Bioinformatics, 31(18), 2015, 2989–2998
doi: 10.1093/bioinformatics/btv325
Advance Access Publication Date: 21 May 2015
Original Paper

OXFORD

Gene expression

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 and [2]Department of Mathematics, Technische Universität München 85748 Garching, Germany

*To whom correspondence should be addressed.
†Present address: European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK
Associate Editor: Ziv Bar-Joseph

Received on July 31, 2014; revised on March 25, 2015; accepted on May 18, 2015

Abstract

Motivation: Single-cell technologies have recently gained popularity in cellular differentiation studies regarding their ability to resolve potential heterogeneities in cell populations. Analyzing such high-dimensional single-cell data has its own statistical and computational challenges. Popular multivariate approaches are based on data normalization, 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 pseudotemporal 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 datasets (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 and are advantageous over other techniques for non-linear dimension reduction such as t-distributed Stochastic Neighbour Embedding for preserving the global structures and pseudotemporal ordering of cells.
Availability and implementation: 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
Supplementary information: Supplementary data are available at *Bioinformatics* online.

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 synchronizing cells is experimentally challenging. Moreover, stem cell populations that have been functionally characterized often show

heterogeneity in their cellular and molecular properties (Dykstra *et al.*, 2007; Huang, 2009; 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 (Bandura *et al.*, 2009; Chattopadhyay *et al.*, 2006), 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 non-linear continuous process (Bendall *et al.*, 2014; Buettner and Theis, 2012), 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 subpopulations, they usually do not preserve the continuous trajectories of differentiation data. A more recently proposed algorithm Wanderlust (Bendall *et al.*, 2014) incorporates the non-linearity and continuity concepts but provides a pseudotemporal ordering of cells only if the data comprise a single branch. Furthermore, in gene expression measurement techniques, there 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 analyzing 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 non-linear structure of data as a continuum and are robust to noise. Furthermore, with density normalization, diffusion maps are resistant to sampling density heterogeneities and can capture rare as well as abundant populations. As a non-linear 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 datasets 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 Ω. 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 normalization 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 Ω as follows:

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

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

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

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

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

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:

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

Each row of $\tilde{P}^t$ can be viewed as a vector, which we represent as $p^t(\mathbf{x},\cdot)$ 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:

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

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

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

The corresponding eigenvector to the largest eigenvalue λ_0 is a constant vector $\psi_0 = \mathbb{1}$. 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 visualization we select these eigenvectors and instead of the mathematical notation ψ, 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 (non-linear) 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 visualization 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 (Buettner *et al.*, 2014; McDavid *et al.*, 2013). 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-normalized 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 $M_1 + \sigma$ normalized 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-normalization constraint):

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

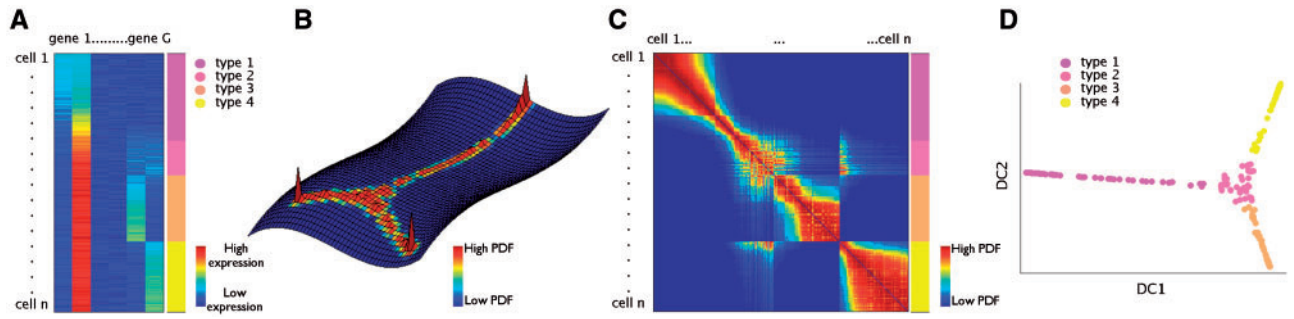

Fig. 1. Schematic overview of diffusion maps embedding. (**A**) The $n \times G$ matrix representation of single-cell data consisting of four different cell types. The last column on the right side of the matrix (colour band) indicates the cell type for each cell. (**B**) Representation of each cell 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

