HELMHOLTZ
ASSOCIATION
Alliance on Systems Biology
HelmholtzZentrum münchen
German Research Center for Environmental Health

TECHNISCHE
UNIVERSITÄT
MÜNCHEN

Geometric Diffusions for Reconstruction of Cell Differentiation Dynamics

Laleh Haghverdi

September 2016

TECHNISCHE UNIVERSITÄT MÜNCHEN

Helmholtz Zentrum München Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH)

Lehrstuhl M12 (Mathematische Modellierung biologischer Systeme)

Geometric Diffusions for Reconstruction of Cell Differentiation Dynamics

Laleh Haghverdi

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

Univ.-Prof. Dr. Silke Rolles

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Dr. Fabian J. Theis
2. Univ.-Prof. Dr. Massimo Fornasier
3. Univ.-Prof. Dr. Mauro Maggioni (nur schriftliche Beurteilung), Johns Hopkins University, USA

Die Dissertation wurde am 19.09.2016 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 07.12.2016 angenommen.

Preface and acknowledgement

This dissertation is a work on the interface of mathematics and biology. It is an instance of mathematics coming to help our understanding of a biological problem, namely cell differentiation process. My main goal in this work was to a design a mathematical/computational tool that can in practice add to the biological insight we gain from single-cell differentiation data. It is fortunate that our effort has proved useful through several positive feedbacks in various conferences and our publications on the topic have been well received in the computational biology society.

Any interchange between natural sciences and mathematics, not only benefits the natural science by aiding to answer for their questions, but also mathematics. Natural sciences have always served as a genuine source of initiation and motivation for development of several mathematical fields. Architecting a bridge between these two—often decoupled—worlds needs contemplation for both the mathematical concepts and the natural sciences problem. Specialising in only one field does not meet the needs of today's science and technology, and there is a rising demand for scientists who have insight in both subjects. I am glad of achieving a contribution —although small—for such an interchange of knowledge between mathematics and biology; a small bridge which is hopefully well-founded enough to allow coherent biological interpretations and further developments.

I would like to thank my supervisor Prof. Fabian Theis and Dr. Florian Buettner (back then Helmholtz Zentrum Munich and more recently European Bioinformatics Institute) for motivating this project and the trust and flexibility they offered to me in conducting it, giving room to my abilities and covering my weaknesses by their supportive advise.

I thank Prof. Massimo Fornasier, who very kindly first told me anything about diffusion maps (the groundwork for this dissertation) when I approached him with a question after his lecture.

I am grateful to my external supervisor Prof. Mauro Maggioni (back then Duke University and more recently Johns Hopkins University) for his constant support and helpful advise.

I also would like to thank all my coauthors. I thank Lukas Simon and F. Alexander Wolf for the proofreading of my dissertation. Additional thanks to Prof. Fabian Theis for gathering Institute of Computational Biology at Helmholtz Zentrum Muenchen as a friendly and lively place to work at, and all my nice colleges at ICB for their friendly support.

Abstract

Conventional techniques for transcriptional profiling quantify average ribonucleic acid (RNA) abundance levels in large populations of cells. Emerging technologies for singe-cell profiling offer unique opportunities for understanding core cellular processes such as cell differentiation, which cannot be resolved using the conventional ensemble measurements. Typical studies of differentiating cells involve profiled cells form multiple time-points of the differentiation process, thus resulting in snapshot-data of largely unsynchronised cells. To gain insights into the transcriptional dynamics that drive cell differentiation we propose adapting diffusion maps for single-cell data analysis. Our first motivation is benefiting from general properties of diffusion maps such as robustness to noise and taking non-linearities into account when generating a latent space. Second, cell differentiation is a diffusion-like process where starting from a pluripotent stage, cells move smoothly within the transcriptional landscape towards more differentiated states, with some stochasticity along their path. Thus diffusion maps are especially relevant for the analysis of such data intrinsically generated from a biological diffusion-like process. Third, it was previously not clear how to handle missing values and technical uncertainties inherent in the experimental technique such as detection limits or low sensitivities. We encourage application of density-normalized diffusion maps, hence accounting for the non-uniform density of sampled cells in the gene expression space. Moreover, we account for measurement noise and missing values. We further extend the geometric diffusions approach for pseudotime ordering of cells rather than mere data embedding and visualisation via diffusion maps. Our proposed method for diffusion pseudotime ordering of cells (DPT) can also separate several cell fates (branchings) in the data. We demonstrate the application and capabilities of DPT on several single-cell differentiation experimental data sets from various labs. We also clarify

the relationship between pseudotime and actual time measurements through the notion of universal time. That is the deterministic part of the dynamics all cell of the same fate take in the transcriptional space despite their asynchrony and stochasticity in the course of development.

Zusammenfassung

Herkömmliche Techniken zur Messung des Transkriptoms messen die durchschnittliche Konzentration von Ribonukleinsäuren (RNA) in großen Populationen von Zellen. Neue Technologien der Einzelzell-Messung bieten erstmalig Möglichkeiten für ein Verständnis von zellulären Prozessen, wie Zelldifferenzierung, die mit herkömmlichen Messtechnicken nicht adressiert werden können.

Typische Einzelzell-Messungen von differenzierenden Zellen enthalten das Transkriptom von Zellen aus unterschiedlichen Stadien des Differenzierungsprozesses. Es liegt dann eine 'Momentaufnahme' von weitgehend unsynchronisierten Zellen vor. Um Einblicke in die Transkriptionsdynamik während der Zelldifferenzierung zu gewinnen, schlagen wir die Anwendung der 'diffusion map' Methode vor und leiten deren Anpassung für die Einzelzell-Datenanalyse ab. Die primäre Motivation für die Anwendung von diffusion maps liegt in deren allgemeinen Eigenschaften: sie sind robust bei Rauschen und sie berücksichtigen Nichtlinearitäten bei der Erzeugung eines latenten Raums. Eine weitere Motivation stammt aus folgender Überlegung. Die Zelldifferenzierung ist ein diffusionsartiger Prozess, bei dem, ausgehend von einem pluripotenten Zustand, sich Zellen innerhalb der Transkriptionslandschaft stochastisch in Richtung differenzierter Zustände bewegen. Somit sind diffusion maps besonders relevant für die Analyse von Daten, die von einem biologischen diffusionsartigen Prozess erzeugt werden. Schließlich war es bisher nicht klar, wie fehlende Werte und technische Unsicherheiten zu behandeln sind, die in Einzeldaten relevant sind, weil Messtechniken Nachweisgrenzen oder niedrige Empfindlichkeit haben. Wir verwenden Dichte-normierte diffusion maps um heterogenen Samplingbedinungen Rechnung zu tragen, die nur auf experimentelle Bedingungen zurückzuführen sind. Weiterhin berücksichtigen wir Messrauschen und fehlende

Werte. Wir erweitern diffusion maps um Zellen anhand einer ‘Pseudozeit‘ ordnen zu können und führen dazu ein Ähnlichkeitsmaß —’diffusion pseudotime’ (DPT) —ein. Zuvor waren diffusion maps als bloßes Dateneinbettungsverfahren und zur Visualisierung in Verwendung. Mit unserem DPT basiereten Verfahren können wir Verzweigungen des Differenzierungsprozess in mehrere Zellsubtypen auflösen. Wir zeigen die Anwendung und die Möglichkeiten von DPT auf mehreren Datensätzen zur Einzeldifferenzierung. Wir klären auch die Beziehung zwischen Pseudozeit und der tatsächlichen Zeit durch den Begriff der ’universal time’. Diese kann mit dem deterministischer Teil der Zelldynamik in Verbindung gebracht werden, der durch DPT oder ’universal time’ trotz Asynchronität und Stochastizität abgeleitet werden kann.

List of contributed articles

i) M. Bongini, M. Fornasier, F. Fröhlich, and L. Haghverdi, **Sparse stabilization of dynamical systems driven by attraction and avoidance forces.** *NHM, 9(1), pp.1-31(2014).*

ii) V. Moignard, S. Woodhouse, L. Haghverdi, A. J. Lilly, Y. Tanaka, A. C. Wilkinson, F. Buettner, I. C. Macaulay, W. Jawaid, E. Diamanti, S.-I. Nishikawa, N. Piterman, V. Kouskoff, F. J. Theis, J. Fisher, and B. Göttgens, **Decoding the regulatory network of early blood development from single-cell gene expression measurements.** *Nature Biotechnology, 33(3), pp.269-276 (2015).*

iii) L. Haghverdi, F. Buettner, and F. J. Theis, **Diffusion maps for high-dimensional single-cell analysis of differentiation data.** *Bioinformatics, 31(18), pp.2989-2998 (2015).*

iv) A. Ocone, L. Haghverdi, N. S. Mueller, and F. J. Theis, **Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data.** *Bioinformatics, 31(12), pp.i89-i96 (2015).*

v) P. Angerer, L. Haghverdi, M. Büttner, F. J. Theis, C. Marr, and F. Buettner, **Destiny: diffusion maps for large-scale single-cell data in R.** *Bioinformatics, 32(8), pp.1241-1243 (2016).*

vi) L. Haghverdi, M. Büttner, F.A. Wolf, F. Buettner, F.J. Theis, **Diffusion pseudotime robustly reconstructs lineage branching.** *Nature Methods, 13(10), pp.845-848 (2016).*

Contents

Chapter 1

Introduction

In this dissertation the goal was to find the hierarchy of developmental stages in cell differentiation. We wanted to reconstruct the trajectories that cell lineages take in transcriptional space from the pluripotent stem cell to their fully differentiated fate stage. For this, we proposed the use and extension of geometric diffusions. In this chapter characteristics of single-cell differentiation data and existing methods (ranging from preprocessing to clustering and differential gene expression analysis) for their analysis are described as well as our motivation for adaption and application of geometric diffusions.

1.1 Single-cell differentiation data and its analysis

The life of several multicellular organisms including plants and animals starts from solely a single cell called zygote. The zygote divides several times to produce more and more cells which constitute the embryo. Going through multiple divisions and a hierarchy of several developmental stages, developing cells acquire their specific functional specialisation to make up the complex system of tissues and cell types in a mature organism. Such a process that roots in the same single cell (called the pluripotent cell) and results in several cells each with a specific fate, is called differentiation. Throughout the differentiation process, the whole genome of the original pluripotent cell is passed to all the later progeny cells. The rate of accumulating

mutations through cell proliferation (in healthy cells) is too low to affect anything on the gene expression level. It is thus merely by changing the gene expression profile that different cells acquire different fates.

Differentiation is generally known as a directed process. Differentiated cells do not naturally revert to earlier stages of development (with few exceptions [1, 2]). An important property of cell differentiation is the asynchrony of development among individual cells. Reaching the same state of development can take long for one cell and short for the other. This makes conventional ensemble gene expression measurement techniques such as bulk RNA-Seq and bulk qPCR measurements as used in [3] and [4] inefficient, as these techniques average out all the heterogeneity of expression present among single cells [5, 6]. It was only in the last decade that the advent of new technologies enabled gene expression measurements at the single-cell level [7, 8], thus the heterogeneity of expression in a population of cells can be resolved. Typical single-cell measurement techniques provide gene expression profiles for several cells (the number of cells and genes measurement capacity depends on the specific technique being used) that are captured on one or few time points, hence providing one or few snapshot data sets. A challenging property of single-cell data is the high level of noise. Because of low copy numbers of messenger RNA (mRNA) and proteins in single cells, such measurements are very imprecise and suffer high levels of noise and missing values as usually several amplifications of the gene product are required for detecting its expression [9, 10, 11]. This amount of inaccuracy and noise is specific to single-cell measurements does not exist in classical bulk measurements, as there the expression from many cells is pooled together resulting in sufficient amount of molecules for detection. In single-cell data however, it is quite common that several genes expression levels fall below the sensitivity and thus the limit of detection of the single-cell measurement technique, leading to prevalence of undetected and missing values.

Single-cell measurement technologies fall into five major categories described below:

i. Microscopy and cell imaging [12, 13]

   By integration of genetically encoded fluorescent proteins to specific genes of interest, gene expression level can be measured by fluorescent intensities observed by cell microscopy. A popular cell imaging technique, fluorescent in situ

hybridyzation (FISH) [13] uses fluorescently tagged oligonucleotide probes to mark expressed mRNA molecules in single cells. Microscopy and cell imaging technologies provide the opportunity to measure gene expression of single cells in the course of actual time hence obtaining expression time-series. Microscopy is in fact the only single-cell technology in which the actual time resolution of measurements is not lost. In certain types of experiments cells are not fixed in their location. To obtain the expression time-series of a lineage thus requires computational tools for image processing and cell tracking. Cell imaging techniques are currently very limited in the number of proteins they can monitor (maximum four proteins). Cell tracking [14, 15] is also a computationally challenging task especially in presence of cell proliferation. Nonetheless, the field of cell tracking methodologies is developing rapidly. Furthermore, the asynchronous development during cell differentiation necessitates further computational effort for analysis of time-lapse data from imaging techniques since naive averaging over time points is not suitable (see the Suppl. Note 6 in [16]).

ii. Flow (and mass) cytometry [17, 18]

In these technologies cell surface proteins (i.e. proteins that are bound to the cell membrane) are marked, either by fluorescent antibodies (in flow cytometry) or isotopically pure heavy elements (in mass cytometry). The abundance of the surface protein in a cell is then inferred from the signal intensity of each marker . An immediate limitation of cytometry technologies is that only surface proteins can be probed. The number of markers that can be measured with flow cytometry is rather limited in a single experiment (typically below 20) due of interference of signals from different fluorescence channels [17]. Mass cytometry does not suffer this kind of signal leakage and can therefore measure a larger number of markers (around 100). The number of single cells measured by this technique can be up to millions.

iii. Single-cell quantitative polymerase chain reaction (sc-qPCR) [19]

This technique amplifies copies of pre-specified nucleic acids (gene products such as messenger RNAs and microRNAs) across several orders of magnitude, which facilitates the measurement of relative gene expression on the single-cell level. The number of genes probed with this technique is typically below a hundred, because pre-targeting of the sequence of nucleic acids to be amplified is needed

which is time consuming and rather expensive. The number of cells that can be profiled with this technique ranges between hundreds to few thousands of cells.

iv. Single-cell RNA sequencing (scRNA-Seq) [20]

Single-cell RNA sequencing technologies include a wide range of protocols and teqniques such as Fluidigm [21], inDrops [22], SMART-Seq [23], MARS-Seq [24], etc. In scRNA-Seq, the whole mRNA content of single cells is amplified using PCR to reach the level of detection necessary for the subsequent sequencing. Each sequenced read is then aligned to the genome and gene expression levels are inferred from this alignment [25]. RNA-Seq techniques enable measurement of thousands of genes as no it a priori selection of genes is needed. The general drawback of scRNA-Seq approaches is the low accuracy and prevalence of non-detected genes (usually termed drop-outs). The accuracy as well as cell number capacity is quite variable among several protocols. For a discussion on computational challenges of scRNA-Seq analysis see [26, 27].

Despite the variation in sensitivity, accuracy and capacity in the number of cells and genes they can monitor, all the techniques listed above (except single-cell microscopy which can track back the ancestry tree for cell and thus provide measurements on the course of actual time) supply a n by G matrix (n being the number of cells and G being the number of probed genes) data, for which the actual time development history of each cell is lost.

Figure 1.1 shows a general scheme for the analysis of single-cell snapshot data. First, a preprocessing of the noisy expression profiles is required [27]. [28] and [29] are instances of single-cell data preprocessing which apply control measures for the amount of noise in each measurement and therefore identify very noisy genes to be excluded from the downstream computational analysis. Algorithms such as scLVM [30] and OEFinder [31] are commonly used for removing confounding sources of variation (cell-cycle, batch effect between several days of measurement, etc.) from the data. Other statistical methods such as BASiCS [32] provide normalisation methods for single-cell data. After preprocessing, a clustering algorithm is applied to resolve the heterogeneity of gene expression among the cells monitored in the snapshot(s). Application of hierarchical clustering algorithms [33], k-means [34], partitioning around medoid (PAM) [35] on low-dimensional embeddings of data

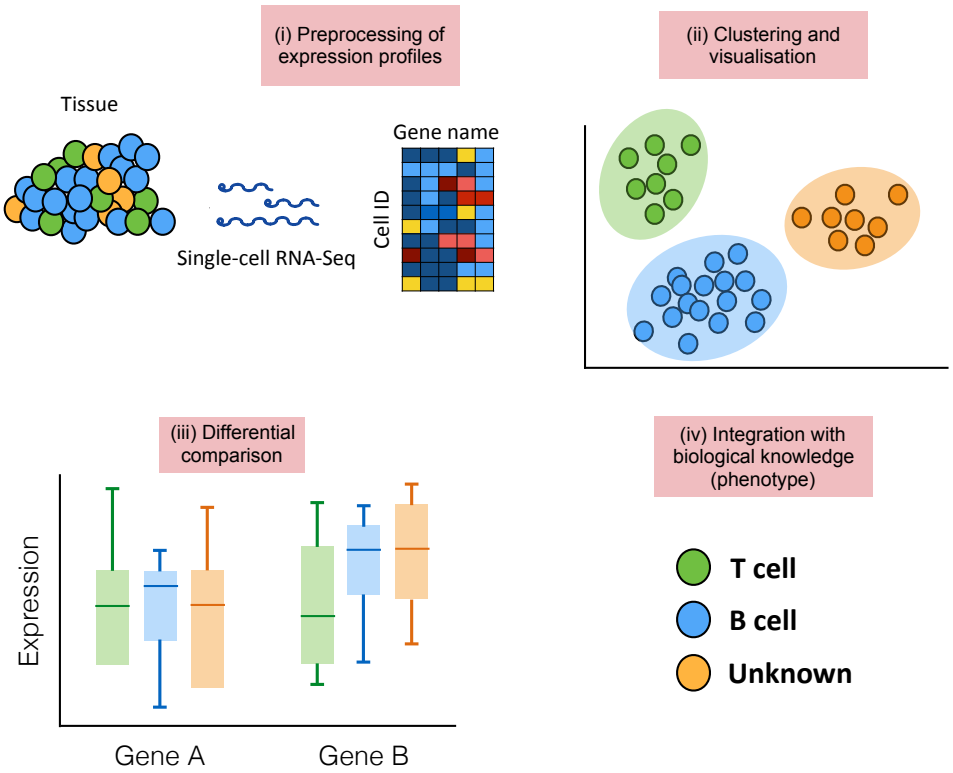


Figure 1.1: A general scheme of single-cell data analysis. A snapshot of a heterogeneous collection of cells (e.g. from a tissue) is captured providing the expression profiles (e.g. single-cell RNA-Seq) of the sampled cells. Preprocessing is essential for bringing experimental measurements into a useful and interpretable form. Preprocessing includes a variety of algorithms for sequence alignment, data normalisations, denoising etc. In the next step data visualisation and clustering is applied. Next, the set of differentially expressed genes between each pair of clusters is identified and matched to the phenotype those genes are associated with. The phenotype might include a wide range of biological knowledge such as genes function or disease association. With this biologists can address the biological functionality of each of the identified clusters (for example T cells, B cells, etc.).

obtained by low-dimensional embedding tools such as principle component analysis (PCA) [36], t-SNE [37] and gaussian process latent variable models (GPLVM) [38] are commonly used for this purpose. Once several clusters of cells are identified, differential gene expression analyses can elucidate the set of genes whose expression characterizes each cluster [39]. Finally, by studying the functionality (or phenotypic associations) of the differentially expressed genes, biologists can infer differences in function between cell clusters. Typically gene ontology and annotation data bases [40, 41] are used for understanding such phenotypical association.

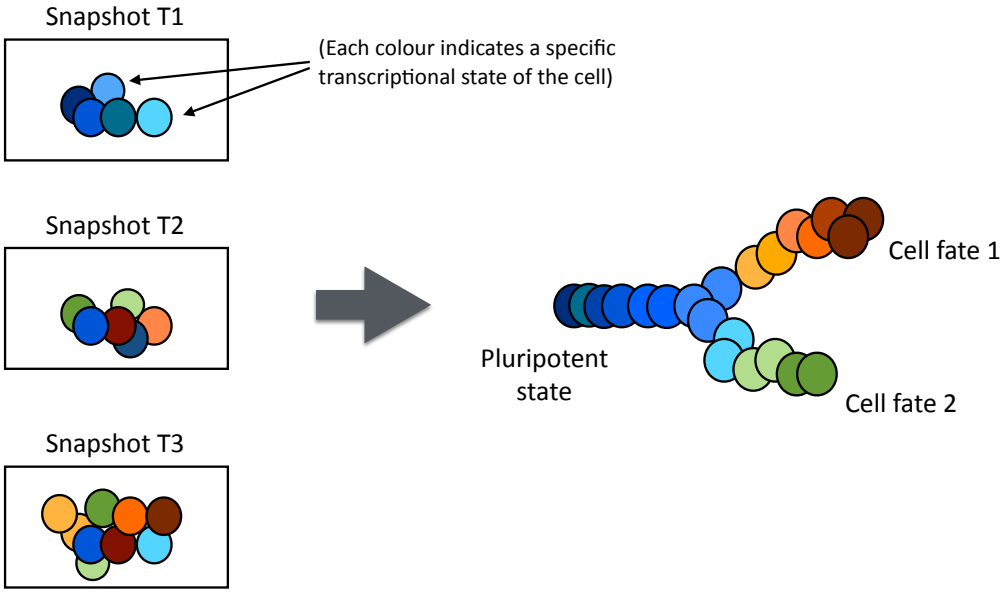


Figure 1.2: Pseudotime ordering of cells from snapshot single-cell expression data. Each time point snapshot measurement consists of a highly heterogeneous collection of cells with respect to their progression along the differentiation hierarchy (each colour indicates a specific transcriptional state of the cell) and hence their expression profiles. The heterogeneity at each time point is due to the asynchronous development among individual cells as well as branching events. As a result, there is little biological relevance in the capture time of each cell (i.e. at which time point snapshot it was collected). Instead, it is desirable to arrange the cells according to the progression stage providing a pseudotime ordering that also recovers the data branching events.

This work focuses on the clustering/visualisation step. Given the continuous nature of developmental processes including differentiation, one might assume the existence of a continuous gene expression manifold as opposed to separate clustering. Thus our goal was to arrange each cell present in the snapshot data (as if a piece of a puzzle) in its correct place with respect to other measured cells in the hierarchy of development. As a result we wish to get a complete and correct picture of the differentiation tree through such ordering, commonly termed *pseudotime* ordering (Figure 1.2).

1.2 Previous methods for cluster analysis and pseudotime ordering of single-cell differentiation data

In the recent years several methods have been proposed for cluster analysis of single-cell differentiation data [42, 43, 44, 45, 46, 47]. Comparison to all proposed methods is not in the scope of this work. For reviews over several methods please see [26, 27, 48]. We compared the proposed method of this work in [49] (Appendix A) and [16] (Appendix B) to the most popular ones which are discussed as follows.

The first method which was proposed for cluster analysis and pseudotime ordering of single-cell differentiation data is SPADE [50]. It was originally developed for the analysis of single-cell flow cytometry data of 13 surface markers in 2011. SPADE uses a hierarchical clustering on cells in the high-dimensional space and then constructs a minimum spanning tree on the clustered data. Hierarchical clustering however, has a highly varying random component which changes in each run of the algorithm. Furthermore, putting a minimum spanning tree on the few clusters in the high-dimensional transcriptional space is very frail and prune to curse of dimensionality effects [51]. In fact the higher the number of dimensions in the data, the more badly affected SPADE is by curse of dimensionality effects. Thus SPADE fails to perform reliably and robustly on high dimensional data to the extent that each run of the algorithm results in a completely different differentiation tree (see Fig. S17 in [49] (Appendix A) and Fig. M14 in [52] for instances).

viSNE [53] is another algorithm applied for analysis of single-cell data published in 2013. It is an extension of the t-distributed stochastic neighbours embedding (t-SNE) [37] to deal with large cell numbers as obtained by several single-cell technologies (e.g. cytometry, scRNA-Seq). t-SNE is a powerful clustering algorithm, however it does not preserve the global structure of the data which is definitely of interest in the case of differentiation data where one wishes to detect the (continues) developmental trajectories that several cell lineages have taken. t-SNE instead breaks the data into separate clusters for which the geometrical/developmental relations are not clear. See Figs. 4-8 in [49] (Appendix A).

In 2014 Monocle [54] was published for the analysis of few hundreds of single cells' RNA-Seq expression. Monocle first performs independent components analysis (ICA) for dimensionality reduction. Next, it builds a minimum spanning tree and computes the PQ tree [55] for the reduced dimensions. Pseudotime order and the branchings of the data are then reported by the organization of the cells along the PQ tree. Unlike SPADE, Monocle overcomes the curse of dimensionality effects by building its minimum spanning tree on the reduced dimensions. However the dimension reduction method it uses (ICA) is linear, whereas nonlinearities are crucial for reconstruction of differentiation paths especially with relatively few measured genes (e.g. sc-qPCR data). Furthermore, as we discuss in [16] (Appendix B), minimum spanning trees are not robust to noise. On the computational aspect, Monocle can only handle approximately (depending on the complexity of the hierarchy structure that needs to be captured by the PQ tree) 1000 cells. These features make Monocle inappropriate for analysis of several single-cell technologies data with rather few probed genes or large cell numbers.

Another method introduced in 2014 for analysis and pseudotime ordering of single cells is Wanderlust [52]. It first builds the nearest neighbours graph on the data in high dimensions and then tries to compute a geodesic distance (by summing the distances on the nn-graph) of cells with respect to a pre defined root cell. To build this geodesic distance, Wanderlust relies on the assumption of non-branching data. For the analysis of branching data, Wanderlust was extended to Wishbone [56] in 2016. In contrast to the coherent structure of Wanderlust, Wishbone is a complex collage of several unrelated algorithmic steps (including denoising by diffusion maps, pseudotime ordering by Wanderlust, branch finding through normalised cut graph segmentation [57] on the disagreement of Wanderlust pseudotimes from several runs, refinement of pseudotime after branch finding and visualizing on t-SNE plots), which in the end leaves the algorithm ad-hoc and error prone. We show the comparison of results from our method (DPT) to Wishbone in Suppl. Note 7 [16](Appendix B). Suppl. Note 7 in [16] also includes detailed methodology comparisons of Wishbone and Monocle to DPT.

As discussed in this section, none of the discussed methods deals with all features of single-cell differentiation data (e.g. high level of noise, non-linearity, continuity of trajectories, etc.) properly. Although these methods have shown to be useful for

specific data sets, they lack the general applicability to other data sets.

1.3 Geometric diffusions approach for cluster analysis and pseudotime ordering of single-cell differentiation data

Because of the demand for a clustering method which deals with single-cell differentiation data properly, in 2015 we published [49] proposing that diffusion maps provide a suitable dimension reduction and data visualisation tool for such data (see Appendix A). Diffusion maps are based on random walks (diffusion), a concept that is intrinsically related to the biological process of differentiation, where starting from the pluripotent state cells follow a stochastic (directed) random-walk like trajectory until the fully differentiated fate. Thus, Euclidean distances on the diffusion map provide a biologically relevant measure for the affinities cells experience between several states in the transcriptional space. An important feature of diffusion maps which makes them suitable for single-cell differentiation data analysis is the robustness to noise. While single-cell data is in general too noisy for several machine learning algorithms (minimum spanning tree, principal component analysis, etc.), diffusion maps provide considerable robustness to noise by considering several random walk paths between each pair of cells. In the main and supplementary text of [49] (Appendix A) we demonstrate superiority of diffusion maps to several other approaches including other Laplacian eigenmaps. Since then diffusion maps and other Laplacian eigenmaps have been implemented in several later appearing methods e.g. [56] and [58].

In 2016 we extended our application of diffusion maps from just mapping and visualisation of data to pseudotime ordering and branch identification [16]. By considering all possible connecting paths between each pair of cells, we built a robust on-manifold distance (we called it the diffusion pseudotime metric, dpt), which was used for pseudotime ordering, branch finding and data visualization (the dpt distance has the same low-dimensional embedding as diffusion maps, see Suppl. Note 1.2 in [16]). We also clarified the relation of the dpt distance, hence our pseudotime,

to actual time measurements ([16] online methods), whereas such a relationship was not clear for previous methods.

Chapter 2

Methodology

Laplacian eigenmaps set up the groundwork for this dissertation, on which the diffusion maps is built. Thus, this chapter begins with theoretical basis for Laplacian eigenmaps with highlighting the properties that are important for this specific work. Then, it is described how we can use a diffusion metric for learning the geometry and branching events of a data manifold. We also describe our definition of universal time [16] which makes a connection between actual time measurements and the pseudotime we infer using geometric diffusions.

2.1 The Laplacian and transition matrices

Consider the directed connected weighted graph $G = (V, E)$, where V and E denote the nodes and the edges of the graph. The adjacency matrix W is defined such that W_{xy} shows the directed weight of edge xy. The laplacian matrix L is defined as:

$$L = D - W \tag{2.1}$$

where the degree matrix D is diagonal with $D_{xx} = \sum_y W_{xy}$. One can define a normalised version of the laplacian matrix in two ways:

i. The random-walk normalized Laplacian matrix defined as: $L^{rw} = D^{-1}L = I - D^{-1}W$

ii. The symmetric normalized Laplacian matrix defined as: $L^{sym} = D^{-1/2}LD^{-1/2} = I - D^{-1/2}WD^{-1/2}$

We also define the respective transition matrices $T^{rw} = D^{-1}W$ and $T^{sym} = D^{-1/2}WD^{-1/2}$ for later use in the diffusion maps frame work. As different only by the unity matrix, it is clear that the transition matrices share the same eigenvectors with their respective normalised Laplacian matrix and their eigenvalues (λ_i) are related to the eigenvalues of the normalised Laplacian matrix (γ_i) through: $\gamma_i = 1-\lambda_i$. Relation of graph Laplacians with the Fokker-Planck equation allows their application for describing the probability distribution p of a random walk on the set of the graph nodes [59]. The Fokker-Planck equation describes the probability distribution p of a random walk on the set of the graph nodes . The Fokker-Planck equation has the following general form:

$$\frac{\partial}{\partial t}p = \nabla \cdot (\nabla \frac{1}{\beta}p + p\nabla U) = \frac{1}{\beta}\Delta p + \nabla \cdot (p\nabla U) \tag{2.2}$$

where β is a thermal factor describing the swiftness of diffusion and U is the potential energy of the nodes. The first term describes the diffusion of the probability distribution p, and the second term corresponds to the (deterministic) forces acting on p. In presence of potential energies (i.e. second term), Equation 2.2 defines a directed flow of probabilities and generates an asymmetric transition (as well as Laplacian) matrix. Coifman *et al.* [59] defined geometric diffusions using the relation

$$\frac{\partial}{\partial t}p = pL^{rw} \tag{2.3}$$

as the asymmetric L^{rw} defines a proper random walk on the graph G with conservation of (outflow) probabilities. In contrast to L^{rw}, the symmetric normalised Laplacian L^{sym} does not define a random walk with conserved probabilities as the row of T^{sym} do not sum to one. However it still defines a time evolution process which allows creation and annihilation of probability p' on the graph, thus:

$$\frac{\partial}{\partial t}p' = p'L^{sym} = L^{sym}p' \tag{2.4}$$

Eventhough L^{sym} does not preserve the outflow probabilities, as we will see is section 2.3, it brings a level of clarity in the mathematics and interpretations of geometric

diffusions. Also in cases where keeping the probability conservation property is a big concern, one can still keep the symmetric form of T^{sym} and only compensate the deviance of the row (and column) sums from one on its diagonal as potential energy of the nodes.

It is interesting to note that L^{rw} and L^{sym} are related to each other by a rotation:

$$L^{sym} = D^{-1/2}LD^{-1/2} = D^{1/2}(D^{-1}L)D^{-1/2} = D^{1/2}(L^{rw})D^{-1/2} \tag{2.5}$$

This shows that L^{rw} and L^{sym} have the same eigenvalues. However, unlike the symmetric L^{sym}, the set of right and left eigenvectors differ from each other for L^{rw}. This implies slight variation for some of the calculations when using either form. We take care of such differences with detailed discussion in such necessary occasions.

2.2 Laplacian eigenmaps as a valid data embedding tool

Laplacian eigenmaps use the Laplacian matrix of a graph built on a set of uniformly sampled data from a manifold for providing an approximation to the Laplace-Beltrami operator. It has been shown that transformation of data to the first eigenvectors of the Laplacian provides a valid data embedding tool [60]. Later in [59] Coifman *et al.* showed that Euclidean distances between data points in the coordinates of the first eigenfunctions, indeed show us the low-dimensional (i.e. embedded) approximation of a metric they defined as *diffusion distance* (see [59] or [49] (Appendix A) for details). In this section we overview the optimal embeddings proof from [60]:

Consider the connected wighted graph $G = (V, E)$ with edge weights given by the matrix W again. If we want to map the graph nodes to $\mathbf{f}$ in k dimensions, a valid objective function (which penalizes falling apart of close-by nodes on the mapped space) would be:

$$\sum_{x,y}(\mathbf{f}_x - \mathbf{f}_y)^2 W_{xy} \tag{2.6}$$

where $\mathbf{f}_x$ represents the map of node x. The objective function in 2.6 implies that the pair of nodes with larger transition weight, stay closer to each other in the map, hence will have smaller distance $||\mathbf{f}_x - \mathbf{f}_y||$. The objective function can be written as $2\mathbf{f}^T L\mathbf{f}$, because:

$$\sum_{x,y}(\mathbf{f}_x - \mathbf{f}_y)^2 W_{xy} = \sum_{x,y}(\mathbf{f}_x^2 + \mathbf{f}_y^2 - 2\mathbf{f}_x\mathbf{f}_y)W_{xy} =$$
$$\sum_x \mathbf{f}_x^2 D_{xx} + \sum_y \mathbf{f}_y^2 D_{yy} - 2\sum_{x,y}\mathbf{f}_x\mathbf{f}_y W_{xy} = 2\mathbf{f}(D - W) = 2\mathbf{f}^T L\mathbf{f} \tag{2.7}$$

Given that all weights W_{xy} are non-negative, the relation $\sum_{x,y}(\mathbf{f}_x - \mathbf{f}_y)^2 W_{xy} = 2\mathbf{f}^T L\mathbf{f}$ proves that L is positive semidefinite. To remove an arbitrary scaling factor in the embedding, an additional constraint $\mathbf{f}^T D\mathbf{f} = 1$ is needed. This constraint together with equation 2.7 reduces the minimization of the objective function to solving for $\mathbf{f}_i$ in:

$$L\mathbf{f}_i = \gamma_i D\mathbf{f}_i \tag{2.8}$$

Thus the right eigenvectors of $L^{rw} = D^{-1}L$ (call them ψ_i corresponding to the ordered eigenvalues $0 = \gamma_0 \leq \gamma_1 \leq ... \leq \gamma_{n-1}$) provide a valid data embedding. We only need to take especial care for the zeroth eigenvector which corresponds to the trivial eigenvalue $\gamma_0 = 0$. This eigenvector is a constant vector equal to $\mathbf{1}$. Thus it is non-informative and we need to excluded it from the minimization. In fact this is why we start counting the eigenvalues/eigenvectors from zero; the zeroth eigenvector $\boldsymbol{\psi}_0$ has to be excluded from several calculations through this work as it is noninformative.

Diffusion maps [59, 61] are a special case of Laplacian eigenmaps where the adjacency matrix on the graph is built using Gaussian kernels. Furthermore in [59] the authors also show how density normalisation can be applied to diffusion maps to approximate the Laplace-Beltrami operator in spite of non-uniformly sampled data from a manifold.

2.3 The kernel in the Laplacian and transition matrix

In [49] (Appendix A), based on that the Gaussian kernel is the product of two Gaussian distributions, we generalised the form of the kernel K on which the Laplacian and transition matrices are built to more general forms keeping the distributions product structure:

$$K(\mathbf{x}, \mathbf{y}) = \int_{-\infty}^{\infty} Y_{\mathbf{x}}(\boldsymbol{x'}) Y_{\mathbf{y}}(\boldsymbol{x'}) \mathrm{d}\boldsymbol{x'} \tag{2.9}$$

where $Y_{\mathbf{x}}$ and $Y_{\mathbf{y}}$ can be any (even distinct from each other) distribution functions around the position of cells $\mathbf{x}$ and $\mathbf{y}$ in the gene expression space with the condition that:

$$\int_{-\infty}^{\infty} Y_{\mathbf{x}}^2(\boldsymbol{x'}) \mathrm{d}\boldsymbol{x'} = 1 \text{ and } \int_{-\infty}^{\infty} Y_{\mathbf{y}}^2(\boldsymbol{x'}) \mathrm{d}\boldsymbol{x'} = 1.$$

This normalisation condition is chosen such to satisfy $K(\mathbf{x}, \mathbf{x}) = 1$ for self-transition weights. While Gaussian kernels are in general a suitable choice because of their localised property (e.i. exponential drop) and the computational advantages, implementation of the general form of a kernel in the construction of the Laplacian and transition matrices can provide advantages in some cases for instance in dealing with missing values as we proposed in [49] (Appendix A). There we propose integration of a uniform distribution over the range of all possible expression values whenever any missing value is encountered. One may also consider any other a priori estimated distribution for the missing values. An alternative sensible choice for the missing values might be a Gaussian distribution with its mean and variance calculated from a group of nearest neighbours of the cell (with the missing value for a gene) in the genes space. Furthermore in [16] Suppl. Note 1.1 we use the general kernel form of equation 2.9 to allow interference (i.e. product) of two Gaussian distributions with distinct kernel widths, hence constructing a locally scaled Laplacian and transition matrix.

2.4 Construction of diffusion distances

Coifman *et al.* [59, 61] used T^{rw} with the corresponding right and left set of eigenvectors $\boldsymbol{\psi}_i$ and $\boldsymbol{\phi}_i$, $i = 0, .., n-1$, to define a metric called diffusion distance. In their work the relation $\boldsymbol{\psi}_i(z) = \boldsymbol{\phi}_i(z)/\boldsymbol{\phi}_0(z)$ holds between the right and the left eigenvectors because:

$$\sum_i a_i e^{-\lambda_i t}\boldsymbol{\phi}_i = e^{-U}\sum_j e^{-\lambda_j t} b_j \boldsymbol{\psi}_j, \tag{2.10}$$

a_i and b_j being constant coefficients, which means $\boldsymbol{\phi}_i = e^{-U} b_i \boldsymbol{\psi}_i$ up to a normalisation constant and $e^{-U} = \lim_{t\to\infty} p = \boldsymbol{\phi}_0$ is the steady state solution of equation 2.3. The normalisation of the eigenvectors is chosen such that $\sum_{z=1}^{n} \boldsymbol{\phi}_i^T(z)\boldsymbol{\psi}_i(z) = 1$. Consequently Coifman *et al.* define a distance measure in such a way that it can be expressed by the right eigenvectors of T^{rw}. This distance measure depends on a time scale parameter (or length of a random walk) t. Using the spectral decomposition of T^{rw},

$$(T^{rw})^t(x,y) = \sum_{i=0}^{n-1} \lambda_i^t \boldsymbol{\psi}_i(x)\boldsymbol{\phi}_i^T(y) \tag{2.11}$$

and the diffusion distance D_t^{rw} is defined as:

$$\begin{aligned}(D_t^{rw})^2(x,y) &= ||(T^{rw})^t(x,.) - (T^{rw})^t(y,.)||^2_{1/\boldsymbol{\phi}_0} = \sum_{z=1}^{n} \frac{(T^{rw})^t(x,z) - (T^{rw})^t(y,z)}{\boldsymbol{\phi}_0(z)} \\ &= \sum_{z=1}^{n} \frac{\boldsymbol{\phi}_i(z)^2}{\boldsymbol{\phi}_0(z)} \sum_i \lambda_i^{2t}(\boldsymbol{\psi}_i(x) - \boldsymbol{\psi}_i(y))^2 = \sum_{z=1}^{n} \boldsymbol{\phi}_i(z)\boldsymbol{\psi}_i(z) \sum_{i=0}^{n-1} \lambda_i^{2t}(\boldsymbol{\psi}_i(x) - \boldsymbol{\psi}_i(y))^2 \\ &= \sum_{i=0}^{n-1} \lambda_i^{2t}(\boldsymbol{\psi}_i(x) - \boldsymbol{\psi}_i(y))^2, \end{aligned} \tag{2.12}$$

which reduces to

$$(D_t^{rw})^2(x,y) = \sum_{i=1}^{n-1} \lambda_i^{2t}(\boldsymbol{\psi}_i(x) - \boldsymbol{\psi}_i(y))^2 \tag{2.13}$$

because $\boldsymbol{\psi}_0(x) = 1$ for all x.

Similarly, it is possible to build a version of diffusion distance based on the T^{sym} matrix, for which $\boldsymbol{\psi}_i(z) = \boldsymbol{\phi}_i(z)$. In this case we do not need the $1/\boldsymbol{\phi}_0$ normalisation of the norm anymore as now we simply have $\sum_{z=1}^{n} \boldsymbol{\phi}_i^2(z) = 1$:

$$(T^{sym})^t(x,y) = \sum_{i=0}^{n-1} \lambda_i^t \boldsymbol{\phi}_i(x) \boldsymbol{\phi}_i^T(y) \tag{2.14}$$

$$\begin{aligned}(D_t^{sym})^2(x,y) = ||((T^{sym})^t(x,.) - (T^{sym})^t(y,.)||^2 &= \sum_{z=1}^{n} (T^{rw})^t(x,z) - (T^{rw})^t(y,z) \\ &= \sum_{z=1}^{n} \boldsymbol{\phi}_i(z)^2 \sum_i \lambda_i^{2t} (\boldsymbol{\phi}_i(x) - \boldsymbol{\phi}_i(y))^2 \\ &= \sum_{i=0}^{n-1} \lambda_i^{2t} (\boldsymbol{\phi}_i(x) - \boldsymbol{\phi}_i(y))^2. \end{aligned} \tag{2.15}$$

In [49] we used the random-walk normalized Laplacian matrix as proposed by Coifman *et al.* [59]. Later on (as in [16]) we recognised it is worth to simplify the calculations and help clarity of interpretations from our method by using the L^{sym} version. The transition matrix we define on a graph of snapshot cells is merely based on the geometrical positions of cells and their proximity in the genes space and thus is completely symmetrical. In other words establishing a random walk interpretation and a Fokker-Planck equation on the graph, based on our symmetric adjacency matrix is completely redundant (See Suppl. Note 1.1 in [16] for more details). Furthermore using L^{sym} for the current geometrically built symmetric adjacency matrix allows us to reserve the random-walk version for the true temporal biological process of cell development to be described by an appropriate Fokker-Planck equation considering cell differentiation as an un-equilibrium process generating a directed (asymmetric) adjacency matrix.

Similar to the calculations in equation 2.15, where we have built a distance measure on the symmetric version of the transition matrix, in [16] we define another distance measure which we call dpt. In short, we first remove the zeroth eigenvector from T^{sym} and call it $\tilde{T}^{sym}$, then we sum $(\tilde{T}^{sym})^t$ over all possible time scales t to make

the so called *accumulated transition matrix* M. The dpt distance is then defined as:

$$\mathrm{dpt}^2(x,y) = ||M(x,.) - M(y,.)||^2 = xM^2x + yM^2y - 2xM^2y \tag{2.16}$$

which can be expressed in terms of the eigenvalues and eigenvectors of T^{sym} (see Suppl. Note 1.2 in [16]):

$$\mathrm{dpt}^2(x,y) = \sum_{i=1}^{n-1}(\frac{\lambda_i}{1-\lambda_i})^2(\boldsymbol{\phi}_i(x) - \boldsymbol{\phi}_i(y))^2 \tag{2.17}$$

Interestingly, using the first power of M instead of M^2 in equation 2.16 would provide another distance measure known as "commute time" [62, 63]. Unlike the dpt distance, when the number of nodes in the graph is large, commute time does not provide information about the geometry of the graph but only about the node densities as shown by von Luxburg *et al.* in [64]. Therefore commute time is not a useful metric for pseudotime ordering and we put forward dpt for this purpose. To read more about dpt distance please refer to Suppl. Note 1.2 in [16].

2.5 Learning the geometry of the data

Learning the structure of cell differentiation hierarchy was on of major goals of this dissertation. A proper metric on the differentiation data manifold (as defined in the previous section) almost solves the problem. However, we still would need to separate different branching events in the data. Having an on-manifold metric in hand, the triangle inequality in Euclidean space can easily be generalised to distances on a manifold . Imagine two cells x and y on the data manifold. If a third cell is picked at the same branch connecting x and y, the sum of the its distances from x and y is a constant. If the third cell is picked on a separate branch than the one connecting x and y, sum of its distances from x and y is greater than that constant. (see Suppl. Note 1.3 and Suppl. Fig. N1 in [16]). Thus the triangle inequality can be used for separation of several branches of a manifold. With the scattered (noisy) pattern of single-cell data however, instead of a sharp constant we get a distribution around it. This requires thresholding for telling it apart from the sum of distances on other branches. To avoid such a thresholding parameter which might be tricky

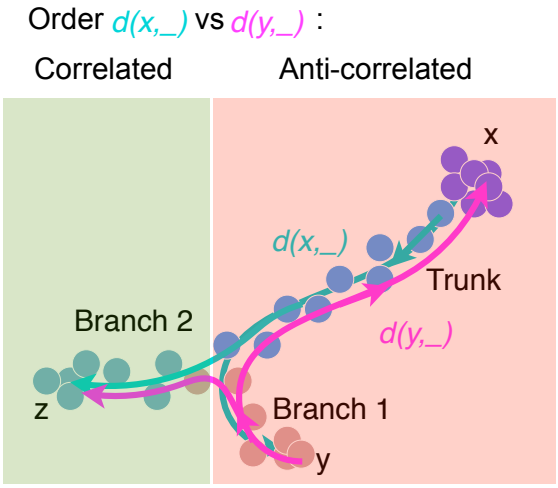


Figure 2.1: Branching points is defined as where the correlation between cell distances from x ($d(x, \_)$ in turquoise) with cell distances from y ($d(y, \_)$ in magenta) changes to anticorrelation.

to choose, we turned to correlation based branch assignment as illustrated in 2.1 (and Fig. 1a in [16]). For more details of the branch separation method see Suppl. Note 1.3 in [16] (Appendix B).

