GPseudoClust: deconvolution of shared pseudo-trajectories at single-cell resolution

Magdalena E Straußᵃ[1,*], Paul DW Kirk[2], John E Reid[2], and Lorenz Wernisch[2]

[1] *Wellcome Sanger Institute, Hinxton, UK*
[2] *MRC Biostatistics Unit, University of Cambridge, Cambridge, UK*
[*] *Corresponding author: ms58@sanger.ac.uk*

3rd March 2019

Abstract

Motivation: A large number of methods have been developed to cluster genes on the basis of their changes in mRNA expression over time, using bulk RNA-seq or microarray data. These methods cannot be directly applied to single-cell data, since the temporal order of the cells is unknown. One way to address this challenge is to first use pseudotime methods to order the cells, and then apply standard clustering techniques. However, pseudotime estimates are subject to high levels of uncertainty, and failing to account for this uncertainty is liable to lead to erroneous and/or over-confident gene clusters.

Results: The proposed method, GPseudoClust, is a novel approach that both clusters genes for pseudotemporally ordered data and quantifies the uncertainty in cluster allocations arising from the uncertainty in the pseudotime ordering. GPseudoClust combines a recent method for pseudotime inference with nonparametric Bayesian clustering methods, efficient MCMC sampling, and novel subsampling strategies. For branching data, GPseudoClust identifies differences in dynamic patterns for different branches. In an application to stimulated dendritic cells, we show that it categorises genes in a way consistent with known biological function. Furthermore, it integrates data from different cell lines, batches or experimental protocols in a principled way.

Availability: An implementation is available on GitHub:
https://github.com/magStra/nonparametricSummaryPSM and
https://github.com/magStra/GPseudoClust.

Contact: ms58@sanger.ac.uk

1 Introduction

During response to stimulations or development, gene expression undergoes significant changes for many genes. For bulk-measurements of gene expression these changes can be understood in terms of a trajectory mapping time points to expression measurements. Genes can be clustered in terms of the similarities of their trajectories. Eisen *et al.* (1998) found that similar expression dynamics of genes are related to biological function. Clustering genes together with similar changes in expression over time can identify genes likely to be co-regulated by the same transcription factors (Cooke *et al.*, 2011). McDowell *et al.* (2018) emphasise that using clustering to identify shared response types helps reduce the complexity of the response, and allows the exploration of regulatory mechanisms underlying the shared response types.

However, the methods proposed in the publications above were developed for bulk-measurements of gene expression, and not for single-cell data, and there is a need for effective clustering algorithms for genes for single-cell data as well, given that single-cell technologies have enabled us to obtain response and developmental trajectories with a much better resolution; see, for example, Griffiths *et al.* (2018); Kunz *et al.* (2018); Nestorowa *et al.* (2016). Single-cell RNA-seq data often follow processes of development, differentiation or immune response, and the order of cells in terms of their progression can be inferred using pseudotemporal ordering, see Ahmed *et al.* (2018); Campbell and Yau (2016); Haghverdi *et al.* (2016); Ji and Ji (2016); Qiu *et al.* (2017); Reid and Wernisch (2016); Strauß *et al.* (2018); Welch *et al.* (2016) among many others. For each gene the ordered gene expression measurements are noisy observations of an underlying latent trajectory characterising the response of the gene to a stimulant or the dynamics of its expression during development. Importantly, the inferred latent trajectory depends on the pseudotime ordering of the cells. Single-cell data are also characterised by higher levels of noise, including dropout effects; see, among others, Stegle *et al.* (2015); Vallejos *et al.* (2015). In addition, the number of cells in single-cell data sets typically exceeds by orders of magnitude that of time points for bulk measurements.

A number of algorithms have been developed specifically for the clustering of *cells* for scRNA-seq data, for instance Kiselev *et al.* (2017); Lin *et al.* (2017) and Wang *et al.* (2017), the latter method using multiple kernel learning. There has been far less progress on the development of clustering methods for *genes* for this specific type of data. Commonly used general clustering algorithms, including mixtures of Normals (e.g. mclust, Fraley and Raftery, 2002; Scrucca *et al.*, 2017), k-medoids clustering (PAM, Kaufman and Rousseeuw, 2008) as implemented in the *cluster* R package (Maechler *et al.*, 2017)), and hierarchical clustering, all fail to account for the pseudotemporal nature of the data.

