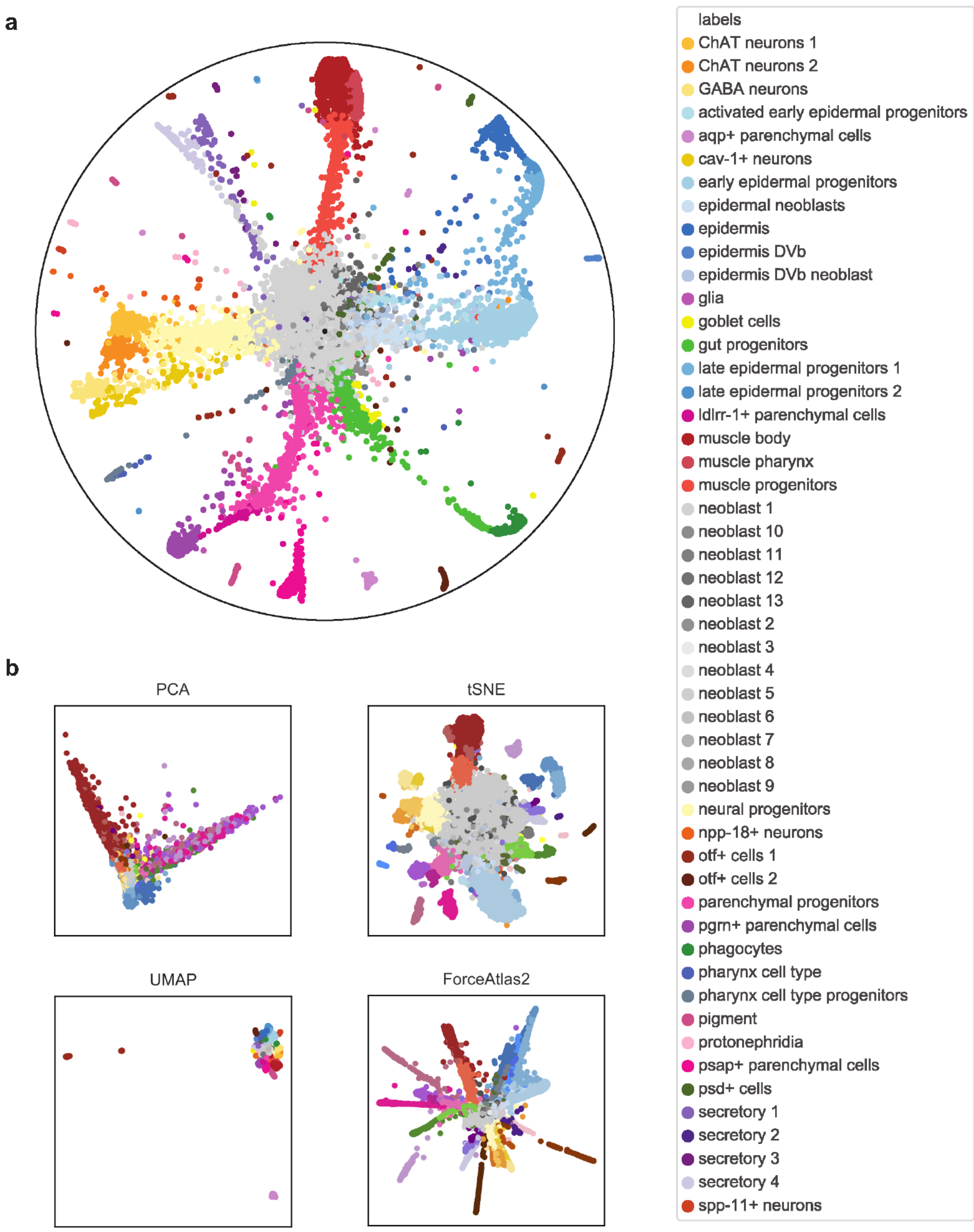
Supplementary Figure 4. Comparison of various embeddings for a synthetic model of myeloid progenitors differentiation (4 distinct branches) with two additional Gaussian clusters. We additionally labeled intermediate states from the simulations. **(a)** Raw Poincaré map. **(b)** Rotation of the Poincaré map with respect to the known root. **(c)** Benchmark methods.