2.6 Universal time

In [16] (Appendix B) we introduce the concept of *universal time* which clarifies the relation of pseudotime to actual time trajectories measurements. As described in section 1.1, cell differentiation is a largely asynchronous process. Even if we consider a single-fated lineage, due to the stochastic nature of the system, a heterogeneous population of cells coexists at any given time. Although each single cell takes a different trajectory in actual time (due to the stochasticity in the differential equation), all these trajectories lie on a common manifold in the transcriptional space ($C \subset \mathbb{R}^G$), where G denotes the number of genes measured. The manifold C presents the average or deterministic part of the developmental trajectory and (if one dimensional) can be parametrized by the arc length s along C. For a single cell trajectory along this manifold we can assign a velocity $\boldsymbol{v}(t)$ to each time point t that is approximately tangent to the manifold C. If we consider an equidistant temporal sampling of the single cell trajectory, the tangent velocity is inversely proportional

to the density $\rho(t)$ of the cell states on the trajectory at that time point, that is $|\boldsymbol{v}(t)| = 1/\rho(t)$. In other words, the longer the time points of the single cell trajectory happen to be in a region of $\mathbb{R}^G$, the slower the single cell has passed through that region. Because $\boldsymbol{v}(t)$ is tangent on C we can write

$$\mathrm{d}s = |\boldsymbol{v}(t)|\mathrm{d}t = \frac{1}{\rho(t)}\mathrm{d}t. \tag{2.18}$$

Integrating ds, starting at the root cell, along C up to actual time t yields the arc length, which we refer to as *universal time*

$$s(t) = \int_{C:[s(0),s(t)]} \mathrm{d}s = \int_0^t |\boldsymbol{v}(t')|\mathrm{d}t' = \int_0^t \frac{1}{\rho(t')}\mathrm{d}t'. \tag{2.19}$$

This assigns a universal time $s(t)$ to every actual time single cell trajectory as measured in time-lapse microscopy. However, for snapshot data there is no relevant time measurement and the time integration step is not possible. Instead in the context of snapshot data we can calculate the pseudotime. In the ideal case, pseudotime can also be defined as an arc length measure over the reconstructed manifold in $\mathbb{R}^G$ of differentiation. While some algorithms like Wanderlust [52] try to measure the pseudotime as the arc length in $\mathbb{R}^G$ (with the cost of nonbranching data assumption however), because of the high level of noise in single-cell data, the common practice is to first map the data to a new space, where noise is diminished and thereby the manifold becomes more pronounced. Then in general one can define pseudotime as the distance (arc length) to the root cell on some mapped manifold C'. In our computation of pseudotime DPT (if one chooses to keep all diffusion dimensions) this is a mapping from $\mathbb{R}^G$ to $\mathbb{R}^{n-1}$, where n is the number of cells, and distances on C' are characterized by the dpt metric. For more details about universal time see Suppl. Note 6 in [16]. Thus, we established a unified framework which can be used to bring time-lapse microscopy data and single-cell snapshot expression data together and make them comparable. The connection of universal time with pseudotime as established here is only valid if cells from different developmental stages are present in a snapshot sample. However, we do not make any assumption of stationary sampling (as e.g. used in [65]). This is especially helpful in the context of single-cell snapshot data where sampling densities are usually far from any stationary state and influenced by cell division rates, noise, and the design of the experiment.

Chapter 3

Summary of contributed articles

This chapter provides a summary of all my contributed articles during the period of doing my doctoral studies. The publications are sorted by date of publication so that the history and sequential connection between them becomes comprehensible. A more detailed summary is provided for my (two) first-author contributed articles. Also please note that the first listed publication is not related to the general theme of this dissertation.

i) Bongini, M., Fornasier, M., Fröhlich, F. and Haghverdi, L. **Sparse stabilization of dynamical systems driven by attraction and avoidance forces.** *NHM, 9(1), pp.1-31(2014).*

   Summary: (This publication is not related to the general theme of this dissertation.) We studied sufficient conditions for sparse control of a system of agents interacting with each other by attraction and avoidance forces.

   My contribution: I contributed to literature screening and several discussions for this article.

ii) Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A.J., Tanaka, Y., Wilkinson, A.C., Buettner, F., Macaulay, I.C., Jawaid, W., Diamanti, E. and Nishikawa, S.I. **Decoding the regulatory network of early blood development from single-cell gene expression measurements.** *Nature Biotechnology, 33(3), pp.269-276 (2015).*

Summary: In this article we studied the development of early blood cells applying diffusion maps to single-cell data for the first time. A branching event was detected in the data and a boolean regulatory network was synthesized according to nearest neighbour state transitions in the data, which made computational simulation of gene perturbation experiments possible. The predictions from the synthesized network concerning the role of Sox and Hox genes in early blood development were validated experimentally.

My contribution: I performed the nearest-neighbours diffusion maps for this data set which enabled the detection of the endothelial versus epithelial branches.

iii) Haghverdi, L., Buettner, F. and Theis, F.J. **Diffusion maps for high-dimensional single-cell analysis of differentiation data.** *Bioinformatics, 31(18), pp.2989-2998 (2015).*

Summary: In this article we describe in detail the advantages of diffusion maps over several other machine learning methods for single-cell data analysis. Our two major points are that first, developing cells take a random-walk like path in the transcriptional space. As diffusion maps are based on random-walks they are relevant to the biological process of development under study. This gives diffusion maps superiority for single-cell differentiation data analysis over other more general machine learning embedding/clustering approaches. Second, diffusion maps also accommodate the nonlinearity of differentiation manifold. Third, diffusion maps are very robust to noise and which proved to play an important role in the analysis of single-cell data. Arguing that the effects of sampling density in single-cell experiments on the embedding are undesirable, we encourage the use of density normalised transition matrix which is facilitated in the Laplace-Beltrami framework of diffusion maps [59]. As the Gaussian kernel can be decomposed to multiplication of two Gaussian distributions, we propose a new approach for integration of missing values that are ubiquitous in single-cell data. The idea being that any prior distribution (as indicated by our biological knowledge about the missing value gene) can be used for constructing the kernel in the same distributions multiplication framework. We also propose heuristics for determination of the kernel width in diffusion maps based on the calculated intrinsic dimensionality of the data manifold. Our intrinsic dimensionality calculation in this article takes an averaging approach

because of computational considerations. Therefore our proposed kernel width in this article is global and not locally adjusted.

My contribution: First author.

iv) Ocone, A., Haghverdi, L., Mueller, N.S. and Theis, F.J. **Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data.** *Bioinformatics, 31(12), pp.i89-i96 (2015).*

Summary: This article is one of the earliest demonstrations on how pseudotime ordering can be used for inference of the gene regulatory network. There we use the reconstructed pseudotime series of gene expressions for an ODE based model selection over several possible regulatory network structures.

My contribution: I performed the pseudotime ordering and contributed to several discussions.

v) Angerer, P., Haghverdi, L., Büttner, M., Theis, F.J., Marr, C. and Buettner, F. **Destiny: diffusion maps for large-scale single-cell data in R.** *Bioinformatics, 32(8), pp.1241-1243 (2016).*

Summary: In this article we developed an efficient R package of diffusion maps adaptation for large-scale single-cell data. The package includes several useful features for single-cell data applications including data projection, use of cosine distances, handling of missing values, preselection of a set of less noisy genes, etc.

My contribution: I contributed to development of the code and several discussions.

vi) Haghverdi, L., Büttner, M., Wolf, F.A., Buettner, F., Theis, F.J. **Diffusion pseudotime robustly reconstructs lineage branching.** *Nature Methods, 13(10), pp.845-848 (2016).*

Summary: In this article we propose a new method DPT (diffusion pseudotime ordering tool) for pseudotime ordering of cells along the differentiation manifold. We define a new metric dpt based on geometric diffusions and used it to calculate the distance of cells from a specified root cell on the data manifold. We demonstrate that such an on-manifold distance is appropriate for pseudotime ordering of cells even in presence of branching events and identifying the cells at

the tip of each branch. We then propose a new solution for separating branching events using the correlation versus anticorrelation relations of dpt distances from several cells at the tip of several branches. DPT also allows identification of metastable states. That is states on the manifold in which cells spend more more before escaping to the next developmental state. Metastable states are thus distinguished by relatively high densities of cells under the assumption of unbiased or close to stationary state cell sampling conditions. In this work we applied DPT to simulated data as well as several experimental data sets. Through detailed differential gene expression analysis we demonstrate that DPT successfully finds pseudotime and branching events in all cases. In an sc-qPCR data set (with 42 monitored genes) from mouse early blood development we also studied the sequential activation/deactivation of genes through the course of pseudotime. Such activation/deactivation study was not possible for the sc-RNA-Seq data sets due to the large number of genes and higher level of noise in sc-RNA-Seq techniques. For those however, we identified the set of important genes in several developmental stages i.e. over pseudotime by differential gene expression analysis. Furthermore we show how our pseudotime is related to actual time measurements such as those provided by time-lapse microscopy measurements. That through the deterministic part of the dynamics all cell of the same fate take in the transcriptional space despite their asynchrony and stochasticity in the course of development. We call the deterministic path and its parametrisation respectively *universal path* and *universal time*.

The supplementary notes of this article include detailed mathematical backgrounds of the DPT method. For example in Suppl. Note 1 we show how one can use locally adjusted kernel width in the construction of the diffusion transition matrix and describe the use the symmetric version of the transition matrix instead of the asymmetric random-walk version that was classically introduced in [59]. In Suppl. Note 6 we describe in detail our proposition of *universal time*.

My contribution: First author.

Chapter 4

Discussion and perspectives

In this dissertation, we showed the application and adaptation of diffusion maps for analysis of single-cell differentiation data and further adapted its diffusion approach for pseudotime ordering and branch detection. The method DPT has several advantages over other existing algorithms namely its robustness to noise, scalability for application to large number of cells and efficiency of calculation because of the mathematical closed form which does not require any simulations. We introduced a new metric dpt, in which we omit the time (random-walk length) dependence that was present in the diffusion distance introduced by Coifman *et al.* [59] by summing the diffusion distances over all time scales. Furthermore, we have eliminated the redundant time direction in diffusion maps by using the symmetric Laplacian (and transition) matrix instead of the random-walk (asymmetric) versions. This way, we reserve the directed transition matrix for the true cellular developmental process for later work. This allows us to make distinction between the directed (naturally) irreversible dynamics of differentiating of cells towards more differentiated cell states and the artificial diffusion dynamics we construct by the merely geometrically built transition matrix which is used in calculation of the diffusion maps or the diffusion pseudotime. This work opens up new questions and perspectives some of which we discuss in the following sections. In [66], [67] and [16] we also inferred for each especial case, several regulatory relations between the genes by using the reconstructed dynamics we obtain from pseudotime ordered data.

4.1 Actual directed dynamics of cell differentiation versus the geometrically constructed diffusion transitions

The gene regulatory network governing cell differentiation dynamics produces a potential landscape (also known as Waddington ladscape) in the transcriptional space [68, 69], in which cells are pushed towards more differentiated states. The directed cell differentiation dynamics over this landscape can be described by a Fokker-Planck equation consisting of diffusion and potential terms. One might also consider additional source and sink term in this equation corresponding to the biological process of cell proliferation and cell death. Such an actual biological dynamics creates a manifold in the genes space which we tried to learn in this work using geometric diffusions. The Laplacian we constructed in this work explains the same manifold only by considering geometrical diffusions or transition probabilities among neighbouring cells. In other words, it misses the potential part of the Fokker-Planck equation in any biologically meaningful sense. This Laplacian is thus different from the actual time dynamics Laplacian which would be far from symmetric because of the directed nature of cell differentiation. It is quite intuitive that different types of time evolution dynamics can take place on the same manifold, giving rise to "equivalent Laplacians". For more mathematical details about equivalent Laplacians, i.e. different Laplacians explaining the same manifold, please see section 4.3 in [70] and [71]. In several differentiation data sets —for example in [22, 72] —snapshots of single cells are available in more than one experimental time point. Until now in this project, we always disregarded this temporal information in the data by appending all the data from several measurement time points into a single input data matrix. It is though interesting to use such time information as well for reconstruction of the actual time Laplacian. Consequently one could learn the potential landscape and sink/source terms (i.e. cell death, cell proliferation rates) over it. A crucial question to address for doing this study is how finely and over what number of cell states would it be possible to resolve the potential and the sink/source terms in the corresponding Fokker-Planck equation, given only a few measurement time points (typically 3 to 5) as it is usually given in single-cell differentiation experimental data sets.

4.2 Integration of time-lapse and snapshot data

One possible approach to gain insight about the actual time dynamics and transition rates, is combination and integration of pseudotimes with actual time-lapse microscopy measurements. The current technology for time-lapse microscopy does not allow simultaneous measurement of more than three or four proteins which is too few for gaining insight over the full dimension dynamics. However, the integration of time-lapse microscopy data with pseudotime obtained from high-dimensional snapshot data may resolve this issue. Up to date, we did not have access to a differentiation system where snapshot data and time-lapse measurements were both available. However, we have facilitated doing such data integration through introduction of universal time (see Online methods and Suppl. Note 6 in [16]) which can be applied once such data is at hand.

4.3 Quality control and uncertainity of pseudotime

As any other physical measure, calculation of pseudotime also includes estimation of the error and uncertainty in it. As we discuss in Suppl. Note 7.5.5 of [16] (Appendix B), pseudotime ordering of cells in the metastable states (i.e. potential wells in the gene expression landscape where cells tend to spend a longer time before escaping those states), is less meaningful and the uncertainty of pseudotime is expected to be larger in these regions. Bootstrapping of samples is a possible way for estimating this error. For demonstration, we calculated the pseudotime for 1000 sampled cells from a toy regulatory gene network (see Fig. S5 in [49]). Next we performed pseudotime calculation on 100 runs of bootstrap samples. For the pseudotimes to be comparable, we scaled the pseudotime from each run between zero and one. Panel A in Figure 4.1 shows the gene expression over pseudotime on the complete set of 1000 sample cells. Panel B shows the standard deviation of pseudotime for each cell over the multiple bootstrap samples. We clearly observe that the standard deviation (i.e. uncertainty) of pseudotime varies along the x axis, indicating two pronounced metastable state in the beginning and at the end of the pseudotime.

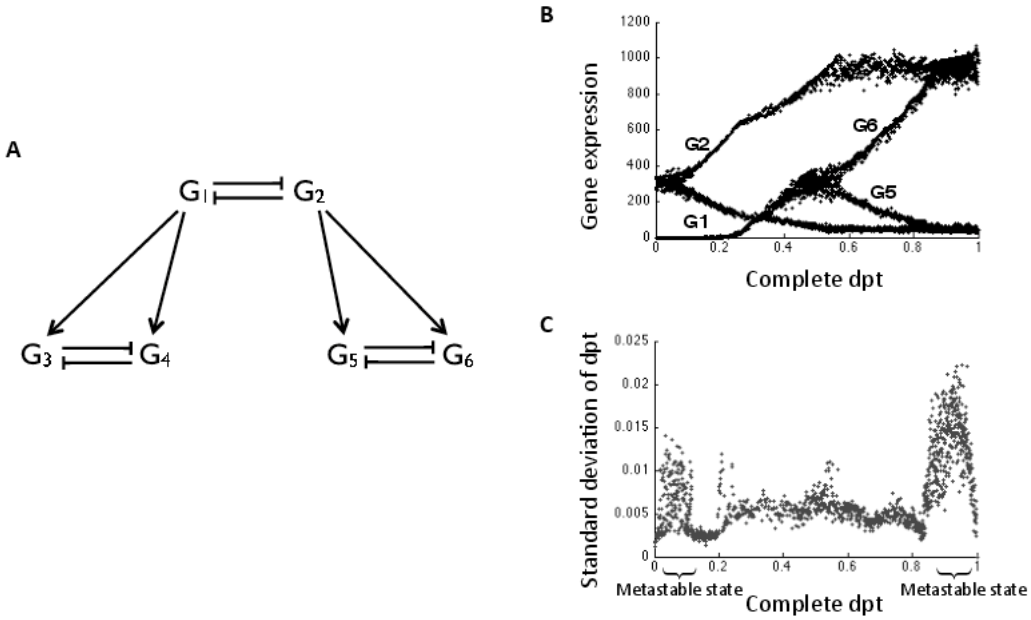


Figure 4.1: A) A toy regulatory gene network (see Fig. S5 in [49] and Fig. N8 in [16] for details about the toy model). B) Gene expression versus diffusion pseudotime of a single fate corresponding to $G_2^+ G_6^+$ branch in A. C) Standard deviation of diffusion pseudotime across 100 bootstrap samples, implies two pronounced metastable states marked by curly brackets on the pseudotime axis.

Other measures might also be proposed for quality control of pseudotime. Whereas pseudotime measurements are usually obtained through a specific objective function, it would be useful to fix a series of biological or statistical tests for the quality control of the obtained pseudotime. Recently, Bayesian methods have been applied as an alternative for assessing the uncertainty of pseudotime [73]. Comparison of results from Bayesian approaches to the bottom-up approaches of pseudotime construction (e.g. DPT, Monocle, Wanderlust) which try to reconstruct possible trajectories of cell development through a specific objective function would be interesting.

4.4 Inference of gene regulatory network

One of the main goals of pseudotime ordering is observation of how gene expression profiles change over the hierarchy of stages in cell development. Thus pseudotime can be useful for inferring the gene regulatory network (GRN) governing the development. Nevertheless GRN inference based on pseudotime is a challenging task due to several reasons. The highly noisy expression over pseudotime is one to name. Moreover, it is very common that several important genes of the regulatory network are missing in the measurements. This makes the reconstruction of the complete GRN challenging if not impossible. On the contrary, in experiments where (almost) all genes are measured, the complexity of the big gene network is the challenging part. In [66] the inference of regulatory network of the genes was synthesised on one-gene Boolean state-changes assumption for transitions in neighbouring cells. For a qPCR data set from mouse early blood development system in [16] (Appendix B) we ordered the genes by their activation/deactivation on-set over the pseudotime, hence suggesting the sequential role of genes in cell differentiation. In [67] we proposed a solution for inference of small GRNs using pseudotime using model selection from systems of ordinary differential equations (ODE). Proposition of a more general and efficient approach to tackle this problem requires a deeper study which remains for future work.

4.5 Building the transition matrix on other noise models

Classically, the diffusion transition matrix is built using a Gaussian kernel. In section 2.1 of [49] (Appendix A), we showed that the Gaussian kernel implies a multiplication of Gaussian noise models. There we argued that the same multiplication pattern can be adapted to non-Gaussians models. We also used the uniform distribution as a model for missing gene expression values (section 2.2 in [49]) and demonstrated that the kernel created by such distribution can provide a proper diffusion map in presence of missing expression values (see for example Fig. 6 and Fig. S1 in [49]). Gaussian distribution provides a good approximation for several other distributions

including binomial or Poisson distributions [74] around the centre (i.e. peak of the distribution). Especially, when a large number of sampled cells are at hand, one can cut the distribution at a certain distance from the peak to keep the approximation local and valid. However in several cases (in particular when the number of sampled cells is so few to impose a wide Gaussian for obtaining a connected graph) Gaussian distributions are not the optimal noise model choice. It is known that negative binomial distributions explain the expression data quite well [75]. Therefore it would be worth to adapt e.g. a negative binomial distribution in the construction of the kernel and compare its performance to the classical Gaussian kernel construction through quality control of the obtained pseudotimes. One advantage of approximating the noise by isotropic Gaussians is the reduced computational cost of integration in R^G which reduces to a one dimensional integration (in isotropic polar coordinates). Use of anisotropic noise models requires taking a complete G dimensional integral for building up of the kernel. Because of this high computational cost, in practice adaptation of anisotropic noise models might be limited only to relatively low dimensional data sets (e.g. single-cell qPCR, cytometry) rather than high dimensional data with large number of measured genes (scRNA-Seq techniques).

Bibliography

[1] D. Stocum, "Amphibian regeneration and stem cells," in *Regeneration: Stem Cells and Beyond*, pp. 1–70, Springer, 2004.

[2] K. Giles, "Dedifferentiation and regeneration in bryophytes: a selective review," *New Zealand Journal of Botany*, vol. 9, no. 4, pp. 689–694, 1971.

[3] J. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh, "McKusick's online Mendelian inheritance in man (OMIM®)," *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D793–D796, 2009.

[4] O. T. Avery, C. M. MacLeod, and M. McCarty, "Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III," *The Journal of Experimental Medicine*, vol. 79, no. 2, pp. 137–158, 1944.

[5] M. J. McConnell, M. R. Lindberg, K. J. Brennand, J. C. Piper, T. Voet, C. Cowing-Zitron, S. Shumilina, R. S. Lasken, J. R. Vermeesch, I. M. Hall, *et al.*, "Mosaic copy number variation in human neurons," *Science*, vol. 342, no. 6158, pp. 632–637, 2013.

[6] Y. Marcy, C. Ouverney, E. M. Bik, T. Lösekann, N. Ivanova, H. G. Martin, E. Szeto, D. Platt, P. Hugenholtz, D. A. Relman, *et al.*, "Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth," *Proceedings of the National Academy of Sciences*, vol. 104, no. 29, pp. 11889–11894, 2007.

[7] C. Gawad, W. Koh, and S. R. Quake, "Single-cell genome sequencing: current

state of the science," *Nature Reviews Genetics*, vol. 17, no. 3, pp. 175–188, 2016.

[8] R. Sandberg, "Entering the era of single-cell transcriptomics in biology and medicine," *Nature Methods*, vol. 11, no. 1, pp. 22–24, 2014.

[9] N. Friedman, L. Cai, and X. S. Xie, "Linking stochastic dynamics to population distribution: an analytical framework of gene expression," *Physical Review Letters*, vol. 97, no. 16, p. 168302, 2006.

[10] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, "Stochastic gene expression in a single cell," *Science*, vol. 297, no. 5584, pp. 1183–1186, 2002.

[11] J. M. Levsky and R. H. Singer, "Gene expression and the myth of the average cell," *Trends in Cell Biology*, vol. 13, no. 1, pp. 4–6, 2003.

[12] T. Schroeder, "Long-term single-cell imaging of mammalian stem cells," *Nature Methods*, vol. 8, no. 4s, pp. S30–S35, 2011.

[13] A. M. Femino, F. S. Fay, K. Fogarty, and R. H. Singer, "Visualization of single RNA transcripts in situ," *Science*, vol. 280, no. 5363, pp. 585–590, 1998.

[14] C. Marr, M. Strasser, M. Schwarzfischer, T. Schroeder, and F. J. Theis, "Multi-scale modeling of GMP differentiation based on single-cell genealogies," *FEBS Journal*, vol. 279, no. 18, pp. 3488–3500, 2012.

[15] H. M. Eilken, S.-I. Nishikawa, and T. Schroeder, "Continuous single-cell imaging of blood generation from haemogenic endothelium," *Nature*, vol. 457, no. 7231, pp. 896–900, 2009.

[16] L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, and F. J. Theis, "Diffusion pseudotime robustly reconstructs lineage branching," *Nature Mathods*, vol. 13, no. 10, pp. 845–848, 2016.

[17] H. M. Shapiro, *Practical flow cytometry*. John Wiley & Sons, 2005.

[18] D. R. Bandura, V. I. Baranov, O. I. Ornatsky, A. Antonov, R. Kinach, X. Lou, S. Pavlov, S. Vorobiev, J. E. Dick, and S. D. Tanner, "Mass cytometry:

technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry," *Analytical Chemistry*, vol. 81, no. 16, pp. 6813–6822, 2009.

[19] K. Taniguchi, T. Kajiyama, and H. Kambara, "Quantitative analysis of gene expression in a single cell by qPCR," *Nature Methods*, vol. 6, no. 7, p. 503, 2009.

[20] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, *et al.*, "mRNA-Seq whole-transcriptome analysis of a single cell," *Nature Methods*, vol. 6, no. 5, pp. 377–382, 2009.

[21] J. Liu, C. Hansen, and S. R. Quake, "Solving the "world-to-chip" interface problem with a microfluidic matrix," *Analytical Chemistry*, vol. 75, no. 18, pp. 4718–4723, 2003.

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

[23] D. Ramsköld, S. Luo, Y.-C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, *et al.*, "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells," *Nature Biotechnology*, vol. 30, no. 8, pp. 777–782, 2012.

[24] D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, *et al.*, "Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types," *Science*, vol. 343, no. 6172, pp. 776–779, 2014.

[25] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[26] O. Stegle, S. A. Teichmann, and J. C. Marioni, "Computational and analytical challenges in single-cell transcriptomics," *Nature Reviews Genetics*, vol. 16, no. 3, pp. 133–145, 2015.

[27] R. Bacher and C. Kendziorski, "Design and computational analysis of single-cell RNA-sequencing experiments," *Genome Biology*, vol. 17, no. 1, p. 1, 2016.

[28] A. McDavid, G. Finak, P. K. Chattopadyay, M. Dominguez, L. Lamoreaux, S. S. Ma, M. Roederer, and R. Gottardo, "Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments," *Bioinformatics*, vol. 29, no. 4, pp. 461–467, 2013.

[29] S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson, "Quantitative single-cell RNA-seq with unique molecular identifiers," *Nature Methods*, vol. 11, no. 2, pp. 163–166, 2014.

[30] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells," *Nature Biotechnology*, vol. 33, no. 2, pp. 155–160, 2015.

[31] N. Leng, J. Choi, L.-F. Chu, J. A. Thomson, C. Kendziorski, and R. Stewart, "Oefinder: a user interface to identify and visualize ordering effects in single-cell RNA-seq data," *Bioinformatics*, p. btw004, 2016.

[32] C. A. Vallejos, J. C. Marioni, and S. Richardson, "BASiCS: Bayesian analysis of single-cell sequencing data," *PLoS Comput Biol*, vol. 11, no. 6, p. e1004333, 2015.

[33] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[34] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl, *et al.*, "Constrained k-means clustering with background knowledge," in *ICML*, vol. 1, pp. 577–584, 2001.

[35] L. Kaufman and P. J. Rousseeuw, "Partitioning around medoids (program pam)," *Finding groups in data: an introduction to cluster analysis*, pp. 68–125, 1990.

[36] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[37] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[38] F. Buettner and F. J. Theis, "A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst," *Bioinformatics*, vol. 28, no. 18, pp. i626–i632, 2012.

[39] G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, *et al.*, "MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data," *Genome Biology*, vol. 16, no. 1, p. 1, 2015.

[40] F. D. Gibbons and F. P. Roth, "Judging the quality of gene expression-based clustering methods using gene annotation," *Genome Research*, vol. 12, no. 10, pp. 1574–1581, 2002.

[41] G. O. Consortium *et al.*, "The gene ontology (GO) database and informatics resource," *Nucleic Acids Research*, vol. 32, no. suppl 1, pp. D258–D261, 2004.

[42] J. Shin, D. A. Berg, Y. Zhu, J. Y. Shin, J. Song, M. A. Bonaguidi, G. Enikolopov, D. W. Nauen, K. M. Christian, G.-l. Ming, *et al.*, "Single-cell RNA-seq with waterfall reveals molecular cascades underlying adult neurogenesis," *Cell Stem Cell*, vol. 17, no. 3, pp. 360–372, 2015.

[43] E. Marco, R. L. Karp, G. Guo, P. Robson, A. H. Hart, L. Trippa, and G.-C. Yuan, "Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape," *Proceedings of the National Academy of Sciences*, vol. 111, no. 52, pp. E5643–E5650, 2014.

[44] N. Leng, L.-F. Chu, C. Barry, Y. Li, J. Choi, X. Li, P. Jiang, R. M. Stewart, J. A. Thomson, and C. Kendziorski, "Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments," *Nature Methods*, vol. 12, no. 10, pp. 947–950, 2015.

[45] J. E. Reid and L. Wernisch, "Pseudotime estimation: deconfounding single cell time series," *bioRxiv*, p. 019588, 2015.

[46] J. D. Welch, A. J. Hartemink, and J. F. Prins, "SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data," *Genome Biology*, vol. 17, no. 1, p. 1, 2016.

[47] Z. Ji and H. Ji, "TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis," *Nucleic Acids Research*, p. gkw430, 2016.

[48] C. Trapnell, "Defining cell types and states with single-cell genomics," *Genome Research*, vol. 25, no. 10, pp. 1491–1498, 2015.

[49] L. Haghverdi, F. Buettner, and F. J. Theis, "Diffusion maps for high-dimensional single-cell analysis of differentiation data," *Bioinformatics*, vol. 31, no. 18, pp. 2989–2998, 2015.

[50] P. Qiu, E. F. Simonds, S. C. Bendall, K. D. Gibbs Jr, R. V. Bruggner, M. D. Linderman, K. Sachs, G. P. Nolan, and S. K. Plevritis, "Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE," *Nature Biotechnology*, vol. 29, no. 10, pp. 886–891, 2011.

[51] D. L. Donoho *et al.*, "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math Challenges Lecture*, pp. 1–32, 2000.

[52] S. C. Bendall, K. L. Davis, E.-A. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, and D. Pe'er, "Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development.," *Cell*, vol. 157, no. 3, pp. 714–725, 2014.

[53] E.-a. D. Amir, K. L. Davis, M. D. Tadmor, E. F. Simonds, J. H. Levine, S. C. Bendall, D. K. Shenfeld, S. Krishnaswamy, G. P. Nolan, and D. Pe'er, "viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia," *Nature Biotechnology*, vol. 31, no. 6, pp. 545–552, 2013.

[54] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.," *Nature Biotechnology*, vol. 32, pp. 381–6, 4 2014.

[55] K. S. Booth and G. S. Lueker, "Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms," *Journal of Computer and System Sciences*, vol. 13, no. 3, pp. 335–379, 1976.

[56] M. Setty, M. D. Tadmor, S. Reich-Zeliger, O. Angel, T. M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, and D. Pe'er, "Wishbone identifies bifurcating developmental trajectories from single-cell data," *Nature Biotechnology*, 2016.

[57] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[58] K. Campbell, C. P. Ponting, and C. Webber, "Laplacian eigenmaps and principal curves for high resolution pseudotemporal ordering of single-cell RNA-seq profiles," *bioRxiv doi: 10.1101/027219*, 9 2015.

[59] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 21, pp. 7426–7431, 2005.

[60] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering.," in *NIPS*, vol. 14, pp. 585–591, 2001.

[61] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[62] C. D. Meyer, Jr., "The role of the group generalized inverse in the theory of finite markov chains," *SIAM Review*, vol. 17, pp. 443–464, 7 1975.

[63] F. Fouss, A. Pirotte, J. Renders, and M. Saerens, "A novel way of computing dissimilarities between nodes of a graph, with application to collaborative filtering and subspace projection of the graph nodes (technical report no. iag wp 06/08)," *Université catholique de Louvain*, 2006.

[64] U. Von Luxburg, A. Radl, and M. Hein, "Hitting and commute times in large random neighborhood graphs.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1751–1798, 2014.

[65] R. Kafri, J. Levy, M. B. Ginzberg, S. Oh, G. Lahav, and M. W. Kirschner, "Dynamics extracted from fixed cells reveal feedback linking cell growth to cell cycle," *Nature*, vol. 494, no. 7438, pp. 480–483, 2013.

[66] V. Moignard, S. Woodhouse, L. Haghverdi, A. J. Lilly, Y. Tanaka, A. C. Wilkinson, F. Buettner, I. C. Macaulay, W. Jawaid, E. Diamanti, S.-I. Nishikawa, N. Piterman, V. Kouskoff, F. J. Theis, J. Fisher, and B. Göttgens, "Decoding the regulatory network of early blood development from single-cell gene expression measurements," *Nature Biotechnology*, vol. 33, no. 3, pp. 269–76, 2015.

[67] A. Ocone, L. Haghverdi, N. S. Mueller, and F. J. Theis, "Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data," *Bioinformatics*, vol. 31, no. 12, pp. i89–i96, 2015.

[68] C. H. Waddington, "The strategy of the genes," *London: Allen*, vol. 86, 1957.

[69] J. Wang, K. Zhang, L. Xu, and E. Wang, "Quantifying the Waddington landscape and biological paths for development and differentiation," *Proceedings of the National Academy of Sciences*, vol. 108, no. 20, pp. 8257–8262, 2011.

[70] P. Blanchard and D. Volchenkov, *Mathematical analysis of urban spatial networks*. Springer Science & Business Media, 2008.

[71] S. Butler, "Interlacing for weighted graphs using the normalized laplacian," *Electronic Journal of Linear Algebra*, vol. 16, no. 90-98, p. 87, 2007.

[72] G. Guo, M. Huss, G. Q. Tong, C. Wang, L. L. Sun, N. D. Clarke, and P. Robson, "Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst," *Developmental Cell*, vol. 18, no. 4, pp. 675–685, 2010.

[73] K. Campbell, C. Yau, C. P. Ponting, and C. Webber, "Bayesian gaussian process latent variable models for pseudotime inference in single-cell RNA-seq data," *bioRxiv doi: 10.1101/026872*, 9 2015.

[74] D. B. Peizer and J. W. Pratt, "A normal approximation for binomial, F, beta, and other common, related tail probabilities, I," *Journal of the American Statistical Association*, vol. 63, no. 324, pp. 1416–1456, 1968.

[75] D. Grün, L. Kester, and A. van Oudenaarden, "Validation of noise models for single-cell transcriptomics," *Nature Methods*, vol. 11, no. 6, pp. 637–40, 2014.

Appendix A

This is a pre-copyedited, author-produced PDF of an article accepted for publication in Bioinformatics following peer review. The version of record
L. Haghverdi, F. Buettner, and F. J. Theis, **Diffusion maps for high-dimensional single-cell analysis of differentiation data.** *Bioinformatics, 31(18), pp.2989-2998 (2015).*
is available online at:
http://bioinformatics.oxfordjournals.org/content/31/18/2989.full

BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1–10

Diffusion maps for high-dimensional single-cell analysis of differentiation data

Laleh Haghverdi [1,2], Florian Buettner [1] *†and Fabian J. Theis [1,2]*

[1]Institute of Computational Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany
[2]Department of Mathematics, Technische Universität München, 85748 Garching, Germany

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor:

ABSTRACT

Motivation: Single-cell technologies have recently gained popularity in cellular differentiation studies regarding their ability to resolve potential heterogeneities in cell populations. Analysing such high-dimensional single-cell data has its own statistical and computational challenges. Popular multivariate approaches are based on data normalisation, followed by dimension reduction and clustering to identify subgroups. However, in the case of cellular differentiation, we would not expect clear clusters to be present but instead expect the cells to follow continuous branching lineages.

Results: Here we propose the use of diffusion maps to deal with the problem of defining differentiation trajectories. We adapt this method to single-cell data by adequate choice of kernel width and inclusion of uncertainties or missing measurement values, which enables the establishment of a pseudo-temporal ordering of single cells in a high-dimensional gene expression space. We expect this output to reflect cell differentiation trajectories, where the data originates from intrinsic diffusion-like dynamics. Starting from a pluripotent stage, cells move smoothly within the transcriptional landscape towards more differentiated states with some stochasticity along their path. We demonstrate the robustness of our method with respect to extrinsic noise (e.g. measurement noise) and sampling density heterogeneities on simulated toy data as well as two single-cell quantitative polymerase chain reaction (qPCR) data sets (i.e. mouse haematopoietic stem cells and mouse embryonic stem cells) and an RNA-Seq data of human pre-implantation embryos. We show that diffusion maps perform considerably better than Principal Component Analysis (PCA) and are advantageous over other techniques for non-linear dimension reduction such as t-distributed Stochastic Neighbour Embedding (t-SNE) for preserving the global structures and pseudo-temporal ordering of cells.

Availability: The Matlab implementation of diffusion maps for single-cell data is available at https://www.helmholtz-muenchen.de/icb/single-cell-diffusion-map.

Contact: fbuettner.phys@gmail.com, fabian.theis@helmholtz-muenchen.de

*To whom correspondence should be addressed

†Current address: European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, CB10 1SD, UK

1 INTRODUCTION

The advantages of single-cell measurements to various biological research fields have become obvious in recent years. One example is the stem cell studies for which population measurements fail to reveal the properties of the heterogeneous population of cells at various stages of development. Purifying for a certain cell type or synchronising cells is experimentally challenging. Moreover, stem cell populations that have been functionally characterised often show heterogeneity in their cellular and molecular properties (Huang, 2009; Dykstra *et al.*, 2007; Stingl *et al.*, 2006). To overcome these barriers, on the one hand researchers conduct continuous single-cell observation using time-lapse microscopy (Park *et al.*, 2014; Rieger *et al.*, 2009; Schroeder, 2011), accompanied by single-cell tracking and analysis tools. However this approach is still limited as the expression of very few genes (typically one to three) could be observed. On the other hand, with the advent of new technologies, such as single-cell qPCR (Wilhelm and Pingoud, 2003) or RNA-Seq (Chu and Corey, 2012) and flow or mass cytometry (Chattopadhyay *et al.*, 2006; Bandura *et al.*, 2009), it is now possible to measure hundreds to thousands of genes from thousands of single cells at different specific experimental time-points (time course experiments). However, several single cells measured at the same experimental time point may be at different developmental stages. Therefore, there is a need for computational methods which resolve the hidden temporal order that reflects the ordering of developmental stages of differentiating cells.

While differentiation has to be regarded as a nonlinear continuous process (Buettner and Theis, 2012; Bendall *et al.*, 2014), standard methods used for the analysis of high-dimensional gene-expression data are either based on linear methods such as Principal Component Analysis (PCA) and Independent Components Analysis (ICA) (e.g. used as part of the monocle algorithm, (Trapnell *et al.*, 2014)) or they use clustering techniques that groups cells according to specific properties. Hierarchical clustering methods as used in SPADE (Qiu *et al.*, 2011) and t-SNE (Van der Maaten and Hinton, 2008) as used in viSNE (Amir *et al.*, 2013) are examples of clustering methods. However, as these methods are designed to detect discrete sub-populations, they usually do not preserve the continuous trajectories of differentiation data. A more recently proposed algorithm Wanderlust (Bendall *et al.*, 2014) incorporates the nonlinearity and continuity concepts but provides a pseudo-temporal ordering of cells only if the data comprise a single branch. Furthermore, in gene expression measurement techniques, there

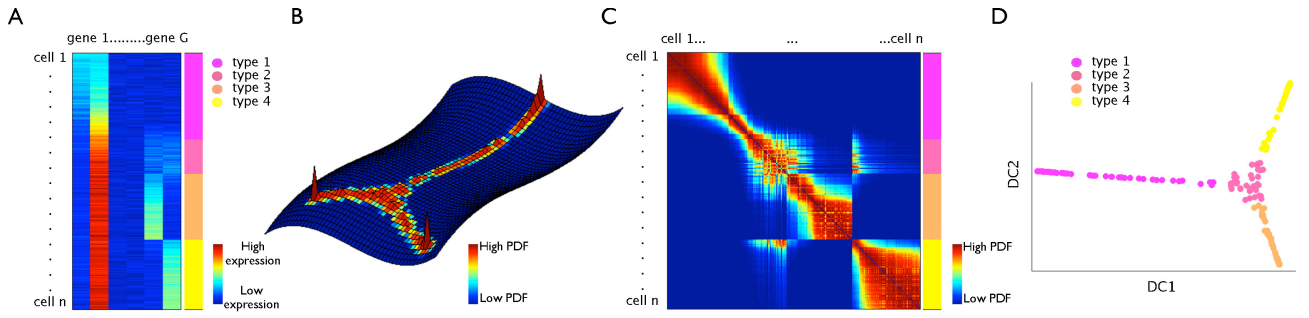


Fig. 1: Schematic overview of diffusion maps embedding. A) The $n \times G$ matrix representation of single-cell data consisting of four different cell types. The last column on the right side of the matrix (colour band) indicates the cell type for each cell. B) Representation of each cell by a Gaussian in the G-dimensional gene space. Diffusion paths (continuous paths with relatively high probability density) form on the data manifold as a result of interference of the Gaussians. The Probability density function is shown in the heat map. C) The $n \times n$ Markovian transition probability matrix. D) Data embedding on the first two eigenvectors of the Markovian transition matrix (DC1 and DC2) which correspond to the largest diffusion coefficients of the data manifold. The embedding shows the continuous flow of cells across four cell types, however it does not suggest the putative time direction.

is usually a detection limit at which lower expression levels and non-expressed genes are all reported at the same value. Buettner *et al.* (2014) suggested the use of a censoring noise model for PCA, whereas for the other methods it is unclear how these uncertain or missing values are to be treated. A variety of other manifold learning methods including (Hessian) Locally-Linear Embedding (HLLE) (Donoho and Grimes, 2003) and Isomap (Tenenbaum *et al.*, 2000) exist in the machine learning community and are discussed in detail in the discussion and conclusion section.

Here, we propose diffusion maps (Coifman *et al.*, 2005) as a tool for analysing single-cell differentiation data. Diffusion maps use a distance metric (usually referred to as diffusion distance) conceptually relevant to how differentiation data is generated biologically, as cells follow noisy diffusion-like dynamics in the course of taking several differentiation lineage paths. Diffusion maps preserve the nonlinear structure of data as a continuum and are robust to noise. Furthermore, with density normalisation, diffusion maps are resistant to sampling density heterogeneities and can capture rare as well as abundant populations. As a nonlinear dimension-reduction tool, diffusion maps can be applied on single-cell omics data to perform dimension-reduction and ordering of cells along the differentiation path in a single step, thus providing insight to the dynamics of differentiation (or any other concept with continuous dynamics). In this article, we

- propose an adaptation of diffusion maps for the analysis of single-cell data which is less affected by sampling density heterogeneities and addresses the issues relating to missing values and uncertainties of measurement,
- propose a criterion for selecting the scale parameter in a diffusion map,
- evaluate the performance of the diffusion map and its robustness to noise and density heterogeneities using a toy model that mimics the dynamics of differentiation,
- apply the adapted diffusion map algorithm to two typical qPCR and one RNA-Seq data sets and show that it captures the differentiation dynamics more accurately than other algorithms.

2 METHODS

2.1 Diffusion maps

Let n be the number of cells and let G be the number of genes measured for each cell. Denote the set of all measured cells by $\mathbf{\Omega}$. We allow each cell $\mathbf{x}$ to diffuse around its measured position $\boldsymbol{x} \in \mathbb{R}^G$ through an isotropic Gaussian wave function,

$$Y_{\mathbf{x}}(\boldsymbol{x'}) = \left(\frac{2}{\pi\sigma^2}\right)^{1/4} \exp\left(-\frac{||\boldsymbol{x'}-\boldsymbol{x}||^2}{\sigma^2}\right) \tag{1}$$

The normalisation of $Y_{\mathbf{x}}(\boldsymbol{x'})$ is such that $\int_{-\infty}^{\infty} Y_{\mathbf{x}}^2(\boldsymbol{x'})\mathrm{d}\boldsymbol{x'} = 1$. The Gaussian width σ^2 determines the length scale over which each cell can randomly diffuse. The transition probability from cell $\mathbf{x}$ to cell $\mathbf{y}$ is then defined by the interference of the two wave functions $Y_{\mathbf{x}}$ and $Y_{\mathbf{y}}$. One can easily show that this interference product is another Gaussian (all prefactors cancel out):

$$\int_{-\infty}^{\infty} Y_{\mathbf{x}}(\boldsymbol{x'})Y_{\mathbf{y}}(\boldsymbol{x'})\mathrm{d}\boldsymbol{x'} = \exp\left(-\frac{||\boldsymbol{x}-\boldsymbol{y}||^2}{2\sigma^2}\right) \tag{2}$$

Hence, we can construct the $n \times n$ Markovian transition probability matrix P for all pairs of cells in $\mathbf{\Omega}$ as follows:

$$P_{\mathbf{xy}} = \frac{1}{Z(\boldsymbol{x})} \exp\left(-\frac{||\boldsymbol{x}-\boldsymbol{y}||^2}{2\sigma^2}\right) \tag{3}$$

$$Z(\boldsymbol{x}) = \sum_{\mathbf{y}\in\mathbf{\Omega}} \exp\left(-\frac{||\boldsymbol{x}-\boldsymbol{y}||^2}{2\sigma^2}\right) \tag{4}$$

At the position of each cell, $Z(\boldsymbol{x})$ is the partition function which provides an estimate of the number of neighbours of $\mathbf{x}$ in a certain volume defined by σ. Hence it can be interpreted as the density of cells at that proximity.