One way of clustering pseudotemporal single-cell gene trajectories is a two-step approach (Macaulay *et al.*, 2016): first use a pseudotime ordering method such as SLICER (Welch *et al.*, 2016) or DeLorean (Reid and Wernisch, 2016); then cluster genes using a method for time-stamped bulk data, such as GPClust (Hensman, 2013; Hensman *et al.*, 2015). The two-step approach is unable to integrate the uncertainty of inferred pseudotimes into the modelling of cluster structures. In contrast, the method proposed here, *GPseudoClust*, samples from a full posterior distribution of cluster allocations, which depends on a posterior distribution of pseudotime orders sampled jointly with the cluster allocations. A two-step approach is also implemented in Monocle 2 (Qiu *et al.*, 2017), which uses PAM on a distance measure between smoothed pseudotime trajectories.

The shapes of the trajectories and their uncertainties change depending on the order of the cells. GPseudoClust addresses this additional challenge by modelling the orders of cells and cluster allocations of genes jointly, thereby accounting for dependencies between the orders of the cells and cluster allocations of the genes. We previously developed a method to capture the uncertainty of pseudotime (Strauß *et al.*, 2018), which is now combined with Bayesian clustering using Dirichlet process mixtures of hierarchical GPs (Hensman, 2013; Hensman *et al.*, 2015).

2 System and methods

2.1 Cell orderings and pseudotime

We assume we have preprocessed log-transformed gene expression data in the form $\mathbf{y}_j$ of gene $j = 1, \ldots, n_g$, where $\mathbf{y}_j$ is a vector of length T, the number of cells. We start with a vector of pseudotime

points $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_T)$ and and define an ordering of cells as a permutation $\mathbf{o} = (o_1, \ldots, o_T)$, $o_i \in \{1, \ldots, T\}$, $o_i \neq o_j$ for $i \neq j$, where o_i is the index of the cell assigned to pseudotime τ_i in the ordering.

Orders $\mathbf{o} = (o_1, \ldots, o_T)$ are mapped to pseudotimes $\boldsymbol{\tau}(\mathbf{o}) = (\tau_1(\mathbf{o}), \ldots, \tau_T(\mathbf{o}))$ using approximate geodesic distances (Tenenbaum *et al.*, 2000) between the ordered cells. The mapping of orders to pseudotimes is required to allow changes in scale for the underlying biological development, while allowing the computational benefit of sampling from a slightly smaller, if still large, set of possible orders rather than an even much larger one of possible high-dimensional pseudotime-vectors in $[a, b]^T$, where $[a, b]$ is an interval in $\mathbb{R}^+$. For details concerning the sampling of the cell orderings and the geodesic mapping, see Strauß *et al.* (2018).

2.2 Hierarchical GPs for pseudotemporal data

A Gaussian process (GP, Rasmussen and Williams (2006)) is a distribution over functions that is specified using a mean function μ and a covariance function Σ. For an input vector $\boldsymbol{\tau}(\mathbf{o}) = (\tau_1, \ldots, \tau_T)$ of pseudotime points depending on orders $\mathbf{o}$, $\mu(\boldsymbol{\tau}(\mathbf{o}))$ is a vector of T function evaluations of the mean function μ and $\Sigma(\boldsymbol{\tau}(\mathbf{o}))$ is a $T \times T$ matrix of covariance function evaluations of Σ. The distribution of functions $f \sim GP(\mu(\mathbf{o}), \Sigma(\mathbf{o}))$ is described by stating that, for any vector of pseudotime points $\boldsymbol{\tau}(\mathbf{o}) = (\tau_1(\mathbf{o}), \ldots, \tau_T(\mathbf{o}))$, evaluations $f(\tau_i(\mathbf{o}))$ follow a multivariate Normal $(f(\tau_1(\mathbf{o})), \ldots, f(\tau_T(\mathbf{o}))) \sim \mathcal{N}_T(\mu(\boldsymbol{\tau}(\mathbf{o})), \Sigma(\boldsymbol{\tau}(\mathbf{o})))$. Here we use a squared exponential covariance function for Σ:

$$\left[\Sigma(\boldsymbol{\tau}(\mathbf{o}); \sigma_w^2, l)\right]_{i,j} = \sigma_w^2 \exp\left(-\frac{(\tau_j - \tau_i)^2}{2l^2}\right) \tag{1}$$

where σ_w^2 is a scale parameter and l a length scale, and $[.]_{i,j}$ refers to the element in row i and column j of a matrix.