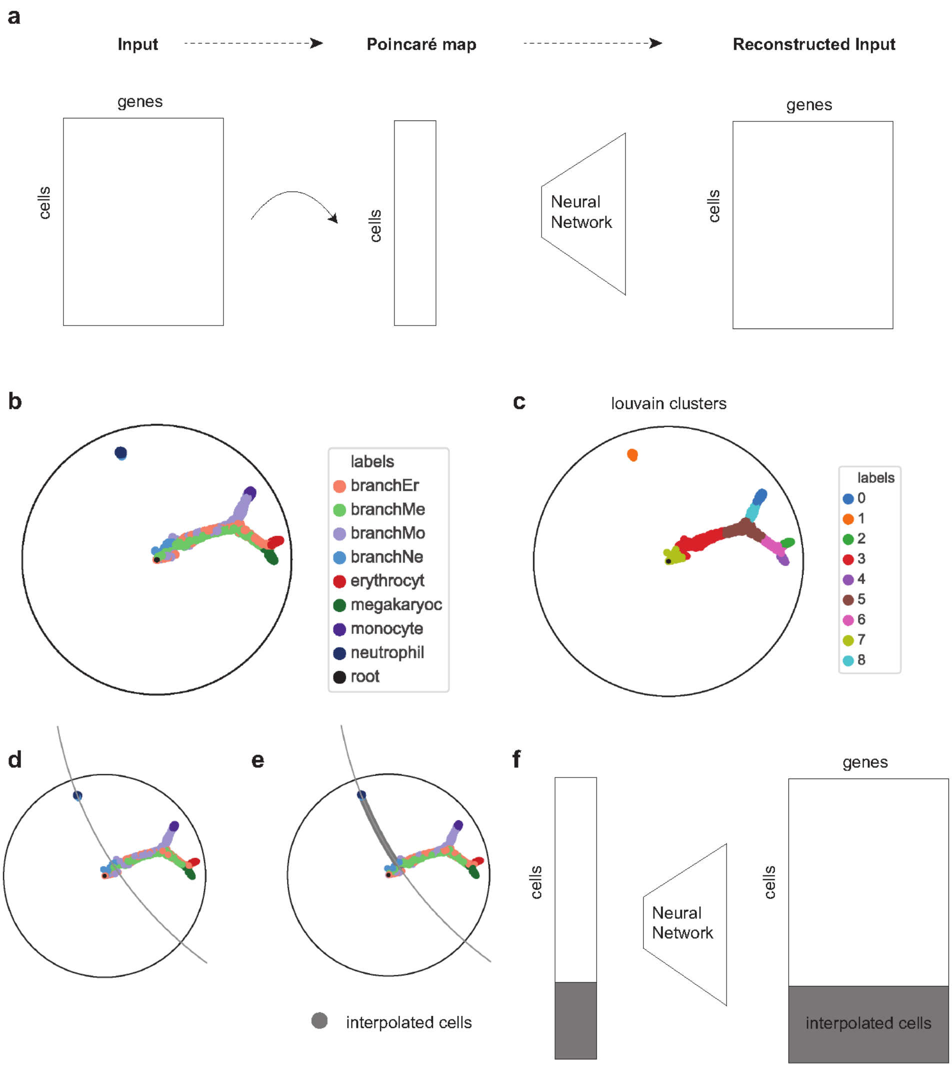
Supplementary Figure 5. Comparison of various embeddings for the scRNAseq dataset of mouse myelopoesis (Olsson et al.). **(a)** Canonical hematopoetic cell lineage tree. Colored circles correspond to the population colors from the dataset. **(b)** Poincaré map rotated with respect to the root. **(c)** Benchmark methods.



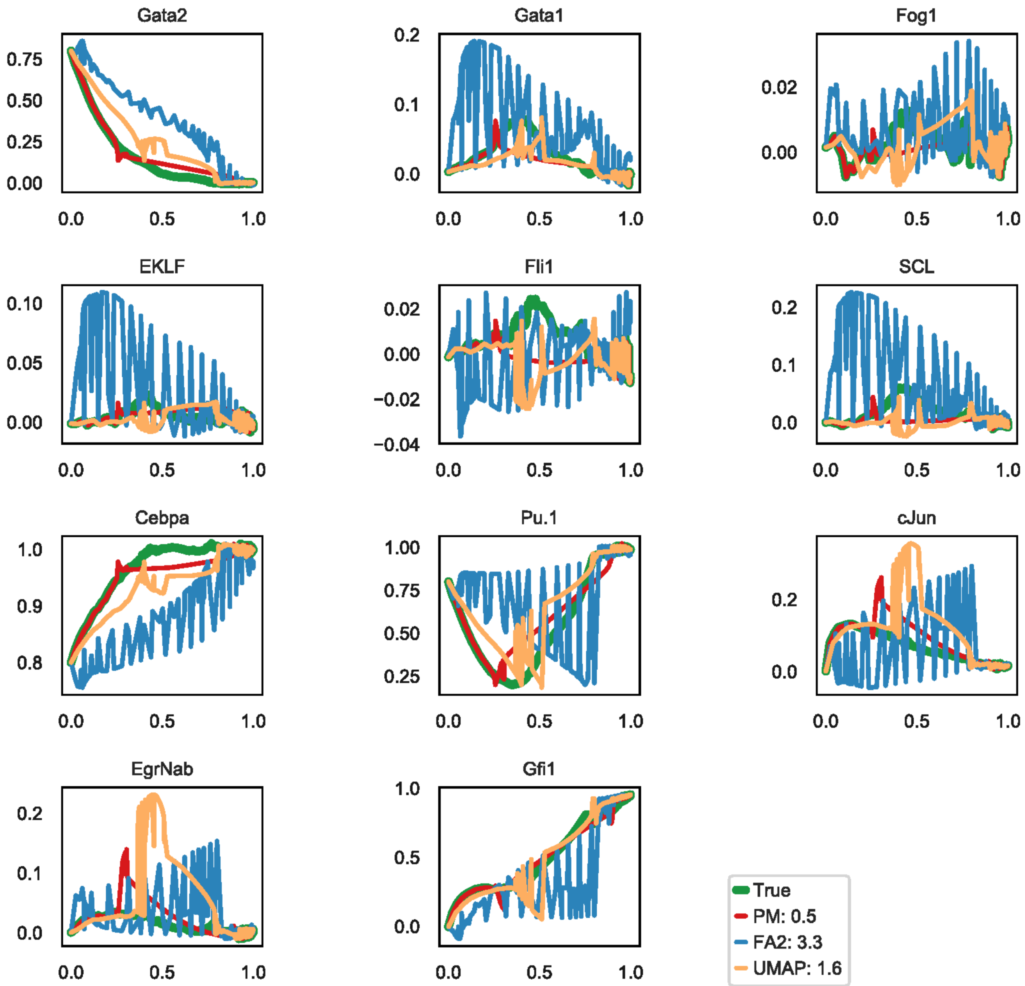
Supplementary Figure 6. Comparison of various embeddings for the mouse myeloid progenitors MARS-seq dataset (Paul et al.). (a) Canonical hematopoetic cell lineage tree. Colored nodes correspond to the population colors from the dataset. White nodes correspond to intermediate annotated states. **(b)** Rotated Poincaré map with respect to the root (medoids of MEP and GMP cluster). **(c)** Hierarchical relationships suggested by the Poincaré map. **(d)** Benchmark methods. To reproduce the Monocle 2 tree, the lymphoid cluster was removed.