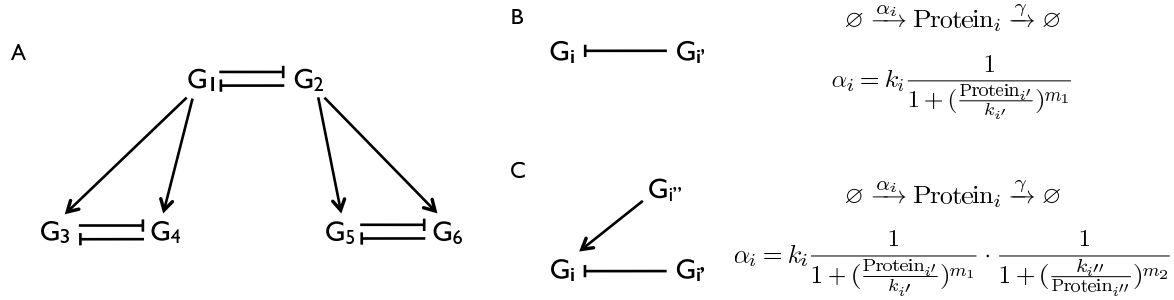


Fig. 2: A) Toy model of a differentiation regulatory network consisting of three pairs of antagonistic genes simulated by the Gillespie algorithm. The arrows show activation or inhibition interactions between genes. The toy model employs two classes of gene regulation: B) G_i is connected to an inhibitor, its production rate α_i is proportional to a Hill function of the concentration of the inhibitor $\mathrm{Protein}_{i'}$, C) G_i is connected to an inhibitor $G_{i'}$ and an activator $G_{i''}$, its production rate α_i is proportional to product of an inhibiting and an activating Hill function. The degradation rate γ is constant for all proteins.

Consequently, we redefine the density normalised transition probability matrix $\tilde{P}$ as:

$$\tilde{P}_{\mathbf{xy}} = \frac{1}{\tilde{Z}(\boldsymbol{x})} \frac{\exp\left(-\frac{||\boldsymbol{x}-\boldsymbol{y}||^2}{2\sigma^2}\right)}{Z(\boldsymbol{x})Z(\boldsymbol{y})} \quad , \tilde{P}_{\mathbf{xx}} = 0 \tag{5}$$

$$\tilde{Z}(\boldsymbol{x}) = \sum_{\mathbf{y}\in\boldsymbol{\Omega}/\mathbf{x}} \frac{\exp\left(-\frac{||\boldsymbol{x}-\boldsymbol{y}||^2}{2\sigma^2}\right)}{Z(\boldsymbol{x})Z(\boldsymbol{y})} \tag{6}$$

Because we are only interested in the transition probabilities between cells and not the on-cell potentials imposed by local densities, we set the diagonal of $\tilde{P}$ to zero and exclude $\mathbf{y} = \mathbf{x}$ from the sum in the partition function $\tilde{Z}$. For a large enough σ, the matrix $\tilde{P}$ defines an ergodic Markovian diffusion process on the data and has n ordered eigenvalues $\lambda_0 = 1 > \lambda_1 \geq ... \geq \lambda_{n-1}$ with corresponding right eigenvectors $\boldsymbol{\psi}_0 ... \boldsymbol{\psi}_{n-1}$.

The t-th power of $\tilde{P}$ will present the transition probabilities between cells in a diffusion (random walk) process of length t. Noting that $\tilde{P}^t$ has the same eigenvectors as $\tilde{P}$, one can show that this transition probability can be represented as follows:

$$\tilde{P}^t_{\mathbf{xy}} = \sum_{i=0}^{n-1} \lambda_i^t \boldsymbol{\psi}_i(\boldsymbol{x}) \boldsymbol{\psi}_i(\boldsymbol{y}) \tilde{Z}(\boldsymbol{y}) \tag{7}$$

Each row of $\tilde{P}^t$ can be viewed as a vector, which we represent as $\boldsymbol{p}^t(\mathbf{x}, \cdot)$ and consider as the feature representation (Shawe-Taylor and Cristianini, 2004) for each cell $\mathbf{x}$. By computing the weighted L^2 distance in the feature space, the diffusion distance D_t^2 between two cells $\mathbf{x}$ and $\mathbf{y}$ is defined as follows:

$$D_t^2(\mathbf{x}, \mathbf{y}) = ||\boldsymbol{p}^t(\mathbf{x}, \cdot) - \boldsymbol{p}^t(\mathbf{y}, \cdot)||^2_{1/\tilde{Z}} = \sum_{\mathbf{z}} \frac{(\tilde{P}^t_{\mathbf{xz}} - \tilde{P}^t_{\mathbf{yz}})^2}{\tilde{Z}(\boldsymbol{z})} \tag{8}$$

This diffusion distance can be expressed in terms of the eigenvectors of $\tilde{P}$ such that:

$$D_t^2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n-1} \lambda_i^{2t} (\boldsymbol{\psi}_i(\boldsymbol{x}) - \boldsymbol{\psi}_i(\boldsymbol{y}))^2 \tag{9}$$

The corresponding eigenvector to the largest eigenvalue λ_0 is a constant vector $\boldsymbol{\psi}_0 = \mathbb{1}$. Therefore, it only contributes a zero term to D_t^2 and is excluded from the spectral decomposition of D_t^2 in Equation 9. That means the Euclidean distance of the cells in the first l eigenvector space represents an approximation of their diffusion distance D_t^2. Moreover, the eigenvalues of $\tilde{P}$ determine the diffusion coefficients in the direction of the corresponding eigenvector. As real data usually lie on a lower dimensional manifold than the entire dimensions of space G, these diffusion coefficients drop to a noise level other than a few first (l) prominent directions. Therefore, if there is a significant gap between the l-th and $(l+1)$-th eigenvalue, the sum up to the l-th term usually determines a good approximation for diffusion distances. Thus, for data visualisation we select these eigenvectors and instead of the mathematical notation $\boldsymbol{\psi}$, we call them Diffusion Components (DCs).

Figure 1 presents a summary of diffusion map embedding. Each cell is represented by a Gaussian wave function in the G-dimensional gene space. On an adequate Gaussian width, the wave functions of neighbouring cells interfere with each other and form the diffusion paths along the (nonlinear) data manifold in the high-dimensional space. Hence, we construct the Markovian transition probability matrix, the elements of which are the transition probabilities between all pairs of cells. The eigenfunctions of the Markovian transition probability matrix (DC1 and DC2) are then used for low-dimensional representation and visualisation of data.

2.2 Accounting for missing and uncertain values

The data generated from qPCR, RNA-Seq or cytometry experiments are very often prone to imperfections such as missing values or detection limit thresholds. It is important to properly treat such uncertainties of data (McDavid *et al.*, 2013; Buettner *et al.*, 2014). Our probabilistic interpretation of diffusion maps allows a straightforward mechanism of handling missing and uncertain data measurements. First, we have to decompose the kernel into G components. Then, instead of a Gaussian, we can use any other wave function that best represents our prior knowledge on the probability distribution of the missing or uncertain values, which then should be square-normalised to ensure equal contribution of the G components. For example, for missing values and non-detects (measurements below the limit of detection), one might choose a uniform distribution over the whole range of possible values.

In the following we describe how to account for the uncertainty of non-detect measurements in qPCR data. The statistical subtleties of non-detect values in qPCR experiments have been systematically studied by McDavid *et al.* (2013) for univariate models. In addition, for a multivariate PCA analysis, Buettner *et al.* (2014) proposed that different kernels be allowed in each dimension. For the diffusion map implementation, we assume any value between the detection limit (M_0) and a completely non-expressed (off) state of genes valued as M_1, is equally possible for the non-detect measurements. Considering the kernel width formulated in the diffusion map wave functions, we assume an indicator wave function between $M_0 - \sigma$ and

$M_1 + \sigma$ normalised by $(M_1 - M_0 + 2\sigma)^{-1/2}$. Thus, we have to calculate three different kinds of interference of wave functions:
The interference of two cells with definite measured values for gene g is the standard Gaussian kernel (see section 2.1):

$$\int_{-\infty}^{\infty} Y_x(x'_g)Y_y(x'_g)\mathrm{d}x'_g = \exp\left(-\frac{(x_g - y_g)^2}{2\sigma^2}\right),$$

the interference of two cells both with non-detect values for gene g is 1 (due to the square-normalisation constraint):

$$\int_{-\infty}^{\infty} Y_x(x'_g)Y_y(x'_g)\mathrm{d}x'_g = 1,$$

the interference of a missing (non-detect) value to a definite measured value x_g is:

$$\int_{-\infty}^{\infty} Y_x(x'_g)Y_y(x'_g)\mathrm{d}x'_g =$$

$$\int_{M_0-\sigma}^{M_1+\sigma} \frac{1}{\sqrt{M_1 - M_0 + 2\sigma}}\left(\frac{2}{\pi\sigma^2}\right)^{1/4} \exp\left(-\frac{(x'_g - x_g)^2}{\sigma^2}\right)\mathrm{d}x'_g$$

$$= \frac{1}{\sqrt{M_1 - M_0 + 2\sigma}}\left(\frac{\pi\sigma^2}{8}\right)^{1/4}$$

$$\cdot\left(\mathrm{erfc}\left(\frac{M_0 - \sigma - x_g}{\sigma}\right) - \mathrm{erfc}\left(\frac{M_1 + \sigma - x_g}{\sigma}\right)\right).$$

For data with missing or uncertain values, we need to check the pairwise interference of the wave functions for each gene. The computation time is thus proportional to the number of genes G for a fixed number of cells n. Therefore, it might be preferable (especially in the case of large G) to choose the wave function of the missing (or uncertain) value also in the form of a Gaussian such that the multiplication of the G components of interference can be expressed as the sum of the exponents and the exponentiation step can be performed only once at the end of the algorithm for computation of the transition matrix. An implementation of this fast version of the censoring algorithm is also provided in the codes package. Figure S1 in the supplement provides an illustration of our approach for accounting for missing and uncertain values.

2.3 Determination of Gaussian kernel width

The parameter σ in Equation 1 determines the scale at which we visualise the data. If σ is extremely small, most elements of the transition probability matrix $\tilde{P}$ will tend to be zero and we do not get an overall view of a connected graph structure. In fact, when σ is too small, the number of degenerate eigenvectors with eigenvalue equal to one, indicates the number of disconnected segments that $\tilde{P}$ defines on the data. For too large σ however, the transition probability sensitivity on the distance between the cells blurs. There is a certain range of σ variations for which $\tilde{P}$ defines an ergodic diffusion process on the data as a connected graph and still the diffusion distances between the cells are informative.

The un-normalised density at each cell ($Z(\boldsymbol{x})$ in Equation 3) is proportional to the number of cells in a fixed volume in its neighbourhood and depends on σ. At scales of σ close to zero, cells do not have any neighbours and their average density is 1 (because of the 1s on the diagonal of P). By increasing σ, the average density gradually increases as more cells find other cells in their neighbourhood. There is a density saturation point where σ reaches the system size and all cells form part of one neighbourhood. At this point, for every cell $\mathbf{x} \in \boldsymbol{\Omega}$, the density $Z(\boldsymbol{x})$ will be equal to the entire system size n.

Assuming that the density gradient is not extremely sharp along the data manifold, the number of neighbours of cell $\mathbf{x}$ in the neighbourhood σ will be proportional to the volume of a hypersphere of radius σ, hence:

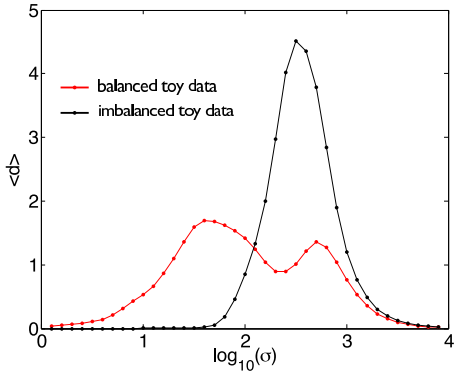


Fig. 3: The average dimensionality of the data $\langle d \rangle$ as a function of $\log_{10}(\sigma)$ for the balanced and imbalanced toy data sets.

$$Z(\boldsymbol{x}) \propto \sigma^{d(\boldsymbol{x},\boldsymbol{\sigma})} \tag{10}$$

where $d(\boldsymbol{x}, \boldsymbol{\sigma})$ is the dimensionality of data manifold at the position of cell $\mathbf{x}$ and at the scale σ. By differentiating both sides with respect to $\log(\sigma)$, we find that the average dimensionality of the manifold can be estimated by the slope of the log-log plot of the number of neighbours versus the length scale:

$$\langle d(\sigma)\rangle_x = \frac{\partial \langle \log(Z(\boldsymbol{x}))\rangle_x}{\partial \log(\sigma)} \tag{11}$$

where we compute the average of $\log(Z(\boldsymbol{x}))$ with consideration of density heterogeneities such that:

$$\langle \log(Z(\boldsymbol{x}))\rangle_x = \frac{\sum_x (\log(Z(\boldsymbol{x})) \cdot (1/Z(\boldsymbol{x})))}{\sum_x (1/Z(\boldsymbol{x}))} \tag{12}$$

It is worth noting that this average density underestimates the real dimensionality of the structure due to the contribution of the cells lying on the surface of the manifold. However, this does not affect our heuristic since the variation of $\langle d \rangle$ is our main interest rather than $\langle d \rangle$ itself.

Each time $\langle d \rangle$ reaches its maximum and starts to decrease, one can deduce that an intrinsically lower-dimensional structure is emerging from the noise-enriched distributed cells in the original high-dimensional space. Therefore several characteristic length scales of the data manifold (i.e. width of its linear parts, radius of its curves, etc.) give rise to several local maxima in $\langle d \rangle$. Such characteristic scales indeed make our choice for the Gaussian width σ since they indicate the scale at which the Euclidean distances used in the Gaussian kernel are sensible in an assumed Euclidean tangent space to the manifold. Although Euclidean distances are also valid for smaller σs than the characteristic length scale, they are not recommended because smaller kernel width would mean less connectivity in the cells graph which in turn results in an increased sensitivity to noise. Figures S2 and S3 in the supplement illustrate the resulting diffusion map on optimal kernel width and several other kernel widths values for a U-shaped toy data. Also the performance of diffusion map at the optimal kernel width when there is no distinguishable pattern in the data (e.g. normally distributed data in all dimensions or sparse data) is illustrated in the supplementary Figure S4.

2.4 Toy model for differentiation

As toggle switches are known to play a role in differentiation branching processes (Orkin and Zon, 2008), we designed a regulatory network of three pairs of toggle genes to evaluate the performance of our method on a toy data set that mimics a differentiation tree (Krumsiek *et al.*, 2011). Assuming a genetic regulatory module as presented in Figure 2A, we simulated the stochastic differentiation process by the Gillespie algorithm (Gillespie, 1977) with the reactions as shown in 2B and 2C (Strasser *et al.*,

2012). More details about the chemical reactions and the reaction rates used in the Gillespie algorithm model can be found in the supplement (Figure S5 A and B). Genes G_1 and G_2 are antagonistic to each other through an inhibiting Hill function. Therefore, starting from an initial undifferentiated state where G_1 and G_2 are both in a very low expression level, single samples may end up in either of the states where G_1 or G_2 is exclusively expressed. At this stage, the next pair of toggle genes in the differentiation hierarchy is activated (through an activating binding Hill function), which are again antagonistic to each other. This model generates four different types of fully differentiated cells in the six-dimensional space of genes.

To establish a steady state in the cell population, once a cell hits the end of each branch, we remove it from the population and initiate a new cell at the original undifferentiated state. This approach maintains the population size of cells. After an extended simulation run, the steady state of the population is established and resembles the haemostatic state of (e.g. haematopoietic) stem cells in natural organisms.

We sampled cells from this toy model in two different sets, a balanced toy data set, wherein 600 samples serve as a a snapshot of the steady state of the system with no additional extrinsic noise, and an imbalanced toy data set, wherein 1800 sample are derived from a non-steady-state density distribution with heavier sampling density on the $G_1^+G_3^+$ branch. We also added an extrinsic Gaussian noise with a variance of 25% maximum expression to each gene. The gene expression plot for a simulated single cell as it proceeds from the initial pluripotent state to a fully differentiated state is presented in the supplement (Figure S5 C and D).

2.5 Experimental data

2.5.1 qPCR data of mouse haematopoietic stem cells. We calculated a diffusion map embedding for the haematopoietic and progenitor stem cells data set from Moignard *et al.* (2013). In this experiment, 597 cells from five different haematopoietic cell types, namely, haematopoietic stem cell (HSC), lymphoid-primed multipotent progenitor (LMPP), megakaryocyte-erythroid progenitor (PreMegE), common lymphoid progenitor (CLP) and granulocyte-monocyte progenitor (GMP) were gated by FACS sorting. Single-cell qPCR expression level measurement was then performed for 24 genes. Housekeeping genes were only used for cell-cycle normalisation, where for each cell, all expression values were divided by the average expression of its housekeeping genes. Furthermore we excluded the five housekeeping genes, as well as *c-Kit*, which is a stem-cell receptor factor expressed on the surface of all analysed cells, from the diffusion map analysis.

2.5.2 qPCR data of mouse stem cells from zygote to blastocyst. To understand the earliest cell fate decision in a developing mouse embryo, Guo *et al.* (2010) conducted a qPCR experiment for 48 genes in seven different developmental time points. The gene expression levels were normalised to the endogenous controls *Actb* and *Gapdh*. The authors also identified four cell types, namely, inner cell mass (ICM), trophectoderm (TE), primitive endoderm (PE) and epiblast (EPI) using characteristic markers. The total number of single cells used in the diffusion map analysis was 429.

2.5.3 RNA-Seq of human preimplantation embryos. For the data set published by Yan *et al.* (2013), RNA-Seq analysis was performed on 90 individual cells from 20 oocytes and embryos. The sequenced embryos were picked at seven crucial stages of preimplantation: metaphase II oocyte, zygote, 2-cell, 4-cell, 8-cell, morula and late blastocyst at the hatching stage.

3 RESULTS

In this section we evaluate the performance of the diffusion map on each of the data sets described in the Methods section and compare it to the performance of two other dimension-reduction methods PCA and t-SNE. Data embeddings with several other methods including ICA, SPADE, kernel-PCA (Schölkopf *et al.*, 1998), isomap and Hessian Locally-Linear Embedding (HLLE) are provided in the supplementary Figures S16-S20.

3.1 Diffusion maps cope with high noise level and sampling density heterogeneity for toy data

3.1.1 Gaussian width determination of the toy data. We demonstrate the heuristic determination of σ on balanced and imbalanced toy data sets. The average dimensionality of the structure of some chosen characteristic length scale can be estimated by Equation 11. Figure 3 shows the average dimensionality $\langle d \rangle$ for balanced toy data (red) and imbalanced toy data (black) as a function of $\log(\sigma)$. The balanced set exhibits two maxima. The first one arises at the length scale of the thickness of the differentiation branches which include only a few cells. At this σ several subpopulations form at the more densely populated stages of the steady state. The second maximum appears at a larger length scale when several subpopulations become visible to each other and the continuous branches form. We picked the σ at the second maximum for visualisation (data visualisation at the first maximum is provided in the supplementary Figure S6). For the imbalanced set, however, due to the high noise level, the first maximum vanished and we only detected one maximum which we then used for the visualisation.

3.1.2 Performance of the diffusion map on the toy data as compared to the other methods. Definition of diffusion distance (Equation 8) based on probability of transition between cells through several paths renders diffusion maps very robust to noise. Figure 4 presents a comparison between the performance of the diffusion map and the other two methods PCA and t-SNE on the balanced toy data set. The eigenvalues of the diffusion map (Figure 4D) suggest that there are four leading dimensions that explain the data structure and the higher dimensions present noise rather than the intrinsic structure of the data manifold. The complete set of two-by-two projections up to the fourth eigenvector can be found in the supplementary Figure S7. PCA of this data set generated results that were similar to the diffusion map, where all four branches of the data could be distinguished. However, standard t-SNE did not preserve the data structure continuity. Visualisation using t-SNE with non-standard perplexity values are also provided in the supplementary Figure S8. To determine how additional extrinsic noise and density heterogeneities affect each method, we also applied the three methods on imbalanced toy data (Figure 5). The eigenvalues plot of the diffusion map in this figure suggests the same order of significance for the third and fourth eigenvectors as λ_4 almost equals λ_3 and that the higher-order eigenfunctions mostly present noise. We chose two projections (DC, DC2, and DC3) and (DC1, DC2, and DC4) for illustration in Figure 5. The complete set of two-by-two projection can be found in the supplementary Figure S9. From Figure 5A, one can infer the same size for all four branches of differentiation despite different sampling densities. This figure also suggests that the diffusion map clearly shows four branches of the imbalanced toy data, whereas PCA and t-SNE produce noisier visualisation and represent the two rarer branches as smaller. For additional t-SNE visualisations with non-standard perplexity values for the imbalanced toy data see Figure S10 in the supplement.

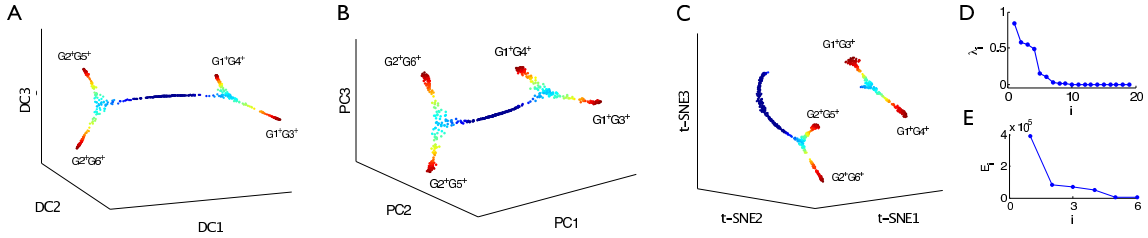


Fig. 4: Visualisation of the balanced toy data on A) the first three eigenvectors of the diffusion map, B) PCA and C) t-SNE. The colours (heat map of blue to red) indicate the maximum expression among all genes. Eigenvalues sorted in decreasing order for D) diffusion map and E) PCA.

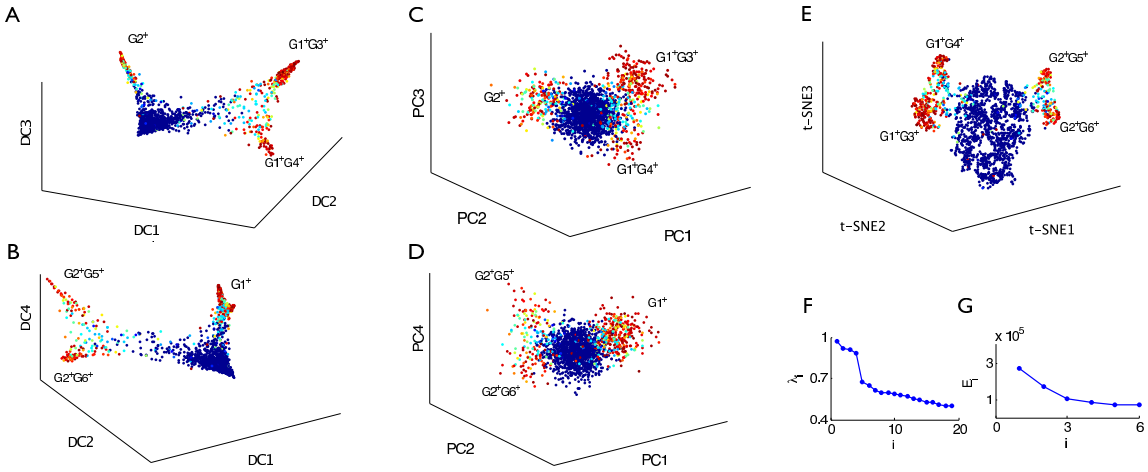


Fig. 5: Visualisation of the imbalanced toy data on A) the first three eigenvectors of the diffusion map, B) the first, second and fourth eigenvectors of the diffusion map, C) the first three components of the PCA D) the first, second and fourth components of PCA and E) t-SNE. The colours (heat map of blue to red) indicate the maximum expression among all genes. Eigenvalues sorted in a decreasing order for F) diffusion map and G) PCA.

3.1.3 Refinement of the transition matrix by density normalisation, zero diagonal and accounting for missing values. In order to adapt the standard diffusion map algorithm to the properties of single-cell gene expression parameters, we refined the transition matrix in different ways. First, we set the diagonal of the transition matrix to zero (Equation 5) since the (non-zero version) diagonal carries information about local sampling densities. Unlike many other applications where the information about local densities has some value, the sampling density in the context of single-cell data is somewhat arbitrary (e.g. only specific cell types are monitored, different proliferation rates in several stages of differentiation alters the sampling density, outlier cells show lower density, etc.). For a demonstration of how zero diagonal improves the quality of the diffusion map see supplementary Figure S11. Second, we refined the Markovian transition matrix by density normalisation (Equation 5) since the number of diffusion paths between two cells depends on the density of cells connecting them and more densely sampled regions of the data would seem to have smaller diffusion distance to each other on a diffusion map without density normalisation. Supplementary Figure S12 demonstrates how density normalisation improves the quality of the diffusion map. The third refinement that we used in our implementation of diffusion maps is accounting for missing and non-detect values (section 2.2). Generally speaking as the proportion of missing and non-detect values increases, there is a decrease in the quality of the diffusion map. However the magnitude of this effect depends highly on the architecture of the gene regulatory network and the role of the corresponding gene in the network. For example, for a toggle switch, low expression of a gene would always imply high expression of the other gene. Therefore, increasing the detection threshold (i.e. increasing number of non-detects) does not have a major influence on the analysis, as the information is still present in the other gene with high expression. We evaluate the performance of diffusion map in several proportions of missing values for the balanced toy data in supplementary Figure S13.

3.2 Diffusion map allows identification of differentiation trajectories on experimental data

3.2.1 Performance on haematopoietic stem cells qPCR data as compared to the other methods. The diffusion map embedding for the haematopoietic stem cells (Figure 6A) indicates a major

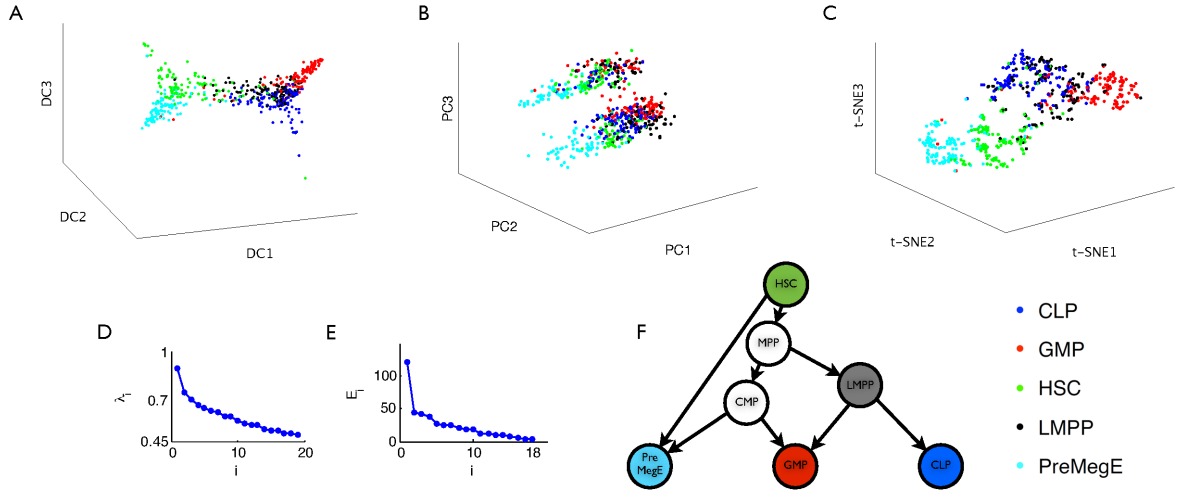


Fig. 6: Visualization of haematopoietic stem cells data on the first three eigenvectors of A) diffusion map, B) PCA and C) t-SNE. Eigenvalues sorted in a decreasing order for D) diffusion map and E) PCA. F) The hierarchy of haematopoietic cell types.

branching of HSCs to PreMegE and LMPP cell types and a further branching of LMPPs to CLP and GMP cells. The branching structures are less clear in the PCA plot (Figures 6B). Moreover, PCA produces artificial planes of data in the embedding because of the non-detect measurements in the qPCR data. The t-SNE plot (Figure 6C) almost separated the cell types (except for LMPPs) into different clusters. However, the notion of temporal progress is less clear compared to the diffusion map embedding. In addition, since uncertainties in the values of non-detects were not considered, a widening within the clusters is observed. Detailed visualisation using the three methods and the Gaussian width determination for diffusion map embedding are provided in the supplementary Figure S14. The ordered eigenvalues plot for the diffusion map and PCA are shown in Figures 6D and 6E. The ordered eigenvalues plot of the diffusion map suggests that there is no clear separation between the eigenvectors of the diffusion map that captures the intrinsic low-dimensional data manifold and those characterising noise for this data set. However, what makes the diffusion map embedding of this data set more plausible is the concordance between the branching structure as suggested by the diffusion map and the recently established hierarchy of haematopoietic cell types (Moignard *et al.*, 2013; Arinobu *et al.*, 2007) illustrated in Figure 6F.

3.2.2 Performance of the diffusion map on mouse embryonic stem cells qPCR data as compared to the other methods. For the mouse embryonic stem cells, diffusion map visualisation using the first three eigenvectors indicated a branching at the early 16-cell stage to the ICM and TE cell types, and further branching of the ICM at the late 32-cell stage into the EPI and PE (Figure 7A). The branching structure is unclear in the PCA and t-SNE plots (Figure 7B and 7C). The ordered eigenvalues plot for the diffusion map and PCA are shown in Figures 7D and 7E. The branching structure indicated by the diffusion map is in agreement with the results of previous studies on this data set (Guo *et al.*, 2010; Buettner and Theis, 2012), which suggests a branching into the two cell types, ICM and TE, after the 8-cell stage and further branching of the ICM into EPI and PE cells (Figure 7F). More information on Gaussian width determination and two-dimensional projections of data on each pair of the first to fourth eigenvectors of the diffusion map are provided in the supplementary Figure S15.

3.2.3 Performance on human pre-implantation embryos RNA-Seq data compared with other methods. The performance of the diffusion map on this RNA-Seq data set is comparable (although slightly sharper with respect to pseudo-time ordering) to the other two methods, PCA and t-SNE (Figure 8). The number of single cells measured in RNA-Seq is currently limited due to high sequencing costs. A low number of sampled cells could not meaningfully indicate a complex structure. Hence, PCA and t-SNE performance is almost as good as that of the diffusion map. However, with the expected development of new and cheaper RNA sequencing technologies, we propose a diffusion map that could be used as a powerful dimension-reduction tool the computation time of which is only linear with respect to the number of genes.

4 DISCUSSION AND CONCLUSION

In this manuscript, we have demonstrated the capabilities of diffusion maps for the analysis of continuous dynamic processes, in particular, differentiation data in a toy data set and a few experimental data sets. Using a biologically relevant distance metric (i.e. diffusion distance), the adapted diffusion map method outperforms other dimension-reduction methods in pseudo-temporal ordering of cells along the differentiation paths and could capture the expected differentiation structure in all cases. Table 1 provides a general comparison of several dimension-reduction methods, detailing capabilities and limitations in application to single-cell omics data. Among these methods, isomap and (H)LLE have not been applied for the analysis of single-cell differentiation data and pseudo-time ordering so far, mainly because they do not meet the specific requirements for the analysis of such data

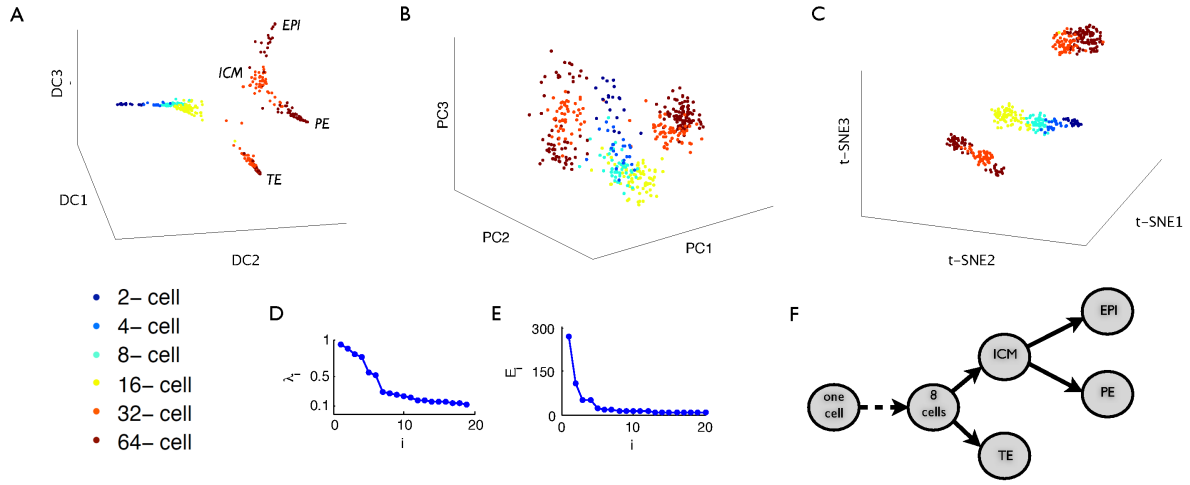


Fig. 7: Visualisation of mouse embryonic stem cells on A) the first three eigenvectors of diffusion map, B) PCA and C) t-SNE. Eigenvalues sorted in a decreasing order for D) diffusion map and E) PCA. F) The hierarchy of cells for mouse embryonic stem cells.

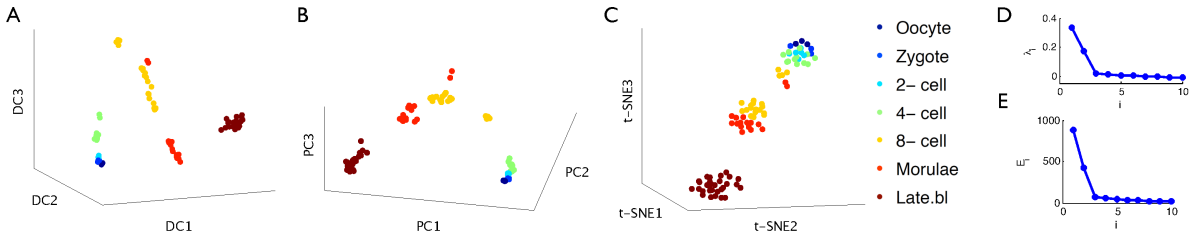


Fig. 8: Visualisation of human preimplantation embryos data on A) the first three eigenvectors of the diffusion map, B) PCA, and C) t-SNE. Eigenvalues sorted in a decreasing order for D) the diffusion map and E) PCA.

including capability to handle high levels of technical noise, sampling density heterogeneities, detection limits and missing values. Figures S19 and S20 in the supplement demonstrate the poor performance of these methods for finding the differentiation manifold in presence of noise and density heterogeneity for our toy data set as well as the three experimental single-cell data sets. For any data set, it is important to consider the advantages and disadvantages of each method with respect to the data properties and the purpose of the analysis, in order to make a suitable choice for applying to that data set.

In our diffusion maps implementation, by performing density normalisation and setting the diagonal of the transition probability matrix to zero, we propose a mapping technique wherein the closeness of cells in the diffusion metric is unaffected by density heterogeneities in data sampling (see supplementary Figures S9 and S10). This feature can be crucial for the detection of rare populations, which is one of the main challenges in the analysis of differentiation data.

By breaking the diffusion kernel (Mohri *et al.*, 2012) to its multiplicand wave functions, we also propose a method in accommodating the uncertainties of measurement and missing values into the wave function. Consequently, we have successfully addressed uncertainties in the value of non-detects in qPCR data.

Tuning the scale parameter σ is also important for generating insights into the structure of the data, for which we proposed a criterion on the basis of the characteristic length scales of the data manifold. Because of computational limitations, for our criterion we compute the average intrinsic dimensionality and hence the average characteristic length scale. However, when density heterogeneities are extremely large, or the data manifold has many sharp changes and several scales, a single σ may not provide a globally optimal scale for data embedding. Therefore, implementation of an efficient and cost-effective method for several locally valid σs determinations, instead of a single global value is of interest.

It is worth noting that the mathematical ergodicity in diffusion maps reached by adequate kernel width selection does not necessarily imply biological ergodicity. If there appears a trace of transitory cells between two clusters, we conclude the two clusters are also biologically connected to each other in an ergodic sense. However this trace might be not present if the transition is too fast or switch-like abrupt, so that no transitory cells have been caught in the finite set of sampled cells of snapshot data. Thus it has to be proven with dedicated biological experiments (e.g. as used by Buganim *et al.* (2012) and Takahashi and Yamanaka (2006)) whether the data is biologically ergodic or not.

	ref.	methodology	linear/ non-linear	structure faithfulness	robustness to noise / density heterogeneities	no. of dims needed for embedding	handles missing / uncertain values?	keeps single-cell resolution?	clustering / keeping continuity	tuning parameters	best performance
PCA	Hotteling, 1993	orthogonal transformation	linear	global	+ / -	depends on eigenvalues	+ (Buettner *et al.*, 2014)	+	- / +	none	linear data subspace
ICA	Stone, 2004	orthogonal transformation	linear	global	+ / -	arbitrary	-	+	- / +	none	linear data subspace, known no. of sources
SPADE	Qui *et al.*, 2011	agglomerative / k-means clustering, minimum spanning trees	non-linear	local and (weak) global	- / +	2D	-	-	+ / +	-outlier density -target density -desired no. of clusters	low noise, desired no. of clusters $\gtrsim O(2^{d*})$
t-SNE	Van der Maaten and Hinton, 2008	attraction / repulsion balance	non-linear	local	+ / ++	2 or 3D	-	+	+ / -	perplexity	clustering to separate groups, presence of noise and density heterogeneities
kernel-PCA	Scholkopf *et al.*, 1998	kernel methods	non-linear	global	+ / -	depends on eigenvalues	+ (Buettner *et al.*, 2014)	+	+ / +	depends on the used kernel	physically relevant kernel
Isomap	Tenenbaum *et al.*, 2000	spectral clustering, geodesic distance	non-linear	global	- / +	depends on eigenvalues	-	+	- / +	no. of nearest neighbours	low noise or a priory known geodesics
(H)LLE	Donoho and Grimes, 2003	weighted linear combination of nearest neighbours	non-linear	global	- / -	arbitrary	-	+	- / +	no. of nearest neighbours	continuos data manifold, low noise, uniform sampling
Diffusion map	Coifman *et al.*, 2005	spectral clustering, diffusion distance	non-linear	global	++ / +	depends on eigenvalues	+ (our implementation)	+	- / +	kernel width	continuos data manifold, presence of noise and density heterogeneity

* d is the intrinsic dimensionality of the data manifold

Table 1: Comparison of several dimension-reduction algorithms in the view of single-cell omics data application.

A possible strategy for enhancing the capacity of capturing details of the structure of rare populations using diffusion maps is to limit the transition possibility of each cell only to its closest neighbours. In this scenario, we could render the diffusion map more local by building the transition matrix $\tilde{P}$ in Equation 6 for k nearest neighbours only. This method, however, might end up with several disconnected sub-graphs of cells when the sampling density along the intrinsic data manifold is extremely heterogeneous. Furthermore, $\tilde{P}$ (without the row normalisation) will not be symmetric any more and we cannot ensure real eigenvalues for the transition probability matrix. However, as long as the graph is connected and eigenvalues are real, we can benefit from a more locally detailed map.

One caveat in the current version of diffusion map is the $n^2 \times G$ computation time which can be prohibitive for large cell numbers ($> 10^4$) as generated from cytometry experiments. Choosing the k nearest neighbours version of diffusion map can therefore be a solution to this problem. Diffusion distances are based on a robust connectivity measure between cells which is calculated over all possible paths of a certain length between the cells. Thus, a diffusion mapping obtained by accounting for a smaller fraction of all possible paths (namely those going through each cells' nearest neighbours) can still provide a good approximation of the diffusion distance between the cells and yet avoid computing all n^2 elements of the transition probability matrix. With such modifications, diffusion maps prevail as a promising method for the analysis of large cell numbers omics data.

Another issue is the number of embedding dimensions. The number of significant dimensions of the diffusion map is determined where a remarkable gap occurs in its sorted eigenvalues plot. This is not intrinsically bound to the conventional visualisable dimensions two or three. In contrast, for some other methods such as t-SNE, one can pre-determine the number of visualisation dimensions for the embedding optimisation to two or three dimensions.

We conclude that diffusion maps are appropriate and powerful for the dimension-reduction of single-cell qPCR and RNA-Seq cell differentiation data as they are able to handle high noise levels, sampling density heterogeneities, and missing and uncertain values. As a result diffusion maps can organise single cells along the nonlinear and complex branches of differentiation, maintain the

global structure of the differentiation dynamics and detect rare populations as well.

ACKNOWLEDGEMENT

We would like to thank Michael Strasser (Institute for Computational Biology, Helmholtz Centre Munich), Victoria Moignard (Cambridge Institute of Medical Research), Berthold Goettgens (Cambridge Institute of Medical Research) and Mauro Maggioni (Department of Mathematics, Duke University) for helpful advice and discussions.

Funding: This study has been funded by The Bavarian Research Network for Molecular Biosystems (BioSysNet) and the European Research Council (ERC starting grant LatentCauses).

REFERENCES

Amir, E.-a. D., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., Shenfeld, D. K., Krishnaswamy, S., Nolan, G. P., and Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature biotechnology*, **31**(6), 545–552.

Arinobu, Y., Mizuno, S.-i., Chong, Y., Shigematsu, H., Iino, T., Iwasaki, H., Graf, T., Mayfield, R., Chan, S., Kastner, P., *et al.* (2007). Reciprocal activation of GATA-1 and PU. 1 marks initial specification of hematopoietic stem cells into myeloerythroid and myelolymphoid lineages. *Cell stem cell*, **1**(4), 416–427.

Bandura, D. R., Baranov, V. I., Ornatsky, O. I., Antonov, A., Kinach, R., Lou, X., Pavlov, S., Vorobiev, S., Dick, J. E., and Tanner, S. D. (2009). Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Analytical Chemistry*, **81**(16), 6813–6822.

Bendall, S. C., Davis, K. L., Amir, E.-a. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., Shenfeld, D. K., Nolan, G. P., and Peer, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, **157**(3), 714–725.

Buettner, F. and Theis, F. J. (2012). A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. *Bioinformatics*, **28**(18), i626–i632.

Buettner, F., Moignard, V., Göttgens, B., and Theis, F. J. (2014). Probabilistic PCA of censored data: accounting for uncertainties in the visualisation of high-throughput single-cell qPCR data. *Bioinformatics*, page btu134.

Buganim, Y., Faddah, D. A., Cheng, A. W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S. L., van Oudenaarden, A., and Jaenisch, R. (2012). Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell*, **150**(6), 1209–1222.

Chattopadhyay, P. K., Price, D. A., Harper, T. F., Betts, M. R., Yu, J., Gostick, E., Perfetto, S. P., Goepfert, P., Koup, R. A., De Rosa, S. C., *et al.* (2006). Quantum dot semiconductor nanocrystals for immunophenotyping by polychromatic flow cytometry. *Nature medicine*, **12**(8), 972–977.

Chu, Y. and Corey, D. R. (2012). RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic acid therapeutics*, **22**(4), 271–274.

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(21), 7426–7431.

Donoho, D. L. and Grimes, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, **100**(10), 5591–5596.

Dykstra, B., Kent, D., Bowie, M., McCaffrey, L., Hamilton, M., Lyons, K., Lee, S.-J., Brinkman, R., and Eaves, C. (2007). Long-term propagation of distinct hematopoietic differentiation programs in vivo. *Cell stem cell*, **1**(2), 218–229.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, **81**(25), 2340–2361.

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

Huang, S. (2009). Non-genetic heterogeneity of cells in development: more than just noise. *Development*, **136**(23), 3853–3862.

Krumsiek, J., Marr, C., Schroeder, T., and Theis, F. J. (2011). Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network. *PloS one*, **6**(8), e22649.

McDavid, A., Finak, G., Chattopadyay, P. K., Dominguez, M., Lamoreaux, L., Ma, S. S., Roederer, M., and Gottardo, R. (2013). Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics*, **29**(4), 461–467.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of machine learning*. MIT Press.

Moignard, V., Macaulay, I. C., Swiers, G., Buettner, F., Schütte, J., Calero-Nieto, F. J., Kinston, S., Joshi, A., Hannah, R., Theis, F. J., *et al.* (2013). Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nature cell biology*, **15**(4), 363–372.

Orkin, S. H. and Zon, L. I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, **132**(4), 631–644.

Park, H. Y., Lim, H., Yoon, Y. J., Follenzi, A., Nwokafor, C., Lopez-Jones, M., Meng, X., and Singer, R. H. (2014). Visualization of dynamics of single endogenous mrna labeled in live mouse. *Science*, **343**(6169), 422–424.

Qiu, P., Simonds, E. F., Bendall, S. C., Gibbs Jr, K. D., Bruggner, R. V., Linderman, M. D., Sachs, K., Nolan, G. P., and Plevritis, S. K. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nature biotechnology*, **29**(10), 886–891.

Rieger, M. A., Hoppe, P. S., Smejkal, B. M., Eitelhuber, A. C., and Schroeder, T. (2009). Hematopoietic cytokines can instruct lineage choice. *Science*, **325**(5937), 217–218.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, **10**(5), 1299–1319.

Schroeder, T. (2011). Long-term single-cell imaging of mammalian stem cells. *Nature methods*, **8**(4s), S30–S35.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Stingl, J., Eirew, P., Ricketson, I., Shackleton, M., Vaillant, F., Choi, D., Li, H. I., and Eaves, C. J. (2006). Purification and unique properties of mammary epithelial stem cells. *Nature*, **439**(7079), 993–997.