GPs have previously been used for pseudotime ordering – see Ahmed *et al.* (2018); Campbell and Yau (2016); Reid and Wernisch (2016); Strauß *et al.* (2018); Welch *et al.* (2017) – as well as for clustering time-stamped bulk gene expression data; see Cooke *et al.* (2011); Hensman (2013); Kirk *et al.* (2012); McDowell *et al.* (2018).

GPseudoClust models both the cluster-specific latent trajectory and a gene-specific latent trajectory deviating from the cluster-wide trajectory to some extent, see Figure 1. This is referred to as a hierarchical GP (see Hensman, 2013; Hensman *et al.*, 2015, for details).

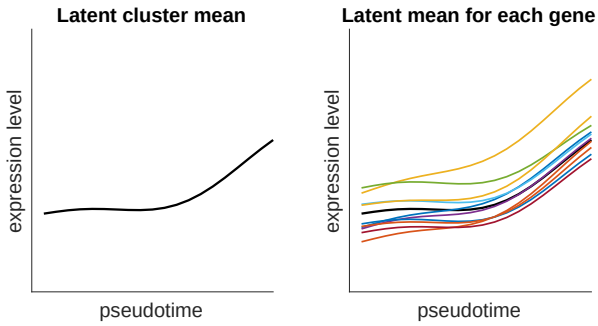


Figure 1: One cluster in the hierarchical GP model. left: cluster-wide latent mean, right: cluster-wide latent mean (black) and latent mean for each gene in the cluster.

We use Dirichlet processes (DPs, Ferguson, 1973) as a Bayesian nonparametric way of performing model-

based clustering. A DP is a distribution over discrete distributions; that is, each draw from a DP is itself a distribution. More precisely, $G \sim DP(\alpha, G_0)$ signifies that for any partition $B_1, \ldots, B_r$ of a parameter space Θ, we have $(G(B_1), \ldots, G(B_r)) \sim \text{Dirichlet}(\alpha G_0(B_1), \ldots, \alpha G_0(B_r))$, where the Dirichlet distribution with r categories and concentration parameters $(\gamma_1, \ldots, \gamma_r)$ is defined as follows: $p(x_1, \ldots, x_r) = \frac{\Gamma(\sum_{k=1}^r \gamma_k)}{\prod_{k=1}^r \Gamma(\gamma_k)} \prod_{k=1}^r x_k^{\gamma_k - 1}$.

3 Algorithm

3.1 Model

Conditional on the order $\mathbf{o}$ of the cells, the allocation of genes to clusters is modelled as a DP mixture model of hierarchical GPs as follows. The latent cluster means $\mu_j, j = 1, \ldots, n_g$ (see black line in Figure 1, n_g is the number of genes) are drawn from a DP with base distribution G_0:

$$G_0 \mid \mathbf{o}, a, \epsilon, L \sim GP(\mathbf{0}, \Sigma(\boldsymbol{\tau}(\mathbf{o}), 3a^2 + \epsilon, L)) \tag{2}$$
$$\alpha \sim \text{Gamma}(2, 4) \qquad G \mid G_0 \sim DP(\alpha, G_0) \qquad \mu_j \mid G \sim G$$

where $\mathbf{0}$ represents the zero mean function, and Σ is defined as in (1). $\boldsymbol{\tau}(\mathbf{o})$ is the vector of pseudotimes corresponding to cell order $\mathbf{o}$ (see Section 2.1). L is the length scale of the GP, $\sigma_W^2 = 3a^2 + \epsilon$ the scale parameter corresponding to σ_w^2 in equation (1). This specific parametrisation of the scale parameter of the latent mean trajectory links it to the scale and noise parameters of the deviations from the cluster-specific mean trajectory of the gene-specific trajectories, see Figure 1 and equation (3) below. The specific parameterisation used for our model is described and discussed in more detail below.

It should be noted that while we draw a mean μ_j for each gene $j = 1, \ldots, n_g$, the DP determines a number $K \ll n_g$ and values $\eta_1, \ldots, \eta_K$ such that for all $j = 1, \ldots, n_g$ there is a $k \in \{1, \ldots, K\}$ such that $\mu_j = \eta_k$. That is, the latent means μ_j only take K distinct values and there are K groups of genes with identical latent means, which form a total of K clusters. The number of clusters is not fixed, but automatically determined at each iteration of the sampler.