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

$$\int_{-\infty}^{\infty} Y_x(x'_g)Y_y(x'_g)\mathrm{d}x'_g = \int_{M_0-\sigma}^{M_1+\sigma} \frac{1}{\sqrt{M_1-M_0+2\sigma}}\left(\frac{2}{\pi\sigma^2}\right)^{1/4}\exp\left(-\frac{(x\prime_g-x_g)^2}{\sigma^2}\right)\mathrm{d}x'_g = \frac{1}{\sqrt{M_1-M_0+2\sigma}}\left(\frac{\pi\sigma^2}{8}\right)^{1/4} \cdot\left(\mathrm{erfc}\left(\frac{M_0-\sigma-x_g}{\sigma}\right)-\mathrm{erfc}\left(\frac{M_1+\sigma-x_g}{\sigma}\right)\right).$$

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. Supplementary Figure S1 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 visualize 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-normalized 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 $\mathbf{x} \in \Omega$, 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:

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

where $d(\boldsymbol{x},\sigma)$ 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. Supplementary Figures S2 and S3 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 dataset 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 Figure 2B and C (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 (Supplementary Figs S5A 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. hematopoietic) stem cells in natural organisms.

We sampled cells from this toy model in two different sets, a balanced toy dataset, wherein 600 samples serve as a snapshot of the steady state of the system with no additional extrinsic noise, and an

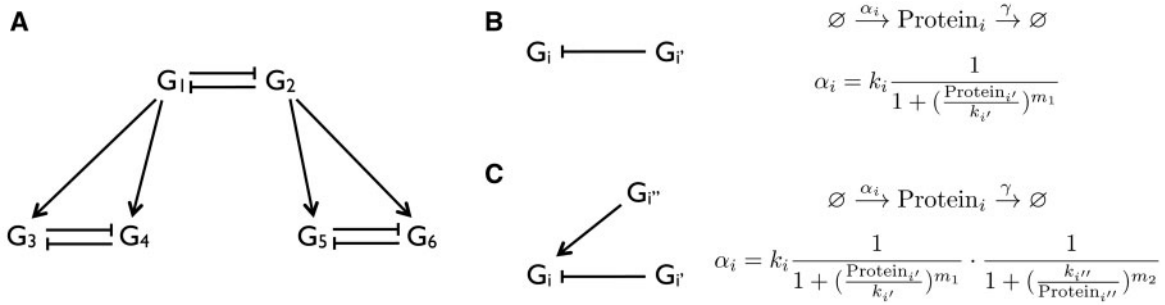

Fig. 2. (**A**) Toy model of a differentiation regulatory network consisting of three pairs of antagonistic genes simulated by the Gillespie algorithm. The arrows show activation or inhibition interactions between genes. The toy model employs two classes of gene regulation: (**B**) G_i is connected to an inhibitor, its production rate α_i is proportional to a Hill function of the concentration of the inhibitor $\text{Protein}_{i'}$ (**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

imbalanced toy dataset, 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 (Supplementary Figs S5C 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 dataset 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 Fluorescence-activated cell sorting (FACS) sorting. Single-cell qPCR expression level measurement was then performed for 24 genes. Housekeeping genes were only used for cell-cycle normalization, 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 analyzed 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 normalized 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 dataset 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 datasets 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 datasets. 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 visualization (data visualization at the first maximum is provided in the Supplementary Fig. 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 visualization.

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 dataset. The eigenvalues of the diffusion map (Fig. 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 dataset 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. Visualization 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 (Fig. 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 visualization and represent the two rarer branches as smaller. For additional t-SNE visualizations with non-standard perplexity values for the imbalanced toy data see Supplementary Figure S10.

3.1.3 Refinement of the transition matrix by density normalization, 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 normalization (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 normalization. Supplementary Figure S12 demonstrates how density normalization 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.

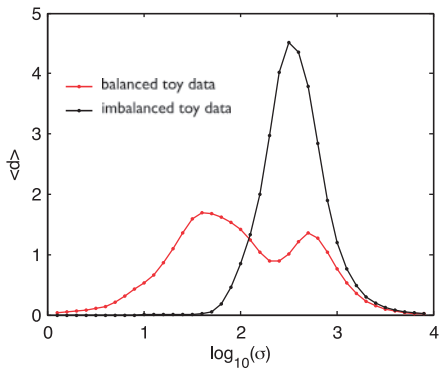

Fig. 3. The average dimensionality of the data $\langle d \rangle$ as a function of $\log_{10}(\sigma)$ for the balanced and imbalanced toy datasets

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 (Fig. 6A) indicates a major 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 (Fig. 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 (Fig. 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 visualization 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 E. 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 characterizing noise for this dataset. However, what makes the diffusion map embedding of this dataset more plausible is the concordance between the branching structure as suggested by the diffusion map and the recently established hierarchy of haematopoietic cell types (Arinobu *et al.*, 2007; Moignard *et al.*, 2013) 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 visualization using the first three eigenvectors indicated a branching at the early