Strasser, M., Theis, F. J., and Marr, C. (2012). Stability and multiattractor dynamics of a toggle switch based on a two-stage model of stochastic gene expression. *Biophysical journal*, **102**(1), 19–29.

Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell*, **126**(4), 663–676.

Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**(5500), 2319–2323.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, **9**(11).

Wilhelm, J. and Pingoud, A. (2003). Real-time polymerase chain reaction. *Chembiochem*, **4**(11), 1120–1128.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., *et al.* (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*.

Supplementary Materials for Diffusion maps for high-dimensional single-cell analysis of differentiation data

Laleh Haghverdi [1,2], Florian Buettner [1], and Fabian J. Theis [1,2]

[1]Institute of Computational Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany
[2]Department of Mathematics, Technische Universität München, 85748 Garching, Germany

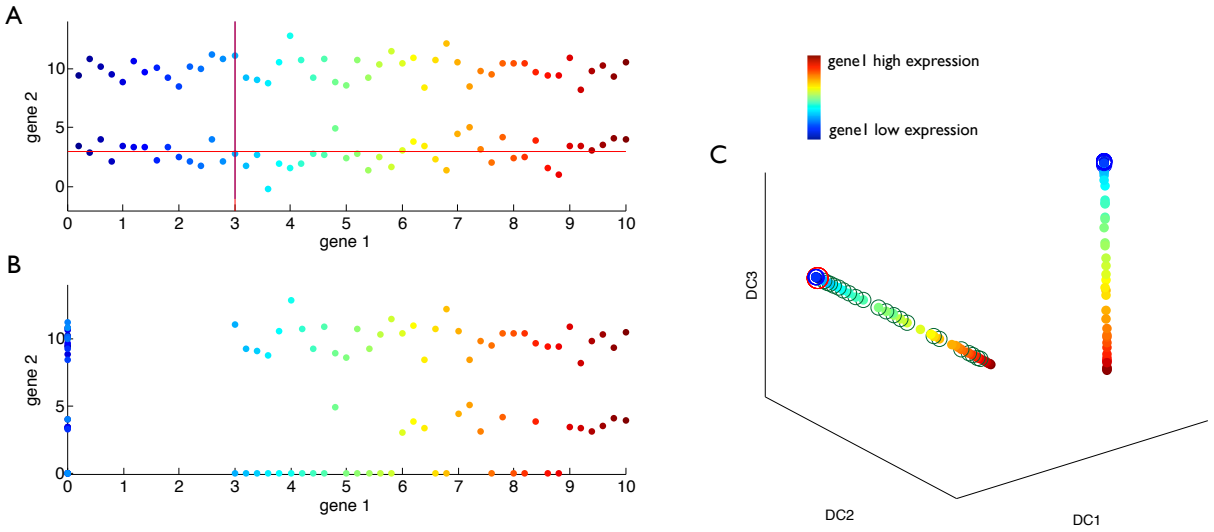


Figure S1: A) A toy "true" data with two genes expression. The limit of detection is assumed at the value 3 for both genes (red lines). B) "Measured" toy data, where any "true" value below the detection limit in A is measured as zero (non-detects). C) Diffusion map obtained by assuming uniform prior distribution between 0 and 3 (limit of detection value) for the non-detects in the "measured" data recovers the two independent clusters of "true" data. Cells with non-detect values for *gene*1 are indicated by a green circle around them and cell with non-detect values for *gene*2 by a blue circle. Red circle indicates the cells with both genes as non-detects (i.e. [0,0] in B).

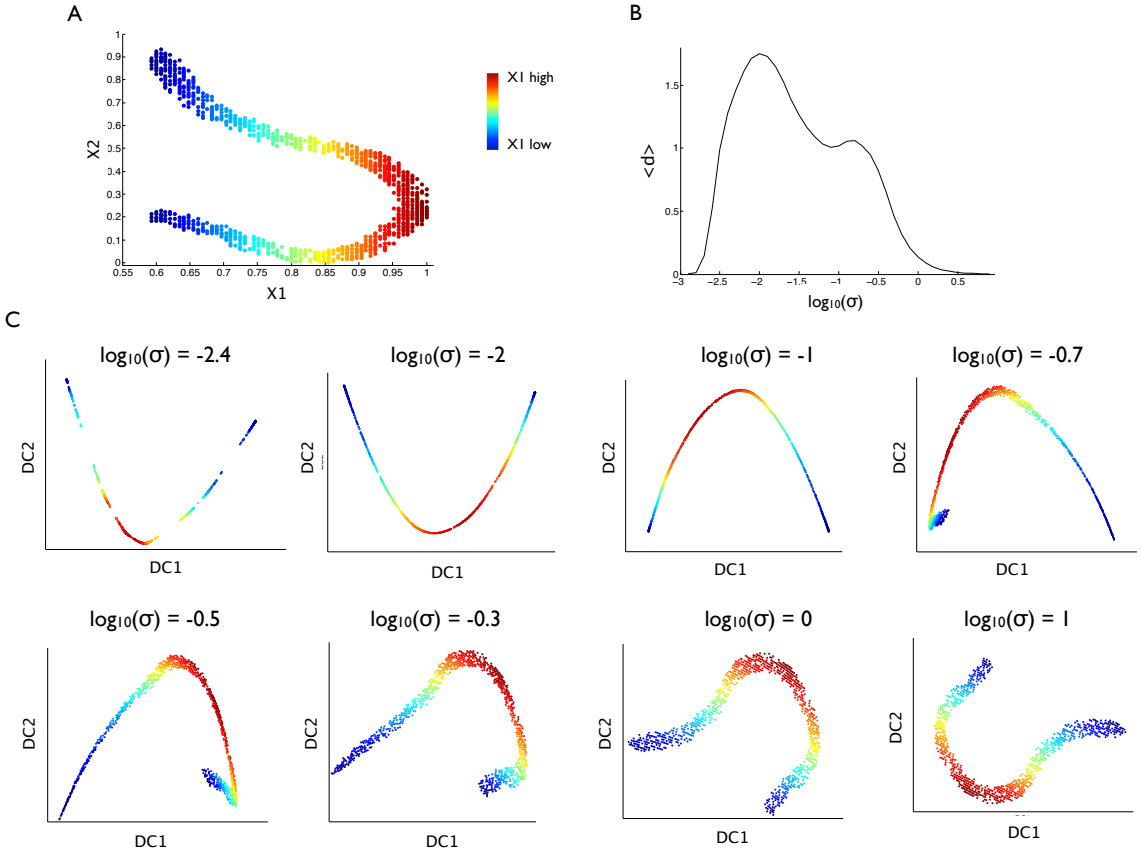


Figure S2: Kernel width determination for a U-shaped data. A) A U-shaped toy data. B) The average intrinsic dimensionality of the data over kernel with variations. The two maximums of the average dimensionality indicate two characteristic length scales of the data, namely the width of the strip and the diameter of the curves of the U (≈ 0.2). C) Diffusion map performance over several choices of the kernel width. At the first optimum of $< d >$ (i.e. $\log_{10}(\sigma) = -2$) the data is embedded as a continuous one dimensional strip. At the second optimum of $< d >$ (i.e. $\log_{10}(\sigma) = -0.7$) several parts of the U become visible to each other.

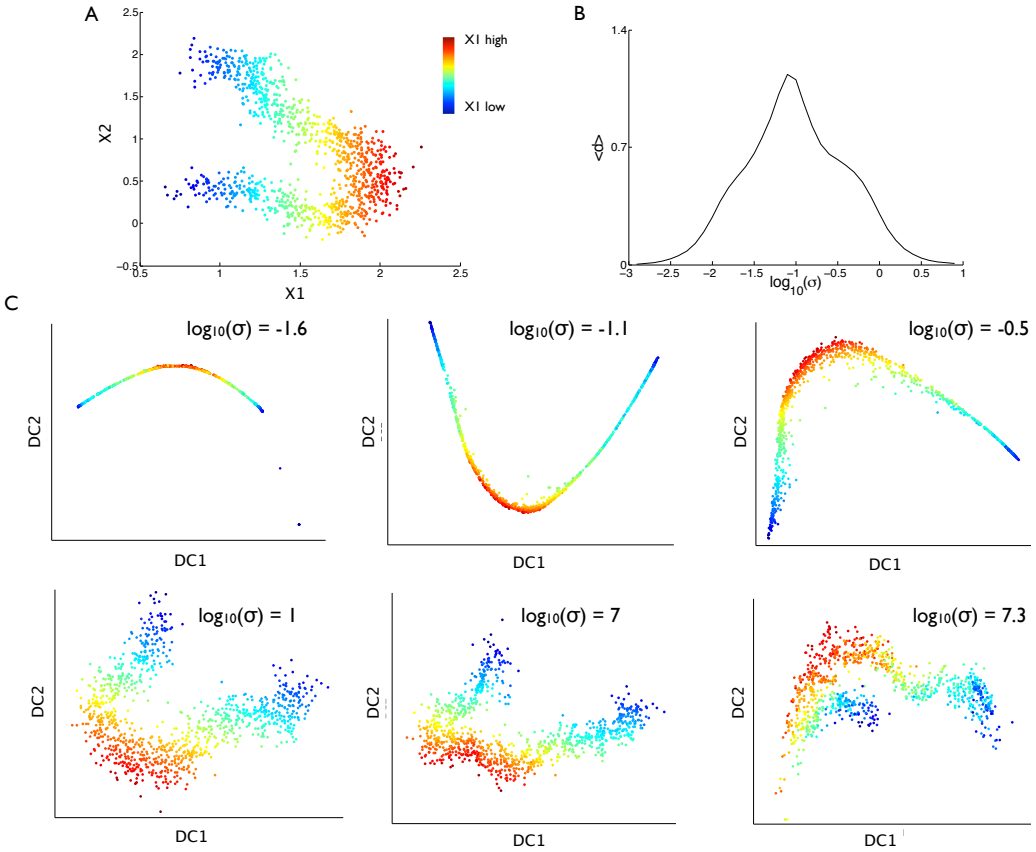


Figure S3: Kernel width determination for a noisy U-shaped data. A) A U-shaped toy data. B) The average intrinsic dimensionality of the data over kernel with variations. The plot suggests an optimum map at $\log_{10}(\sigma) = -1.1$ (i.e.. $\sigma \approx 0.08$) which is about the width of the data. C) Diffusion map with several kernel width. At the optimum σ the data is embedded as a one dimensional strip. The map shows some widening corresponding to the increase of σ but is still robust for a remarkable range of the kernel widths until (for $\log_{10}(\sigma) \geq 7$) it starts distorting from the true structure of the data.

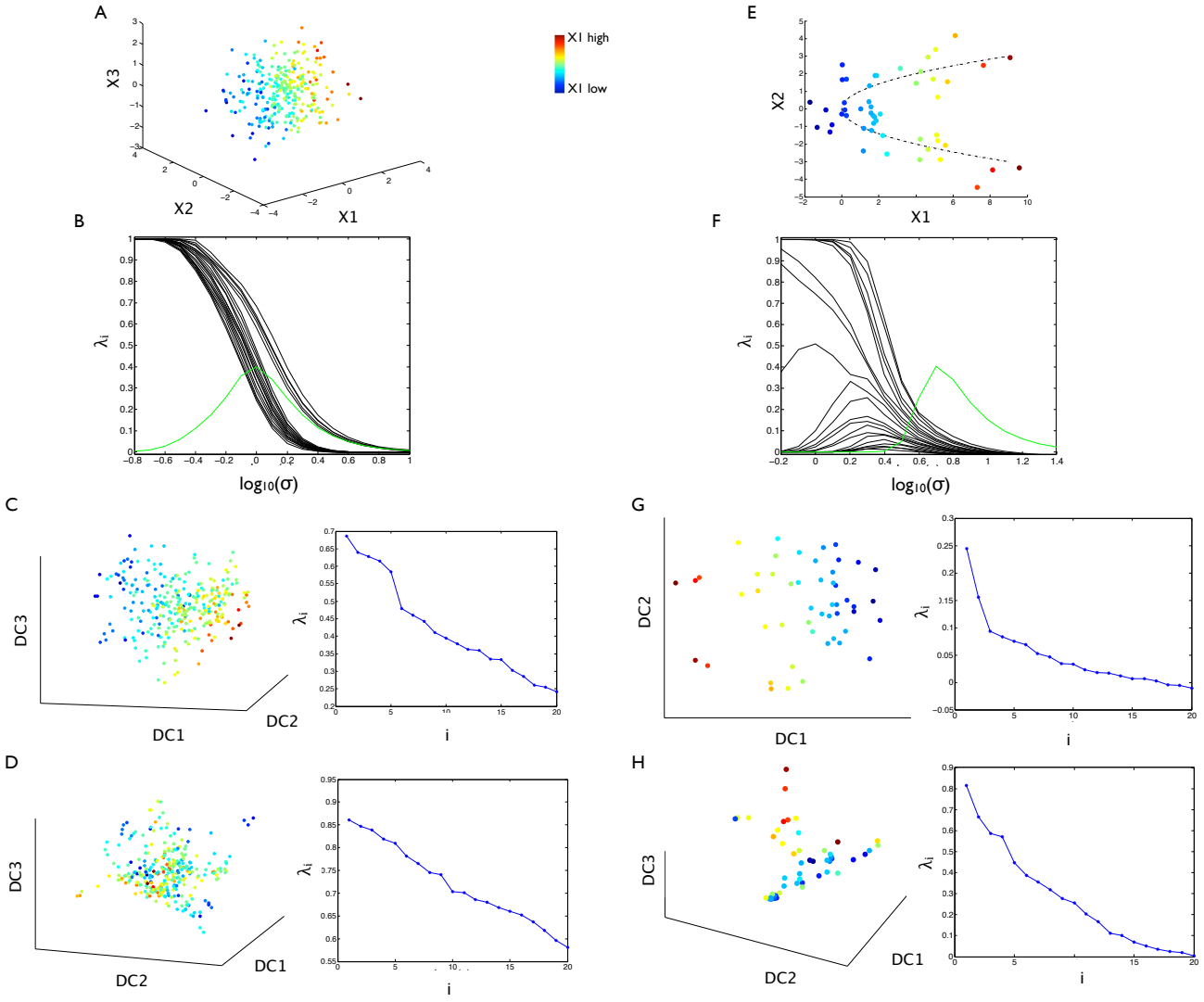
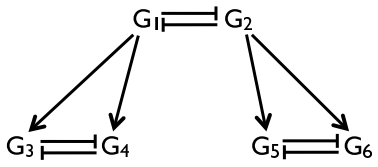

A
X3
X2
X1
Xl high
Xl low
B
λi
log10(σ)
E
X2
X1
F
λi
log10(σ)
C
DC3
DC1
DC2
λi
i
G
DC2
DC1
λi
i
D
DC3
DC2
DC1
λi
i
H
DC3
DC2
DC1
λi
i

Figure S4 *(preceding page)*: A) A normal distribution (zero mean, variance one) toy data with 300 cells and five dimensions (i.e. genes), illustrated on three arbitrary dimensions. B) The diffusion maps eigenvalues plot of the normal distributed toy data set over several kernel widths (σ) (black lines) shows the separation of five signal eigenvalues from the rest noise eigenvalues. The intrinsic dimensionality curve (green) has an optimum in the signal/noise distinction region at $\log_{10}(\sigma) = 0$. C) The diffusion map at the optimum kernel width $\log_{10}(\sigma) = 0$ shows the correct pattern of data. The right plot shows the eigenvalues of the diffusion map at $\log_{10}(\sigma) = 0$. D) The diffusion map at a smaller kernel width $\log_{10}(\sigma) = 0.2$ shows some pattern which is not true, because of the wrongly chosen kernel width. The right plot shows the eigenvalues of the diffusion map at $\log_{10}(\sigma) = 0.2$.
E) A sparse toy data with 50 cells and 100 dimensions (i.e. genes), where two genes (X1 and X2) are centred around the dashed line, and the 98 other genes are from a normal distribution with zero mean and variance one, illustrated on the X1 and X2 dimensions. F) Diffusion maps eigenvalues plot of the sparse toy data set over several kernel widths (σ) (black lines). The intrinsic dimensionality curve (green) has an optimum at a region which indicates the signal/noise eigenvalues are not well separated. G) The diffusion map at the optimum kernel width $\log_{10}(\sigma) = 0.7$ shows the correct pattern of data. The right plot shows the eigenvalues of the diffusion map at $\log_{10}(\sigma) = 0.7$. H) The diffusion map at a smaller kernel width $\log_{10}(\sigma) = 0.3$ shows some pattern which is not true, because of the wrongly chosen kernel width. The right plot shows the eigenvalues of the diffusion map at $\log_{10}(\sigma) = 0.3$.

A



$$\varnothing \xrightarrow{\alpha_i} \text{Protein}_i \xrightarrow{\gamma} \varnothing$$

B

$$\gamma = 5 \times 10^{-5}$$

$$\alpha_1 = 0.05 \frac{1}{1 + (\frac{\text{Protein}_2}{200})^2}$$

$$\alpha_2 = 0.05 \frac{1}{1 + (\frac{\text{Protein}_1}{200})^2}$$

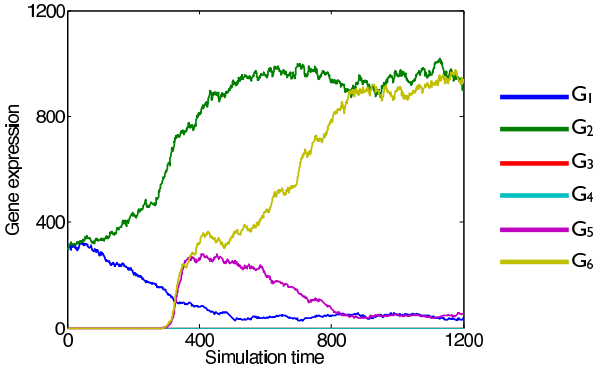
$$\alpha_3 = 0.05 \frac{1}{1 + (\frac{\text{Protein}_4}{200})^2} \cdot \frac{1}{1 + (\frac{700}{\text{Protein}_1})^{20}}$$

$$\alpha_4 = 0.05 \frac{1}{1 + (\frac{\text{Protein}_3}{200})^2} \cdot \frac{1}{1 + (\frac{700}{\text{Protein}_1})^{20}}$$

$$\alpha_5 = 0.05 \frac{1}{1 + (\frac{\text{Protein}_6}{200})^2} \cdot \frac{1}{1 + (\frac{700}{\text{Protein}_2})^{20}}$$

$$\alpha_6 = 0.05 \frac{1}{1 + (\frac{\text{Protein}_5}{200})^2} \cdot \frac{1}{1 + (\frac{700}{\text{Protein}_2})^{20}}$$

C



D

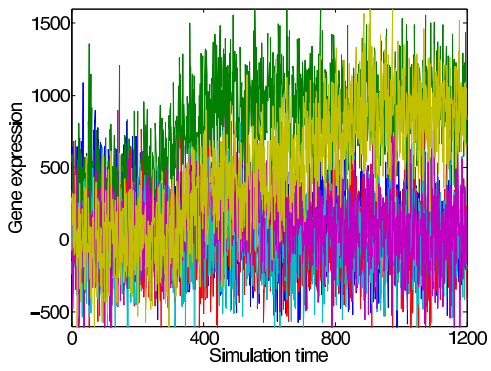


Figure S5: A) The corresponding protein of each of the six genes in the regulatory network the toy model is produced at rate α_i and degraded at a rate γ constant for all proteins. B) The reaction (production and decay) rates for all proteins. The simulated gene expression levels for a sample cell initiated from the pluripotent state at time zero which finally ends up in the $G2^+G6^+$ fully differentiated state from the C) balanced toy data set and D) imbalanced toy data set.

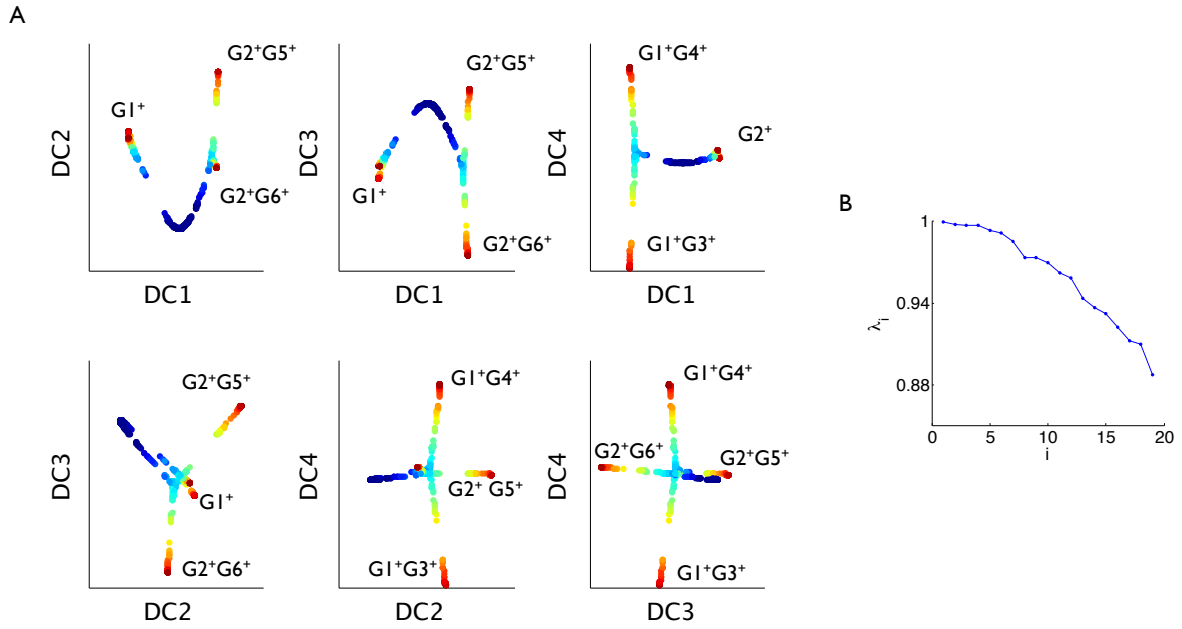


Figure S6: A) Visualisation of the balanced toy data on the first four eigenvectors of diffusion map at the first maximum of $\langle d \rangle$; $\sigma = 10^{1.6}$. Colours (heat map blue to red) indicate the maximum expression among all genes on each cell. B) Eigenvalues sorted in decreasing order for diffusion map.

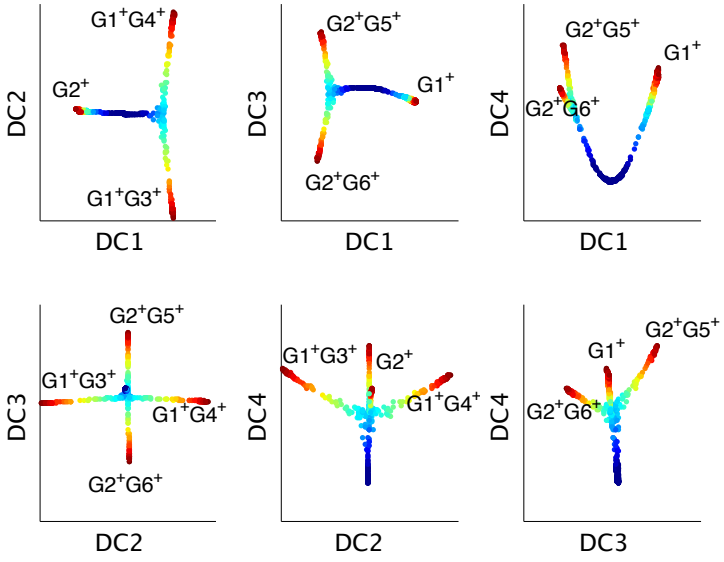


Figure S7: Visualisation of the balanced toy data on the first four eigenvectors of diffusion map at the second maximum of $\langle d \rangle$; $\sigma = 10^{2.7}$. Colours (heat map blue to red) indicate the maximum expression among all genes on each cell.

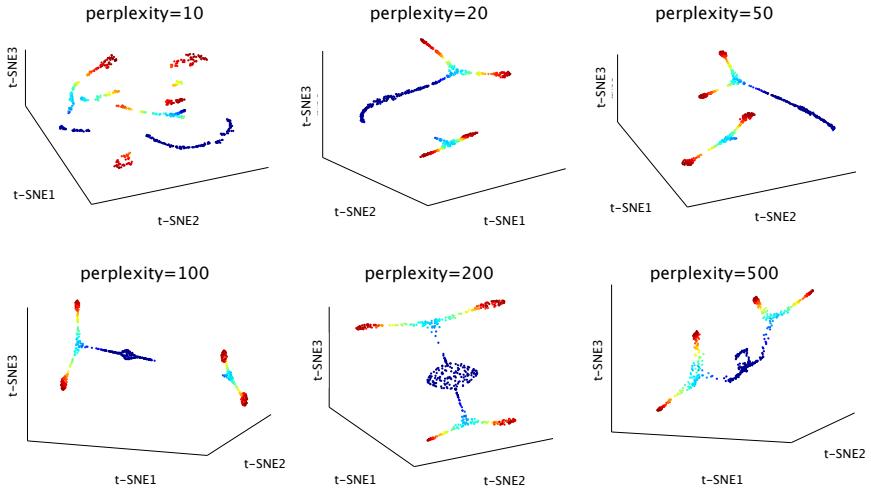


Figure S8: t-SNE with several perplexity values for the balanced toy data. Although the global structure could be kept for simple structures with more careful tuning of the perplexity parameter, t-SNE is not in general an appropriate method for pseudo-time ordering of single-cells.

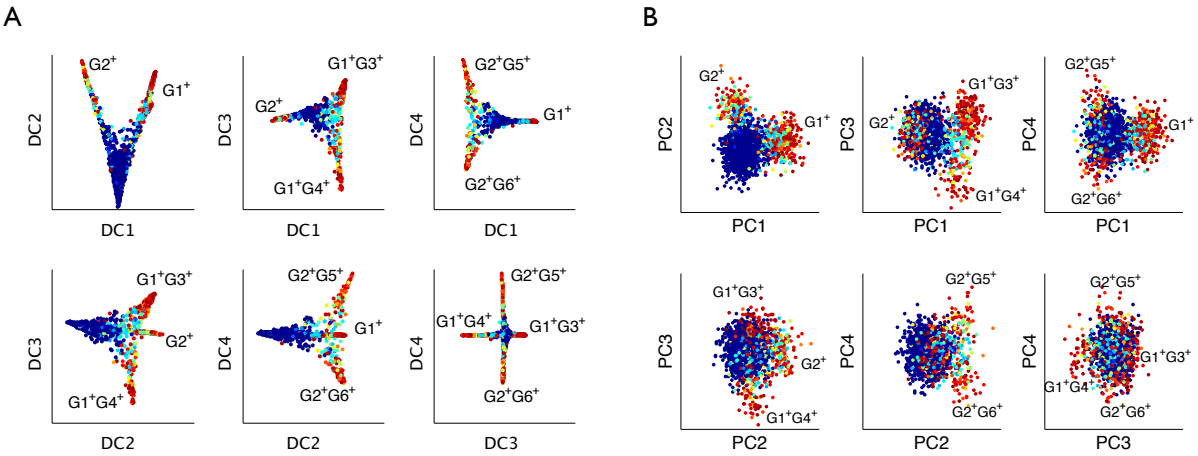


Figure S9: Visualisation of imbalanced toy data on the first four eigenvectors of: A) diffusion map at $\sigma = 10^{2.5}$, B) PCA [3] . Colours (heat map blue to red) indicate the maximum expression of all genes on each cell.

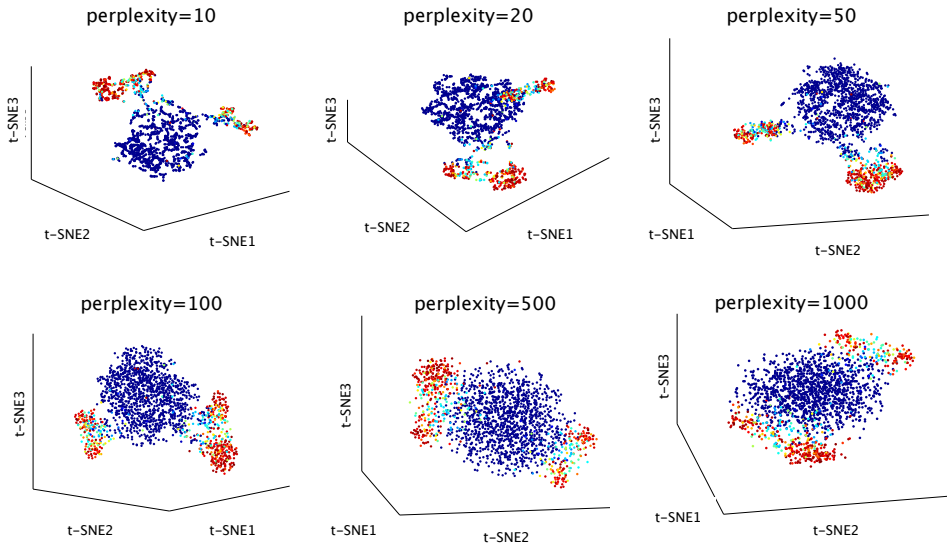


Figure S10: t-SNE with several perplexity values for the imbalanced toy data.

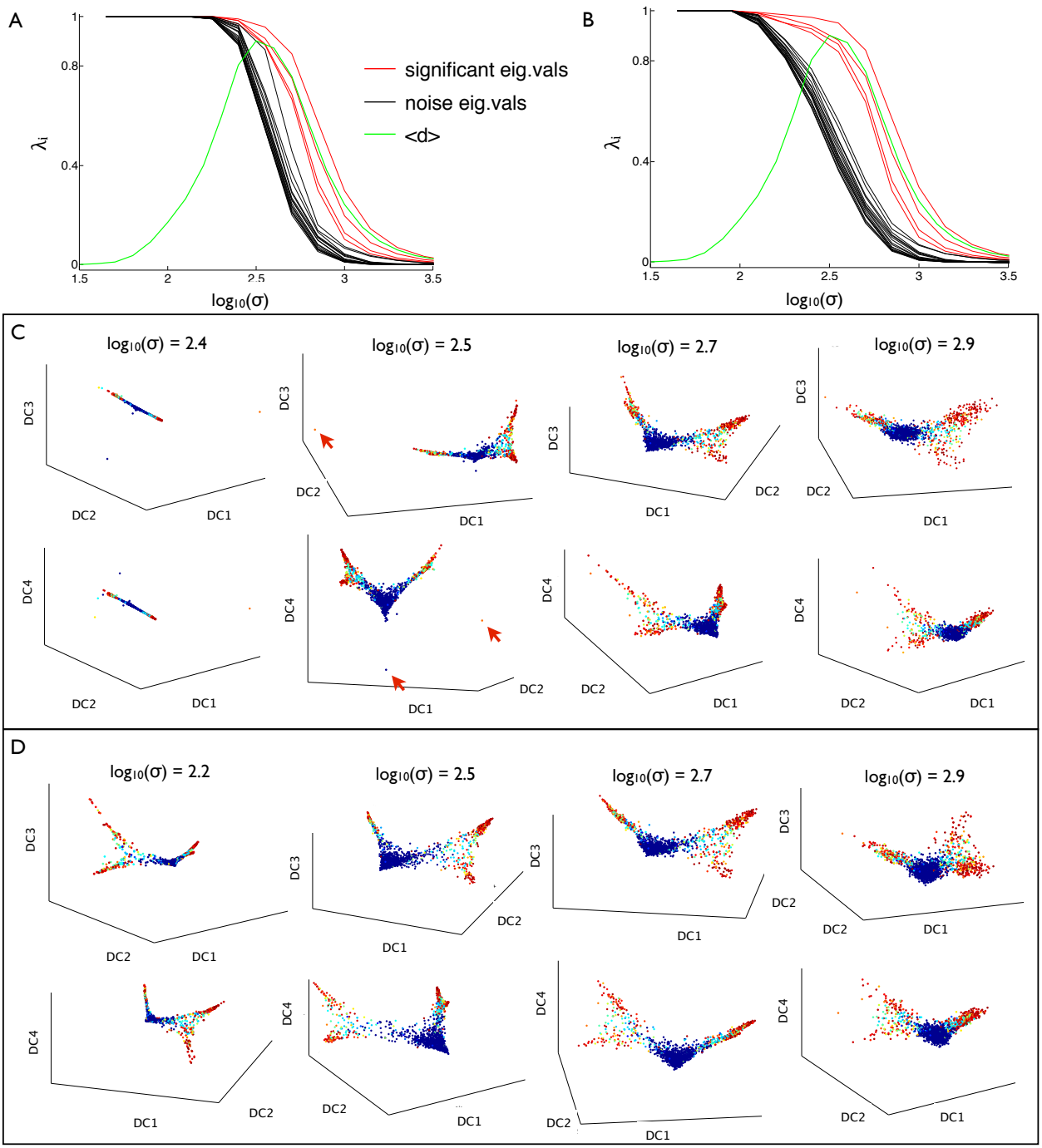

A
B
C
D
significant eig.vals
noise eig.vals
<d>
λi
log10(σ)
log10(σ) = 2.4
log10(σ) = 2.5
log10(σ) = 2.7
log10(σ) = 2.9
log10(σ) = 2.2
DC1
DC2
DC3
DC4

Figure S11 *(preceding page)*: A) Eigenvalues plot over several kernel widths (σ) for the original version of diffusion maps applied on the imbalanced toy data set. B) Eigenvalues plot over several kernel width (σ) for the diffusion maps with zero diagonal transition matrix. The eigenvalue plot of the zero diagonal transition matrix version gives better separation between noise and signal eigenvalues.
C) The original version diffusion map over eigenvalues 1,2,3 (top) and 1,2,4 (bottom) for several σ. For relatively low σ, the map produces some anomalies marked with red arrows.
D) The diffusion map with zero diagonal transition matrix over eigenvalues 1,2,3 (top) and 1,2,4 (bottom) for several σ. The zero diagonal transition matrix does not produce anomalies at low σ and allows more robust diffusion mapping at a wider range of kernel width variation. Zero diagonal becomes important when a large density heterogeneity is present and results in the non-noise eigenvectors being easier to distinguish from the signal eigenvectors. In the original version of diffusion maps, the diagonal of the transition matrix corresponds to the local density at each cell's position. Thus density heterogeneities cause more diffusion barriers (at cells with low density) and a bumpy diffusion space, which has a similar effect to increased noise. While in many applications these local densities are meaningful, sampling density in single-cell data is somehow arbitrary (e.g only specific cell types are monitored, different cell proliferation rates in several stages of differentiation alters the sampling density, etc.). Thus omitting any asymmetries (e.g. by setting the diagonal to zero) among the cells caused by this arbitrary sampling density can improve the quality of the mapping and its robustness in respect of varying kernel width σ.

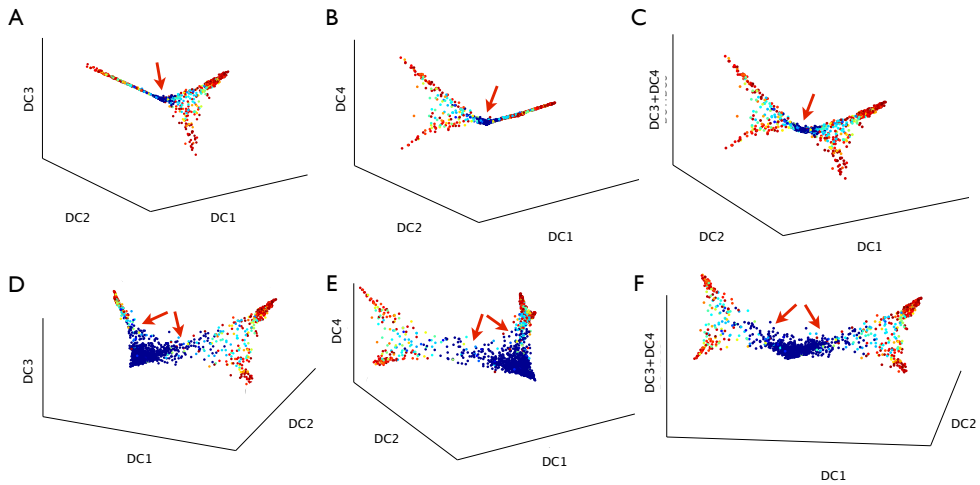


Figure S12: The effect of density normalisation for diffusion maps. A to C) The diffusion map of the imbalanced toy data set obtained without density normalisation is shown on the top row (A, B and C) in three different views of the four dimensional map. Without the density normalisation, diffusion between cells in more densely sampled regions is alleviated simply because more diffusion paths exist between them. Thus, these cells are assigned a smaller diffusion distance to each other (marked with the red arrow) and are placed densely in the same location on the map (crowding problem). D to F) The diffusion map of the imbalanced toy data set obtained by density normalisation is shown on the bottom row in three different views (D, E and F) of the four dimensional map. The cells near the multipotent state are correctly distributed close to each other on the map (region marked with the red arrows), since the effect of heterogeneous sampling density has been cancelled. For a further demonstration of the density normalization effect see [1].

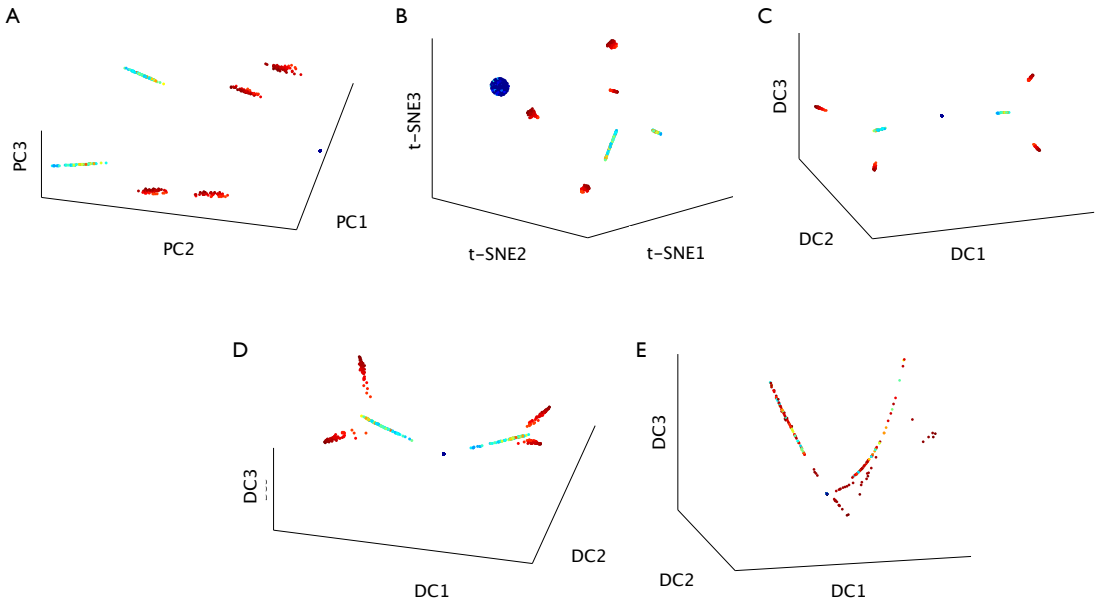


Figure S13: The effect of increasing the proportion of non-detects strongly depends on the architecture of the underlying regulatory network of the genes. While for real qPCR data the fraction of censored values is usually about 30%, diffusion map recovers the data manifold for the balanced toy data (see Figure S5) when up to 83% of the data is censored (all expression values below 800 where censored to zero). However not even diffusion map helps when the fraction of the censored values is too high (94% in this case, where all expression values below 950 where censored to zero).
A) PCA performance on the 83% censored data. B) t-SNE with perplexity= 30 performance on the 83% censored data. C) Diffusion map without accounting for missing values at the optimum kernel width applied on the 83% censored data. D) Diffusion map with accounting for missing values at the optimum kernel width applied on the 83% censored data. E) Diffusion map with accounting for missing values applied on the 94% censored data.
In general, when increasing the proportion of non-detects, the quality of the map depends on whether with the provided measured genes and the prior distributions for the missing genes, a cell can still find the neighbourhood where it originally belonged to or not. If the answer is yes the cell will be put close to them on the diffusion map, irrespective to the number of cells which have missing values imperfections. If there is a degeneracy in a cell's minimum diffusion distance to several neighbourhoods, the cell will be put on a separate cluster on the map.

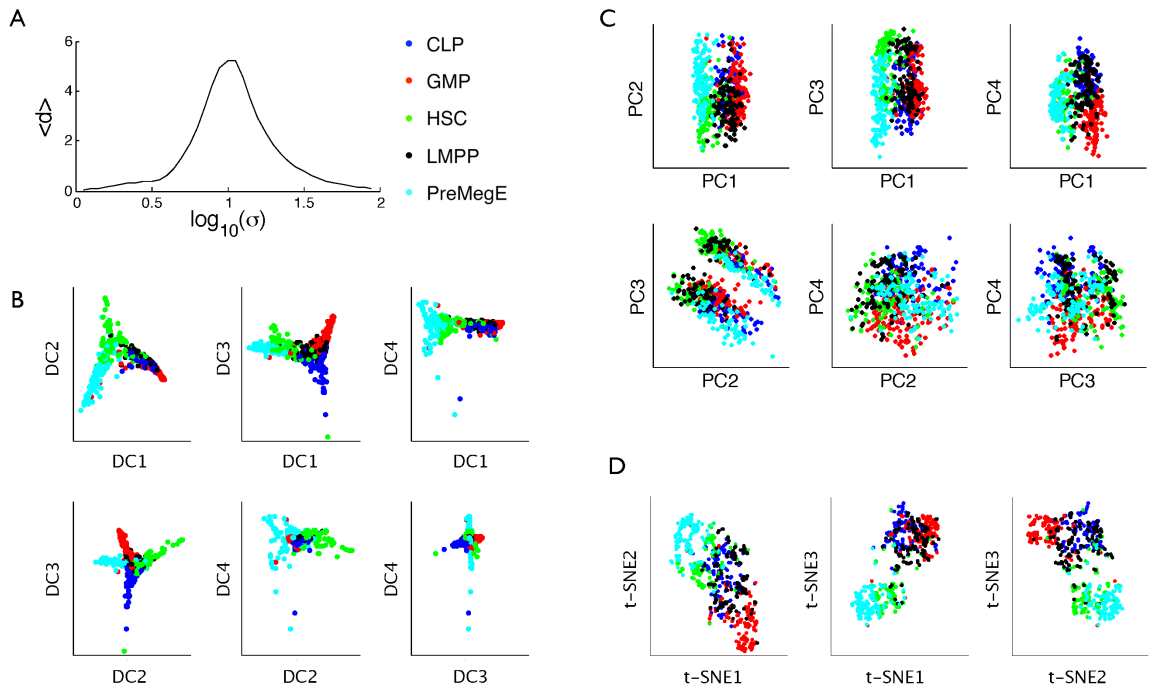


Figure S14: A) The average dimensionality $\langle d \rangle$ for the haematopoietic stem cells data set, B) Visualisation on the four first eigenvectors of diffusion map at $\sigma = 10^{0.9}$, C) Visualisation on the four first eigenvectors of PCA, D) Visualisation on the first three eigenvectors of t-SNE [9].

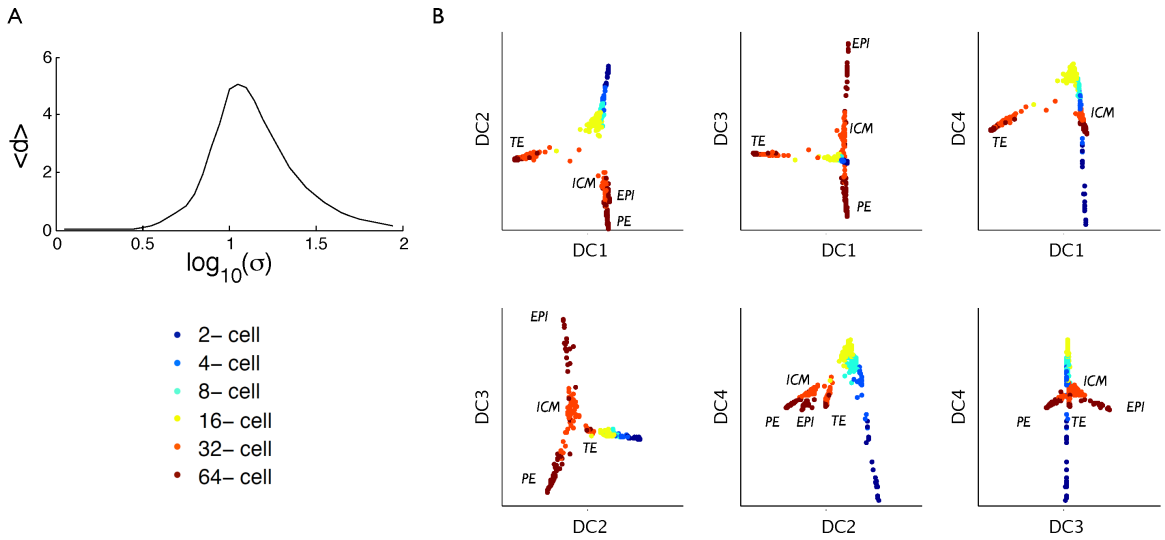


Figure S15: A) The average dimensionality $\langle d \rangle$ for the mouse embryo stem cells data set, B) Visualisation on the four first eigenvectors of diffusion map at $\sigma = 10^{1.05}$.