Shared across clusters, $\mathbf{o}, a$, ϵ and L have the following priors: $\mathbf{o} \sim \text{uniform}(\text{permutations}(\{1, 2, \ldots, T\}))$, $log(L) \sim N\left(\frac{1}{2}, \sigma_L\right)$, $\log(a) \sim N(\sqrt{\frac{1}{2}}, \sigma_a)$, $\log(\epsilon) \sim N(\frac{1}{2}, \sigma_\epsilon)$. Note that the log-Normal distributions guarantee positivity of the parameters. A strong prior for the length scale L is preferable for single-cell data in the context of sampling orders because of their high noise levels. With a vague prior on the length scale the inferred length scale tends to be too short, and the GP tends to overfit. We therefore fix $\sigma_L = 0.01$ for all data sets, as in Strauß *et al.* (2018).

Individual gene trajectories are modelled by GPs with mean $\mu_j, j = 1, \ldots, n_g$ (n_g is the number of genes). GPseudoClust uses as input preprocessed log-transformed gene expression data $y_g(\mathbf{o})$ for gene $g = 1, \ldots, n_g$. Conditional on the pseudotime ordering $\mathbf{o}$ of the cells, the trajectory $\mathbf{y}_j(\mathbf{o})$ of gene j is distributed as $\mathbf{y}_j(\mathbf{o}) \mid \mu_j, a, a_1, \mathbf{o} \sim F$, where

$$F = \text{GP}(\mu_j, \Sigma(\boldsymbol{\tau}(\mathbf{o}), a^2 \cdot a_1, 1) + a^2(1 - a_1)I_T) \tag{3}$$
$$a_1 \sim \text{Beta}(4, 1) \tag{4}$$

Σ is as in equation (1), I_T refers to the T-dimensional identity function. Note that a_1 represents how much variation from the cluster-wide mean is due to stochastic variation from the underlying stochastic

process, while $1-a_1$ represents the proportion of the variation resulting from noise. By equation (3), the ordered gene expression levels $\mathbf{y}_j(\mathbf{o})$ of gene j are noisy representations of individual gene-specific latent means drawn from a GP with cluster-specific mean function, see Figure 1. The black line represents the cluster-specific mean function and the coloured lines represent the gene-specific latent means around the cluster-specific mean.

A strong prior is also used for $\log(a)$ ($\sigma_a = 0.01$), while we use a weaker prior for $\log(\epsilon)$ ($\sigma_\epsilon = 0.1$). The hyperparameter a determines the magnitude of both the deviations of latent means of individual genes (see Figure 1) and noise-related deviations (see equation (3)). Our prior ensures that the method identifies interesting gene clusters whose within-cluster variability is low relative to the between-cluster variability. It also links these deviations to the scale hyperparameter of the cluster-wide GP by setting on it a lower bound which depends on the deviations of the trajectories of the individual genes from the cluster-wide mean trajectory, see equation (2). The raw data are normalised by subtracting the total mean of the expression matrix (not the row-wise mean), and dividing by the total standard deviation. For normalised data the proposed prior on a reflects that for the type of clustering found by GPseudoClust, the deviations from the cluster-wide latent mean account for roughly 50% of their average total variation, unless there is relatively strong evidence in the data for a split into more clusters with smaller deviations or fewer clusters with larger deviations.

3.2 MCMC sampling and block matrix representation

We use Markov Chain Monte Carlo (MCMC, Gilks *et al.* (1996)) sampling for inference of pseudotime orderings and cluster assignments. This allows sampling from a joint probability distribution of clusters, orders and hyperparameters a, L, a_1 and ϵ. For the orders, which are sampled from the discrete space of all possible permutations of cells, we previously developed an efficient sampling strategy (Strauß *et al.*, 2018). To reduce the dimensionality of the inference problem by reducing the number of parameters, we integrate out the cluster-specific mean trajectories, and developed an efficient method for the inversion of the resulting block matrices. For details, see Section S1 of the supplementary materials.

3.3 Subsampling strategies

Sampling orders of cells and clusters of genes simultaneously is a challenging high-dimensional problem, in particular as the posterior distribution of the orders is typically highly complex; see Strauß *et al.* (2018). In addition to the efficient block matrix computation strategies described above we further improve convergence crucially by means of parallel MCMC chains on subsets of cells. The chains are subsequently combined to a summary result approximating the posterior distribution of the cluster allocations.