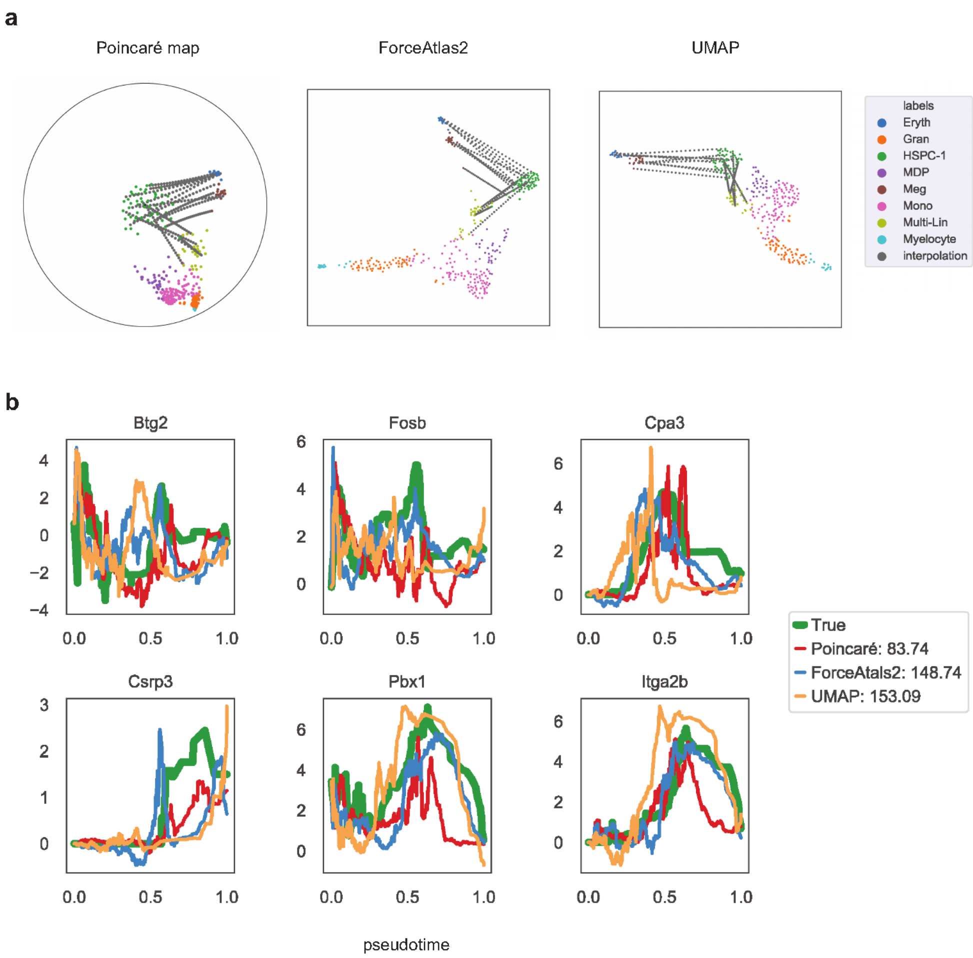
Supplementary Figure 7. Comparison of various embeddings for the planaria Drop-seq dataset (Plass et al.). (a) Poincaré map rotated with respect to the root (medoids of neoblast 1 cluster). **(b)** Benchmark methods.