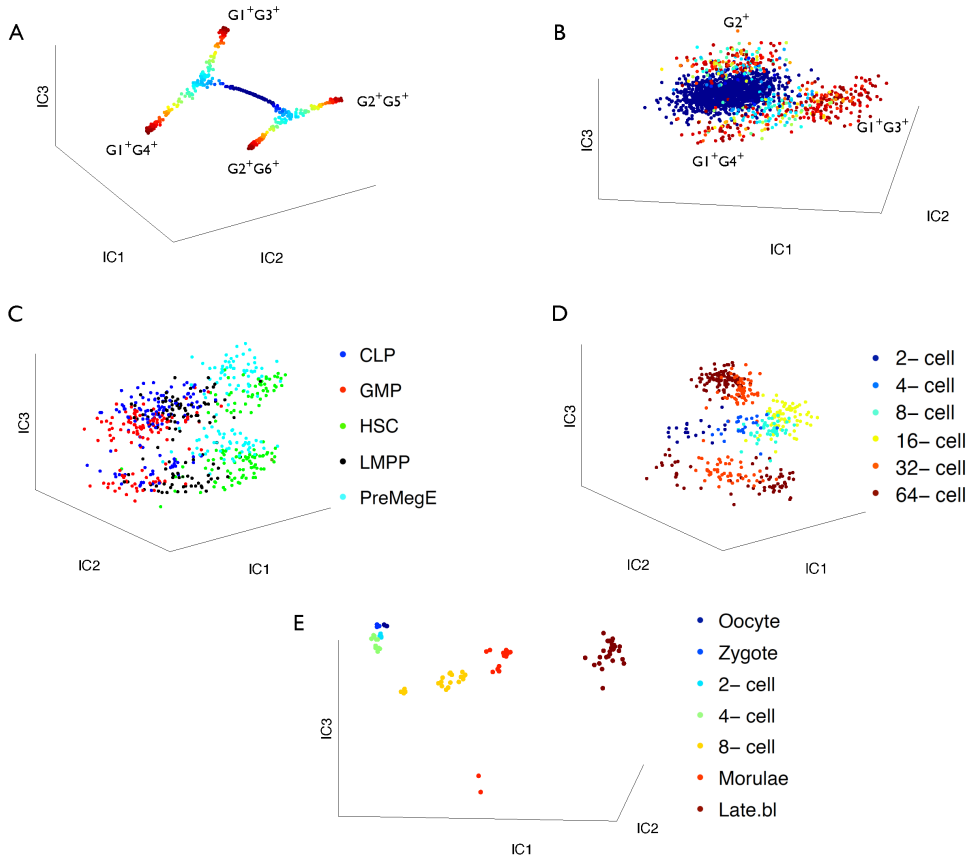


Figure S16: Monocle [8] a recent algorithm for pseudo-temporal ordering of single cells, uses Independent Components Analysis [6] for its dimension reduction step. However ICA as a linear dimension reduction tool performs same as PCA except for rotation and rescaling of the embedding, hence it does not capture the nonlinear trajectories of differentiation. For demonstration here we are representing the ICA embedding for A) balanced toy data set, B) imbalanced toy data set, C) haematopoietic stem cells qPCR data set, D) mouse embryo qPCR data set, F) human preimplantation embryos RNA-Seq data set.

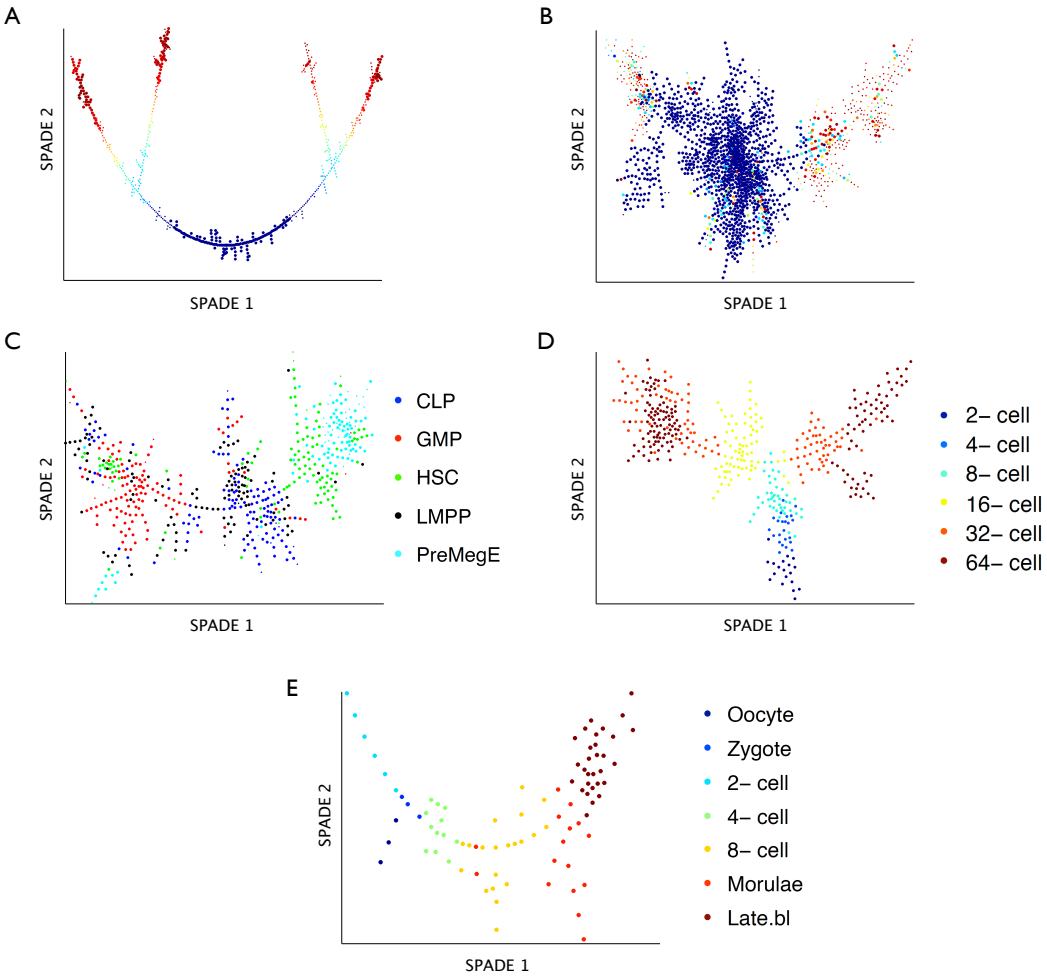


Figure S17 *(preceding page)*: SPADE [4], a clustering algorithm proposed for flow and mass cytometry single-cell data analysis was run on the five differentiation data sets used in this manuscript. In all plots the number of desired clusters was set to the number of cells in each data set as it was possible to do so due to the relatively low number of cells in our data sets and in order to keep single-cell resolution. A) For the balanced toy data set SPADE finds the appropriate tree structure on the data. B) For the imbalanced toy data set the SPADE tree can not properly show all four branches because of high noise level. C) For the haematopoietic stem cells qPCR data set the SPADE tree does not properly match the know hierarchy of cell types (as in Figure 6 F in the main document). D) For the mouse embryo qPCR data set the SPADE performance relatively well. F) For the human preimplantation embryos RNA-Seq data set, the 2-cell and 4-cell states are wrongly separated with the Zygote state in between, since minimum spanning trees (the methodology used by SPADE) are quite sensitive to curse of dimensionality effects in high dimensions and low number of desired clusters (i.e. single cells in this plot).

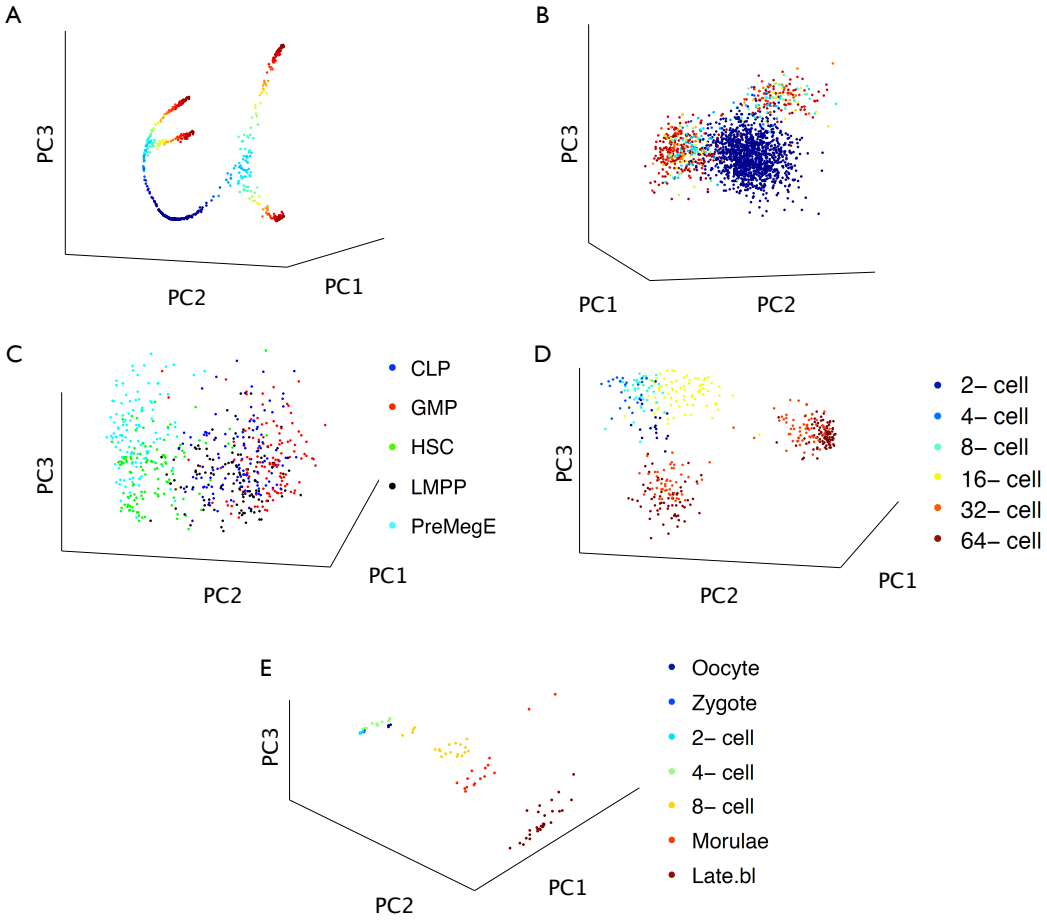


Figure S18: Kernel-PCA [5] with a Gaussian kernel for A) balanced toy data set (Gaussian kernel width=1000), B) imbalanced toy data set (Gaussian kernel width=1000), C) haematopoietic stem cells qPCR data set (Gaussian kernel width=50), D) mouse embryo qPCR data set (Gaussian kernel width=50), F) human preimplantation embryos RNA-Seq data set (Gaussian kernel width=50).

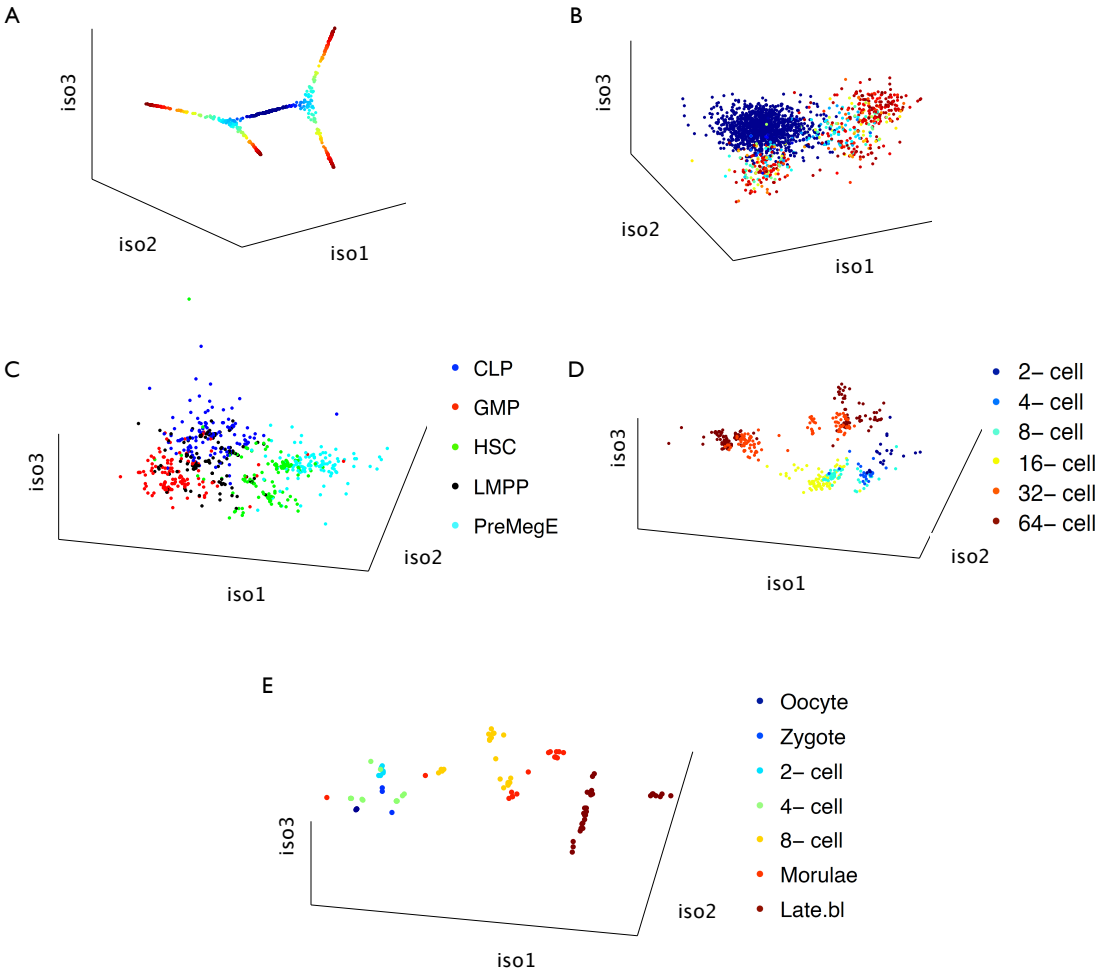


Figure S19: Isomap [7] is a dimension reduction tool based on geodesic distances between the cells, and is highly sensitive to noise and outliers. Isomap embedding for A) balanced toy data set (no. of nearest neighbours=50), B) imbalanced toy data set (no. of nearest neighbours=100), C) haematopoietic stem cells qPCR data set (no. of nearest neighbours=50), D) mouse embryo qPCR data set (no. of nearest neighbours=80), F) human preimplantation embryos RNA-Seq data set (no. of nearest neighbours=50).

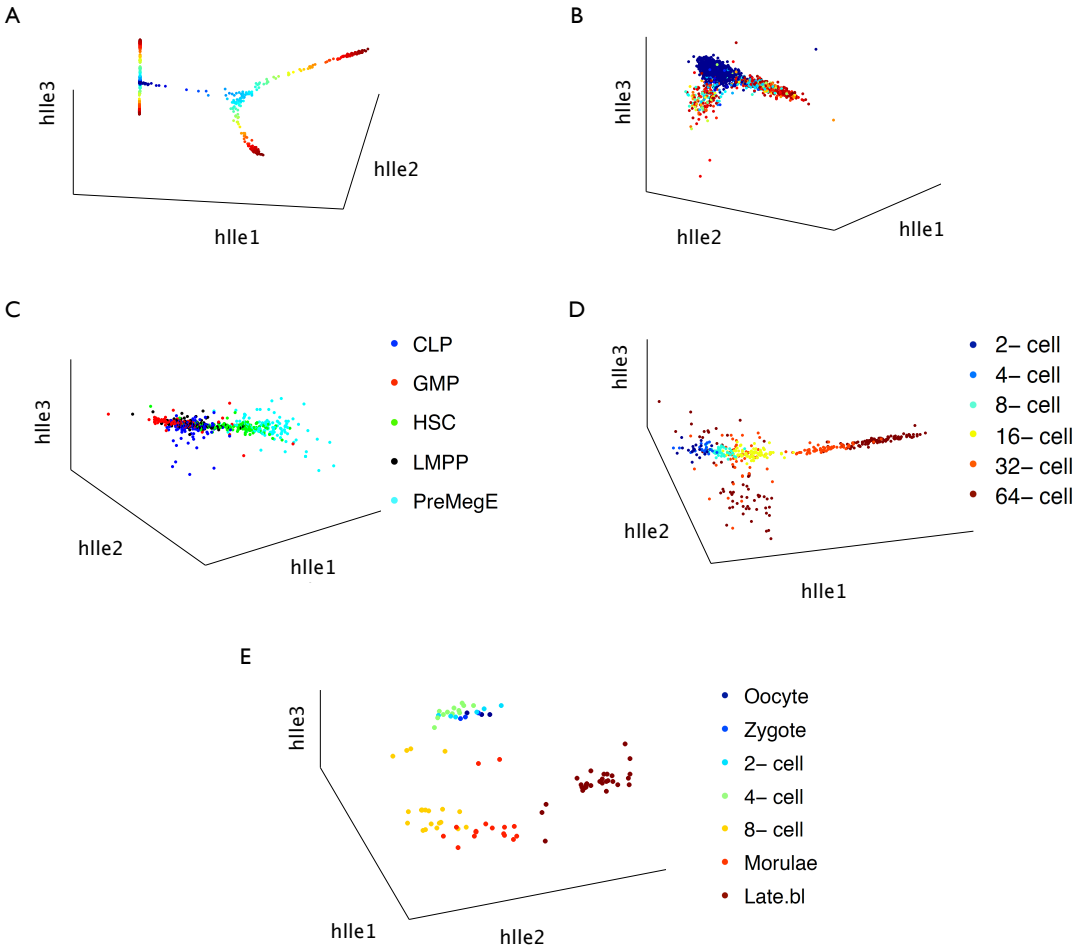


Figure S20: In hessian locally linear embedding (HLLE) [2] the coordinate of each cell is estimated by a linear combination of the coordinates of its neighbours and is especially sensitive to sampling density heterogeneities. HLLE embedding for A) balanced toy data set (no. of nearest neighbours=50), B) imbalanced toy data set (no. of nearest neighbours=100), C) haematopoietic stem cells qPCR data set (no. of nearest neighbours=50), D) mouse embryo qPCR data set (no. of nearest neighbours=100), F) human preimplantation embryos RNA-Seq data set (no. of nearest neighbours=50).

References

[1] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426–7431, 2005.

[2] D. L. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

[3] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[4] P. Qiu, E. F. Simonds, S. C. Bendall, K. D. Gibbs Jr, R. V. Bruggner, M. D. Linderman, K. Sachs, G. P. Nolan, and S. K. Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nature biotechnology*, 29(10):886–891, 2011.

[5] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[6] J. V. Stone. *Independent component analysis*. Wiley Online Library, 2004.

[7] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[8] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 2014.

[9] L. Van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008.

Appendix B

This is a pre-copyedited, author-produced PDF of an article accepted for publication in Nature Methods following peer review. The version of record
L. Haghverdi, M. Büttner, F.A. Wolf, F. Buettner, F.J. Theis, **Diffusion pseudotime robustly reconstructs lineage branching.** *Nature Methods, 13(10), pp.845-848 (2016).*
is available online at:
http://www.nature.com/nmeth/journal/vaop/ncurrent/full/nmeth.3971.html

Diffusion pseudotime robustly reconstructs lineage branching

Laleh Haghverdi[1,3], Maren Büttner[1], F. Alexander Wolf[1], Florian Buettner[1,2], Fabian J. Theis[1,3]

[1]Helmholtz Zentrum München–German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany. [2]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [3]Department of Mathematics, Technische Universität München, Munich, Germany.
Correspondence should be addressed to F.J.T.: fabian.theis@helmholtz-muenchen.de

Single-cell gene expression profiles of differentiating cells encode their intrinsic latent temporal order. We describe an efficient way to robustly estimate this order according to a *diffusion pseudotime*, which measures transitions between cells using diffusion-like random walks. This allows us to reconstruct the cells' developmental progression and to identify transient or metastable states, branching decisions or differentiation endpoints.

Cellular programs are driven by gene-regulatory interactions, which due to inherent stochasticity and external influences often exhibit strong heterogeneity and asynchrony in the timing of program execution. Time-resolved bulk transcriptomics averages over these effects and obscures the underlying gene dynamics. Instead, single-cell profiling techniques allow a systematic observation of a single cell's regulatory state[1] as they capture cells at various developmental stages[2,3]. Since cells are destroyed during measurement, gene dynamics and hence the sequence of cellular programs, have to be inferred from static snapshot data. This is generally achieved by ordering cells according to expression similarity, which is known as `pseudotemporal ordering`. It was initially proposed for bulk expression[4], and has later been extended to single-cell data as measured in RNA-seq and mass cytometry[5-7]. However, existing pseudotime algorithms face problems regarding robustness and scalability when applied to data with branching lineages, which makes a reliable application in many generic experimental settings questionable. The problems are particularly severe in the light of the increasing importance of novel experimental techniques such as Drop-seq[8,9] or MARS-seq[10], which are able to profile tens of thousands of cells, albeit at high noise rates.

To overcome these problems, we introduce a pseudotime measure we call 'diffusion pseudotime' (DPT). Diffusion pseudotime is a random-walk-based distance, which is computed based on simple Euclidian distances in the so-called diffusion map space. The diffusion map is a non-linear method for recovering the low-dimensional structure underlying high-dimensional observations[11]. It organizes data by defining coordinates as dominant eigenvectors of a transition matrix $\boldsymbol{T}$ that describes random walks between data points – here between cells in distinct stages of the differentiation process. A diffusion map strongly reduces noise and is able to represent branching data, but so far has only been used for visualization[12,13]. Our main contribution here is to derive a measure on this space (DPT) that is suitable for recovering the dynamics of biological processes underlying the data, in particular, developmental trajectories from single-cell data. The definition of DPT (Online Methods, eq. (1)) amounts to ordering cells by comparing their probabilities of differentiating towards different cell fates.

Diffusion pseudotime is computed as follows. Given single-cell gene expression data, we build a transition matrix by convolving Gaussians centered at nearby cells; thereby effectively constructing a weighted nearest-neighbor graph of the data, see Figure 1a(1). For each cell, we then determine the probabilities of transitioning to each other cell in the data set using random walks of any length on this graph, which can be seen as a proxy for each cell's probabilities of differentiating towards different fates. We store these probabilities in a vector. The DPT between two cells then is the Euclidian distance between these vectors, Figure 1a(2). Supplementary Note. 1. Provided with a known `root cell`, we measure the developmental progression of each cell in the dataset by computing its DPT with respect to the root cell. Importantly, the definition of DPT benefits from the favorable properties of diffusion maps such as robustness to noise and sampling density[11]. The latter facilitates the application of DPT to experiments where the number of profiled cells varies between different developmental stages (Online Methods). In addition, DPT is computationally efficient as the involved expressions are all available in closed form, so that DPT can be applied to large datasets comprising tens of thousands of cells. DPT does not rely on dimension reduction and thereby accounts for subtle changes in high-dimensional gene expression patterns.

The DPT based analysis then proceeds by identifying branching points that occur after cells progress from a root cell *x* through a common `trunk trajectory`, see Figure 1a(3). We determine the branching of the trunk by measuring the correlation of two pseudotime sequences along trajectories that start from the root cell *x* and from a cell *y* with maximal DPT with respect to *x*. Whereas these sequences are anti-correlated on their direct connection, in a separate branch leading to a third cell *z*, they become correlated, as illustrated in Figure 1a(3). Cells that belong to branching points then are determined as cells for which the two sequences switch from anti-correlated to correlated behavior (see Online Methods and Supplementary. Note 1).

After branch identification, we propose to determine metastable (e.g. quiescent) cells by density analysis: Cell ordering by diffusion pseudotime is not affected by a changed sampling density (see Online Methods). We assume that the progression speed of cells through developmental states is inversely proportional to their density, which means that at pseudotimes where many cells accumulate, progression of cells is slow. Under this assumption, metastable cells are identified as regions of high density when plotting density versus pseudotime.

In a first example, we performed a DPT analysis for single-cell qPCR data focusing on early blood development in mouse[13]. Early hematopoietic cells branch to become either red blood cells or endothelial like cells. DPT ordered cells along their developmental trajectory and identified two branches (Fig. 1b), which correspond to the reported blood (branch 1) and endothelial branches (branch 2)[13]. Plotting gene expression versus pseudotime, we find patterns in the developmental stages that are known to be characteristic for blood progenitors (Fig. 1c,d), namely the hemangioblast-like sequence[14] (subsequent up-regulation of *Cdh1* to *Tal1* and *Cdh5*) in the trunk[13] and the endothelial differentiation route[13] in branch 2 (elevated levels of *Pecam1*, *Erg* and *Sox17* amongst others). In branch 1, we find sequential expression of *Etv2*, *Tal1*, *Runx1* and *Gata1*[15], a sequence of gene activations characteristic for erythroid development. DPT further allows to distinguish early (cf. *Ikaros* expression in Fig. 1c) from late transitions (cf. *Erg* in Fig. 1c) as well as a number of intermediate

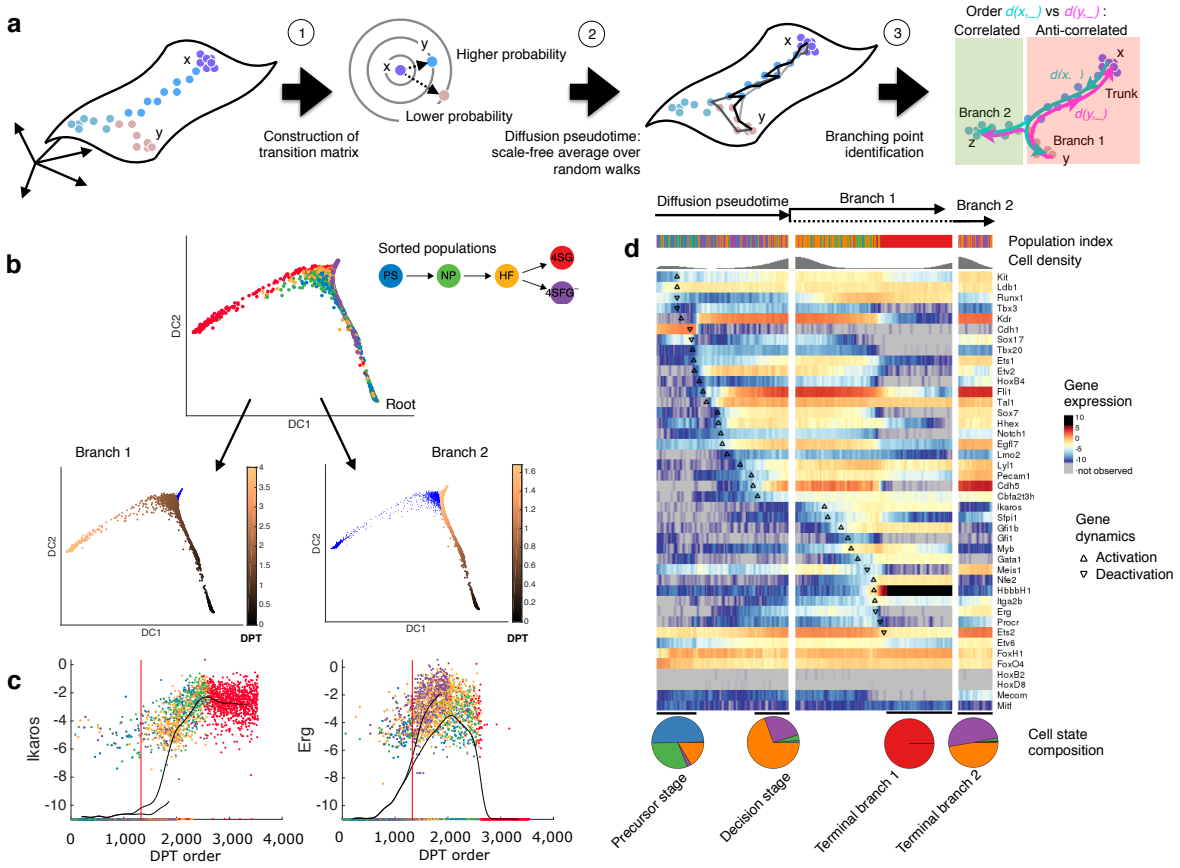


Figure 1:
Diffusion pseudotime reveals temporal ordering and cellular decisions on the single cell level. (**a**) The diffusion transition matrix $\boldsymbol{T}_{xy}$ is constructed by computing the overlap of local kernels at the expression levels of cells *x* and y (1). Diffusion pseudotime dpt*(x,y)* approximates the geodesic distance of *x* and *y* on the mapped manifold (2). Branching points are identified as points where anti-correlated distances from branch ends become correlated (3). (**b**) Application of DPT to single-cell qPCR of 42 genes in 3934 single cells during early hematopoiesis[13], sorted from 5 different populations: primitive streak (PS), neural plate (NP), head fold (HF), four somite GFP negative (4SG-), four somite GFP positive (4SG+). DPT identifies the endothelial branch 1 (4SG-) and the erythroid branch 2 (4SG+). (**c**) Dynamics of genes *Erg* and *Ikaros* in both branches. Black lines show the moving average over 50 adjacent cells. The red vertical line depicts the branching point. (**d**) Heatmap of gene expression (smoothed over 50 adjacent cells), with cells ordered by DPT and branching and genes ordered according to first major change (see Supplementary Note. 2.2). The pie charts (bottom) show the fraction of cells in the four metastable states (metastable state populations are indicated by the black horizontal line above the pie charts).

regulatory events[13] until the onset of *Hbb-bH1* expression (cf. Fig. 1d, black triangles). This information is crucial for the understanding of regulatory interactions: genes that undergo transitions (Supplementary Note 2) earlier than others are candidates for regulators of the differentiation process.

By plotting cell density versus pseudotime, we identify metastable cells via regions of high density (Fig. 1d, top and Supplementary Fig. 1). We found four such states: precursor cells, hemangioblast-like cells at the decision state, erythroid-like and endothelial-like cells. Notably, both decision and precursor states consist of cell mixtures from two or three different stages, stressing the asynchrony of developmental stages that could not be resolved without pseudotemporal ordering.

To identify key decision genes, we quantified expression differences for DPT inferred subgroups and experimentally sorted cells using MAST[16], respectively (Supplementary Fig. 2). Testing differential gene expression for the `decision` versus `precursor` groups inferred by DPT resulted in 32 out of 42 significant genes, including *Cbfa2t3h* and *Pecam1*, which are known to indicate hematopoietic and endothelial development[14], respectively (Supplementary Fig. 2a). Comparing this with differentially expressed genes between the experimental groups head fold (HF) and primitive streak (PS) (Supplementary Fig. 2b) results in a less clear picture: in the latter case, the log-fold change is consistently lower than in the former case (Supplementary Table 1). Also, differential gene expression between HF and 4SG- cells fails to identify endothelial differentiation but brings up erythroid factors (*Runx1*, *Ikaros* and *Gfi1b* amongst others, see Supplementary Fig. 2c-e and Supplementary Table 2). In summary, when comparing differentially expressed genes between metastable states, DPT resolves the developmental patterns more clearly than the experimental sorting of cell populations (Supplementary Fig. 2). This suggests that heterogeneous marker expression across the sorted populations is more likely to be due to cells being in different developmental stages than to mere stochasticity of gene expression.

In a second example, we demonstrate DPT on scRNA-seq combined with droplet barcoding[9]. Klein *et al.* monitored the transcriptomic profiles and heterogeneity in differentiation of mouse embryonic stem (ES) cells after leukemia inhibitory factor (LIF) withdrawal (Fig. 2a). After cell-cycle normalization (Supplementary Figs. 3-4), using DPT we find a single differentiation path from which two small populations branch off. As a striking example, we can resolve upregulation of epiblast markers (*Krt8/18/19*) and downregulation of pluripotency factors (*Nanog*, Fig. 2b) on a sub-day resolution in contrast to Klein *et al.*[9], Figure 7B. The first population (151 cells) that branches off in Figure 2c is enriched in apoptosis-related genes (e.g. *Clu, Cd63*, Supplementary Fig. 5). The second branching event in Figure 2c gives rise to a population (27 cells) with increased primitive endoderm markers (e.g. *Serpinh1*, *Sparc* in Fig. 2c), and a population (67 cells) with increased epiblast markers (*Krt8*, *Actg1* in Fig. 2c, *Krt8*, *Actb* in Fig. 2b). Clustering of the gene expression dynamics (Supplementary Note 3) identified three major clusters, which we observe to differ both in their pseudotemporal behaviors (Fig. 2c and Supplementary Fig. 4) and biological functions (Fig. 2d). For example, the purple cluster consists of pluripotency factors, which we find to be active in early pseudotime and then to decrease gradually. Altogether, DPT analysis leads to an accurate high resolution reconstruction of early embryonic stem cell differentiation events and transcription factor dynamics.

In a third example, a scRNA-seq data set for adult hematopoiesis[10], DPT identifies the dominant branching into different myeloid lineages. It additionally finds a subpopulation of lymphoid outliers and a graded transition reflecting erythroid differentiation, which differs from previously stated cluster sequences[10] (Supplementary Note 4).

Diffusion pseudotime overcomes problems regarding robustness and scalability of previous algorithms[5-7] that prevent the latter to find new biology in many relevant settings (Fig. 2e, Online Methods, Supplementary Note 5), and is in addition more accurate (Fig. 2f). We also clarify the general relation between any pseudotime and actual time measurements (Online

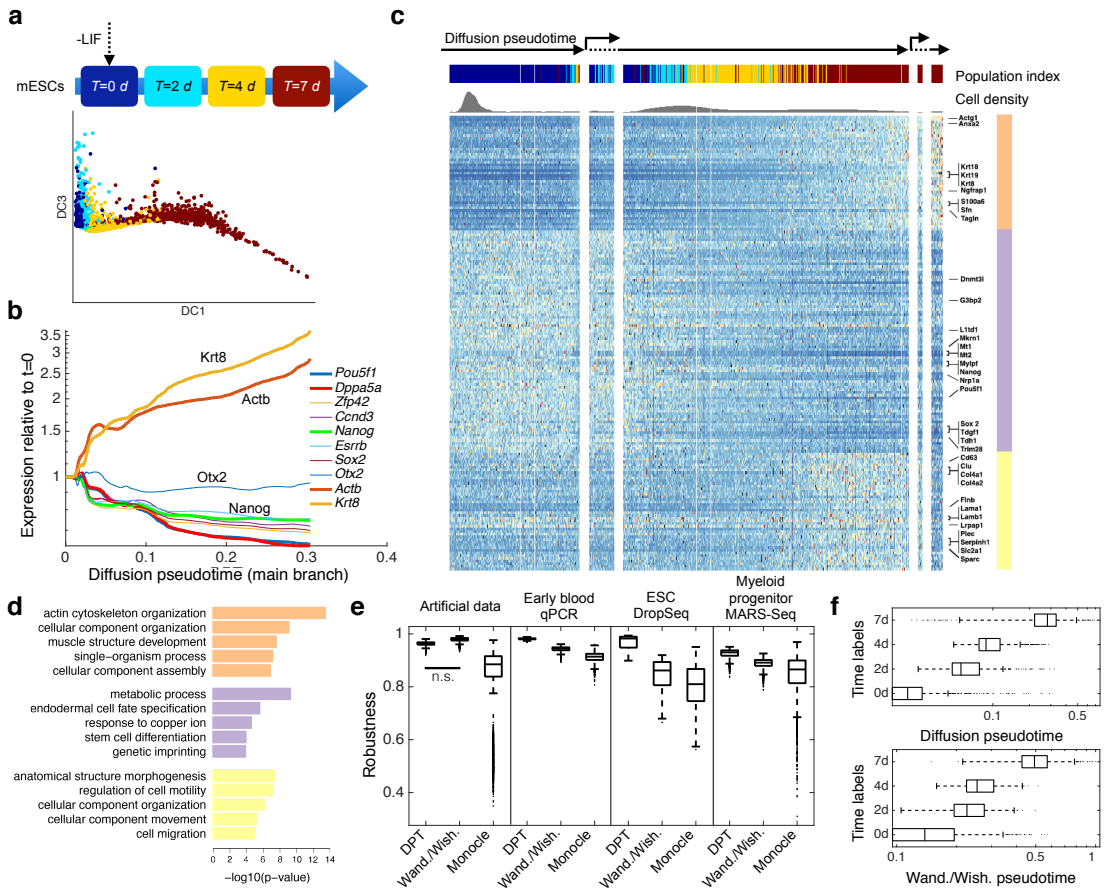


Figure 2:
Diffusion pseudotime identifies differentiation dynamics in droplet-based scRNA-seq experiments[9]. (**a**) Mouse ESCs after LIF withdrawal were harvested at T=0, 2, 4 and 7 days and profiled with the inDrop protocol, yielding 2717 cells with 24175 observed unique transcripts[9]. Visualization using diffusion maps shows temporal dynamics across the four days. (**b**) Pseudotemporal dynamics of the expression of selected genes. Compare with Figure 7B of *Klein et al.*. (**c**) Heatmap of gene expression, with cells ordered by DPT and branches (separated by white vertical bars) and genes ordered according to hierarchical clustering. The heatmap depicts gene dynamics after hierarchical clustering (cf. Supplementary Fig. 4b): The clusters (indicated by color bar on the right) consist of upregulated epiblast markers such as *Krt8/18/19*, *Sfn*, *Tagln* (orange), gradual downregulated pluripotency factors such as *Pou5f1* (*Oct4*), *Sox2, Trim28*, *Nanog* (purple) and upregulated primitive endoderm markers such as *Col4a1/2*, *Lama1/b1*, *Serpinh1*, *Sparc* (yellow). (**d**) Gene ontology enrichment shows a cellular reorganization signature (orange), a metabolic signature for differentiation (purple) and a cell motility signature (yellow). (**e**) Comparisons of robustness of DPT, Wanderlust/Wishbone and Monocle by self-concordance measure on bootstrap samples for several data sets (Supplementary Note 5). DPT consistently shows higher robustness (self-concordance) across all data sets (all 2-sided t-test significance levels $p<0.001$, except for the non-significant ("n.s.") comparison to Wishbone in the artificial data. In the boxplot center line marks the median, edges the first and third quartile, whiskers extend to ±1.5× the interquartile ratio divided by the square root of the number of observations, and single points denote measurements outside this range. (**f**) Boxplots of Kendal rank correlation of pseudotime with experimental days. DPT orders cells (Online Methods) significantly better than pseudotemporal ordering by Wanderlust/Wishbone (Kendal rank correlation 0.77 ± 10^{-3} versus 0.70 ± 10^{-3}). Center line marks the median, edges the first and third quartile, whiskers extend to ±1.5× the interquartile ratio divided by the square root of the number of observations, and single points denote measurements outside this range.

methods and Supplementary Note 6). In the future, robust computation of pseudotimes could allow inferring regulatory relationships without perturbations[13], where DPT allows to scale this to genome-wide models. Recently, pseudotemporal ordering has been applied to

cell morphology to identify cell cycle states[17] – here diffusion pseudotime would allow inclusion of branching for example to identify cells switching into a quiescent state. In summary, diffusion pseudotime provides a robust and scalable tool to infer cellular trajectories from snapshot data in high-dimensional single-cell expression profiles.

Software availability

R and MATLAB implementations of DPT are available on http://www.helmholtz-muenchen.de/icb/dpt .

Accession Codes

The accession number for the mouse early blood qPCR data[13] is Gene Expression Omnibus: GSE61470.

The accession number for the mouse embryonic stem cells inDrop data[9] is Gene Expression Omnibus: GSE65525.

The accession number for the mouse myeloid progenitors MarsSeq data[10] is: GSE72857.

Author contributions

L.H. developed the method and the computational tools, performed the analysis and wrote the paper and the supplement. M.B. contributed to the analysis and biological interpretation of results and wrote the supplement. F.A.W. helped interpret the results and write the supplement, and he wrote the paper. F.B. helped interpret the results. F.J.T. conceived and supervised the study, contributed to the method development and wrote the paper with help from all co-authors.

Acknowledgements

We would like to acknowledge C. Marr, J. Hasenauer, M. Heinig, J. Krumsiek, T. Blasi and P. Angerer for their helpful advice and comments on the manuscript.

M.B. is supported by a DFG Fellowship through the Graduate School of Quantitative Biosciences Munich (QBM). F.A.W. acknowledges support by the "Helmholtz Postdoc Programme", Initiative and Networking Fund of the Helmholtz Association. F.B. is supported by the UK Medical Research Council (MRC) via a Career Development Award (MR/M01536X/1). F.J.T. acknowledges financial support by the German Science Foundation (SFB 1243 and Graduate School QBM) as well as by the Bavarian government (BioSysNet).

Competing financial interests

The authors declare no competing financial interests.

References

1. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498,** 236–240 (2013).
2. Moignard, V. *et al.* Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat Cell Biol* **15,** 363–372 (2013).
3. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509,** 371–375 (2014).
4. Magwene, P. M., Lizardi, P. & Kim, J. Reconstructing the temporal ordering of

biological samples using microarray data. *Bioinformatics* **19,** 842–850 (2003).
5. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32,** 381–386 (2014).
6. Bendall, S. C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157,** 714–725 (2014).
7. Setty, M. *et al.* Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol* 1–14 (2016). doi:10.1038/nbt.3569
8. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015).
9. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161,** 1187–1201 (2015).
10. Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163,** 1663–1677 (2015).
11. Coifman, R. R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 7426–7431 (2005).
12. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31,** 2989–2998 (2015).
13. Moignard, V. *et al.* Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat Biotechnol* **33,** 269–276 (2015).
14. Huber, T. L., Kouskoff, V., Fehling, H. J., Palis, J. & Keller, G. Haemangioblast commitment is initiated in the primitive streak of the mouse embryo. *Nature* **432,** 625–630 (2004).
15. Costa, G., Kouskoff, V. & Lacaud, G. Origin of blood cells and HSC production in the embryo. *Trends in Immunology* **33,** 215–223 (2012).
16. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16,** 278 (2015).
17. Gut, G., Tadmor, M. D., Pe'er, D., Pelkmans, L. & Liberali, P. Trajectories of cell-cycle progression from fixed cell populations. *Nat Meth* **12,** 951–954 (2015).
18. Angerer, P. *et al.* destiny - diffusion maps for large-scale single-cell data in R. *Bioinformatics* btv715 (2015). doi:10.1093/bioinformatics/btv715
19. Luxburg, von, U. A tutorial on spectral clustering. *Stat Comput* **17,** 395–416 (2007).
20. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* **33,** 155–160 (2015).

Online Methods

Details of the diffusion pseudotime analysis

Overview of algorithmic steps in the diffusion pseudotime analysis.

0. Initialization using the following user-provided parameters:
    a. A data matrix of dimension *number of cells* times *number of genes*.
    b. The index or indices of one or several root cells.
    c. Diffusion maps options "classic" with a single parameter *kernel width* or "locally scaled" with a single parameter *number of nearest neighbors* that we

use for adjusting a local kernel width for each cell.

1. Computation of the transition matrix $\boldsymbol{T}$.
2. Computation of the accumulated transition matrix $\boldsymbol{M}$ and of diffusion pseudotime with respect to the specified root cells.
3. Iterative assignment of cells to branches.
4. Identification of metastable states.

Definition of diffusion pseudotime. At the core of diffusion pseudotime is a transition matrix $\boldsymbol{T}$ that approximates the dynamic transitions of cells through stages of the differentiation process. This transition matrix is computed using a nearest neighbor graph whose edge weights have a Gaussian distribution with respect to Euclidian distance in gene expression space: transition probabilities correspond to edge weights. The eigenvectors of $\boldsymbol{T}$ are known as *diffusion components*[11] and have been used in *diffusion maps* for visualizing single cell RNA-seq data[12,18]. While using only few diffusion components yields interpretable visualizations and amounts to a common dimensionality reduction method, important information may be lost by removing the remaining components. Consequently, DPT is based on the full rank $\boldsymbol{T}$ rather than a low rank approximation, i.e. we do not use diffusion maps as a dimensionality reduction method but as a method for *representing* and *organizing* the single-cell data.

We define *diffusion pseudotime* dpt(*x,y*), our novel distance measure, for two cells with index *x* and *y* as

$$\mathrm{dpt}(x,y) = \|\boldsymbol{M}(x,.) - \boldsymbol{M}(y,.)\| \quad , \quad \boldsymbol{M} = \sum_{t=1}^{\infty} \widetilde{\boldsymbol{T}}^t. \qquad (1)$$

Here, $\|\ldots\|$ denotes the L^2 norm. Instead of the probability $(\boldsymbol{T}^t)_{xy}$ for a random walk of fixed length[11] t from x to y, in Eq. (1), we compute the accumulated transition probability $(\boldsymbol{M})_{xy}$ of visiting y when starting from x over random walks of all lengths *t* by summing over *t*. This is done using the modified transition matrix $\widetilde{\boldsymbol{T}}$, which is defined as $\boldsymbol{T}$ without the first eigenspace. The first eigenspace can be associated with the steady state, and the transition matrix $\widetilde{\boldsymbol{T}}$ can be thought of encoding information about how this steady state is approached. We note that the main contribution to $\boldsymbol{M}$ is the *pseudoinverse* of $\boldsymbol{T}$ (it's inverse being not defined), which is a standard object in the theory of Markov Processes, but also in spectral clustering[19]. It is a strong computational advantage that $\boldsymbol{M}$ can be obtained in closed form; it reads

$$\boldsymbol{M} = \sum_{t=1}^{\infty} \widetilde{\boldsymbol{T}}^t = (I - \widetilde{\boldsymbol{T}})^{-1} - I \quad \text{where} \quad \widetilde{\boldsymbol{T}} = \boldsymbol{T} - \psi_0 \psi_0^T,$$

where ψ_0 denotes the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{T}$ (see Supplementary Note 1)**.** Fixing a known root cell *x* as start of the biological process of interest, the diffusion pseudotime of cell *y*, with respect to the root cell *x*, is dpt*(x,y)*.