3.3.1 Posterior similarity matrices

A central step is the computation of posterior similarity matrices (PSMs) for each of the chains on subsets of cells. The PSM is the symmetric positive semidefinite (see Lemma 4 in Section S2.2 of the supplementary materials) matrix whose entry in the ith row and jth column is the frequency with which gene i and gene j are clustered together among the samples drawn from the posterior distribution of cluster allocations. This estimates the posterior probability of the two respective samples being in the same cluster (Fritsch and Ickstadt, 2009).

3.3.2 Obtaining summary clusterings from PSMs

While the uncertainty of the cluster allocations obtained for single-cell data sets does not always justify a single summary clustering, it can nevertheless sometimes be useful to compute summary clusterings for validation and comparison purposes. In addition, the methods presented below to find weights for combining the PSMs obtained from the individual subsampled MCMC chains into one joint PSM also require summary clusterings of individual PSMs. To obtain a summary clustering from a PSM, we apply hierarchical clustering to the columns of the PSM using $1-\text{PSM}$ as the distance matrix (Medvedovic *et al.*, 2004). The optimal number of clusters is determined by a method maximising the posterior expected ARI (PEAR) between the inferred summary clustering and the unknown true clustering structure (Fritsch and Ickstadt, 2009). The ARI (adjusted Rand index, Hubert and Arabie (1985); Rand (1971), see also Section 3.4 and Section S3 of the supplementary materials) is a measure of agreement between two clusterings. The PEAR is therefore a measure of how well the inferred summary clustering is expected to agree with the unknown true clustering.

3.3.3 Combining PSMs

The following methods for combining the PSMs from the individual MCMC chains on subsampled data to obtain a joint overall PSM are proposed here:

Method 'mean psm' The first method proposed to obtain a joint PSM is to compute the element-wise unweighted arithmetic mean of the PSMs of the individual chains. This method is referred to as 'mean PSM' here.

Methods 'PY and PEAR', 'DPM and PEAR' As noise levels tend to differ between subsamples of cells, an unweighted average of the PSMs may not always be the best representation of the overall posterior distribution. We propose new methods to obtain a final PSM as a weighted average of the PSMs of the individual subsampled chains. The proposed methods are based on the following ideas. DP or Pitman-Yor (PY, Ishwaran and James (2001); Pitman and Yor (1997)) mixture models can be extended to perform feature selection. Pitman-Yor processes are a generalisation of DPs, for which the number of clusters is a priori larger than for the DP, see Section S2.1 of the supplementary materials. Using the two different processes here allows us to assess better the robustness of the method. We propose to use DP and PY mixture models with variable selection to identify features which are informative of the clustering, and to discard features that are not. In our case the features are the subsampled MCMC chains. We obtain weights for the PSMs of the subsampled chains as follows: First we obtain a summary clustering from each PSM. Then we use a DP or PY mixture model for discrete input data to model the summary clusterings, and this gives us weights, which are inclusion probabilities of features we obtain from the feature selection process. We refer to the two methods as 'PY and PEAR', 'DPM and PEAR', respectively. For details see Section S2.1 of the supplementary materials.

Method 'lmkk' The differences in noise for different subsampled chains may be gene-specific; to address this, this method applies localised multiple kernel k-means (lmkk, Gönen and Margolin (2014)) to obtain a summary clustering from the set of PSMs for the different chains. lmkk was first used to obtain summary clusterings from consensus clustering matrices in Cabassi and Kirk (2019b,a). Unlike the other methods proposed in this section, the 'lmkk' method does not aim to provide a full estimate of the

overall posterior similarity matrix, but it is an optimisation method to find a summary clustering from multiple PSMs. The method proposed in this paper also finds weights for an overall summary matrix representation of posterior cluster allocation probabilites. For details on our approach, see Section S2.2 of the supplementary materials.

3.4 Alternative clustering methods and assessment

The following other clustering methods are applied to the simulated and Shalek data sets: mclust, PAM, hierarchical clustering and SIMLR. In addition, we applied the following two-step methods (first pseudotime ordering of cells, then clustering of genes in a second step): SLICER (Welch *et al.*, 2016) and DeLorean (Reid and Wernisch, 2016) combined with GPclust (Hensman, 2013; Hensman *et al.*, 2015), and Monocle 2. Generally, standard settings are used. For SLICER the number of edges of the nearest neighbours graph in the low dimensional space is set to 5. For the initialisation of the noise and variance parameters for GPclust a number of different values were tried in an attempt to achieve a good clustering solution. The method turned out to be sensitive to initial conditions. For those methods which do not determine the number k of clusters automatically the average silhouette width (Rousseeuw, 1987), a standard criterion, is used to determine the optimal number of clusters. For the Shalek data we assume a minimum of four clusters, to distinguish between at least four different shapes of response trajectories, including early and late response and different levels of response.