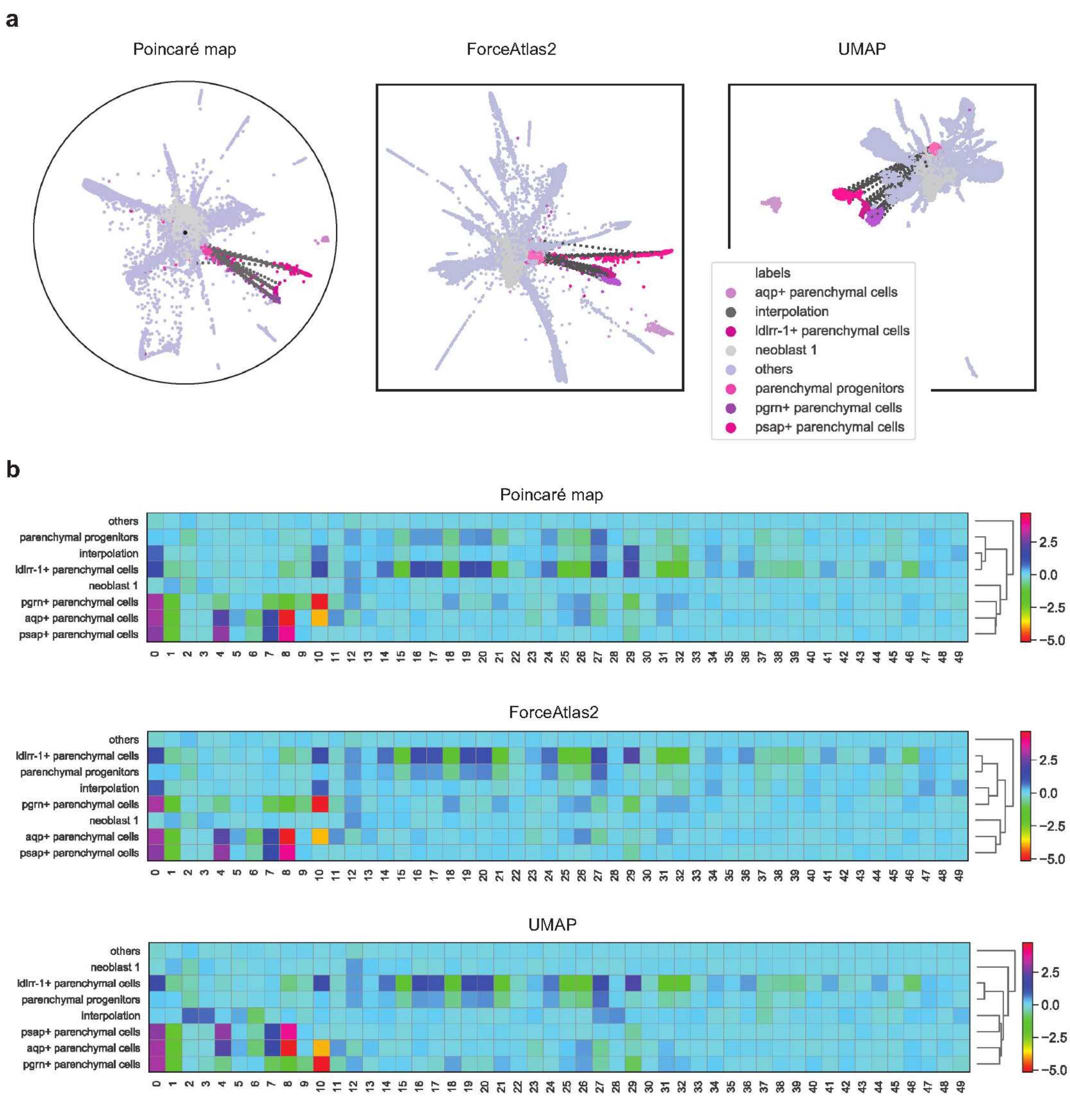
Supplementary Figure 8. Tutorial on the interpolation using Poincaré maps on a simple example of a synthetic dataset of Myeloid progenitors. We remove the majority of neutrophil progenitors "branchNe" with the goal of predicting their gene expression values. **(a)** The perturbed dataset was embedded into the Poincaré disk. We train a neural network to reconstruct the original features from their position in the Poincaré map. **(b)** Poincaré map of the perturbed dataset. **(c)** Louvain clustering of the perturbed dataset. **(d)** Geodesic between two closest points in clusters 1 and 3. **(e)** Points sampled along the geodesic. **(f)** We use the trained neural network to predict values in the original gene expression space for the interpolated cluster.