We point out that diffusion pseudotime is a measure of distance over random walks of arbitrary length, and can hence be considered ``scale-free" in contrast to *diffusion distance*, as introduced by Coifman *et al.*[11]. Characterizations of distance using diffusion maps have mostly relied on *diffusion distance*, which involves as scale parameter *t*, the fixed length of random walks that occur in the definition of *diffusion distance*. In DPT, by summing over all

random walk lengths, *t* is no longer present. DPT therefore has contributions from random walks of all scales. Further mathematical details on the definition of diffusion pseudotime are given in Supplementary Note 1.

Effects of sampling density. We point out that the diffusion pseudotime ordering of cells is (almost) independent of cell sampling density. Coifman *et al.*[11] show that this can be achieved by normalizing the transition matrix $\boldsymbol{T}$ with a proxy for sampling density. The result is that one can sample more (or less) cells in a specific region of the data manifold, but distances on the manifold, such as DPT or diffusion distance, do not change. We provide numerical evidence for this in Supplementary Note 7.5.3 for simulated data. For this, we compute DPT using the original simulated data (related to a toggle switch), for which the sampling density versus DPT is strongly inhomogeneous, i.e. cells accumulate at certain DPT intervals (Supplementary Fig. N13 A). We then subsample from this original data set with the constraint of enforcing an (almost) homogeneous sampling density of cells with respect to DPT. As shown in Supplementary Figure N13 A, the sampling density of this subsampled data differs strongly from the original data, whereas the DPT ordering, as shown in Supplementary Figure N13 B, stays almost the same.

Metastable states. Although diffusion pseudotime ordering is (almost) independent under sampling density changes, our definition of metastable states is not. As explained in the main text, our definition of metastable states is based on the assumption that sampling densities reflect how fast cells are passing through differentiation states. Hence, factors such as cell creation and death processes (e.g. when having a transit-amplifying set of cells) or a subjective choice of cells that are screened in the experiment, influence our identification of metastable states.

Branching points. Branching points are determined by comparing two independent diffusion pseudotime orderings over cells, one starting at the root cell *x* and the other at its maximally distant cell *y*. The two sequences of pseudotimes are anticorrelated until the two orderings merge in a new branch, where they become correlated. This criterion robustly identifies branching points as we illustrate for simulation data for which the ground truth is known (Supplementary Fig. N9). The procedure is sketched in Figure 1a(3). One can repeat the procedure of branch finding iteratively in each of the found branches to identify further (i.e. more than three) sub-branches in the data. Further details are provided in Supplementary Note 1.

Comparison of diffusion pseudotime to previous algorithms

Numerical experiments. When applying Monocle[5] and Wishbone[7] to the qPCR data from our first example, both fail to identify the endothelial branch (Supplementary Fig. N10). Similarly, for the scRNA-seq data from our second example, neither of the two methods identifies the important split between epiblast and primitive endoderm (Supplementary Fig. N11). Also for our third example, without extensive preprocessing, only DPT identifies the dominant branching (Supplementary Fig. N12). A detailed simulation study on robustness (Fig. 2e, Supplementary Note 5) and comparisons on artificial data (Supplementary Fig. N9) further confirm the superior robustness of DPT. Finally, even when previous algorithms are able to obtain qualitatively meaningful results, we observe DPT to yield a higher quantitative accuracy of e.g. the ESC differentiation process in the scRNA-seq data from our second

example in the main text: while the Kendall rank correlation of the DPT ordering with the true time ordering is τ = 0.77±10^{-3}, the one of Wanderlust/Wishbone is 0.70±10^{-3}, see Figure 2f. Kendall rank correlation τ is defined as:

$$\tau = 2\,(m - m')/n(n-1)$$

where *m* is the number of concordant pair of cells (between pseudotime and experimental day), *m'* is the number of disconcordant pair of cells and *n* is the total number of cells.

Methodological comparison. The high robustness of DPT can theoretically be understood from its random-walk based formulation. In comparison to Monocle, which is based on a minimum spanning tree approach, DPT's average over random walks [eq. (1)] is significantly more robust (Fig. 2e). Furthermore, DPT is scalable to high cell numbers whereas Monocle is not. In comparison to Wanderlust and Wishbone, which rely on an approximate and computationally costly sampling of shortest paths on graphs, DPT's average over random walks [eq. (1)] is exact and computationally cheap. Regarding in particular the recently published Wishbone: Although Wishbone – in contrast to Wanderlust – is able to identify branching events, it does this using a number of preprocessing steps and independent algorithms (preprocessing by dimension reduction, manual selection of components in the dimension reduction). Furthermore, its pseudotime distance is based on a shortest-path distance between so-called `way-point` cells, which in the presence of branching, has to be computed in a rather complicated, iterative way and then does not constitute a good proxy for geodesic distance any more. By contrast, DPT consists in a single clean definition of a pseudotime measure and a simple algorithm for its computation. We believe that this fundamental difference in method design is responsible for the practical advantages of DPT over Wishbone, as discussed in Supplementary Note 7.

Data analysis and experiments

Detecting transcriptional changes. To identify the succession of switch-like transcriptional changes revealed by the pseudotemporal order in qPCR data, we computed an approximate derivative of the smoothed gene expression level along branch 1. A switch-like change is defined as the maximum in the derivative (details in Supplementary Note 2.2).

Differential expression analysis. We employed a generalized linear model that allows to quantify the proportion of cells expressing a certain gene as well as the mean expression level, a modified Hurdle model[16]. Briefly, the model has two parts: A discrete part to decide whether a gene is expressed and a continuous part using a normal distribution to quantify expression of a gene. Then, a likelihood ratio test is used to identify differentially expressed genes (details in Supplementary Notes 2.3 and 3.4 and Finak et al[16]).

First example: early blood development data (ESC qPCR). We analyzed a single-cell qPCR dataset (normalized version with 3934 cells, 42 genes) focusing on early blood development[13]. For each gene, the limit of detection (LOD) was the average Ct value for the last dilution at which all six replicates had positive amplification. The overall LOD of 25 for the gene set was the median Ct value across all genes. Raw Ct values and normalized data can be found in supplemental table 7 of Moignard et al[13]. Gene expression was subtracted from the limit of detection and normalized on a cell-wise basis to the mean expression of the

four housekeeping genes (*Eif2b1*, *Mrpl19*, *Polr2a* and *Ubc*) in each cell. Cells that did not express all four housekeeping genes were excluded from subsequent analysis, as were cells for which the mean of the four housekeepers was ±3 s.d. from the mean of all cells. A dCt value of −14 was then assigned where a gene was not detected. 85–90% of sorted cells were retained for further analysis. *Gata2* did not amplify correctly and *HoxB3* was not expressed in any cells, so these factors have been excluded from the analysis. The analyses were done on the dCt values for all transcription factors and marker genes, but not housekeeping genes. For more details, see Supplementary Note 2

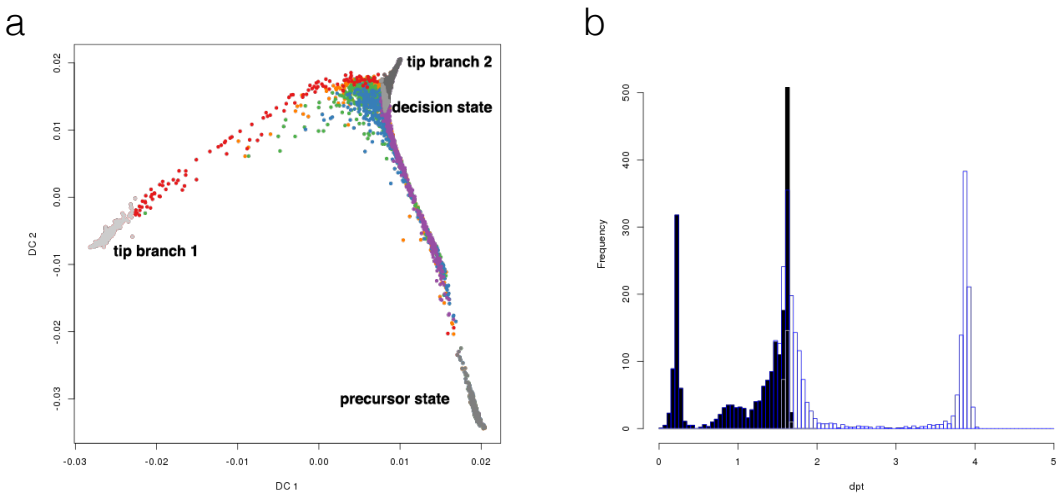
Second example: InDrop differentiation data. We analyzed a single-cell RNA-seq data set using the inDrop protocol from Klein et al[9]. Here, single cells along with a set of uniquely barcoded primers were captured in tiny droplets and sequenced. The capabilities of this technique were demonstrated using an undirected differentiation process of mouse embryonic stem cells upon leukemia inhibitory factor (LIF) withdrawal. The data set is publicly available under the GEO accession number GSE65525. Count data were normalized by library size and $\log_{10}$ transformed (see Supplementary Note 3.1). We corrected for cell-cycle and batch effects using scLVM[20] on 2047 highly variable genes (see supplemental table 3 in Klein et al[9]). Hierarchical clustering was performed in R (http://www.r-project.org/) using the *hclust* package on quantile-normalized data (Supplementary Note 3.2) and displayed with *ComplexHeatmap* package, where the distance was defined as 1 – correlation between all samples (Supplementary Note 3.3). In addition, we performed a rank sums test on the first side branch that identified apoptotic genes as being significantly differently expressed as compared to the initial pluripotent and the late epiblast-like cells (Supplementary Note 3.4). For more details, see Supplementary Note 3.

Relation of pseudotime in snapshot data to actual time

In contrast to measurements in actual time, e.g. from time-lapse microscopy measurements, high-throughput snapshot experiments only encode the progression stage of development, but not the stochastic trajectories of single cells. We measure this progression stage using the geodesic distance on the data manifold and refer to it as 'universal time'. Here, the assumption is that the data manifold is representative for the deterministic program underlying stochastic cellular processes. For time-lapse data, universal time can be constructed by estimating the velocity $v(t)$ tangential to the manifold (C) from each single-cell trajectory. Universal time, i.e. the geodesic distance

$$s(t) = \int_{C:[s(0),s(t)]} ds = \int_0^t dt' \; |\boldsymbol{v}(t')| \approx \int_0^t dt' \frac{1}{\rho(t')}$$

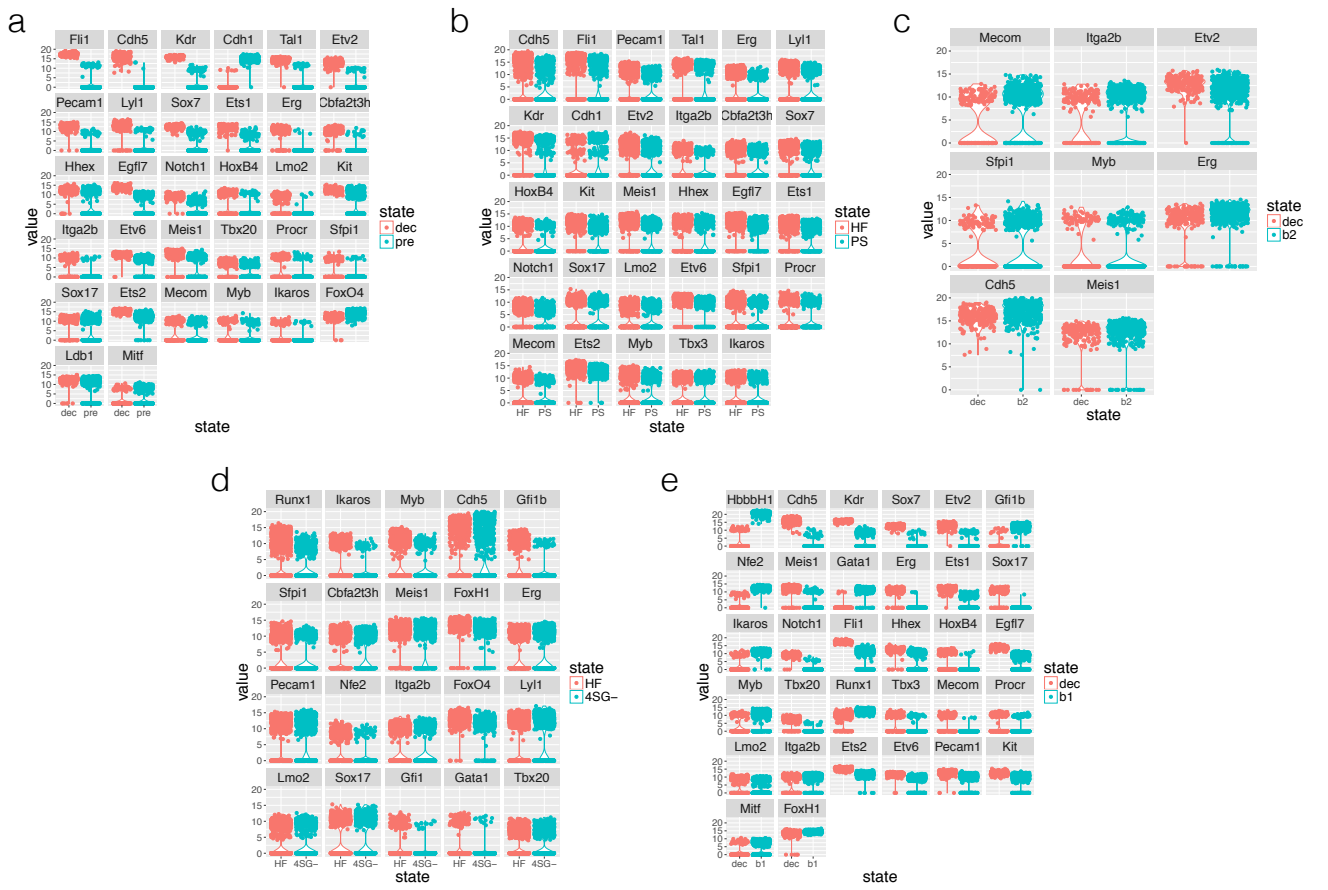
then quantifies the arc length along the manifold, where $\rho(t)$ denotes the local density of samples on a single cell trajectory (see Supplementary Note. 6). Pseudotimes are proxies for universal time (Supplementary Figs. N6-N8). Our proposed DPT approximates universal time better than other pseudotime schemes as it does not involve dimension reduction, and better than *diffusion distance*[11] as it accounts for random walks on all length scales.



Supplementary Figure 1

metastable states of mouse early blood development qPCR data

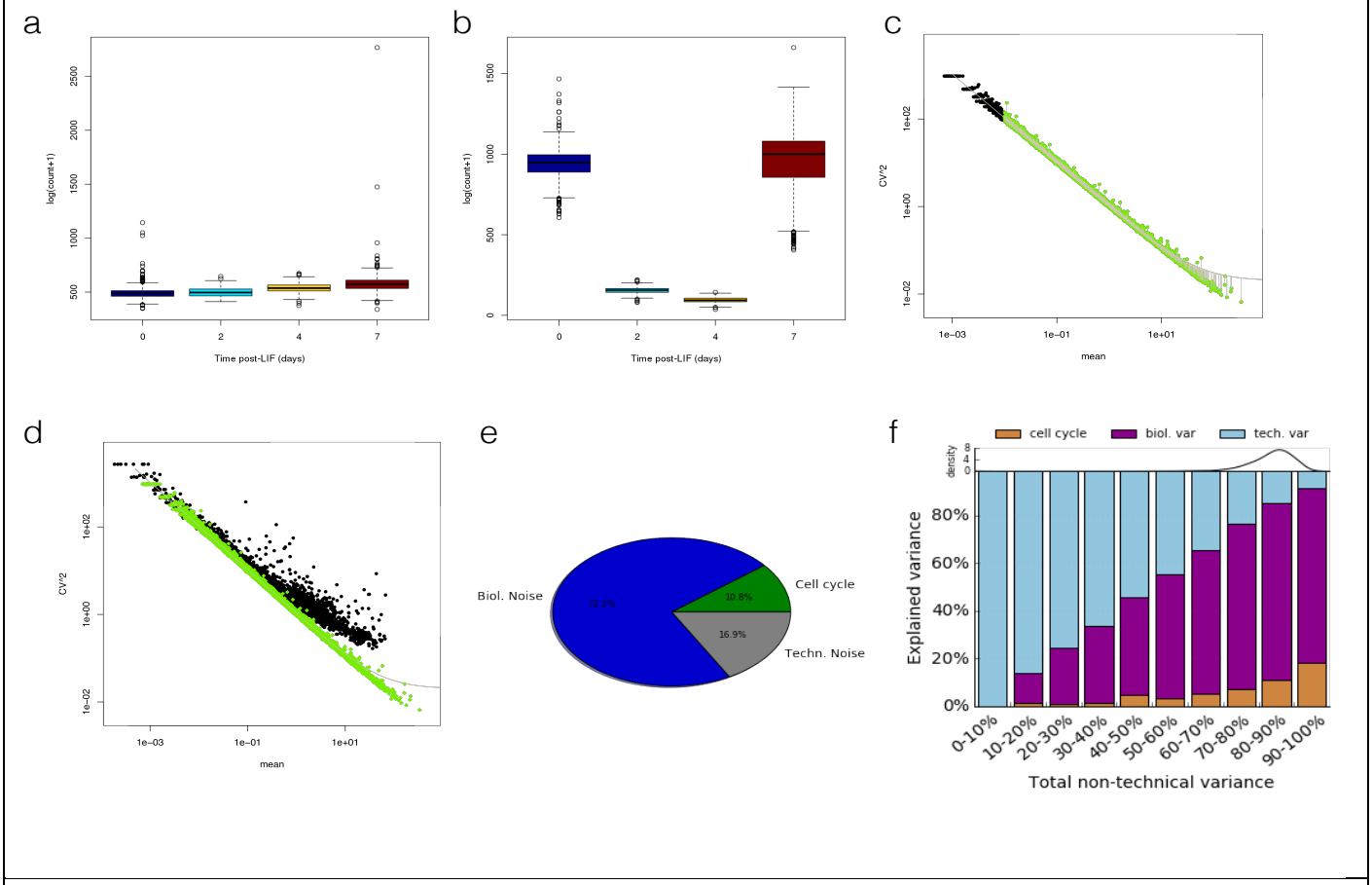
a) Diffusion map plot illustrating four metastable states along pseudotemporal ordering. Lower right: Precursor state. Left: Tip branch 1. Upper right: Decision state (light gray) and tip branch 2 (dark gray). b) Histogram plot of the cell density along the branches. Blue bars: branch 1, black bars: branch 2. Both branches share the precursor branch up to the decision state (gray bars).



Supplementary Figure 2

Differential expression analysis using MAST on mESC inDrop data

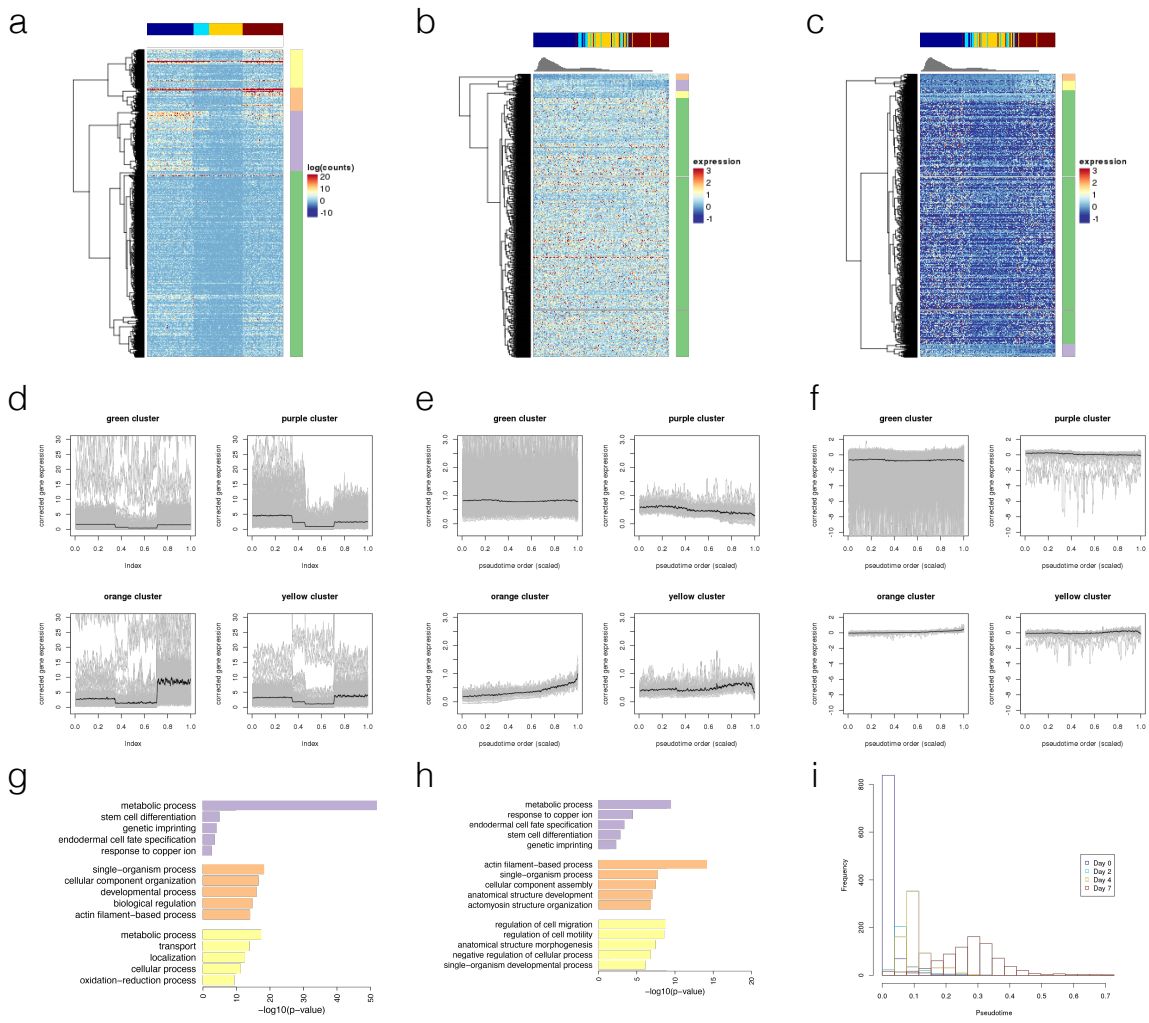
Log-fold change (lfc) analysis of the DPT inferred ‘decision’ group vs. all other groups (a,c,e) and head fold cells vs. primitive streak and 4SG- cells (d, e). The displayed genes were filtered for an lfc > 1 and a Bonferroni-adjusted p-value< 0.01. Plots are ordered by absolute lfc between the states. a) Decision area (red) vs. Precursor area (blue), b) Head fold (red) vs. Primitive streak (blue), c) Decision area (red) vs. branch 2 end point (blue), d) Head fold (red) vs. 4SG negative cells (blue), e) Decision area (red) vs. branch 1 end point (blue).



Supplementary Figure 3

Influence of cell-cycle correction on data clustering and GO enrichment

a,b) The total count of transcripts from 2047 heterogeneous genes per day. a) log-normalized counts before cell-cycle correction. b) log-normalized counts after cell-cycle correction. c) Fit the CV2-mean relation according to Brennecke et al [11] to a pure RNA control and d) superimpose these technical genes with endogenous genes. e) Variance decomposition according to the identified latent variables. f) Detailed variance decomposition sorted by technical noise contribution.

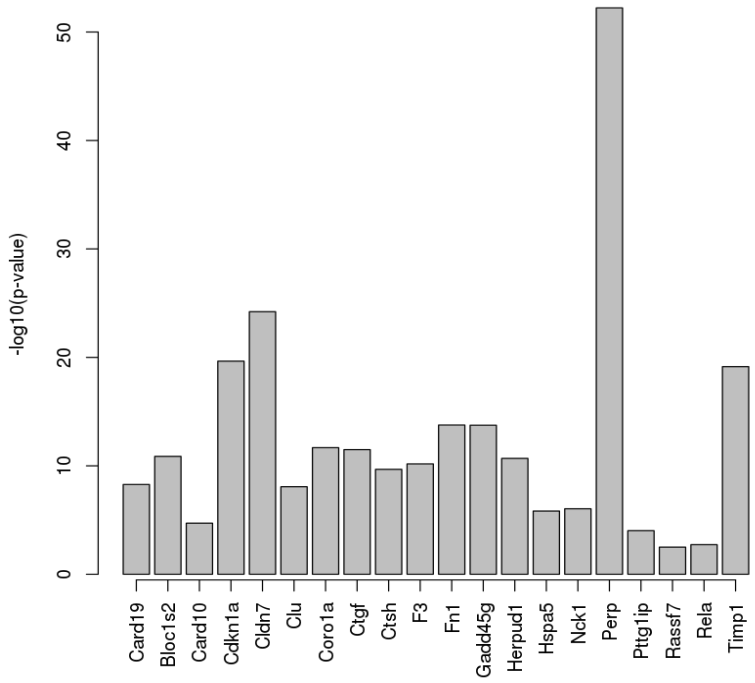


Supplementary Figure 4

Expression profiles of highly variable genes before and after cell-cycle correction and pseudotime ordering of mESC inDrop data

Heatmap displaying the expression profiles of 2047 highly variable genes before a) and after cell-cycle correction and pseudotime ordering (b,c), time courses of gene expression along batch (d) and pseudotime (e,f), GO enrichment analysis of the clusters in (a,c). The colored top bar (a-c) indicates the time after LIF withdrawal (dark blue: day 0, light blue: day 2, yellow: day 4, red: day 7). a) Gene expression with strong day-to-day variability. b) Cell-cycle corrected gene expression and additional quantile normalization. c) Cell-cycle corrected gene expression and additional Z-score normalization. Pseudotemporal ordering is indicated by mixed colors in the top annotation bar. In the time courses, the respective genes are indicated in grey, the black curve is the smoothed mean. d) log-

transformed gene expression counts. e) Cell cycle correction, log transformed gene expression counts, quantile normalization (cf. Fig. 2d in main text). f) As in E), with Z-score normalization. All clusters share the same temporal behavior. The green cluster GO terms are not shown. For each cluster, five representative GO terms are displayed. g) GO terms before cell-cycle correction, h) after cell-cycle correction and Z-score normalization. i) Distribution of cells along pseudotime labeled by time after LIF withdrawal.



Supplementary Figure 5

p-values of Wilcoxon rank sum test applied to the first population that branches off the main branch in mESC inDrop data

Shown are the 20 apoptosis-related genes (GO:0006915) among the 108 genes identified by Wilcoxon rank sum test. The test compares cells from the early state population (see text) with cells from the first population that branches off the main branch.

Supplementary Notes for:
Diffusion pseudotime robustly reconstructs lineage branching

Laleh Haghverdi, Maren Büttner, F. Alexander Wolf, Florian Buettner, Fabian J. Theis

Supplementary Note 1: DPT formulation

1.1 Locally scaled transition matrix (and diffusion map)

Some of us proposed diffusion maps as a visualization and dimensionality reduction method for single-cell data in Ref. [1]. For the present work, we developed the approach further, introducing a more robust version of diffusion maps. We refer to the original version of [1] as "classic" and to the new version as "locally scaled". The improved "locally scaled" version of diffusion maps is *not* a necessary basis for the construction of *diffusion pseudotime*, which can be used with the "classic" version as well. Nevertheless we recommend using the improved version. In the following, we provide a self-contained presentation of diffusion maps, and note differences between the "classic" and the "locally scaled" version. The major change amounts to using a local Gaussian kernel width for each cell, estimated as the cell's distance to its κth nearest neighbor, instead of a fixed global Gaussian kernel width.

To motivate this major change, it is worth noting that in gene expression space, Euclidean distance is not necessarily a useful quantification of similarity between cells. In some parts of gene expression space even very small Euclidean distances might imply a large biological dissimilarity while in some other parts large Euclidean distance is merely an artifact of human designed measurements. In many cases viewing the data in a different scale (e.g. using a log transformation, *arcsinh* transformation) is helpful, still a log (or *arcsinh*) transformation is often not adequate. Instead, we assume that cell adjacency relations are a better measure of biological similarity at least when a sufficiently large sample of single cells is considered. Adjusting the kernel width for each cell to the cell's distance to its κth nearest neighbor can be intuitively thought of accounting for each cell's "accessible space".

In our previous work [1], we used a globally fixed kernel width and suggested a method for choosing it. This global kernel width selection method can be useful for mono-scale data. However, differentiation in general can have several branching events with several scales, possibly including both densely and sparsely sampled branches. A locally adjusted kernel width as we present here is able to resolve such data better.

In Ref. [1] we suggested an interpretation of the Gaussian kernel in terms of interfering wave functions. In other words, the Gaussian kernel can be decomposed into its multiplicand wave functions. This interpretation turns out to be useful when the diffusion wave function varies (because of varying noise models) at the position of each cell. Concretely, we assume the Gaussian kernel width ($\sigma_{\mathbf{x}}$) is different for each cell $\mathbf{x}$, and given by the cell's distance to its κth nearest neighbor. The Gaussian wave function associated with cell $\mathbf{x}$ reads

$$Y_{\mathbf{x}}(\boldsymbol{x'}) = \left(\frac{2}{\pi\sigma_{\mathbf{x}}^2}\right)^{1/4} \exp\left(-\frac{||\boldsymbol{x'}-\boldsymbol{x}||^2}{\sigma_{\mathbf{x}}^2}\right). \tag{1}$$

The interference of two cells at $\boldsymbol{x}$ and $\boldsymbol{x'}$ allows to define a locally-scalled kernel K, which reads

$$K(\mathbf{x},\mathbf{y}) = \int_{-\infty}^{\infty} Y_{\mathbf{x}}(\boldsymbol{x'})Y_{\mathbf{y}}(\boldsymbol{x'})\mathrm{d}\boldsymbol{x'} = \left(\frac{2\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}}{\sigma_{\mathbf{x}}^2+\sigma_{\mathbf{y}}^2}\right)^{1/2} \exp\left(-\frac{||\boldsymbol{x}-\boldsymbol{y}||^2}{2(\sigma_{\mathbf{x}}^2+\sigma_{\mathbf{y}}^2)}\right). \tag{2}$$

We proceed by normalizing K with a proxy for the sampling density of cells $Z(\mathbf{x})$ at position $\mathbf{x}$ [2], to obtain a matrix W

$$W_{\mathbf{xy}} = \frac{K(\mathbf{x},\mathbf{y})}{Z(\boldsymbol{x})Z(\boldsymbol{y})} \tag{3}$$

$$Z(\boldsymbol{x}) = \sum_{\mathbf{y}\in\boldsymbol{\Omega}} K(\mathbf{x},\mathbf{y}) \tag{4}$$

Using the "row normalization"

$$\tilde{Z}(\boldsymbol{x}) = \sum_{\mathbf{y}\in\boldsymbol{\Omega}} W(\mathbf{x},\mathbf{y}), \tag{5}$$

one can define a (right-stochastic) transition matrix $T^{\mathrm{asym}}_{\mathbf{xy}}$ as

$$T^{\mathrm{asym}}_{\mathbf{xy}} = \frac{1}{\tilde{Z}(\boldsymbol{x})} W_{\mathbf{xy}}. \tag{6}$$

Here, rows sum to one and $T^{\mathrm{asym}}_{\mathbf{xy}}$ can be interpreted as the probability of transitioning from $\mathbf{x}$ to $\mathbf{y}$.

The right eigenvectors of T^{asym} are referred to as diffusion components, and taken together, they constitute a diffusion map [2]. As described in [1] for the case of single-cell data, the lowest eigenvectors (those associated with the largest eigenvalues) provide a low dimensional map of the original data manifold that is suitable for visualization.

The original diffusion map publication [2], but also literature on spectral clustering [3], pointed out that there exists a matrix T^{sym} that has the same eigenvalues as T^{asym} and whose eigenvectors relate in a simple way to T^{asym}: λ is an eigenvalue of T^{asym} with eigenvector u if and only if λ is an eigenvalue of T^{sym} with right eigenvector w with elements $w_{\mathbf{x}} = \tilde{Z}(\mathbf{x})^{\frac{1}{2}} u_{\mathbf{x}}$. T^{sym} is defined as follows

$$T^{\mathrm{sym}}_{\mathbf{xy}} = \tilde{Z}(\boldsymbol{x})^{-1/2} W_{\mathbf{xy}} \tilde{Z}(\boldsymbol{y})^{-1/2}. \tag{7}$$

Whereas the eigenvectors of T^{asym} are asymmetrical (left and right eigenvectors differ), the eigenvectors of T^{sym} are symmetric (left and right eigenvectors are equal).

For the present work, we choose to use $T^{\mathrm{sym}}_{\mathbf{xy}}$ together with its eigenvectors as a basis for all further analysis instead of the "traditional" diffusion components (the right eigenvectors of T^{asym}). With slight abuse of notation, we use the term "diffusion components" also for the eigenvectors of $T^{\mathrm{sym}}_{\mathbf{xy}}$, and refer to $T^{\mathrm{sym}}_{\mathbf{xy}}$ as the "transition matrix", usually dropping the superscript $^{\mathrm{sym}}$, when there is no danger of confusion.

We note that regarding interpretation of results, it is not of importance whether the symmetric $T^{\mathrm{sym}}_{\mathbf{xy}}$ or the asymmetric T^{asym} version is used (this is known in the literature, see e.g. [3]). There is though a simplification when interpreting $T^{\mathrm{sym}}_{\mathbf{xy}}$: all one learns about data using diffusion maps is about the geometry of the manifold and not about directions on the manifold. It therefore seems unnatural to have asymmetric (directed) transition probabilities, i.e. $p_{\mathbf{x}\to\mathbf{y}} \neq p_{\mathbf{x}\leftarrow\mathbf{y}}$, as in the asymmetric transition matrix T^{asym}. Also, using the symmetric version provides some computational advantages in the eigen decomposition.

1.2 Diffusion pseudotime

The t'th power of the transition matrix T represents a random walk of length t on the data graph. The transition matrix enables us to simulate the time propagation of a wave function (or probability) that has been localized to some specific region of the graph (e.g. a wave function that resides on the pluripotent cells only) at time zero.

The time evolution of a probability density $f(t) \in \mathbb{R}^n$ is described by the graph Laplacian matrix $L = I - T$ as follows:

$$f(t) = f(t-1) + f(t-1)(-L) \tag{8}$$

or in terms of T:

$$f(t) = f(t-1)T = f(0)T^t. \tag{9}$$

To account for the asynchrony of differentiating cells present in snapshot data, one may study the term $\sum_{t=1}^{\infty} f(t)$ which provides the (time independent) "path integral" for reaching each cell from $f(0)$:

$$\sum_{t=1}^{\infty} f(t) = f(0) \sum_{t=1}^{\infty} T^t \tag{10}$$

The sufficient constraint for the convergence of the sum above is that all eigenvalues of T^t are smaller than one. However, a stationary state of f exists for $t \to \infty$ which means that T has an eigenvalue equal to 1. The corresponding eigenvector equals the cells density $\psi_0(x) = \tilde{Z}(x)$. The equal to one eigenvalue implies that the sum in equation (10) diverges. As already mentioned, the stationary state contains information only about the cells' sampling density and not about the consecutive states of temporal evolution (i.e. no pseudotime information). Thus, we can reduce the stationary component of T and perform the sum in equation (10). Subtracting the stationary component (eigenvalue 1 contribution) amounts to the same standard calculation as done when using the pseudoinverse of the transition matrix in the study of Markov processes or spectral clustering (see e.g. Sec. 6.2 of [3]). We call the new matrix M:

$$M = \sum_{t=1}^{\infty} (T - \psi_0 \psi_0^T)^t \tag{11}$$

$$= (I - (T - \psi_0 \psi_0^T))^{-1} - I, \tag{12}$$

which shares the same eigenvectors with T (except for ψ_0):

$$\begin{aligned} M(x,z) &= \sum_{t=1}^{\infty} \left(T(x,z) - \psi_0(x)\psi_0^T(z)\right)^t \\ &= \sum_{t=1}^{\infty} \sum_{i=1}^{n-1} \lambda_i^t \psi_i(x) \psi_i^T(z) \\ &= \sum_{i=1}^{n-1} \frac{\lambda_i}{1-\lambda_i} \psi_i(x) \psi_i^T(z). \end{aligned} \tag{13}$$

If $f(0)$ is chosen localized at cell x (i.e. if $f(0) = \delta(x)$), $f(0)M$ will be a row of M which we present by $M(x,.)$. Moreover, we consider $M(x,.)$ as the feature representation for cell x to define the *diffusion*

pseudotime distance measure $\mathrm{dpt}(x, y)$ as follows:

$$\mathrm{dpt}^2(x, y) = ||M(x, .) - M(y, .)||^2 \tag{14}$$

$$= \sum_z (M(x, z) - M(y, z))^2$$

$$= \sum_{i=1}^{n-1} \left(\frac{\lambda_i}{1 - \lambda_i} \right)^2 (\psi_i(x) - \psi_i(y))^2 \tag{15}$$

The measure dpt is a (weighted L^2 norm) distance metric. Due to the favorable properties of diffusion maps, it is robust to noise and yet does not utilize low-dimensional approximations usually applied for visualization. This robustness allows to apply diffusion pseudotime also in settings, in which Euclidean distances in the original $\mathbb{R}^G$ gene expression space are too much affected by noise.

The diffusion pseudotime of a cell x with respect to a single defined root cell r is:

$$\mathrm{dpt}(r, x) \tag{16}$$

We are also able to define diffusion pseudotime with respect to more than one root cell. This is useful when a root belongs to a population (metastable state) of cells with large variance in their expression state. We then assign a diffusion pseudotime to the cells as follows:

$$||f(0)M - XM||, \tag{17}$$

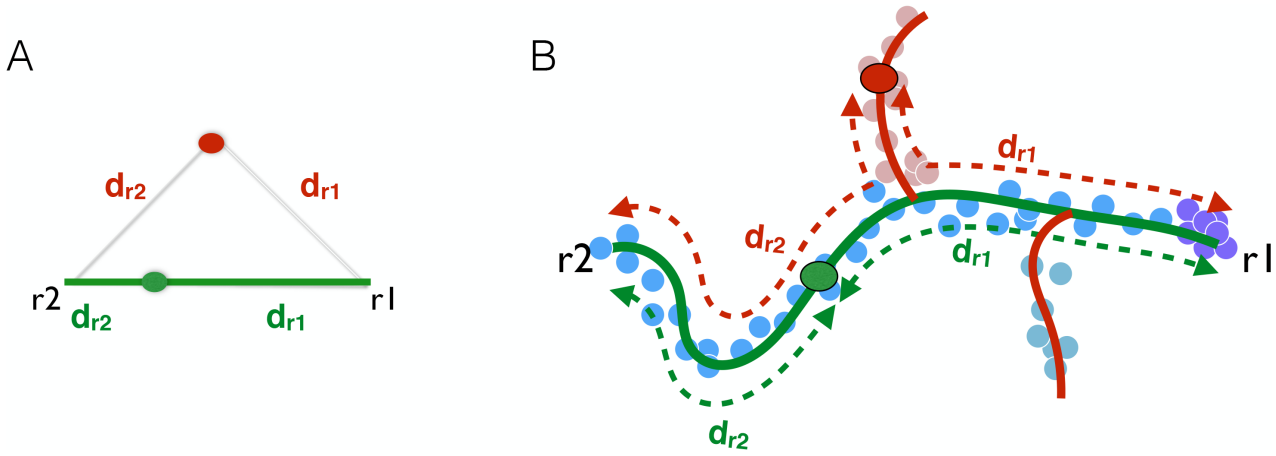
where $f(0)$ is a vector equal to one on all root cells and zero everywhere else, and X being a vector equal to one on cell x and zero everywhere else.

The idea of developing a graph-based distance measure that includes contributions from all length scales of random walks (e.g. average first passage time [4], average commute time) for data clustering has been studied before [2, 5]. In practical comparisons, we found those measures to be inferior to DPT. The average first passage time is not a metric and its symmetrized form, the average commute time, does not provide information about the geometry of the graph, but only about the node densities as shown by von Luxburg et al. in [6].

1.3 Branch assignment

As a robust and biologically relevant (in the context of cell development) metric on the data manifold, DPT can be used to identify the cells at the tip of the branches of data using the triangle inequality. This is illustrated in Supplementary Figure N1. However, using the triangle inequality for scattered manifolds as obtained from single-cell snapshot data requires setting a threshold for the uncertainty of distances. For a robust automatic branch identification algorithm we would like to avoid setting such a threshold. Thus we use an alternative method based on correlation versus anti-correlation of distances on the manifold (Fig. 1a of the main text).

The detailed procedure is as follows. Let us consider a manifold with only three branches. Picking a random cell r_0 we identify the first cell at a tip r_1 which maximizes the dpt distance to r_0. We then identify the second cell at another tip r_2 that maximizes $\mathrm{dpt}(r_1, r_2)$. For any cell residing on the (direct) connecting path between r_1 and r_2, the triangle inequality holds at its lower bound, i.e. $\mathrm{dpt}(r_1, x) + \mathrm{dpt}(x, r_2)$ is equal to (or only slightly higher than) $\mathrm{dpt}(r_1, r_2)$. It is only for cells residing on the third branch that $\mathrm{dpt}(r_1, x) + \mathrm{dpt}(x, r_2)$ becomes significantly bigger than $\mathrm{dpt}(r_1, r_2)$. Thus the



Supplementary Figure N1: A) Triangle inequality in Euclidean space. For any point on the green line $d_{r1} + d_{r2}$ is a constant. For any point off the green line $d_{r1} + d_{r2}$ is larger than that constant. B) Triangle inequality on a manifold. For any point on the green line $d_{r1} + d_{r2}$ is almost a constant, leading to anti-correlation between d_{r1} and d_{r2} on the green line and correlation of d_{r1} and d_{r2} on any of the red branches.

third tip cell r_3 can be identified as the cell maximizing the sum of distances to r_1 and r_2. In brief:

$$r_1 = \arg\max_x \mathrm{dpt}(r_0, x) \tag{18}$$

$$r_2 = \arg\max_x \mathrm{dpt}(r_1, x) \tag{19}$$

$$r_3 = \arg\max_x \Big(\mathrm{dpt}(r_1, x) + \mathrm{dpt}(x, r_2)\Big) \tag{20}$$

Now we can perform a pseudotime ordering where the initial probability $f(0)$ is chosen zero everywhere except at the tip of a branch (either r_1, r_2 or r_3). The ordering on every two branches will correlate with each other only on the third branch and anti-correlate on the two branches themselves (see Fig. 1a(3) in the main text). We use this property to find a cutoff x for each branch. More precisely, to separate branch 1, we first do three independent orderings $O1, O2, O3$ with assigning r_1, r_2 and r_3 as the root of ordering correspondingly. Then, based on Kendall-tau correlations we build a new measure of concordance between the $O2$ and $O3$ orderings from s_1 until x and their anti-concordance for the rest of cells:

$$\begin{aligned} K_{2,3}(x) = &\,\mathrm{Kendall.tau}(O2(r_1 : x), O3(r_1 : x)) \\ &- \mathrm{Kendall.tau}(O2(x+1 : end), O3(x+1 : end). \end{aligned} \tag{21}$$

Finally, we find the cutoff x such that

$$x_{O1} = \arg\max_x \Big(K_{2,3}(x) - K_{2,3}(x-1)\Big). \tag{22}$$

Such finite difference optimization choice is to avoid influence of densities on where the cutoff should be. Note that we used this formulation to enhance clarity. The implementation to compute $K_{2,3}(x)$ uses a more efficient, recursive form: $K_{2,3}(x) = K_{2,3}(x) + \Delta K_{2,3}(x)$ and $x_{O1} = \arg\max_x \Big(\Delta K_{2,3}(x)\Big)$.

After this procedure there usually remains set of cells cannot be assigned to a single branch. We refer to these cells as *undecided cells* or as *cells in the decisions state*. See, for example the group of cells marked in light grey in Supplementary Fig. 1a. Once the three major branches in data have been found as described above, we can repeat the same procedure in each of the three branches to find further sub-branches.