For the simulated data sets the following measures of comparison between the true and the inferred cluster allocations are used: the Adjusted Rand index (ARI) (Hubert and Arabie, 1985; Rand, 1971), the Fowlkes-Mallows Index (FMI), and normalised mutual information (NMI) (Kvalseth, 1987). For all of these measures a score of one signifies perfect agreement between true and inferrred cluster allocations. For a definition of the measures see Section S3 of the supplementary materials.

4 Implementation

4.1 Data sets

Simulated data sets 1 and 2 We simulated two data sets with five clusters each. The specific construction of the two data sets is tailored for the first one to have very clearly separated clusters, and the second one to have clusters that cannot be disentangled using methods ignoring the pseudotemporal structure of the data, see Supplementary Figure S1. scRNA-seq data often consist of large numbers of repeated measurements at a few capture times. To mimic this situation, we assume 3 capture times for the simulated cells: the first 20 cells have capture time 1, cells 21 to 40 have capture time 2, and 41 to 60 capture time 3. We remove information about the true order by applying a random permutation to the order of the cells within each capture time, to mimic the lack of temporal information in applications.

Simulated data sets 1 and 2 were simulated using GPs, but not the same GP model as GPseudoClust. For both the simulated data sets cluster1 contains 8 genes, cluster2 4 genes, cluster3 12 genes, cluster4 16 genes and cluster5 12 genes. For a detailed description of the simulation set-up, see Section S4 of the supplementary materials.

Simulation studies with dropout noise scRNA-seq data are affected by technical noise leading to zero-expression values when the gene is actually expressed in the cell. To study the robustness of the method to technical zero-inflation without the presence of any other confounders, we use one of the data sets which we used to validate the subsampling procedures (simulated data set 2, see paragraph above and Supplementary Figure S1), and set nonzero values to zero at random. Note that while we could have used a dropout rate which depends on the actual gene expression level, with higher expression levels associated with lower probability of dropout (Pierson and Yau, 2015), our way of testing the robustness is more stringent by allowing larger perturbations of the trajectory. This additional simulation study comprises three sets of 100 data sets, to test for robustness of the GPSeudoClust method and all of the proposed subsampling methods (Section 3.3) to different levels of dropout, including a simulation study for which different groups of genes are affected by dropout to different degrees. The three dropout-related simulations were repeated 100 times each. For details, see Section S6.1 of the supplementary materials.

Overview of experimental data sets The validity of the subsampling approach with parallel chains each run on a subset of cells is further validated by applying GPseudoClust both with and without subsampling to a data set with 600 genes and 35 cells (Sasagawa data set below, Sasagawa *et al.* (2013), see below for a description of each of the data sets). An application of GPseudoClust to branching data (Moignard data, Moignard *et al.* (2015)) confirms existing, but also finds new results on differences of cluster structures of genes for different branches. GPseudoClust is also applied to non-branching data (Shalek data, Shalek *et al.* (2014)). Finally, the subsampling method and the combination of weighted PSMs are used to integrate data from different cell lines (Stumpf data, Stumpf *et al.* (2017)).

Moignard data: Moignard *et al.* (2015) applied single-cell RT-qPCR to 3,934 mouse early hematopoietic cells. In an in-vivo experiment cells were captured at four time points between embryonic day 7.0 and 8.5. In Moignard *et al.* (2015); Haghverdi *et al.* (2015, 2016) diffusion maps (Coifman *et al.*, 2005) are used to identify two branches, a blood and an endothelial branch. Here GPseudoClust is used to identify and compare different clustering structures for genes for the different branches. We use the pre-processed (Haghverdi *et al.*, 2016; Moignard *et al.*, 2015) data available as supplementary material to Haghverdi *et al.* (2016). Before the application of GPseudoClust, branches are inferred using diffusion maps, as in Haghverdi *et al.* (2016), which leads to the identification of an endothelial and an erythroid branch. We use diffusion maps for the identification of the branches, but find cluster allocations and their uncertainties using GPseudoClust without prior pseudotime ordering.

Sasagawa data: Mouse embryonic stem cells, cell-cycle related genes. GPseudoClust is also applied to a Quartz-Seq (FPKM normalised) data set of 35 mouse embryonic stem cells (Sasagawa *et al.*, 2013), on cell-cycle related genes. Cell cycle genes were selected by finding genes associated with GO:0007049, as in Liu *et al.* (2017).