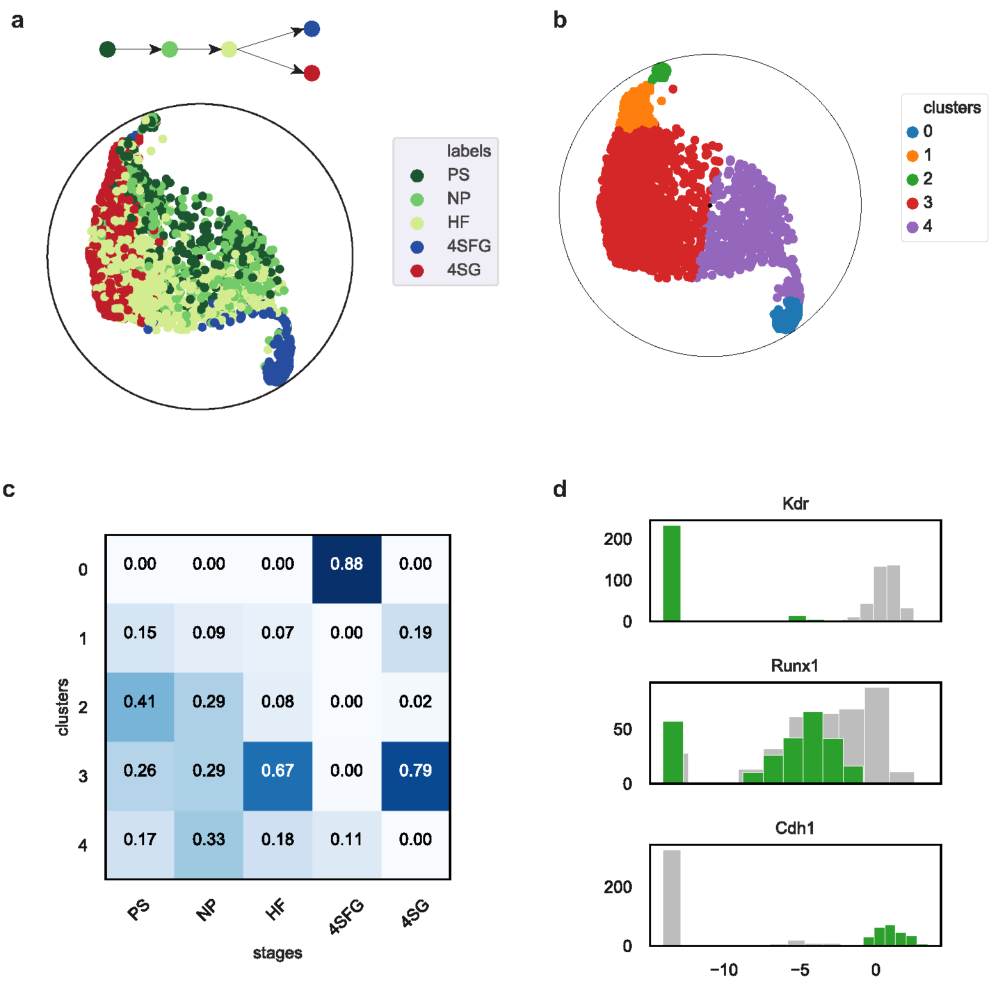
Supplementary Figure 9. Demonstration of interpolation using embeddings on synthetic dataset of myeloid progenitors. The majority of "branchNe" cells were removed with the goal of predicting their values. We predict the cell states interpolating along Poincaré or Euclidean geodesics. The values for the unseen "branchNe" cells were predicted using neural network mapping from the corresponding embedding space to the original gene expression space. We compare the reconstruction of the trajectories using average dynamic time warping measure among individual genes. The ordering of the interpolated lineage is obtained using diffusion pseudotime, and then compared to the diffusion pseudotime of the original dataset.



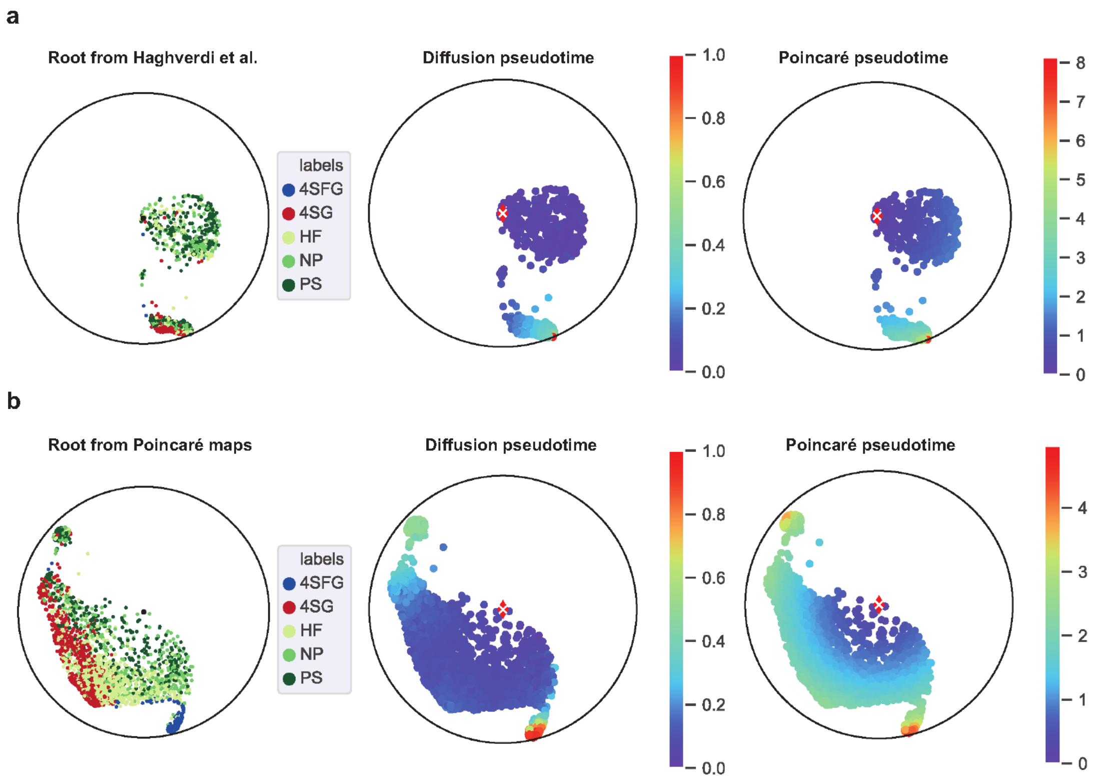
Supplementary Figure 10. Demonstration of interpolation using embeddings on Olsson et al. dataset. The "HSPC-2" cluster was removed completely with the goal of predicting its values. **(a)** We predict the cell states interpolating along the geodesic in Poincaré map, ForceAtlas2 and UMAP. **(b)** We compare the reconstruction of trajectories using average dynamic time warping measure among individual genes. The ordering of the interpolated lineage is obtained using diffusion pseudotime, and then compared to the diffusion pseudotime of the original dataset.