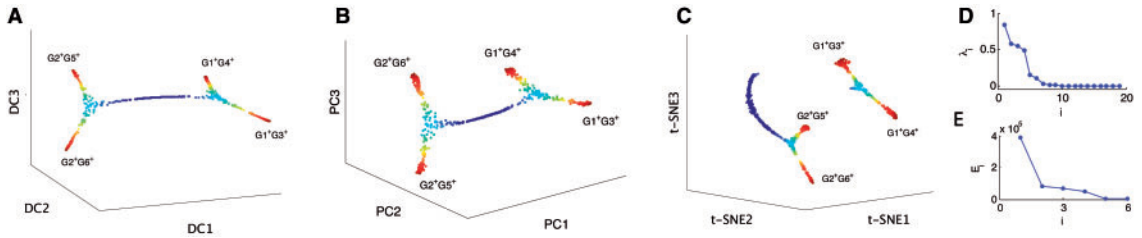

Fig. 4. Visualization of the balanced toy data on (**A**) the first three eigenvectors of the diffusion map, (**B**) PCA and (**C**) t-SNE. The colours (heat map of blue to red) indicate the maximum expression among all genes. Eigenvalues sorted in decreasing order for (**D**) diffusion map and (**E**)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

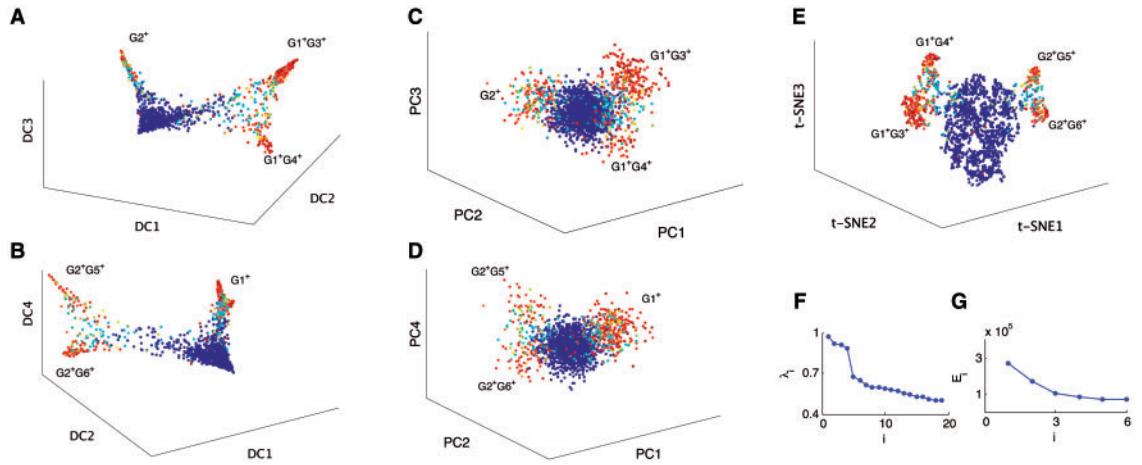

Fig. 5. Visualization of the imbalanced toy data on (**A**) the first three eigenvectors of the diffusion map, (**B**) the first, second and fourth eigenvectors of the diffusion map, (**C**) the first three components of the PCA (**D**) the first, second and fourth components of PCA and (**E**) t-SNE. 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