Shalek data: LPS-stimulated mouse dendritic cells, scRNA-seq. Shalek *et al.* (2014) examined the response of primary mouse bone-marrow-derived dendritic cells in three different conditions using scRNA-seq. GPseudoClust is applied to the 74 genes identified by a previous method (Reid and Wernisch, 2016) as those with the highest temporal variance relative to their noise levels and to the 183 cells from the LPS (Lipopolysaccharide stimulated) condition and capture times 2h, 4h, and 6h, dropping the cells captured at 0h and 1h, to focus on differences between gene expression levels in reaction to the stimulus

rather than before the reaction has set in. The data were log-transformed, and an adjustment for cell size applied, according to Anders and Huber (2010) and Reid and Wernisch (2016).

Stumpf data: Stumpf *et al.* (2017) generated an RT-qPCR data set for 94 genes from two cell lines following the progression of mouse embryonic stem cells along the neuronal lineage, containing 96 cells per capture time (0h, 24h, 48h, 72h, 96h, 120h, 172h). The proposed subsampling methods allow taking subsamples of cells from each cell line separately and combining the chains as described in Section 3.3. For the preprocessing, the steps described in Stumpf *et al.* (2017) were applied to each cell line separately. The raw data are available on Mendeley Data (http://dx.doi.org/10.17632/g2md5gbhz7.1).

Details on numbers of MCMC chains and subsampled cells for the different data sets are provided in Section S5 of the supplementary materials.

4.2 Simulated data: only pseudotemporal methods unravel latent trajectories

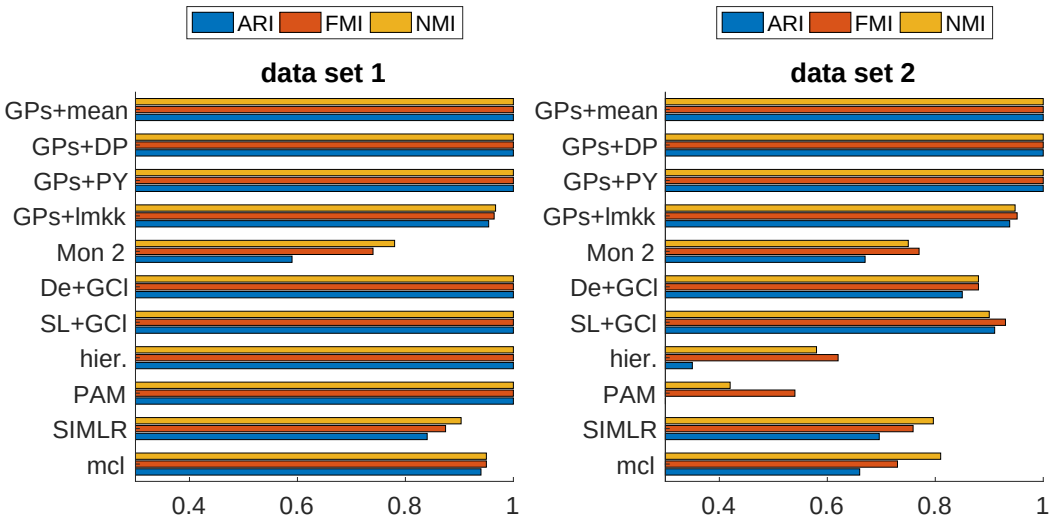


Figure 2: Simulated data sets 1 and 2. Comparison of estimates to true cluster allocations. Methods compared: GPs+mean = GPseudoClust and 'mean psm', GPs+DP = GPseudoClust+'DPM+PEAR', GPs+PY = GPeudoClust and 'PY+PEAR', GPs+lmkk = GPseudoClust method followed by summary clustering using lmkk, Mon 2 = Monocle 2 (2 steps: ordering and then clustering), De+GCl = DeLorean & GPclust (2 steps), SL+GCl = SLICER & GPclust (2 steps), hier = hierarchical clustering, PAM, SIMLR, mcl = mclust. Left: simulated data set 1 (more clearly separated clusters), right: simulated data set 2 (less clearly separated clusters).