Supplementary Figure 11. Demonstration of interpolation using embeddings on Plass et al. dataset. Part (more developed) of the "parenchymal progenitors" cluster was removed with the goal of predicting its values. We used the preprocessed data provided by the authors. Instead of the original gene expression data, the preprocessed dataset is in the form of 50 principal components, which values we attempt to predict in the interpolation task. **(a)** We predict the cell states interpolating along the geodesic in Poincaré map, ForceAtlas2 and UMAP. **(b)** Comparison of the reconstruction of original features in terms of normalized average expression inside each cluster.



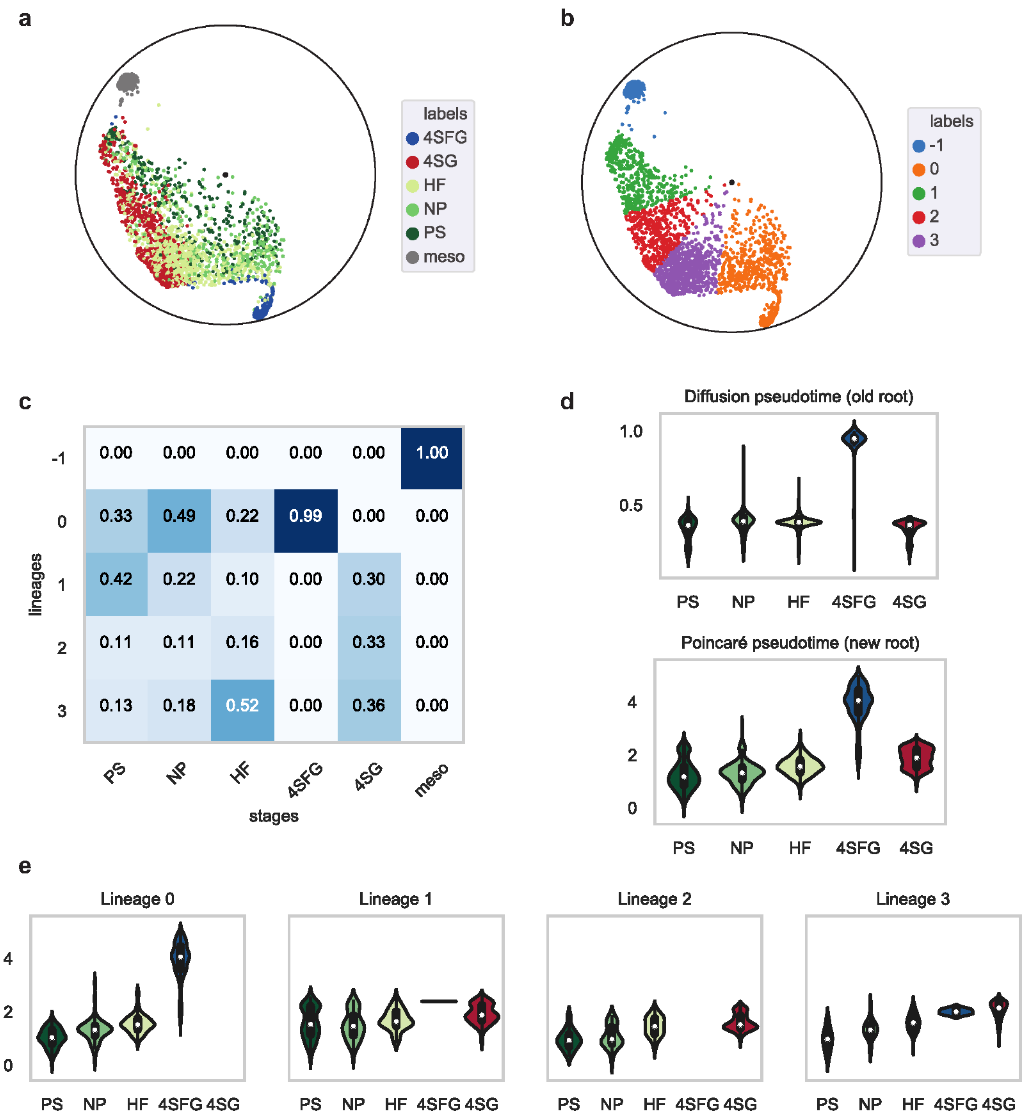
Supplementary Figure 12. **(a)** Poincare map of the Moignard dataset. **(b)** Spectral clustering with Poincaré distances. **(c)** Analysis of stage-composition of the defined clusters. Clusters 1 and 3 most likely represent development of blood cells. Clusters 0 and 4 potentially correspond to endothelial development. Cluster 2 corresponds to the cluster named "mesodermal cells at primitive strike" in the original paper. **(d)** Comparison of the median expression of markers at PS stage for cluster 2 against the rest of PS cells. Cluster 4 consists mostly of Flk1-Runx1-.