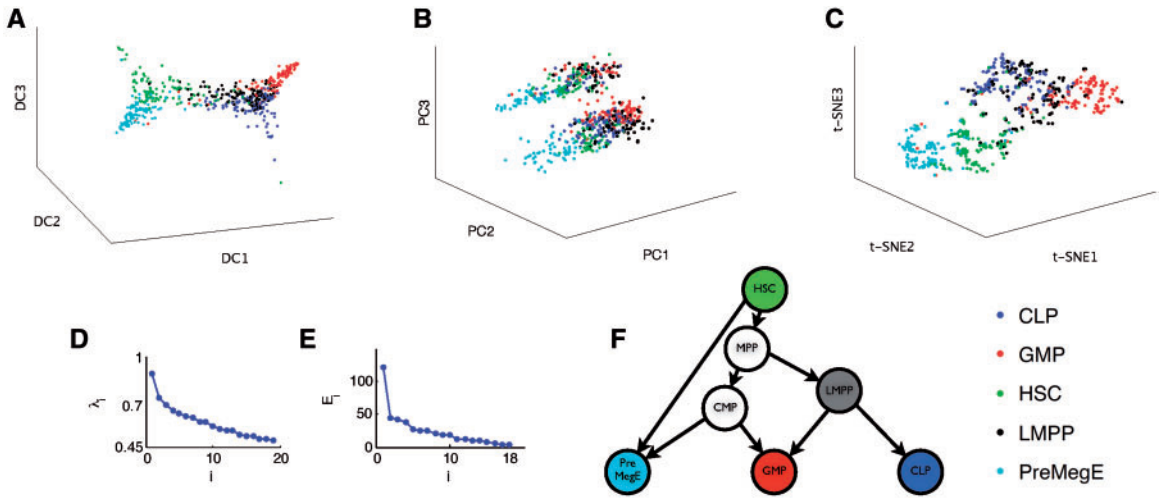

Fig. 6. Visualization of haematopoietic stem cells data on the first three eigenvectors of (**A**) diffusion map, (**B**) PCA and (**C**) t-SNE. Eigenvalues sorted in a decreasing order for (**D**) diffusion map and (**E**) PCA. (**F**) The hierarchy of haematopoietic cell types

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 (Fig. 7A). The branching structure is unclear in the PCA and t-SNE plots (Figs 7B and C). 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 dataset (Buettner and Theis, 2012; Guo *et al.*, 2010), 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 (Fig. 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 dataset is comparable (although slightly sharper with respect to pseudotime ordering) to the other two methods, PCA and t-SNE (Fig. 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 dataset and a few experimental datasets. Using a biologically relevant distance metric (i.e. diffusion distance), the adapted diffusion map method outperforms other dimension-reduction methods in pseudotemporal 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.

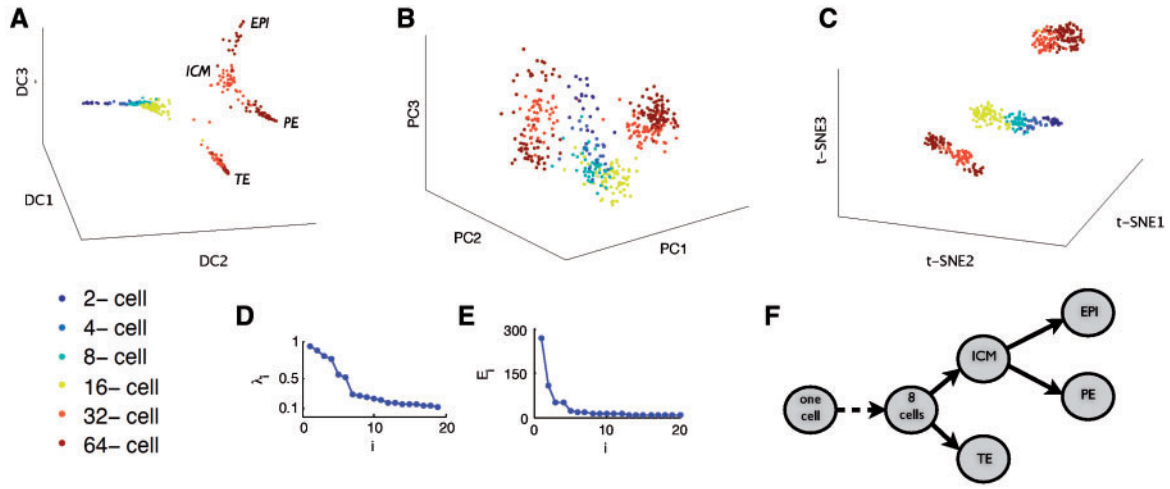

Fig. 7 Visualization of mouse embryonic stem cells on (**A**) the first three eigenvectors of diffusion map, (**B**) PCA and (**C**) t-SNE. Eigenvalues sorted in a decreasing order for (**D**) diffusion map and (**E**) PCA. (**F**) The hierarchy of cells for mouse embryonic stem cells

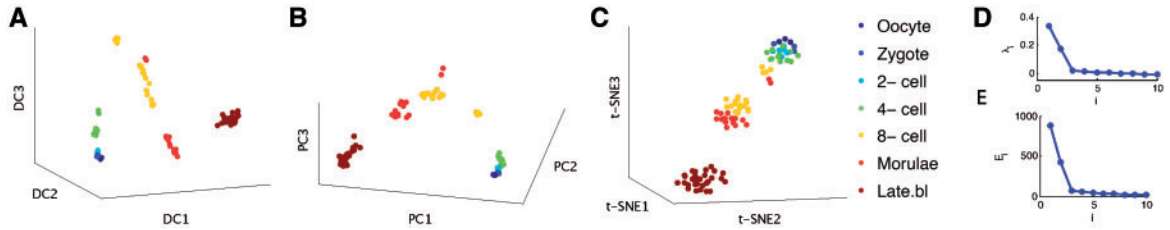

Fig. 8. Visualization of human preimplantation embryos data on (**A**) the first three eigenvectors of the diffusion map, (**B**) PCA and (**C**) t-SNE. Eigenvalues sorted in a decreasing order for (**D**) the diffusion map and (**E**) PCA