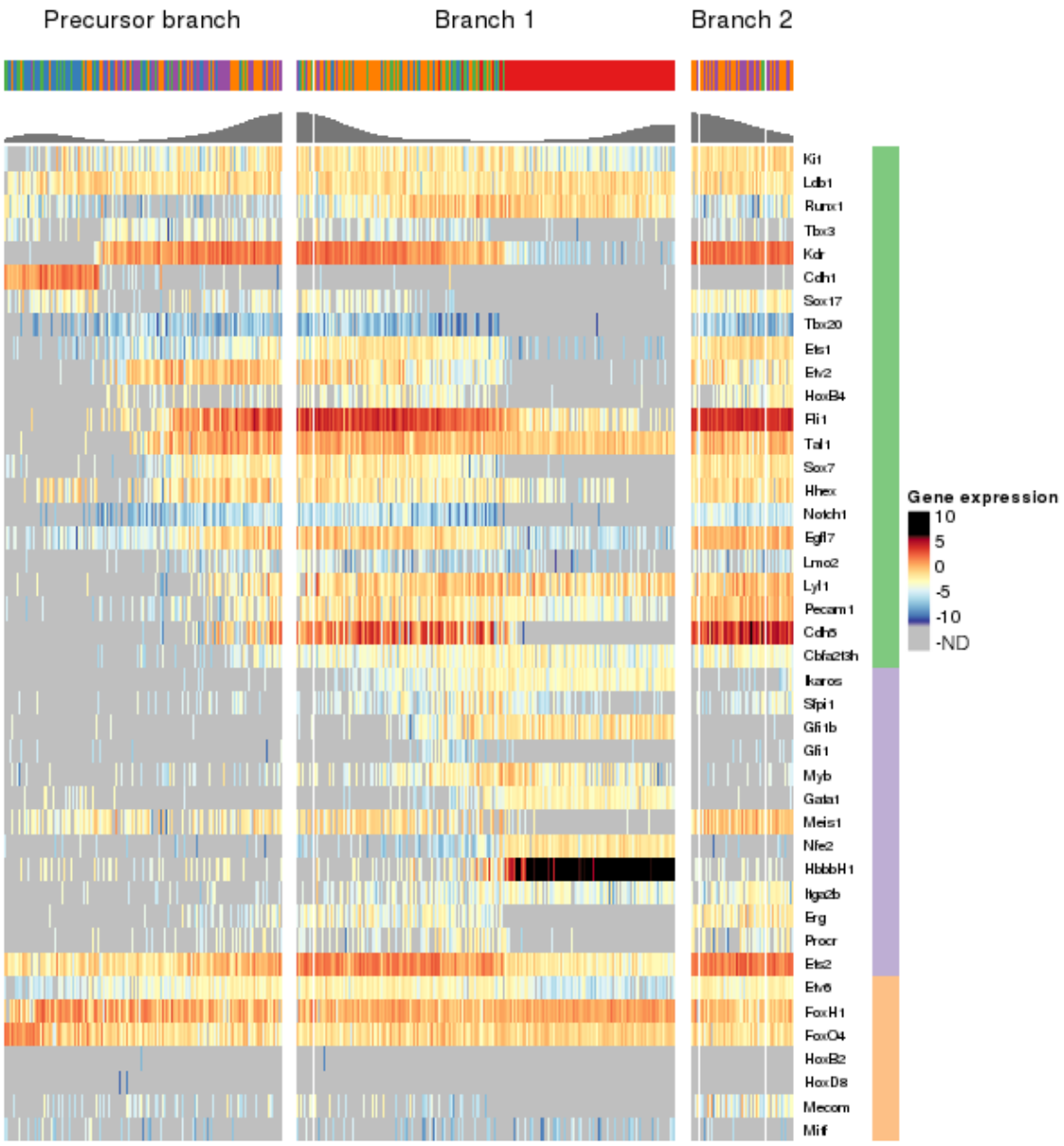
Supplementary Note 2: Mouse early blood development qPCR data

In the main text around Fig. 1, we discuss a DPT analysis of single-cell qPCR data set focusing on early blood development [7]. This supplementary note provides further details of the discussion in the main text. The data is publicly available in GEO with accession number GSE61470.

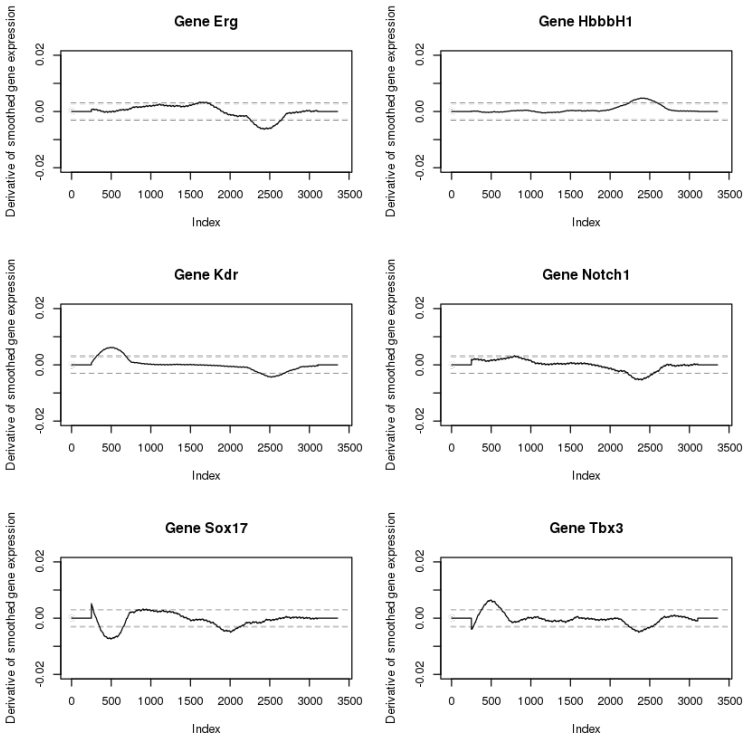
2.1 Pre-processing

We followed the normalization procedure suggested in Ref. [7]. Normalization was done by subtraction of gene expression from the limit of detection and normalization on a cell-wise basis to the mean expression of the four housekeeping genes (*Eif2b1, Mrpl19, Polr2a* and *Ubc*) in each cell. Cells that did not express all four housekeeping genes were excluded from subsequent analysis, as were cells for which the mean of the four housekeepers was ± 3 s.d. from the mean of all cells. A dCt value of -14 was then assigned where a gene was not detected.

Pseudotime ordering allows to stress the succession of different transcription factors in the qPCR data set. For this purpose, we computed the derivative of the expression along branch 1 and detected the most significant changes. In particular, we used the smoothed version of the data: a sliding window on the gene expression of 50 adjacent cells along the respective branch, where non-detected expression values were modeled with a Gaussian distribution with mean -14 and variance 3 (cf. Fig. 1d in the main text, a non-smoothed version is displayed in Supplementary Fig. N2). Then, we computed an adjusted Z-Score of the expression value with cut-off variance of 3 (in order to prevent largely non-detected genes to increase their noise-level, we used this cut-off in concordance to the noise introduced during smoothing). In addition, the derivative was approximated by a linear regression model over 500 values, largely reducing false positive peaks from noise (see Supplementary Fig. N3).



Supplementary Figure N2: The non-smoothed version of Fig. 1d in the main text. The colored top bar indicates the embryonic stage of origin for each cell (blue: PS, green: NP, orange: HF, red: 4SG+, purple: 4SG-). The top histogram bar indicates the cell density (high values correspond to a metastable state, low values correspond to transitions).



Supplementary Figure N3: Determination of switch-like transitions along the pseudotemporal ordering. The first derivative of the pseudotime series is approximately determined by a linear regression coefficient on a sliding window of size 500 along the pseudotime index of the smoothed gene expression. The smoothing is crucial to reduce noise-induced changes of the derivative. Only transitions above the threshold of 0.0028 were considered. A selection of first derivatives is displayed. Notably, there are sharp on- and off-switches (*Kdr, Sox17*) and weak or slow transitions (*Notch* on switching and *Etv6* expression).

2.2 Detecting transcriptional changes

Transcriptomic data sets at the single-cell level are usually accompanied by non-negligible levels of noise. Moreover, the heterogeneity of cell populations shown in bimodal expression patterns needs to be addressed. We employed a two-part, generalized linear model that allows to quantify the proportion of cells expressing a certain gene as well as the mean expression level, a modified Hurdle model [8]. Briefly, let Y_{ig} the gene expression level of gene g in cell i. Then, an indicator variable Z_{ig} determines, whether gene g is expressed in cell i and the expression level of gene g given it is expressed, is determined by a normal distribution:

$$\begin{aligned} \text{logit}\left(\mathrm{P}(Z_{ig}=1)\right) &= X_i\beta_g^D \\ \mathrm{P}(Y_{ig}=y|Z_{ig}=1) &= \mathcal{N}\left(X_i\beta_g^C, \sigma_g^2\right) \end{aligned}$$

We have two regression components, the discrete D and the continuous C component.
In order to compute a likelihood-ratio on two different populations, [8] developed a combined log-fold change defined as follows. For each gene g, let $u(x)$ the expected value of the continuous component, as $u(x) = \langle C, x\rangle$, and let $v(x) = (1+\exp(-\langle D, x\rangle))^{-1}$ the expected value of the discrete component. The log-fold change from population p_1 to population p_2 is then defined as

$$lfc(p_1, p_2) = u(x|x \in p_2) \cdot v(x|x \in p_2) - u(x|x \in p_1) \cdot v(x|x \in p_1).$$

2.3 Differential gene expression analysis

We employed the following analysis strategy. First, we introduced metastable states along the pseudotemporal order of the cells (see Supplementary Fig.1), where highly similar cells have approximately the same distance to the root cell. We separated three distinct states in branch 1 and two states in branch 2 by an appropriate threshold. The decision state is defined by the branch assignment method and is a site of cell accumulation. We highlighted and labeled metastable states in different grey shades (Supplementary Fig. 1).

Having defined these areas of interest, we fit a modified hurdle model to the gene expression data, calculated the log-fold change of the decision state to all other states (Supplementary Fig. 2) and used a likelihood-ratio test statistic to compute the significance level.

Furthermore, we repeated the differential gene expression analysis by the cell types. We compared the gene expression of head fold cells vs. primitive streak cells and 4SG negative cells, respectively. The comparison of head fold and primitive streak is supposed to correspond to the comparison of precursor and decision state. MAST detected 29 differentially expressed genes (cf. Supplementary Table 1) and most of these genes showed bimodal expression, *Cdh1* has even three levels (cf. Supplementary Fig. 2). The log fold-change of the detected genes has the same sign except for *Tbx20* and the absolute values are almost always lower in the cell type comparison than in the metastable state comparison. Comparing the results from the test decision state vs. terminal branch 2 and head fold vs. 4SG- cells gives a more contradictory picture. The differential analysis of head fold cells and 4SG negative cells detected *Cdh5* as a marker of the endothelial lineage but did not find other markers as *Itga2b*, *Mecom* or *Etv2* that were found in the metastable state setting (decision state vs. branch 2) (cf. Supplementary Fig. 2 and Supplementary Table 2). Both sets of differentially expressed genes share only *Cdh5*, *Myb* and *Sfpi1*. The sign in log fold-change is only consistent in *Myb*, whereas *Cdh5* and *Sfpi1* expression is higher in terminal branch 2 than in the decision state (positive *lfc*) and lower in 4SG- cells than in head fold cells (negative *lfc*).

Supplementary Note 3: Mouse embryonic stem cells inDrop data

In the main text around Fig. 2, we discuss a DPT analysis of single-cell RNA-seq data set using the inDrop protocol [9]. This supplementary Note provides further details of the discussion in the main text. In the experiment, single cells along with a set of uniquely barcoded primers were captured in tiny droplets and sequenced. The capabilities of this technique were demonstrated using an undirected differentiation process of mouse embryonic stem cells upon leukemia inhibitory factor (LIF) withdrawal. The data set is available under the GEO accession number GSE65525.

3.1 Pre-processing

There are various sources of variation in single-cell RNA-seq data, beginning with the tiny amounts of RNA molecules to detect to capturing efficiency and amplification bias. The authors in this data set applied both unique molecular identifiers (UMIs) and technical genes to determine a set of 2047 genes with cell-to-cell variance above the technical noise level and normalized by the total amount of transcripts:

$$\hat{m} = m \cdot \frac{\mathrm{E}(M)}{M}, \tag{23}$$

where $M = \sum_i m_i$ is the total amount of UMI-filtered reads m_i per cell and $\mathrm{E}(M)$ is the average of totals over all cells (cf. [9], supplement). We concentrated our analysis on the heterogeneous genes only. A biological source of variance in single-cell transcriptomics is the influence of the cell cycle genes. In particular, differentiating cells are very actively dividing. Recently, [10] introduced the *scLVM* approach to detect the estimate and correct for hidden biological effects as the cell cycle. The method is also capable to reduce batch effects (see Supplementary Fig. 3). We used the *scLVM* method to account for both technical and cell-cycle induced noise (see Supplementary Fig. 3).
First, we fit the noise model according to Brennecke et al [11] to a pure RNA control sample provided in the data set to estimate the technical noise of the protocol (Supplementary Fig. 3). For cell-cycle correction, we used the $\log_{10}(\hat{m}+1)$ expression values of the 2044 highly variable genes in 2717 cells measured at 0, 2, 4 and 7 days after LIF withdrawal and corrected for cell-cycle genes.

3.2 Clustering of genes

To determine the similarity among genes, we computed the gene-to-gene correlation and define $1 - \mathrm{cor}$ as a similarity measure. Next, we performed a hierarchical clustering on this similarity measure and highlighted four clusters (cf. Supplementary Fig. 4). We used three different types of normalization to compare the data. First, we have the $\log_{10}(\hat{m}+1)$ normalization we used for the cell-cycle correction. Unfortunately, we observed a very high variance in gene expression after the correction and decided to regularize the cell-cycle corrected data.

First, we used a quantile normalization for the expression y_{ig} as follows: We compute the 0.02- and 0.98-percentiles $(p_{g,0.02}, p_{g,0.98})$ for each gene g and calculate

$$\tilde{y}_{ig} = \frac{y_{ig} - p_{g,0.02}}{p_{g,0.98} - p_{g,0.02}} \tag{24}$$

Then, all expression values within the $[p_{0.02}, p_{0.98}]$-interval are normalized to the $[0,1]$-interval and outliers are found outside this interval.

Second, we applied a Z-Score-normalization with zero mean and unit variance for the expression y_{ig}. Both normalizations regularize the cell-cycle corrected gene expression. Though, the quantile normalization is more robust to outliers and alleviates their detection. First, batch and cell cycle correction decreased the day-to-day variability of the samples and by pseudotemporal ordering, the

differentiation process was resolved in greater detail. We are able to spot a single differentiation path in this data set as well as different subpopulations (cf. Supplementary Figs. 3 and 4). The diffusion pseudotime embedding resolves the heterogeneity of the measurement days (top annotation in Supplementary Fig.4 A-C and I). As we consider pseudotime as a measure of differentiation in this case, small pseudotimes correspond to a low degree of differentiation and a high degree of pluripotency, respectively. We observe an increasing degree of heterogeneity with time passed since LIF withdrawal and as reported in [9], we observe large variability at day 7 ranging from pluripotent cells to strongly differentiating cells. A detailed interpretation regarding gene expression patterns is conducted upon gene clusters.

3.3 GO enrichment analysis

To assess the reasonability of the hierarchical clustering, we performed a GO enrichment analysis using Genomatix software suite (*www.genomatix.de*) before and after cell-cycle correction (Supplementary Fig. 4). For illustration, we picked five GO terms for each cluster. Clustering of the non-cell-cycle corrected data revealed strong differences compared to the GO enrichment after cell-cycle correction. Indeed, the GO terms of the Z-score and quantile normalized data are very similar in all clusters. To demonstrate the differences arising with cell-cycle and batch correction, we compared GO terms given in the purple cluster (Supplementary Fig. 4). The p-value of GO:0008152 (metabolic process) is $1.37 \cdot 10^{-52}$ indicating a strong metabolic signature. The other displayed terms range among the top 400 GO terms in this list where we also have key factors of pluripotency promoting endodermal cell fate specification (GO:0001714) (*Pou5f1, Sox2, Nanog*) with a p-value $3.5 \cdot 10^{-4}$. However, in the yellow cluster is also a strong metabolic signature (p-value $4.08 \cdot 10^{-18}$), but deviates from the terms in the cell-cycle corrected data as we do not find GO enrichment for cell migration (GO:0016477) or regulation of cell motility (GO:2000145). Hence, in order to recover cellular events with GO enrichment, we need to consider a robust data normalization.

3.4 Differential gene expression analysis

We found several branches corresponding to subpopulations of differentiating ES cells. We found a first population branching off mainly consisting of day 2 cells and at the late state another split into epiblast-like and primitive endoderm-like cells. To test which genes are differentially expressed in the first side branch, we performed rank sums tests of 250 cells from the early state cells, the side branch cells and the epiblast-like cells, respectively (see Fig. 2b of the main text). We only considered those genes that have approximately the same expression level in the early state and epiblast-like state cells and a significantly different level in the side branch cells (three tests between the three groups, p-values Bonferroni adjusted, Supplementary Fig. 5 shows the p-value for test between the early state cells and the side branch cells). Considering p-values < 0.01, we identified 108 genes with a strong variation along the pseudotemporal order which were relevant in apoptosis (GO:0006915) and single-organism developmental process (GO:0044767).

Supplementary Note 4: Mouse myeloid progenitors MARS-Seq data

In addition to the data sets discussed in the main text, we here perform a DPT analysis for the data of Paul *et al.* [12], who combined index FACS sorting and transcriptomic profiling of single-cells to assess heterogeneity in myeloid progenitors. Data sets are publicly available in GEO with accession numbers GSE72857, GSE72858 and GSE72859.

4.1 Pre-processing and clustering

We focused on sorted c-Kit$^+$ Sca1$^-$ lineage (Lin)$^-$ bone marrow cells, a data set of 2730 single cells with 3461 informative genes. The selection of genes and batch-correction is described in detail in the supplement of [12]. DPT was performed on the $\log_{10}$-transformed data (adding pseudocount of 1).

Paul et al performed an elaborate clustering approach to identify 19 distinct progenitor classes with different degrees of differentiation. Roughly, cluster 1 to 6 represent the erythroid fate, clusters 7 to 10 correspond to the common myeloid progenitor (CMP), cluster 11 reflects the dendritic cell fate, clusters 12 to 18 exhibit a granulocyte-macrophage progenitor (GMP) fate. Cluster 19 is a lymphoid outlier class with 31 cells. In our analysis, we pick up the distribution of cells into branches and their association with the previously defined clusters.

For visualizing expression profiles, we normalized the count data through dividing by the 0.98-percentile of each gene and converted the result to the log-scale (adding pseudocount of 0.01), similarly to [12]. Then, a sliding window mean of 20 adjacent data points was used to account for the large number of displayed cells.

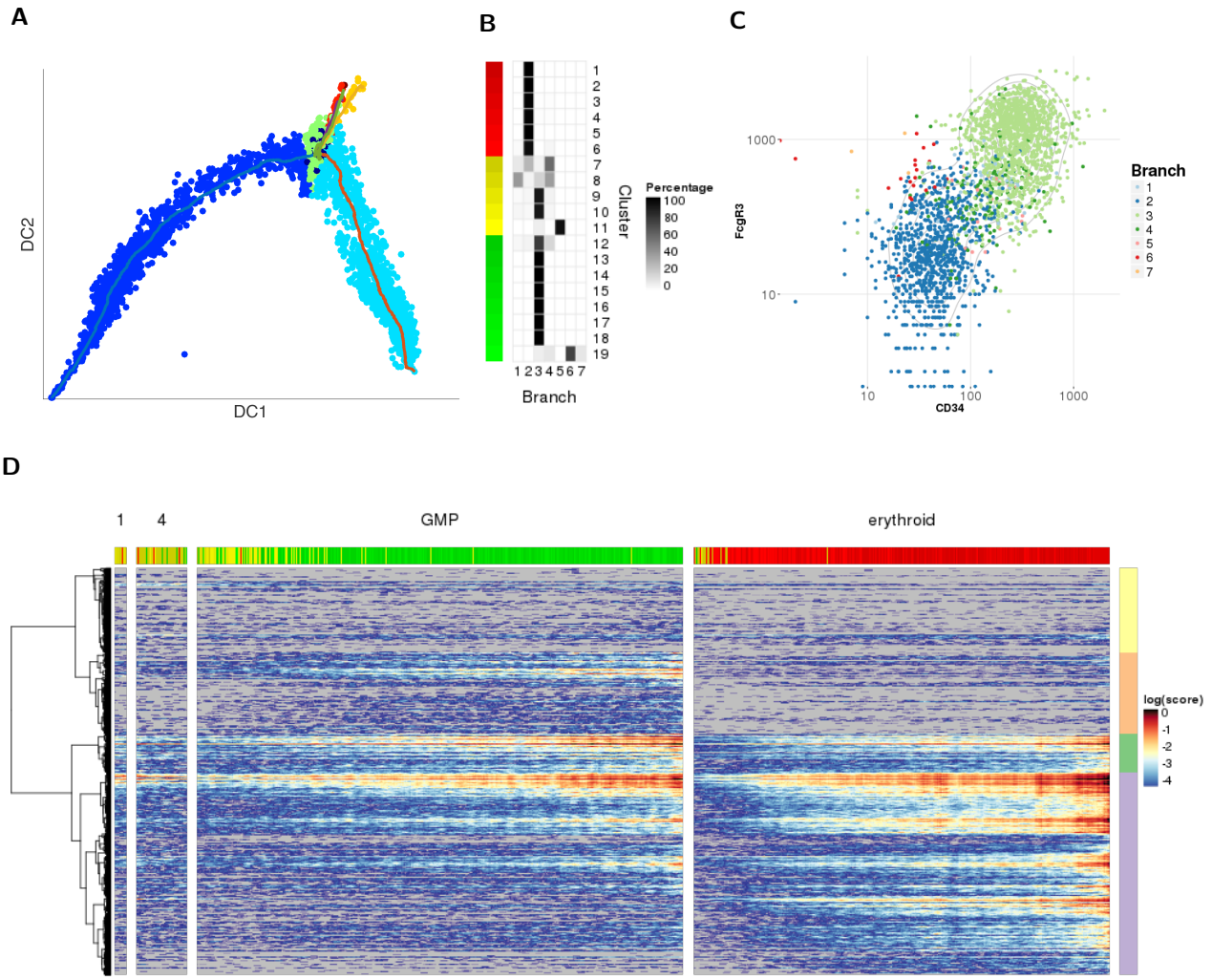
4.2 Data analysis

DPT identifies seven different, clear branches with a start cell in one of the CMP clusters (Supplementary Fig. N4A). DPT was run twice to find the branches (further iterations would only reveal noise, i.e. branches with negligible cell numbers). In Supplementary Fig. N4B, we show that two major branches correspond to the most common lineages with 41.8 % of cells (branch 2, erythroid lineage, marked in red) and 48.7 % of cells (branch 3, GMP fate, marked in light green). Most common myeloid progenitors (CMP) distribute in branch 1 and branch 4 with 1.9 % and 5.6 % of cells and reflect the group of undecided cells (in Supplementary Fig. N4B, each column is associated with a branch, and shows the fractions of cells in different clusters). Branch 5 (28 cells) coincides with cluster 11 and is well separated from the other branches.

Transcriptomic profiling of single cells resolves the heterogeneity of myeloid progenitor cells in contrast to classical FACS gating (Supplementary Fig. N4C). Cells assigned to branches 1 and 4 spread largely in CD34 expression. Using hierarchical clustering of 1−correlation of gene expression with Ward's distance, we identified four major groups in the informative gene set (Supplementary Fig. N4D), for which we performed a gene set enrichment analysis (GSEA, *www.genomatix.de*). The yellow set contains a number of lowly expressed genes with slightly higher levels in branches 1 and 4. Genes of this group are associated with immune response (e.g. *Gab2, Klrk1, Rap1b*) as well as regulation of developmental processes (e.g. *Apoe, Pbx1, Serpina3f*). Also *Gata2* has the highest expression level in branch 1 and 4. The gene expression of *Gata2* is preceding *Gata1* in erythroid development (Supplementary Fig. C). We conclude, that branch 1 and 4 consist of progenitors for both the GMP and erythroid fate, mainly in concordance with the clustering results of Paul et al (Supplementary Figs. N4D top bar and N5A-B).

In the major branches 2 and 3 we observe a continuous increase in metabolic processes characteristic for immune system development (highlighted by green side bar in Supplementary Fig. N4D. For example, *Arpc2, Bst2, Calr, Sec61a1*) are substantially lower expressed in branch 1 and 4 than in branch 2 or 3. Genes corresponding to erythroid cell fate condense in the purple set (e.g. *Gata1, Hba-a2, Car1/2, Gfi1b*) whereas granulocyte-monocyte specific genes agglomerate in the orange set (e.g. *Mpo, Gfi1, Elane, Runx1, Etv6*).

DPT allows us to draw several conclusions. First, DPT identifies multiple branches without extensive pre-processing. Specifically, we find the outlying groups of dendritic cells and natural killer cells. Their presence in the data does not perturb the branching results - manual removal from the input data set does not change the DPT distances. Also, branching complements the concept of classification as

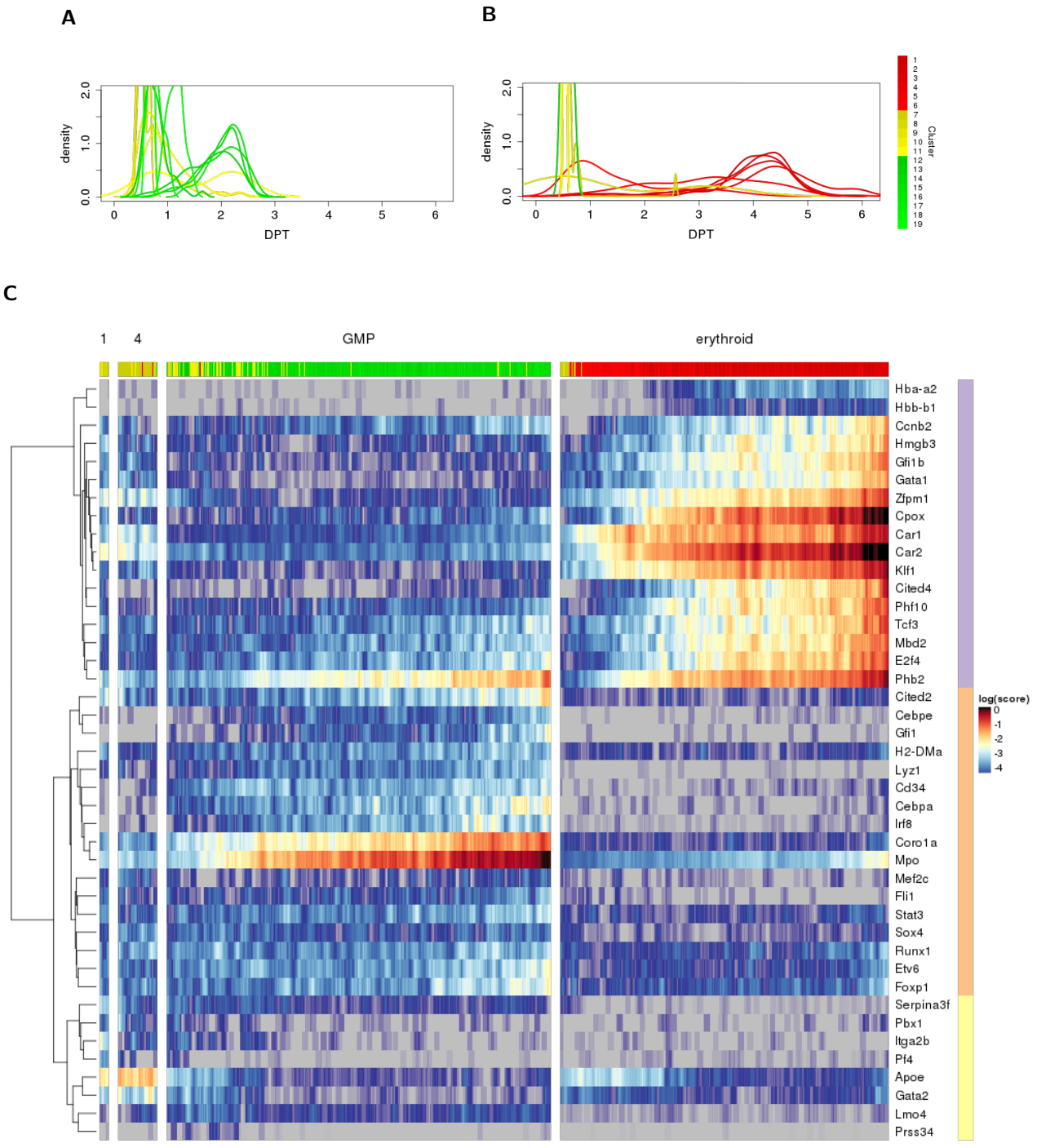


Supplementary Figure N4: A) Diffusion map plot of myeloid progenitors. Branches and DPT trajectories are highlighted. B) Clusters identified in [12] and their distribution into branches identified by DPT given as percentages. Branch 1 is the initial branch, while branch 2 and 3 comprise the erythroid (marked red in the left bar) and GMP cell fate (marked light green in the left bar), respectively. Branch 4 resembles branch 1. Branches 5, 6 and 7 consist of clusters 11 and 19 which where identified as dendritic cell and natural killer like cells [12]. DPT recognized these as loosely related to myeloid progenitor cells. The CMP cells are marked in yellow in the left bar. C) FACS-measured FcgR and CD34 protein expression levels for all cells grouped by branch. Clusters 1 and 4 distribute between GMP-like and erythroid-like populations. D) Gene expression of 3451 informative genes along branches 1 to 4. Genes are sorted via hierarchical clustering. Four groups of genes (side bar) distinguish roughly stem cell genes (yellow side bar), metabolic processes in bone marrow cells (green side bar), GMP cell fate regulators (orange side bar), erythroid cell fate genes (purple side bar).

presented by Paul *et al.*. With DPT, we find a set of cells being "undecided" (branch 1 and 4), whereas all cells on branch 2 and 3 (erythroid or GMP fate) are more likely to develop into the respective fate.

Second, branches do not correspond to a single cell type, but to a set of distinct cells in developmental progression. For both erythroid and granulocyte-monocyte development, we find a single continuous differentiation process by exploiting the transcriptional profiles from this particular snapshot. Clustering of cells does not necessarily lead to groups of homogeneous cells (Supplementary Figs. N5A-B). Paul et al proposed a consecutive pattern of clusters 1 - 7 on the erythroid branch, but the cells mix strongly in DPT. On the other hand, GMP fated cells (clusters 12-19) are summarized in a single branch implying a *single process of differentiation* instead of a *split-up into distinct clusters*. The main evidence for this is that we clearly observe upregulation of lineage specific, but not cell type specific gene sets.

Third, upon using a pseudotime analysis, we observe sequential upregulation of genes (e.g. purple cluster in Supplementary Fig. N5C), which is a promising indicator for regulatory events. As illustration, *Klf1, Zfpm1* and *Gata1* are essential for the development of erythrocytes and they increase successively on the erythroid branch. Lateron, the protein heterodimer Gata1-Zfpm1 represses *Klf1* expression, but the sequential increase starting with *Klf1* could imply a negative feedback loop. In this sense, cell ordering by DPT provides a basis to suggest gene interactions.



Supplementary Figure N5: A,B) Concordance of cluster labels on branch 3 (GMP fate) and branch 2 (erythroid fate). C) Transcriptional expression profiles of key genes along branches 1 to 4. Genes are sorted by hierarchical clustering. Three main gene sets (side bar) are identified and show erythroid marker genes (purple side bar), GMP lineage marker genes (orange side bar) and stem cell marker genes (yellow side bar).

Supplementary Note 5: Simulation study of robustness

We performed pseudotime ordering of 100 bootstrap sets for each data set with all previous algorithms: DPT, Monocle [13] and Wanderlust/Wishbone [14, 15]. Because Monocle fails to run for large numbers of cells, we used a smaller bootstrap sample size (700 cells).

After running each algorithm on all bootstrap sets, we calculated Kendall-tau correlation of two pseudotime ordering runs on the intersection (i.e. common cells in two bootstrap sets) of each pair of bootstrap sets. This resulted in a 100 by 100 (lower triangular) matrix for each method that we call self-concordance. We then calculated the mean μ_i and variance σ_i^2 of the elements across the self-concordance matrix for method i. A 2-sided t-test was performed on the self-concordances to specify the significance of difference in robustness of the Monocle and Wanderlust/Wishbone orderings compared to the diffusion pseudotime (DPT) ordering:

$$\mathrm{t} = \sqrt{\frac{m^2}{2m}} \cdot \frac{|\mu_1 - \mu_2|}{s} \tag{25}$$

where m denotes the number of bootstrap runs with each method (100) and the index i labels the method ($i = 1$ for DPT, $i = 2$ for Monocle or Wanderlust/Wishbone). The weighted variance s is computed as

$$s = \sqrt{\frac{(m-1)\sigma_1^2 + (m-1)\sigma_2^2}{2m-2}}, \tag{26}$$

where $2m$ equals the degrees of freedom (df). The p-values were then computed using the tcdf function in Matlab as $p = 2 \cdot (1 - \mathrm{tcdf}(t, \mathrm{df}))$.

Supplementary Note 6: Actual time, universal time, pseudotime

Cell differentiation is a largely asynchronous process. Even if we consider a single-fated lineage, due to the stochastic nature of the system, a heterogeneous population of cells coexists at any given time. Although each single cell takes a different trajectory in actual time (due to the stochasticity in the differential equation), all these trajectories lie on a common manifold C in the gene expression space ($C \subset \mathbb{R}^G$), where G denotes the number of genes. The manifold C (if one dimensional) can be parametrized by the arc length s along C.

Let us study a single cell trajectory along the manifold. For this, we can assign a velocity $\boldsymbol{v}(t)$ to each time point t that is approximately tangent to the manifold C (Supplementary Fig. N6). If we consider an equidistant temporal sampling of the single cell trajectory, the tangent velocity is inversely proportional to the density $\rho(t)$ of the cell states on the trajectory at that time point, that is $|\boldsymbol{v}(t)| = 1/\rho(t)$. In other words, the more the time points of the single cell trajectory happen to be in a region of $\mathbb{R}^G$ (black circle in Supplementary Fig. N6B), the slower the single cell has passed through that region. Because $\boldsymbol{v}(t)$ is tangent on C we can write

$$\mathrm{d}s = |\boldsymbol{v}(t)|\mathrm{d}t = \frac{1}{\rho(t)}\mathrm{d}t. \tag{27}$$

Integrating ds, starting at the root cell, along C up to actual time t yields the arc length, which we refer to as *universal time*

$$s(t) = \int_{C:[s(0),s(t)]} \mathrm{d}s = \int_0^t |\boldsymbol{v}(t')|\mathrm{d}t' = \int_0^t \frac{1}{\rho(t')}\mathrm{d}t'. \tag{28}$$

This assigns a universal time $s(t)$ to every actual time single cell trajectory as measured in time-lapse microscopy. However, for snapshot data it is often difficult to learn the original manifold C and obtain universal time.

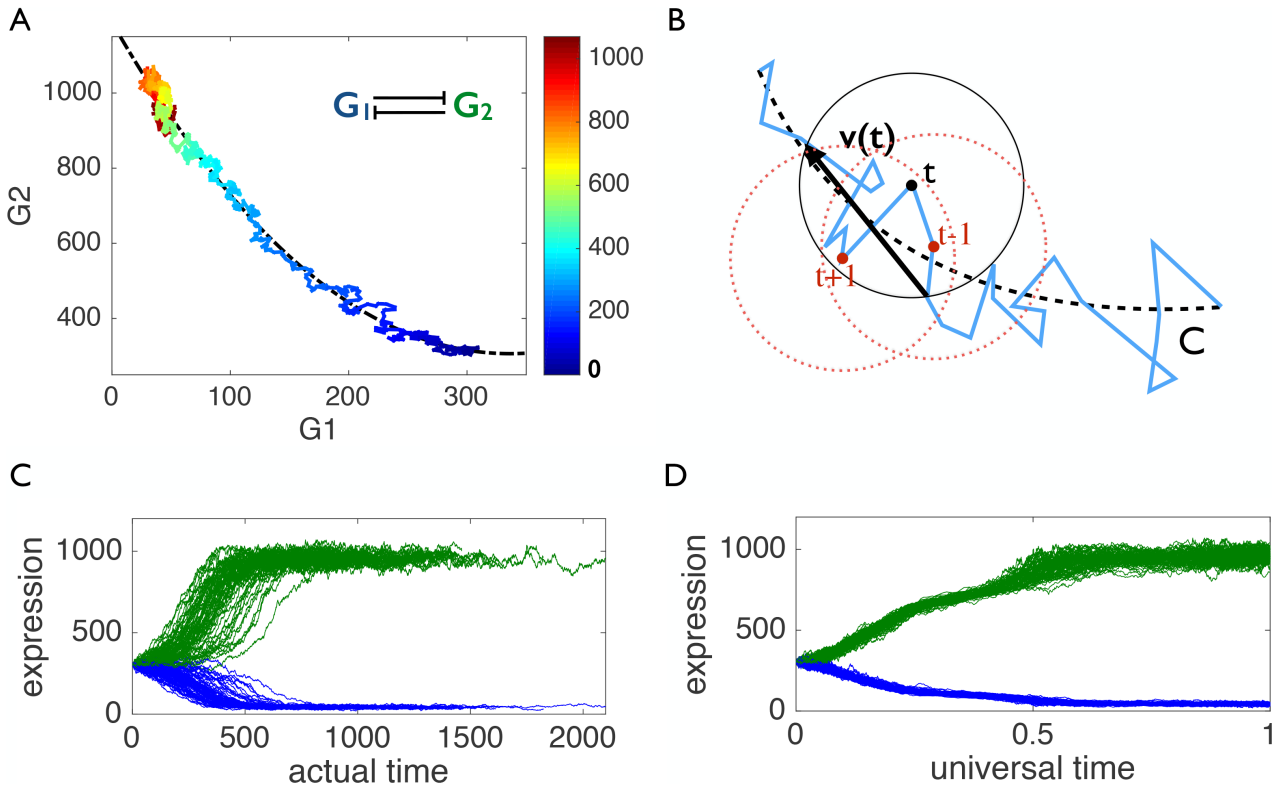
In the context of snapshot data, the common practice is to first map the data to a new space, where noise is diminished and thereby the manifold becomes more pronounced. Then in general one can define pseudotime as the distance (arc length) to the root cell on the mapped manifold (call it C'). Such notion of pseudotime as an arc length has also previously been used in [16, 17, 18] and [14]. In the present work, the mapping is from $\mathbb{R}^G$ to $\mathbb{R}^{n-1}$, where n is the number of cells, and distances on C' are characterized by a new metric we term "diffusion pseudotime".

Supplementary Figure N7 illustrates the three concepts (actual time, universal time and pseudotime) and how they are related to each other. Thus, we established a unified framework which can be used to bring time-lapse microscopy data and single-cell snapshot expression data together and make them comparable (e.g. the data from [19]). The connection of universal time with pseudotime as established here is only valid if cells from different developmental stages are present in a snapshot sample. However, we do not make any assumption of stationary sampling (as e.g. used in [18]). This is especially helpful in the context of single-cell snapshot data where sampling densities are usually far from any stationary state and influenced by cell division rates, noise, and the design of the experiment.

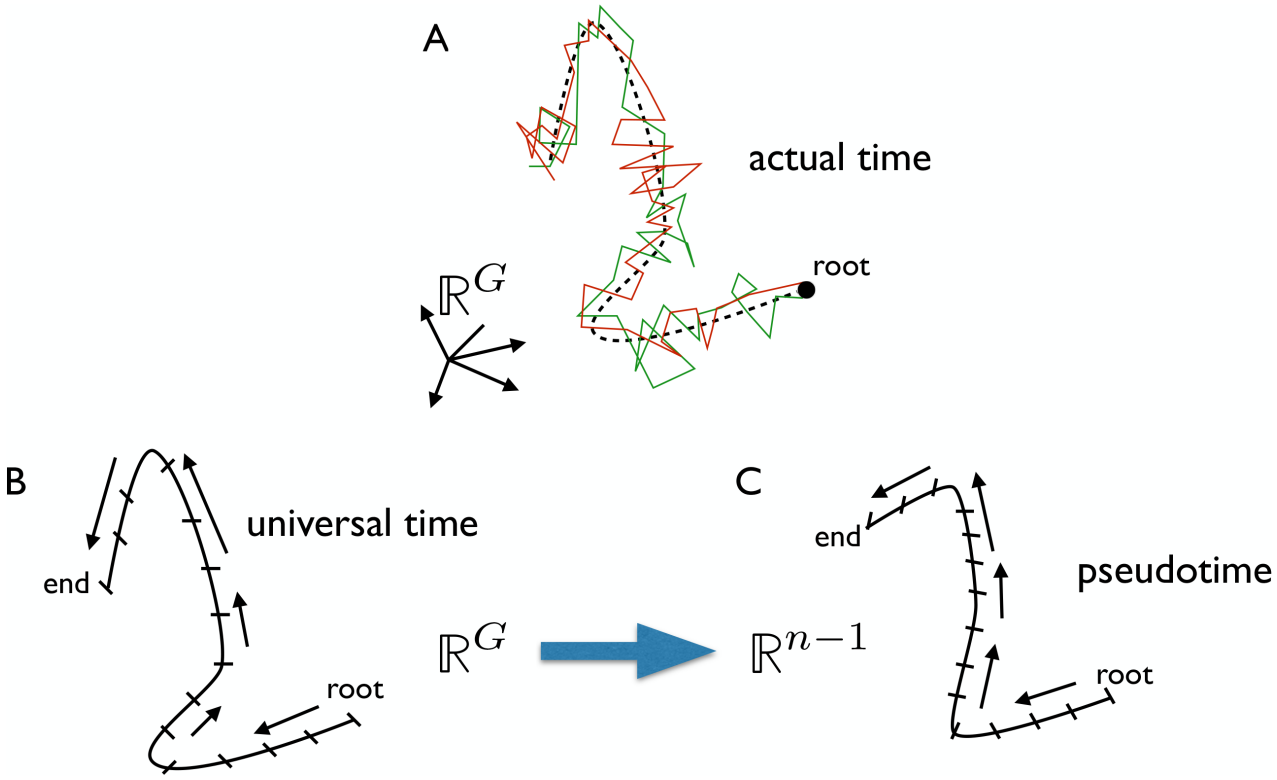
As a demonstration for asynchrony among single cells (even though from a single lineage), we simulated 100 cells (with a gene regulatory network of 6 genes [1]), starting from the same state at time zero (Supplementary Fig. N8A). Supplementary Fig. N8B shows that, when plotting gene expression versus universal time, all expression trajectories of these asynchronous single cells are brought to a unique expression curve, which we refer to as the "universal gene expression trajectory" for that lineage.

Supplementary Figure N8D shows the Wanderlust pseudotime for the toy data and Figures N8E to G) show diffusion pseudotime on several mappings C' depending on the choice of the used diffusion map method and its respective parameter (see caption of Supplementary Fig.N8) .

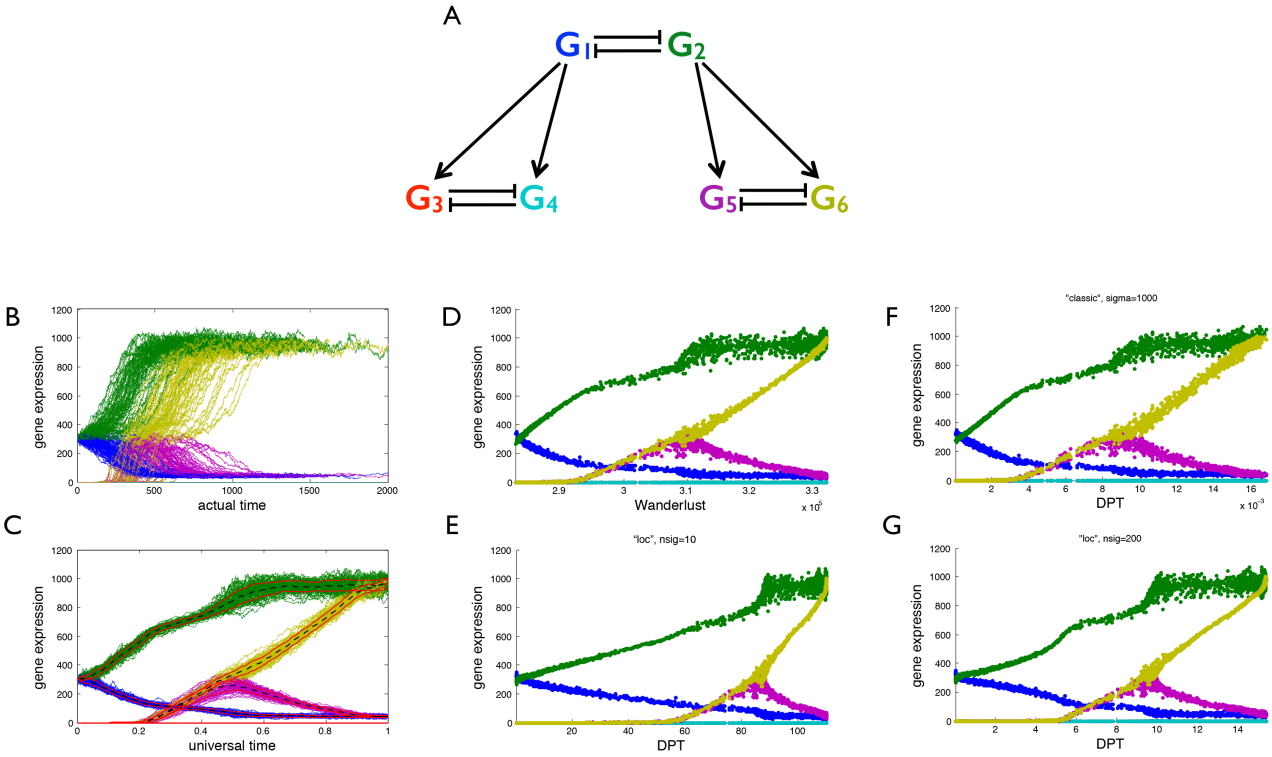
We mention that distances on manifolds have been discussed in several other publications, but only for snapshot data, and without realizing the connection to actual time trajectories as measured in time-lapse microscopy. In Note 7.5.2, we discuss this in detail.