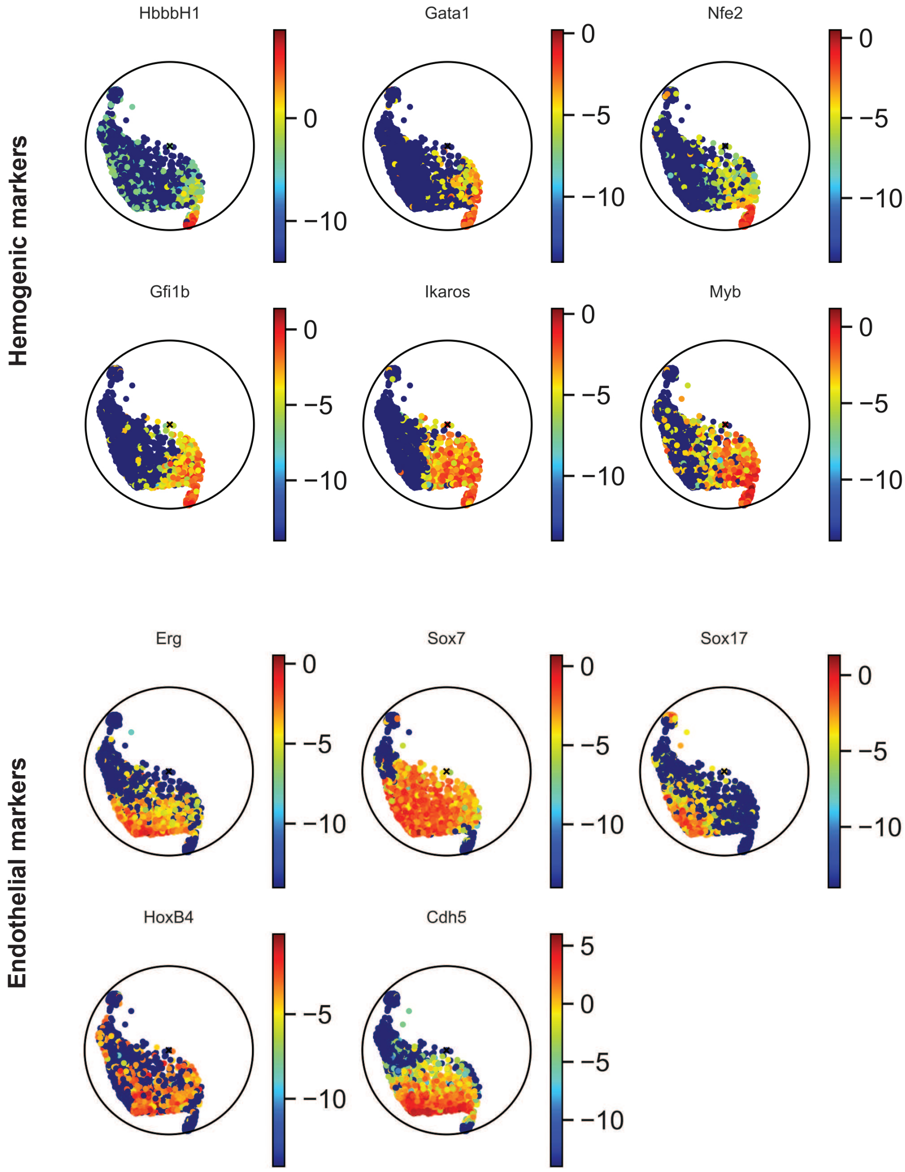
Supplementary Figure 13. Rotation of the Poincaré map, and corresponding pseudotimes with root chosen according to **(a)** Haghverdi et al. **(b)** proposed new root.

Dataset	ToggleSwitch		Myeloid progenitors		MP with blobs	
name	ARS	FMS	ARS	FMS	ARS	FMS
louvain	0.46	0.59	0.58	0.63	0.89	0.91
spectral Poincaré	0.39	0.54	0.63	0.67	0.89	0.91
agglomerative Poincaré	0.49	0.61	0.59	0.64	0.89	0.91
kmedoids Poincaré	0.38	0.53	0.52	0.59	0.54	0.61
spectral raw	0.18	0.39	0.52	0.58	0.28	0.46
agglomerative raw	0.12	0.42	0.54	0.60	0.46	0.64
kmedoids raw	0.19	0.41	0.55	0.61	0.17	0.36
spectral PCA	0.18	0.39	0.53	0.59	0.29	0.41
agglomerative PCA	0.12	0.42	0.48	0.55	0.43	0.60
kmedoids PCA	0.20	0.42	0.49	0.56	0.65	0.70
spectral tSNE	0.47	0.60	0.59	0.64	0.85	0.88
agglomerative tSNE	0.36	0.51	0.49	0.56	0.89	0.90
kmedoids tSNE	0.43	0.57	0.43	0.51	0.71	0.76
spectral UMAP	0.37	0.52	0.42	0.50	0.89	0.91
agglomerative UMAP	0.37	0.52	0.52	0.58	0.89	0.91
kmedoids UMAP	0.31	0.48	0.58	0.63	0.66	0.71
spectral DiffusionMaps	0.55	0.66	0.52	0.58	0.76	0.81
agglomerative DiffusionMaps	0.04	0.39	0.11	0.31	0.12	0.44
kmedoids DiffusionMaps	0.06	0.36	0.42	0.50	0.33	0.50
spectral ForceAtlas2	0.00	0.53	-0.00	0.30	0.00	0.39
agglomerative ForceAtlas2	0.48	0.61	0.56	0.61	0.89	0.91
kmedoids ForceAtlas2	0.42	0.56	0.55	0.61	0.53	0.60

Supplementary Table 1. Comparison of various clustering approaches on the synthetic datasets. Higher values mean better result.