Among these methods, isomap and (H)LLE have not been applied for the analysis of single-cell differentiation data and pseudotime ordering so far, mainly because they do not meet the specific requirements for the analysis of such data including capability to handle high levels of technical noise, sampling density heterogeneities, detection limits and missing values. Supplementary 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 dataset as well as the three experimental single-cell datasets. For any dataset, 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 dataset.

In our diffusion maps implementation, by performing density normalization 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 Figs 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.

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 normalization) will not be symmetric any more and we cannot ensure real eigenvalues for the transition probability matrix. However, as long as the graph is

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

	Reference	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 $\gtrsim O(2d^a)$
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	Continuous 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	Continuous data manifold, presence of noise and density heterogeneity

[a]*d* is the intrinsic dimensionality of the data manifold.

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 visualizable dimensions two or three. In contrast, for some other methods such as t-SNE, one can pre-determine the number of visualization dimensions for the embedding optimization 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 organize single cells along the non-linear and complex branches of differentiation, maintain the global structure of the differentiation dynamics and detect rare populations as well.

Acknowledgements

We 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).

Conflict of Interest: none declared.

References

Amir,E.-a. D. *et al.* (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.*, **31**, 545–552.

Arinobu,Y. *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**, 416–427.

Bandura,D.R. *et al.* (2009). Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.*, **81**, 6813–6822.

Bendall,S.C. *et al.* (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, **157**, 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**, i626–i632.

Buettner,F. *et al.* (2014). Probabilistic PCA of censored data: accounting for uncertainties in the visualisation of high-throughput single-cell qPCR data. *Bioinformatics*.

Buganim,Y. *et al.* (2012). Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell*, **150**, 1209–1222.

Chattopadhyay,P.K. *et al.* (2006). Quantum dot semiconductor nanocrystals for immunophenotyping by polychromatic flow cytometry. *Nat. Med.*, **12**, 972–977.

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

Coifman,R.R. *et al.* (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci. USA*, **102**, 7426–7431.

Donoho,D.L. and Grimes,C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. USA*, **100**, 5591–5596.

Dykstra,B. *et al.* (2007). Long-term propagation of distinct hematopoietic differentiation programs in vivo. *Cell Stem Cell*, **1**, 218–229.

Gillespie,D.T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, **81**, 2340–2361.

Guo,G. *et al.* (2010). Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell*, **18**, 675–685.

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

Krumsiek,J. *et al.* (2011). Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network. *PloS One*, **6**, e22649.

McDavid,A. *et al.* (2013). Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics*, **29**, 461–467.

Mohri,M. *et al.* (2012). *Foundations of Machine Learning*. MIT Press, Cambridge, MA, USA.

Moignard,V. *et al.* (2013). Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nature Cell Biol.*, **15**, 363–372.

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

Park,H.Y. *et al.* (2014). Visualization of dynamics of single endogenous mrna labeled in live mouse. *Science*, **343**, 422–424.

Qiu,P. *et al.* (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.*, **29**, 886–891.

Rieger,M.A. *et al.* (2009). Hematopoietic cytokines can instruct lineage choice. *Science*, **325**, 217–218.

Schölkopf,B. *et al.* (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, **10**, 1299–1319.

Schroeder,T. (2011). Long-term single-cell imaging of mammalian stem cells. *Nat. Methods*, **8**, S30–S35.

Shawe-Taylor,J. and Cristianini,N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Stingl,J. *et al.* (2006). Purification and unique properties of mammary epithelial stem cells. *Nature*, **439**, 993–997.

Strasser,M. *et al.* (2012). Stability and multiattractor dynamics of a toggle switch based on a two-stage model of stochastic gene expression. *Biophys. J.*, **102**, 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**, 663–676.

Tenenbaum,J.B. *et al.* (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319–2323.

Trapnell,C. *et al.* (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnol.*, **32**, 381–386.

Van der Maaten,L. and Hinton,G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

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

Yan,L. *et al.* (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.*, **20**, 1131–1139.