Supplementary Figure N6: A) A single cell trajectory for a toggle switch.The color indicates actual time. The trajectory is adjacent to a manifold C (dashed line) which we hypothesize is the same for all single cell trajectories following the same dynamics model. B) For any single cell trajectory with equidistant sampling of time steps, a velocity $\boldsymbol{v}(t)$ is defined for each time point t that is approximately tangent to the manifold C and $|\boldsymbol{v}(t)|$ is approximately equal to the diameter of a circle centered at time point t divided by the number of time points inside the circle. That is $|\boldsymbol{v}(t)|$ is inversely proportional to the density of trajectory points at t. For each single cell trajectory, we calculate the universal time as $\int_0^t |\boldsymbol{v}(t')|dt'$. C) Expression of G_1(blue) and G_2(green) for several single cell trajectories versus actual time exhibit large asynchrony between the single cells because of the stochasticity in the dynamics model. D) The asynchronous trajectories in C fall on top of each other when plotted against universal time.



Supplementary Figure N7: A) Two (red and green) actual time single cell trajectories in gene expression space($\mathbb{R}^G$). Each jump on a trajectory happens in an (equidistant) unit of actual time. B) Universal time is defined as *arc* length on the data manifold starting from the root. This manifold $C \subset \mathbb{R}^G$ remains the same for several single cell trajectories, as well as for snapshot samples of single cells. C) Pseudotime (in general) is defined as arc length on a more pronounced mapped manifold C' (with respect to noise). In case of diffusion pseudotime this mapping is from $\mathbb{R}^G$ to $\mathbb{R}^{(n-1)}$ where n is the number of sampled cells.



Supplementary Figure N8: A) A toy gene regulatory network with 6 genes. B) Actual time simulation of expression for the single-fated lineage for which G_2 and consecutively G_6 win the toggle-switch competitions. C) Expression time series versus universal time. D) Expression of snapshot sampled data vs. Wanderlust pseudotime. E) Expression of snapshot sampled data vs. diffusion pseudotime (locally rescaled diffusion map, $\kappa = 10$). F) Expression of snapshot sampled data vs. Diffusion pseudotime (classic diffusion map, $\sigma = 1000$). G) Expression of snapshot sampled data vs. Diffusion pseudotime (locally rescaled diffusion map, $\kappa = 200$).

Supplementary Note 7: Comparison of DPT with previous algorithms

We compared the performance of DPT with Monocle [13] and Wanderlust/Wishbone [14, 15] on several generic single-cell data sets;

- artifical data
- early blood development qPCR [7], as discussed in Fig. 1 of the main text and Supplementary Note 2.
- mESC inDrop [9], as discussed in Fig. 2 of the main text and Supplementary Note 3,
- Myeloid progenitors MARS-seq data [12], as discussed in Supplementary Note 4.

We show that DPT finds pseudotime ordering and branching structures across all of the data sets independent of the data size and the experimental technique used for generating them. The performance of Monocle and Wanderlust/Wishbone, by contrast, strongly depends on the dataset: both algorithms yield qualitatively wrong results for the branch detection on all experimental data sets. To assess the performance of Wishbone, we used the default input parameters as set in the Wishbone package [15], except for the number of diffusion map components ("kEigs") which we set to zero, meaning that all diffusion map components are considered, as is the case in DPT. If we set this parameter to another value, we were not able to obtain meaningful results at all. To assess the performance of Monocle, we set the "num-path" parameter in the package [13] to the number of branches we expected for each data set.

7.1 Comparisons on simulated data

For the six dimensional toy model generated by simulation of a stochastic toggle switch with downstream gene activation as shown in Supplementary Figure N8A, we applied a nonlinear transformation of data points to increase the complexity of the data. For this, we applied Eq. (29) to each data point component wise.

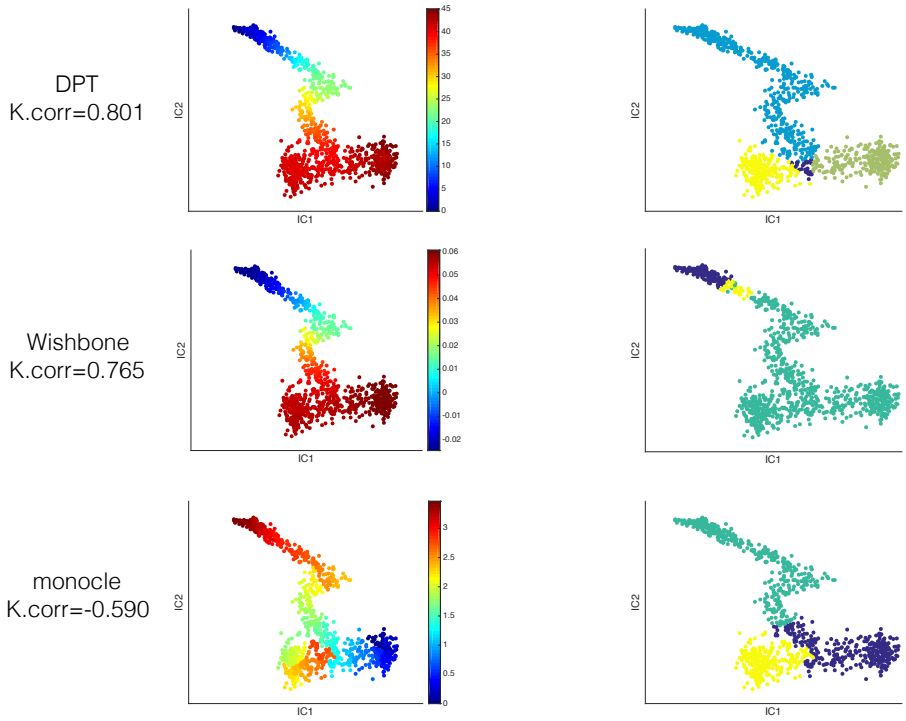
$$y = \frac{1}{1 + \exp(-x)} \tag{29}$$

Then a Gaussian random matrix was multiplied with transformed data matrix in order to add noise and project data into three dimensions.

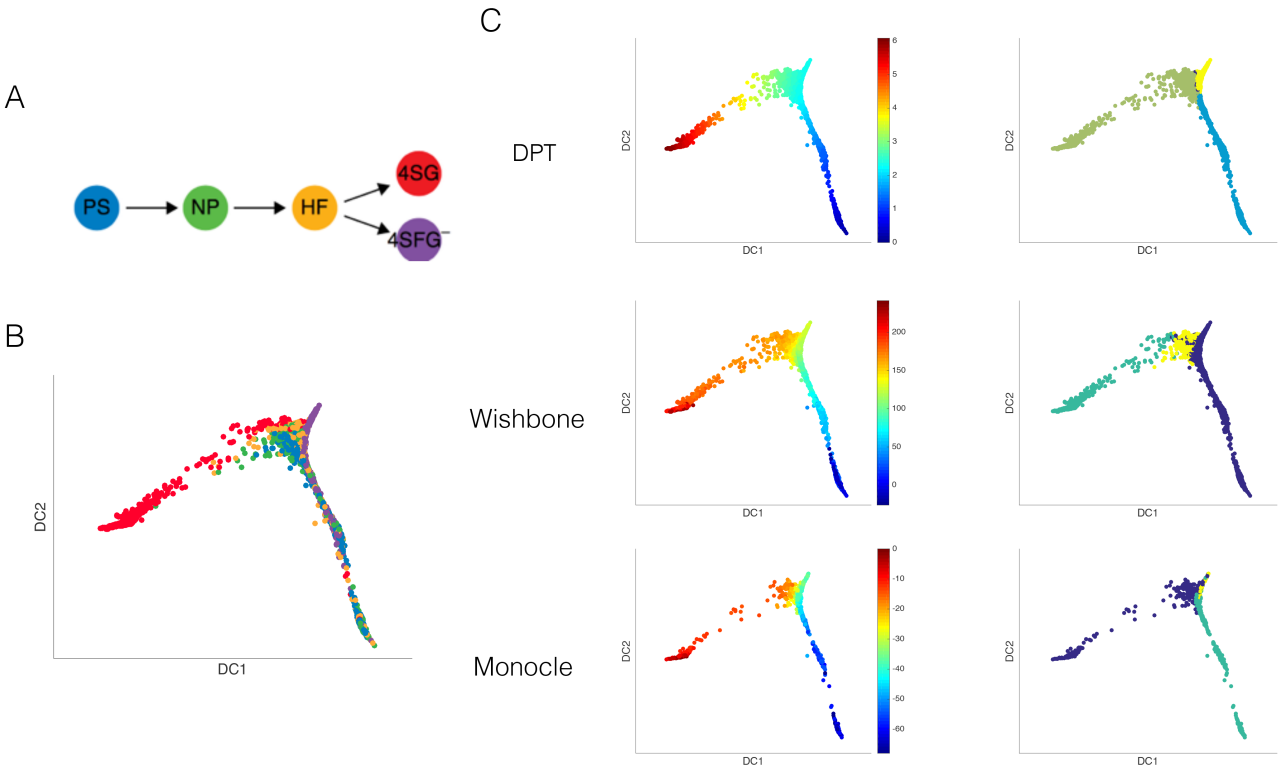
The branching structure of the resulting data is clearly visible on independent component analysis (ICA) plots in Supplementary Fig. N9. We further observe that only DPT is able to identify the branches correctly. DPT also shows the highest concordance of pseudotime ordering to the actual time ordering of the artificial data. Wishbone finds the pseudotime trajectory relatively well but does not find the branching correctly. It identifies a branching event rather in the trunk of the data manifold. A root cell cannot be specified in Monocle which explains the negative Kendall-tau correlation (concordance) of pseudotime with actual time. Whereas for nonbranching data one could simply reverse the pseudotime ordering, for branching structures it requires more effort to put Monocle's order in place. Moreover, the minimum spanning tree approach used in Monocle produces a shortcut on the left branch which leads to wrong pseudotime ordering in that region.

7.2 Comparisons on mouse early blood development qPCR data

Figure N10 shows how DPT, Wishbone and Monocle perform on the early blood cells qPCR data set. Wishbone does not find the correct branching which corresponds to the four somite GFP negative



Supplementary Figure N9: Comparison for artificial data. Shown are pseudotime ordering (left column) and branch detection (right column) of previous algorithms visualized using independent component analysis (ICA) plots. In the left column, the coloring reflects pseudotime values. In the right column, colors correspond to distinct branches. In contrast to Wishbone and Monocle, DPT additionally identifies undecided cells, which we mark in dark blue. The Kendall-tau correlation of pseudotime ordering to actual time ordering is indicated by "K.corr". Only DPT yields a good agreement of pseudotime and actual time orderings (top left panel), and detects the branches correctly (top right panel). Wishbone fails to correctly detect the branches (center right panel), and Monocle fails to recover the true ordering (bottom left panel).

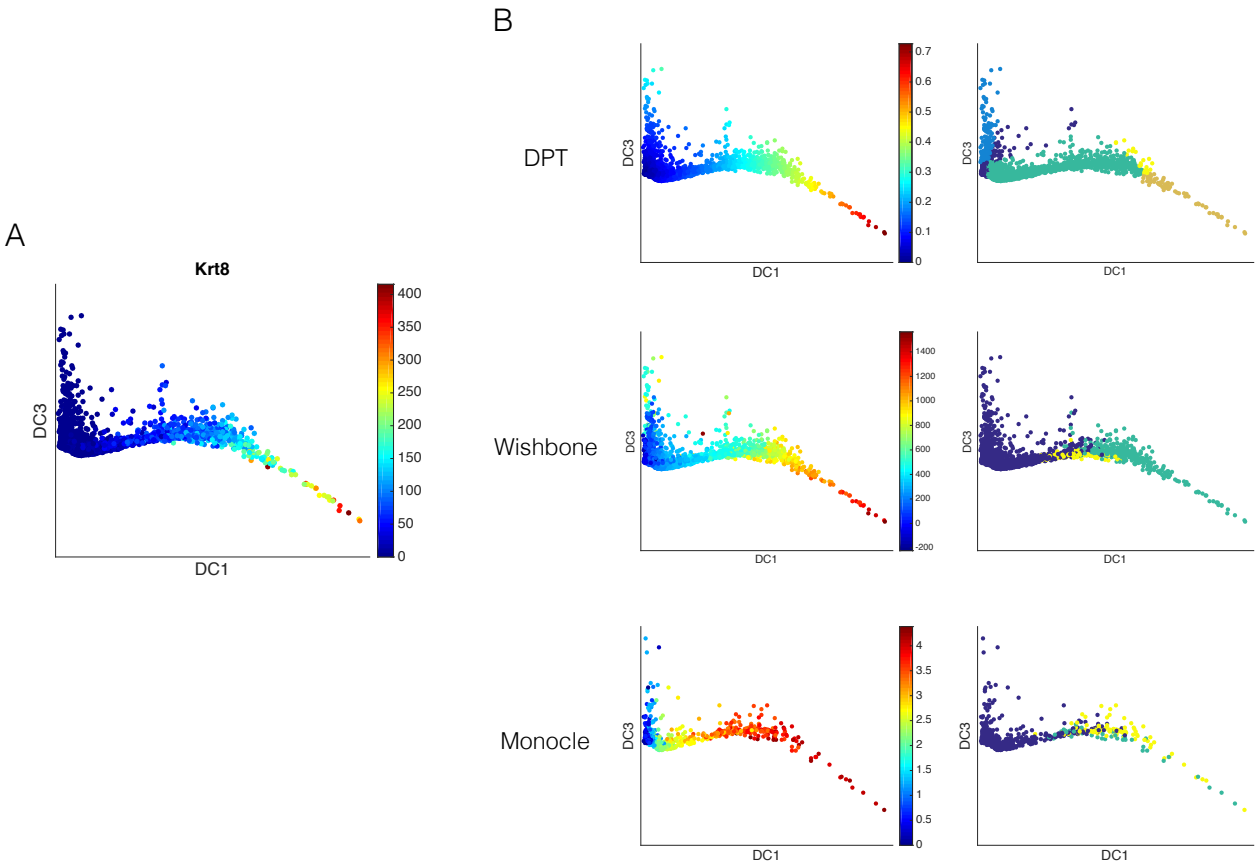


Supplementary Figure N10: Comparison for single-cell qPCR for early hematopoiesis [7]. A) Cells are sorted according to 5 different populations [7]: primitive streak (PS), neural plate (NP), head fold (HF), four somite GFP negative (4SG-), four somite GFP positive (4SG+). B) The 5 populations visualized using the first two diffusion components. Note the endothelial branch in the upper right corner of the panel. C) Comparison of pseudotime ordering (left column) and branch detection (right column) visualized using the first two diffusion components. In the left column, the coloring reflects pseudotime values. In the right column, colors correspond to distinct branches. In contrast to Wishbone and Monocle, DPT additionally identifies undecided cells, which we mark in dark blue (only very few of them are visible here). The pseudotime ordering provided by all three methods is similar. But only DPT detects the endothelial branch, which consists of cells from the 4SG- population (B), whereas Wishbone and Monocle fail to detect it.

(4SG-) population. Instead it finds a branching a region which has no overlap with 4SG- cells. DPT and Monocle find this branch in the correct place. To run Monocle on this data set as well, we had to subsample down to 700 cells.

7.3 Comparisons on mouse embryonic stem cells inDrop data

Comparison for mESC drop-Seq data [9]. We computed the pseudotime of cells by merging the experiments from the four different days (Supplementary Note 3) for all previous algorithms. Figure N11



Supplementary Figure N11: Comparison for mESC drop-Seq data [9]. We computed the pseudotime of cells by merging the experiments from the four different days (Supplementary Note 3). A) Expression of $Krt8$ for each single cell visualized using the first two diffusion components. $Krt8$ is highly expressed in epithelial cells. The plot therefore allows to identify the epithelial branch. B) Comparison of pseudotime ordering (left column) and branch detection (right column) visualized using the first two diffusion components. In the left column, the coloring reflects pseudotime values. In the right column, colors correspond to distinct branches. In contrast to Wishbone and Monocle, DPT additionally identifies undecided cells, which we mark in dark blue. Pseudotime ordering is similar for all three previous algorithms, although DPT shows a higher concordance with the actual time labels (also see Fig. 2f in the main text). But only DPT correctly finds the epithelial branch associated with high expression of $Krt8$, light brown coloring in the upper right panel.

shows how DPT, Wishbone and Monocle perform on inDrop mESC cells. In N11A high expression of $Krt8$ indicates the location of Epithelial cells. While DPT identifies these cells as a branch, Wishbone and monocle do not find this branch correctly.

7.4 Comparisons on Mouse myeloid progenitors MARS-Seq data

We compared the three methods DPT, Wishbone and Monocle on a Myeloid progenitor cells data set [12] gathered by the single-cell technique called MARS-seq [20]. For this data set the authors were able to identify 19 distinct cluster. N12 shows the diffusion map for this data with CMP (corresponding to

clusters 7 and 8), GMP (clusters 9 to 19) and erythroid (clusters 1 to 6) subpopulations. N12 shows the 19 clusters on diffusion maps. We performed DPT and Wishbone analysis on a debatched data matrix of 2730 cells and 3461 informative genes that the authors provide in [12]. For Monocle however we randomly subsampled 700 cells because of the algorithm's cell number limitations. For DPT and Wishbone we picked a random cell from the CMP population as the root cells. Monocle however does not allow inputting a root cell as the start point of the pseudotime.

DPT successfully finds the erythroid and GMP branch as well as two more small branches corresponding to clusters 19 and 11 in [12].

Although authors in [15] find the erythroid and GMP branches, we find that they achieve this by massive preprocessing of data which we were unable to reproduce due to unavailability of their preprocessing code. Thus the Wishbone result we present here shows Wishbone's performance on the whole debatched (2730 cells, 3461 genes) data matrix (the same input for DPT). Wishbone only finds few cells at the tip of each major branch as assigned to the GMP and erythroid branch.

Monocle performs well on this data set separating the GMP and erythroid as well as three more branches (states as called in monocle) when number of paths (an input parameter of monocle) is set to 4.

7.5 Methodological comparison of previous algorithms

In this section we discuss the methodical differences of the previous algorithms (DPT, Monocle [13] and Wanderlust/Wishbone [14, 15]), focussing on several challenges related to single-cell. These theoretical arguments allow to understand the superior performance of DPT, for which we provided numerical evidence in the previous section. Supplementary Table N1 provides an overview of the methodological comparison.

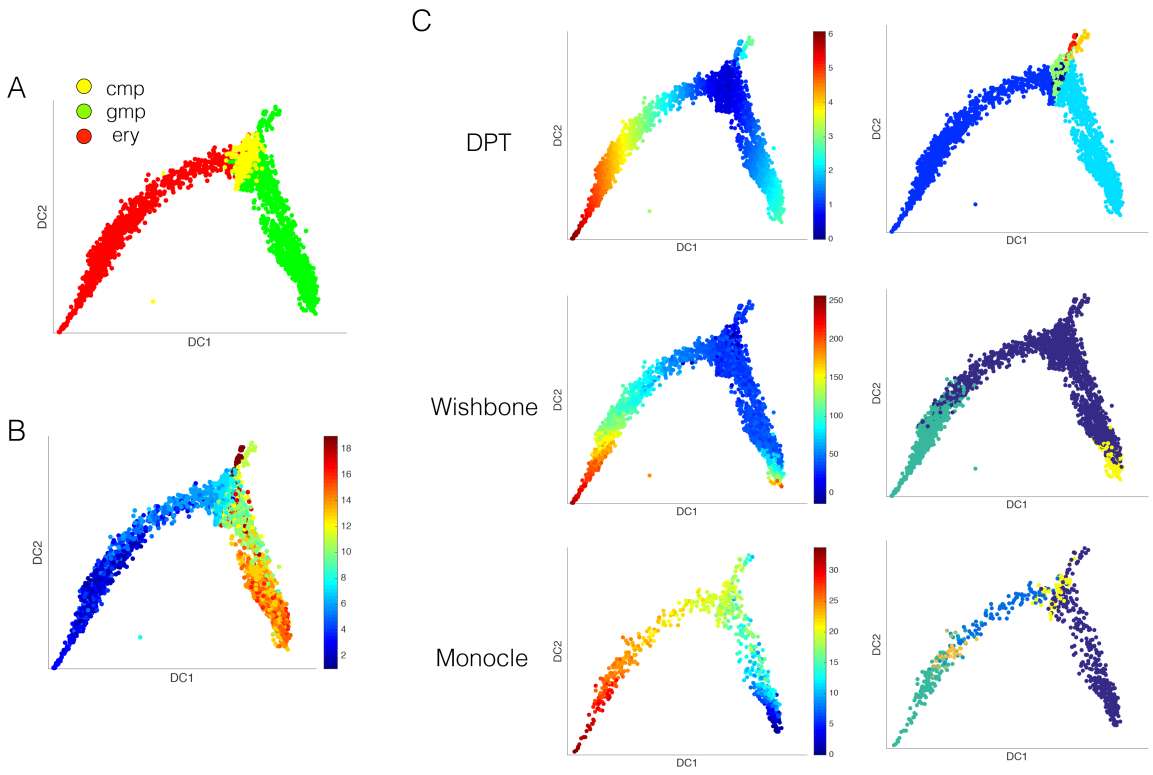
7.5.1 Basic working principle

Here we give a simplified and brief explanation of the Monocle and Wanderlust/Wishbone methods. Monocle orders cells on the Minimum Spanning Tree built on a few Independent Component Analysis (ICA) embedding dimensions. Note that ICA is a linear method and cannot capture the non-linearity of differentiation manifolds present in many experimental data sets. Wishbone is an extended version of Wanderlust for branch finding. Wishbone performs Wanderlust on a few diffusion map components and introduces a new algorithm for finding branches in the data. Wanderlust first builds nearest neighbors graph on cells. Assuming a nonbranching manifold, Wanderlust allows treating relative distances as a scalar (referred to as "displacement"), where addition or reduction of the displacements (i.e. orientation of the 1D manifold with respect to the root cell) is decided by a number of (random) landmarks on the data. It then averages the displacement of each cell relative to the root cell for a finite number of sampled paths on the graph. Wishbone's branch detection is then based on disagreements of the wanderlust distances of cells to several landmark cells (also termed "way-points").

7.5.2 Definition of pseudotime

Monocle defines pseudotime as distances on a single path of a minimum spanning tree that is constructed on an ICA dimension reduced embedding of the data.

If performed on data in the original gene expression space ($\mathbb{R}^G$), Wanderlust aims to recover the geodesic distances from the root cell on the original manifold $C \subset \mathbb{R}^G$, which would exactly be the same as universal time (see Supplementary Figs. N7 and N8C in Note 1.1). However, Wanderlust's path sampling approach hinders the reliable recovery of such geodesics for more complicated manifold structures (e.g. sharp turns, branching) and in presence of large noise in the data. Wanderlust, when



Supplementary Figure N12: Comparison for MARS-seq data for myeloid progenitor cells [12]. A) Common myeloid progenitor (CMP), granulocyte-macrophage progenitor (GMP) and erythroid (ery.) populations with labels from from Ref. [12] visualized using the first two diffusion components. B) The 19 clusters identified in Ref. [12]. C) DPT, Wishbone and Monocle pseudotime orderings (left column) and branch detection (right column). In the left column, the coloring reflects pseudotime values. In the right column, colors correspond to distinct branches. For DPT, we additionally mark undecided cells in light green and root cells in dark blue. Pseudotime ordering is similarly good for all previous algorithms. But, only DPT finds the GMP and erythrocyte branches as well as two smaller branches corresponding to clusters 12 and 19 in (B). Wishbone and Monocle fail to detect the true branch structure.

performed on reduced dimensions as in the case of Wishbone, performs better, though at the cost of neglecting information in higher dimensions.

DPT uses a path integral approach to account for all possible paths between cells. Diffusion pseudotime is finally defined as Euclidean distance to the root in a mapped $\mathbb{R}^{(n-1)}$ space with $\frac{\lambda_i}{1-\lambda_i}\psi_i, i = 1, ..., n-1$ coordinates, as discussed in Supplementary Note 1.2.

7.5.3 Robustness to noise and subsampling

Monocle's pseudotime order is based on only one single connecting path in the minimum spanning tree. This does not consider the possibility of reaching to a state through multiple paths, which is in principle possible in differentiation systems. Whereas a single path might provide a reasonable pseudotime order for clean data, as soon as there is more noise in the data, biologically non-relevant short cuts come into play and biologically relevant information of alternative paths get neglected because the pseudotime measure based on a single path is not robust to noise.

Wanderlust provides considerable robustness to noise by sampling multiple paths to connect each cell to the initial cell (rather than choosing only one single path). However the path sampling approach is dependent on sampling density of several cell subpopulations and the algorithm lacks any correction for density heterogeneity effects on the paths sampling.

DPT's path integral (considering all possible paths) approach renders it quite robust to noise. Furthermore in the implementation of the diffusion map (used by DPT), we correct for density heterogeneity effects [1].

To demonstrate DPT's robustness to sampling density, we performed DPT once on the complete artificial data (see Supplementary Note 7.1). The red curve in Figure N13A shows the nonuniform original sampling density and presence of metastable states in the original data. We then performed DPT on a rather uniformly subsampled set from the original data set (blue bars). Figure N13B shows that the DPT computed from the original data is strongly correlated with DPT computed from the subsampled data in spite of strongly different density distributions.

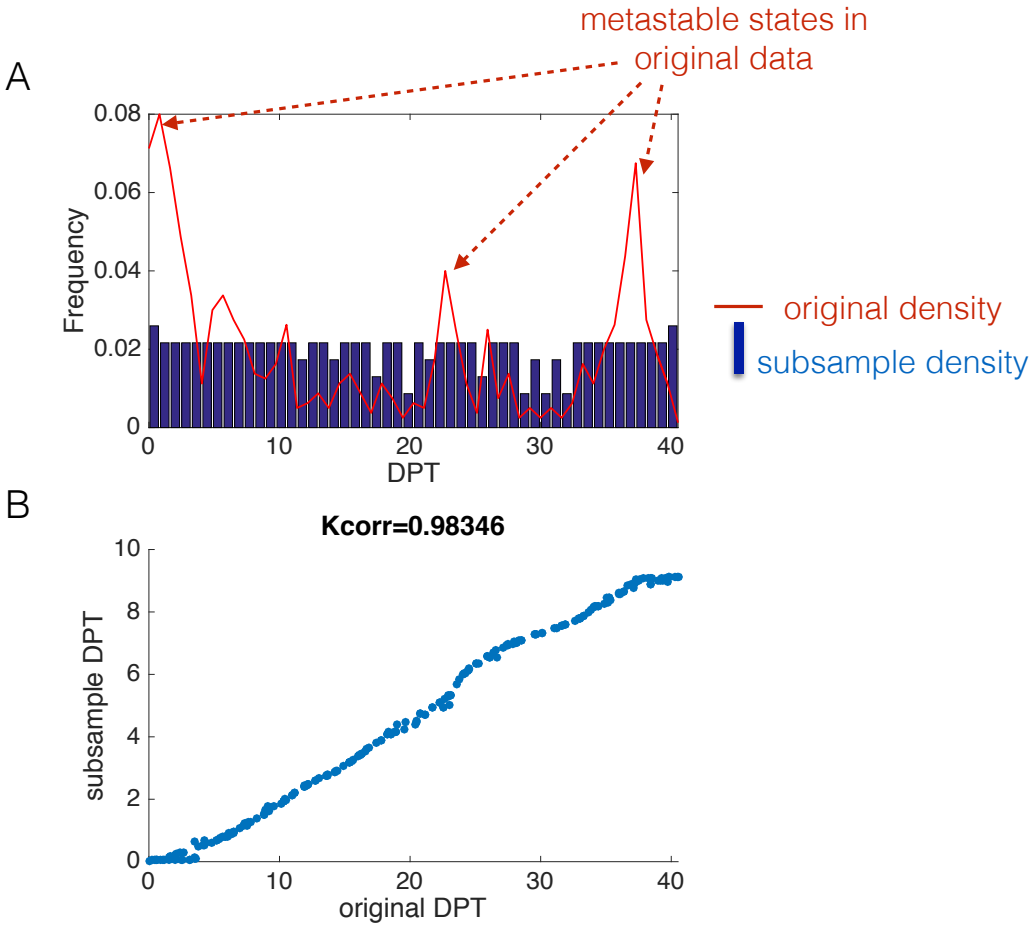
Furthermore we present self concordance of each method's pseudotime ordering over 100 bootstrap sets of data as a measure of robustness of performance. DPT generally shows a higher mean and smaller variance of self concordance across several data sets (see figure 2e in the main text).

7.5.4 Branch finding approach

If a reliable and robust measure of on manifold distance is in hand, to distinguish whether cells lie on a specific path or lie off the path reduces to a simple triangle inequality question (see Note 1.3). However for a scattered manifold as we get from single-cell differentiation data, one should set a threshold for separating the in-road from off-road cells. To avoid such selection of threshold and to be able to find the branches automatically without manual supervision, in DPT we use the correlation versus anti-correlation relations of dpt distances from the branch tips for branch finding. This can be viewed as a milder version of the triangle inequality which does not need defining a threshold (see Note 1.3). Monocle and Wishbone however, in lack of such robust on-manifold distance measure rely on less robust and more randomly affected (either by data sampling structure or random components in the algorithm itself) ways of finding a branch which is PQ-tree [21] and disagreements of distances to the landmarks correspondingly.

7.5.5 Dealing with metastable cell states

There is barely any biological meaning to ordering cells in metastable states. Although in actual time dynamics there is such an order even in the metastable states for a single cell trajectory, the order in a metastable state can be shuffled in another differentiating single cell's trajectory. That is, all the



Supplementary Figure N13: A) Sampling density over diffusion pseudotime for artificial data (Supplementary Note 7.1) for original data (red) and an (almost) uniformly subsampled data set. B) Diffusion pseudotime computed for the original data strongly correlates with diffusion pseudotime of subsampled data (Kendal-tau correlation=0.983) in spite of strongly different density distributions.

states belonging to the same metastable state will have almost the same universal time value. Thus we suggest that pseudotime should be measured as the distance to the initial cell on a reconstruction of differentiation manifold. That is no matter how long in actual time a single cell trajectory is trapped in one of the metastable states, there is almost no progression on neither the original manifold ($C \subset \mathbb{R}^G$) or any mapping of it (C'). Hence one should expect almost the same pseudotime for cells in a metastable state.

Monocle however provides an absolute order of cells based on the constructed MST, which is less reliable in metastable states.

For non-branching data, Wanderlust provides a valid pseudotime measured on the manifold by adding up Euclidean distances of neighboring cells for a sampled path that connect each cell to the initial cell. Wishbone's pseudotime correspondence to an on-manifold distance however, relies on how well it has performed on branch identification level as it refines the pseudotime to its non-branching version (from Wanderlust) after branch identification. Thus if the process of branch finding goes successful in Wishbone, one can assure that it is providing proper pseudotime in metastable states as well.

DPT measures pseudo-time on a mapped manifold such that cells from the same metastable state

are correctly places on the same neighborhood on the manifold and a have a similar distance relative to the root cell.

7.5.6 Applicability to large cell numbers and run time

Monocle fails to find the MST solution (generates an error message) when applied on a large body of cells ($> 10^3$) that do not show a simple structure in the ICA embedding. This tends to happen more often if a very large set of genes are used for ordering. ICA runs in computation time of order Gn and MST's run time scales with the number of edges, thus $O(n^2)$. Thus the total computation time for monocle is $O(n(G+n))$.

With appropriate choice of the tuning parameters Wanderlust preforms well for large cell numbers. The computational time for finding nearest neighbors is $O(\log(n))$. Thus wanderlust performs in $O(\log(n) + n \times \text{num.landmarks})$. Wishbone first calculates the diffusion map and performs Wanderlust on its reduced dimensions. Several more steps also integrated in the Wishbone algorithm thus total performance time is at least $O(nk^2) + O(\log(n) + n \times \text{num.landmarks})$, k being the number of nearest neighbors used in diffusion maps.
The computational time needed for DPT using the complete transition matrix is $O(Gn^2)$. However the transition matrix T is sparse, most entries are close to zero. Therefore constructing the transition matrix on k nearest neighbours graph is very efficient, which we provide as an option in the package to be used with large cell numbers. In the next step, sparse matrix diagonalization of the transition matrix needs a computation time of $O(nk^2)$. Inversion of the transition matrix is then trivial. Further steps only require $O(nk)$.

7.5.7 Allowing for multiple root cells

The heterogeneity of differentiating cell populations can also hold for the pluripotent state. Very often all the pluripotent cells reside on a close neighborhood on the differentiation manifold. In this case even with a single chosen root cell, the pseudotime measured on the manifold would automatically account for the metastable state of pluripotency. However one could expect larger variability among the pluripotent cells or even existence of several sources for pluripotent cells which could consequently take different probabilities for paths of development depending on the state of the root cell. DPT thus provides the possibility of choosing multiple root cells. Neither Monocle or Wishbone provide this feature.

7.5.8 Data embedding and visualization

Monocle visualizes the ICA embedding of data, which is the same reduced dimension space that monocle perform pseudotime ordering . Wanderlust does not provide any specific visualization of the data manifold as the assumption is dealing with one dimensional (i.e. nonbranching) data manifold. Wishbone handles branching and uses t-SNE for data visualization. However this visualization is not intrinsically related to the underlying pseudotime measure used in Wishbone which is Wanderlust. DPT measures pseudotime on a space that shares the same eigenvectors with diffusion maps except for the non-informative (zeroth) eigenvector (see Equation 13). Consistently, the non-informative eigenvector is left out in diffusion maps embedding as well. Thus diffusion map embedding provides a consistent reduced dimension visualization for DPT.

algorithm	reference	methodology	tuning parameters	computation time	handles branching lineages	scalability to large sample numbers	robustness to noise/ density heterogeneity	allows several root (pluripotent) cells	pseudotime is measured on full data dimensions	provides data embedding consistent with its pseudotime measure	handles missing/ uncertain values
Monocle	Trapnell et al[13]	MST on ICA embedding	number of ICA embedding dimensions, root cell	$\mathcal{O}(n^2 + G \cdot n)$	+	−	−	−	−	+	−
Wishbone	Setty et al [15]	sampling paths on nearest neighbors graphs/ distance disagreements for branch finding	k, l, num_graphs (n_g), num_landmarks (n_l), num.diffusion map components, root cell	$\mathcal{O}(nk^2)$+ $\mathcal{O}(n \cdot n_l)$	+	+	+	−	−	−	−
DPT		analytic graph path integral	diffusion parameter (σ or κ), root cell	$\mathcal{O}(nk^2)$	+	+	+	+	+	+	+

Supplementary Table N1: Comparison of several single-cell pseudotime ordering algorithms.

References

[1] L. Haghverdi, F. Buettner, and F. J. Theis, "Diffusion maps for high-dimensional single-cell analysis of differentiation data," *Bioinformatics*, vol. 31, pp. 2989–98, 5 2015.

[2] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[3] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, p. 395, Nov. 2007.

[4] C. D. Meyer, Jr., "The role of the group generalized inverse in the theory of finite markov chains," *SIAM Review*, vol. 17, pp. 443–464, 7 1975.

[5] F. Fouss, A. Pirotte, and M. Saerens, "The application of new concepts of dissimilarities between nodes of a graph to collaborative filtering," *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, p. 550—556, 1 2004.

[6] U. Von Luxburg, A. Radl, and M. Hein, "Hitting and commute times in large random neighborhood graphs.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1751–1798, 2014.

[7] V. Moignard, S. Woodhouse, L. Haghverdi, A. J. Lilly, Y. Tanaka, A. C. Wilkinson, F. Buettner, I. C. Macaulay, W. Jawaid, E. Diamanti, S.-I. Nishikawa, N. Piterman, V. Kouskoff, F. J. Theis, J. Fisher, and B. Göttgens, "Decoding the regulatory network of early blood development from single-cell gene expression measurements," *Nature Biotechnology*, vol. 33, no. 3, pp. 269–76, 2015.

[8] G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, P. S. Linsley, and R. Gottardo, "MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data," *Genome Biology*, vol. 16, no. 1, p. 278, 2015.

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

[10] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells," *Nature Biotechnology*, vol. 33, pp. 155–160, 1 2015.

[11] P. Brennecke, S. Anders, J. K. Kim, A. A. Kołodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, and M. G. Heisler, "Accounting for technical noise in single-cell RNA-seq experiments.," *Nature methods*, vol. 10, pp. 1093–5, 11 2013.

[12] F. Paul, Y. Arkin, A. Giladi, D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, D. Winter, D. Lara-Astiaso, M. Gury, A. Weiner, *et al.*, "Transcriptional heterogeneity and lineage commitment in myeloid progenitors," *Cell*, vol. 163, no. 7, pp. 1663–1677, 2015.

[13] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.," *Nature biotechnology*, vol. 32, pp. 381–6, 4 2014.

[14] S. C. Bendall, K. L. Davis, E.-A. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, and D. Pe'er, "Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development.," *Cell*, vol. 157, no. 3, pp. 714–725, 2014.

[15] M. Setty, M. D. Tadmor, S. Reich-Zeliger, O. Angel, T. M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, and D. Pe'er, "Wishbone identifies bifurcating developmental trajectories from single-cell data," *Nature Biotechnology*, 2016.

[16] K. Campbell, C. P. Ponting, and C. Webber, "Laplacian eigenmaps and principal curves for high resolution pseudotemporal ordering of single-cell RNA-seq profiles," *bioRxiv doi: 10.1101/027219*, 9 2015.

[17] K. Campbell, C. Yau, C. P. Ponting, and C. Webber, "Bayesian gaussian process latent variable models for pseudotime inference in single-cell RNA-seq data," *bioRxiv doi: 10.1101/026872*, 9 2015.

[18] R. Kafri, J. Levy, M. B. Ginzberg, S. Oh, G. Lahav, and M. W. Kirschner, "Dynamics extracted from fixed cells reveal feedback linking cell growth to cell cycle.," *Nature*, vol. 494, no. 7438, pp. 480–483, 2013.

[19] G. Gut, M. D. Tadmor, D. Pe'er, L. Pelkmans, and P. Liberali, "Trajectories of cell-cycle progression from fixed cell populations.," *Nature methods*, vol. 12, no. 10, pp. 951–954, 2015.

[20] D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, *et al.*, "Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types," *Science*, vol. 343, no. 6172, pp. 776–779, 2014.

[21] K. S. Booth and G. S. Lueker, "Testing for the consecutive ones property, interval graphs, and graph planarity using pq-tree algorithms," *Journal of Computer and System Sciences*, vol. 13, no. 3, pp. 335–379, 1976.

Supplementary Table 1: Differential expression results in precursor vs. decision state and PS vs HF cells

	decision state vs. precursor state		HF vs. PS	
Gene name	lfc	p_{adj}	lfc	p_{adj}
Cbfa2t3h	−9.60	$1.78 \cdot 10^{-146}$	−4.26	$8.50 \cdot 10^{-61}$
Cdh1	14.53	$5.07 \cdot 10^{-219}$	5.15	$3.05 \cdot 10^{-58}$
Cdh5	−15.63	$3.28 \cdot 10^{-195}$	−8.10	$2.66 \cdot 10^{-119}$
Egfl7	−7.47	$2.66 \cdot 10^{-234}$	−3.19	$4.21 \cdot 10^{-131}$
Erg	−9.77	$3.31 \cdot 10^{-151}$	−5.66	$1.32 \cdot 10^{-112}$
Ets1	−9.82	$3.95 \cdot 10^{-178}$	−3.18	$1.12 \cdot 10^{-57}$
Ets2	−2.77	$5.66 \cdot 10^{-123}$	−1.89	$3.26 \cdot 10^{-127}$
Etv2	−12.21	$1.44 \cdot 10^{-187}$	−4.81	$5.00 \cdot 10^{-71}$
Etv6	−5.10	$4.33 \cdot 10^{-133}$	−2.36	$1.12 \cdot 10^{-126}$
Fli1	−16.36	$4.61 \cdot 10^{-273}$	−6.99	$6.59 \cdot 10^{-118}$
FoxO4†	1.79	$2.01 \cdot 10^{-46}$	0.60	$4.19 \cdot 10^{-17}$
Hhex	−8.53	$4.82 \cdot 10^{-78}$	−3.41	$1.84 \cdot 10^{-38}$
HoxB4	−6.78	$1.60 \cdot 10^{-68}$	−4.10	$5.27 \cdot 10^{-56}$
Ikaros	−1.95	$8.51 \cdot 10^{-16}$	−1.40	$2.67 \cdot 10^{-8}$
Itga2b	−5.46	$5.85 \cdot 10^{-62}$	−4.34	$1.07 \cdot 10^{-70}$
Kdr	−14.78	$8.21 \cdot 10^{-299}$	−5.29	$3.77 \cdot 10^{-92}$
Kit	−5.97	$8.95 \cdot 10^{-80}$	−4.08	$1.07 \cdot 10^{-120}$
Ldb1†	−1.52	$8.86 \cdot 10^{-8}$	−0.39	$2.58 \cdot 10^{-6}$
Lmo2	−6.46	$1.45 \cdot 10^{-98}$	−2.39	$6.10 \cdot 10^{-23}$
Lyl1	−11.43	$5.32 \cdot 10^{-144}$	−5.42	$1.52 \cdot 10^{-78}$
Mecom	−2.61	$1.83 \cdot 10^{-10}$	−1.91	$5.75 \cdot 10^{-22}$
Meis1	−4.91	$4.26 \cdot 10^{-74}$	−4.05	$2.04 \cdot 10^{-122}$
Mitf†	1.23	$3.20 \cdot 10^{-3}$	0.36	1
Myb	−2.59	$1.15 \cdot 10^{-13}$	−1.55	$1.19 \cdot 10^{-7}$
Notch1	−7.23	$3.76 \cdot 10^{-137}$	−3.13	$1.18 \cdot 10^{-73}$
Pecam1	−11.86	$4.72 \cdot 10^{-185}$	−6.35	$8.32 \cdot 10^{-151}$
Procr	−3.99	$3.83 \cdot 10^{-31}$	−2.16	$2.26 \cdot 10^{-17}$
Sfpi1	−3.30	$2.53 \cdot 10^{-34}$	−2.33	$1.59 \cdot 10^{-20}$
Sox7	−10.75	$1.72 \cdot 10^{-190}$	−4.17	$1.05 \cdot 10^{-65}$
Sox17	−3.09	$8.58 \cdot 10^{-13}$	−2.63	$4.41 \cdot 10^{-24}$
Tal1	−13.80	$4.73 \cdot 10^{-194}$	−5.66	$3.90 \cdot 10^{-76}$
Tbx20	−4.14	$5.58 \cdot 10^{-44}$	1.51	$7.45 \cdot 10^{-9}$

Results of the likelihood-ratio test for the gene expression levels in two different states (decision state and precursor state) and cell types (head fold and primitive streak), respectively. We considered a gene expression significant, if the absolute log fold-change was above 1 and the Bonferroni adjusted p-value was below 0.01. † indicates genes being significantly differential in two metastable state comparison but not in cell type comparison.

Supplementary Table 2: Differential expression results in terminal branch 2 vs. decision state and 4SG- vs HF cells

	decision state vs. terminal branch 2		HF vs. 4SG-	
Gene name	lfc	p_{adj}	lfc	p_{adj}
Cbfa2t3h*	0.54	0.54	−2.50	$4.15 \cdot 10^{-22}$
Cdh5	1.36	$3.46 \cdot 10^{-17}$	−2.66	$1.36 \cdot 10^{-12}$
Erg	1.43	$8.32 \cdot 10^{-13}$	−1.92	$9.36 \cdot 10^{-21}$
Etv2†	2.16	$1.19 \cdot 10^{-17}$	−0.45	0.76
FoxH1*	−0.97	$1.23 \cdot 10^{-5}$	−2.17	$2.97 \cdot 10^{-117}$
Gfi1b*	0.28	1	−2.65	$2.35 \cdot 10^{-40}$
Ikaros*	−0.42	1	−4.02	$1.99 \cdot 10^{-87}$
Itga2b†	2.47	$2.69 \cdot 10^{-12}$	−1.5	$3.36 \cdot 10^{-13}$
Mecom †	2.77	$2.43 \cdot 10^{-12}$	0.77	$5.61 \cdot 10^{-15}$
Meis1	1.13	$7.41 \cdot 10^{-8}$	2.40	$2.06 \cdot 10^{-20}$
Myb	−1.53	$2.50 \cdot 10^{-3}$	−3.29	$5.28 \cdot 10^{-42}$
Runx1*	−0.11	1	−4.20	$2.50 \cdot 10^{-94}$
Sfpi1	1.75	$1.17 \cdot 10^{-6}$	−2.65	$1.02 \cdot 10^{-26}$

Results of the likelihood-ratio test for the gene expression levels in two different states (decision state and terminal branch 2) and cell types (head fold and 4SG- cells), respectively. We considered a gene expression significant, if the absolute log fold-change was above 1 and the Bonferroni adjusted p-value was below 0.01. † indicates genes being significantly differential in two metastable state comparison but not in cell type comparison. $*$ indicates genes being significantly differential in cell type comparison, but not in metastable state comparison.