Supplementary Figure 14. Analysis of stage ordering in different lineages. **(a)** Poincaré map with developmental stages. **(b)** Detected lineages with clustering by angle in the Poincaré disk. **(c)** Lineage composition per stage. **(d)** Average diffusion (from Haghverdi et al.) and Poincaré pseudotime per stage for the whole dataset. **(e)** Average Poincaré pseudotime per stage for in the individual lineage.



Supplementary Figure 15. Expression of main endothelial and hemogenic markers visualized on the Poincaré disk.

Dataset	dpt	pmpt	dpt-pmpt
ToggleSwitch: branch1	0.99	0.99	0.99
ToggleSwitch: branch2	0.98	0.98	0.99
ToggleSwitch: avg	0.99	0.98	0.99
MyeloidProgenitors: erythrocyt	0.89	0.94	0.91
MyeloidProgenitors: megakaryoc	0.94	0.95	0.93
MyeloidProgenitors: monocyte	0.93	0.89	0.98
MyeloidProgenitors: neutrophil	0.91	0.91	0.99
MyeloidProgenitors: avg	0.92	0.92	0.95
MyeloidProgenitors with blobs: Ery	0.89	0.94	0.90
MyeloidProgenitors with blobs: Mk	0.94	0.96	0.93
MyeloidProgenitors with blobs: Mo	0.93	0.89	0.97
MyeloidProgenitors with blobs: Neu	0.91	0.91	0.99
MyeloidProgenitors with blobs: avg	0.92	0.92	0.95

Supplementary Table 2. Comparison of diffusion pseudotime (dpt) and Poincaré pseudtotime (pmpt) against real time on synthetic datasets using Pearson correlation coefficient between. The last column corresponds to the correlation coefficient between diffusion pseudotime and Poincaré pseudtotime. Higher values are better.

Dataset	Myeloid progenitors			Olsson			Plass		
method	no	w=5	w=15	no	w=5	w=15	no	w=5	w=15
Poincaré	0.5	0.3	0.4	301.2	138.8	84.6	482.4	239.3	165.0
ForceAtlas2	3.3	1.8	1.3	440.3	198.0	142.8	584.6	273.2	220.0
UMAP	1.6	1.2	0.8	428.8	206.1	153.6	880.9	456.4	368.9

Supplementary Table 3. Dynamic time warping of interpolations using various embeddings. Smaller values are better.

References

[1] Gromov, M. *Metric structures for Riemannian and non-Riemannian spaces* (Springer Science & Business Media, 2007).

[2] Nickel, M. & Kiela, D. Poincaré embeddings for learning hierarchical representations. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems 30*, 6338–6347 (Curran Associates, Inc., 2017). URL http://papers.nips.cc/paper/7213-poincare-embeddings-for-learning-hierarchical-representations.pdf.

[3] De Sa, C., Gu, A., Ré, C. & Sala, F. Representation tradeoffs for hyperbolic embeddings. *arXiv preprint arXiv:1804.03329* (2018).

[4] Nickel, M. & Kiela, D. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *Proceedings of the Thirty-fifth International Conference on Machine Learning* (2018).

[5] Bonnabel, S. Stochastic gradient descent on Riemannian manifolds. *IEEE Trans. Automat. Contr.* **58**, 2217–2229 (2013). URL http://dx.doi.org/10.1109/TAC.2013.2254619.

[6] Maaten, L. v. d. & Hinton, G. Visualizing data using t-sne. *Journal of machine learning research* **9**, 2579–2605 (2008).

[7] McInnes, L. & Healy, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[8] Haghverdi, L., Buettner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods* **13**, 845 (2016).

[9] Wolf, F. A. *et al.* Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *bioRxiv* (2017).

[10] Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one* **9**, e98679 (2014).

[11] Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nature methods* **14**, 979 (2017).

[12] Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 15 (2018).

[13] Wittmann, D. M. *et al.* Transforming boolean models to continuous models: methodology and application to t-cell receptor signaling. *BMC systems biology* **3**, 98 (2009).

[14] Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in escherichia coli. *Nature* **403**, 339 (2000).

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

[16] Krumsiek, J., Marr, C., Schroeder, T. & Theis, F. J. Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network. *PloS one* **6**, e22649 (2011).

[17] Olsson, A. *et al.* Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature* **537**, 698 (2016).

[18] Murphy, K. & Weaver, C. *Janeway's immunobiology* (Garland Science, 2016).

[19] Paul, F. *et al.* Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).

[20] Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049 (2017).

[21] Plass, M. *et al.* Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, eaaq1723 (2018).

[22] Moignard, V. *et al.* Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature biotechnology* **33**, 269 (2015).