Figure 2 illustrates the importance of using methods modelling the pseudotemporal nature of the data. It includes results for point estimates obtained by combining the GPseudoClust method with the proposed methods to obtain a joint PSM from several subsampled chains (see Section 3.3). Except for lmkk, where the method itself provides a summary clustering, a final summary clustering was obtained from the summary PSM by means of hierarchical clustering and the PEAR criterion. While for data sets with clearly separated clusters most clustering methods will perform satisfactorily (Figure 2, left), this is not the case for data sets where the cluster structure only becomes apparent through modelling the data as a pseudotime series, see Figure 2 (right). In the latter case only methods taking into account

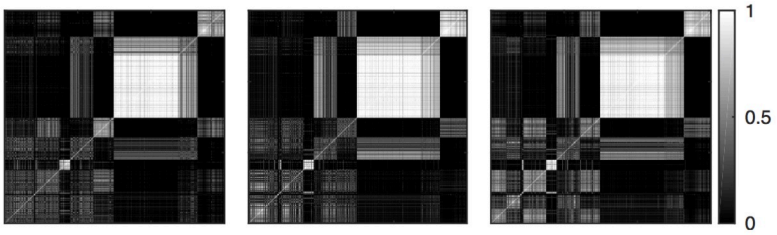
the pseudotemporal dynamics, such as GPclust combined with pseudotime ordering, and GPseudoClust, work well, while mclust and SIMLR perform best among those methods not incorporating the pseudotime structure. It should be noted that, as indicated by the different values for the two ARIs, and also the FMIs and NMIs, in Figure 2 for the two different two-step methods combining GPclust with existing pseudotime methods, the clustering results depend on the chosen pseudotime method.

4.3 Robustness to dropout

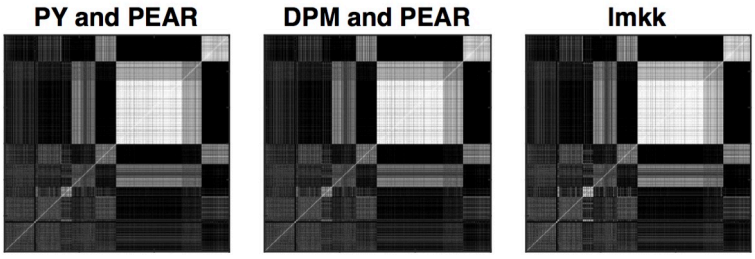
Further simulation studies on a total of 300 data sets (dropout studies 1, 2, and 3, see Section 4.1) with different levels of dropout noise demonstrate the robustness of GPseudoClust. For details, see Section S6 of the supplementary materials. Supplementary Figure S2 shows high ARIs with the true clustering for summary clusterings obtained by means of GPseudoClust and the proposed subsampling methods. While all the subsampling methods have a similar level of robustness to dropout noise when all genes are affected for all cells with equal probabilities (see Supplementary Figure S2), the 'lmkk' method is shown to be the best performing one for the case where there are groups of cells known to be less affected by dropout for a subset of the genes (dropout study 3), see Supplementary Figure S3. Section S6.2 in the supplementary materials also presents comparisons of the summary PSMs obtained using the different subsampling methods, see Supplementary Figures S4 to S12.

4.4 Validating subsampling: Sasagawa data

The Sasagawa data set has only 35 cells, which makes it suitable for comparing the proposed subsampling methods to applying the GPSeudoClust method to all the cells. Figure 3 illustrates good convergence of the GPseudoClust method with and without subsampling. Moreover it demonstrates that the proposed



(a) GPseudoClust without subsampling: PSMs for four chains.



(b) GPseudoClust with subsampling: 36 subsampled chains with 15 cells each.

Figure 3: Sasagawa data: (a) illustrates four PSMs obtained without subsampling, by applying GPseudoClust to all cells for each of the four chains. (b) compares the proposed subsampling methods 'PY and PEAR', 'DPM and PEAR', and 'lmkk'.

subsampling methods 'PY and PEAR' and 'DPM and PEAR' lead to PSMs convincingly similar to the ones obtained without the subsampling, and that a similar matrix is obtained using lmkk.

4.5 Immune response trajectories cluster around functional trajectories

The genes analysed for the Shalek data (see Section 4.1) are from three modules identified in Shalek *et al.* (2014) as 'peaked inflammatory module', which shows a 'rapid, yet transient induction' to LPS stimulation, 'core antiviral module, enriched for annotated antiviral and interferon response genes', and 'sustained inflammatory module; exhibiting continued rise in expression under LPS'. While the analysis proved to be very stable with regard to the number of subsampled chains (Supplementary Figure S27), for the following analysis the PSM obtained using the 'PY + PEAR' method with 96 subsampled chains is used. However, as illustrated by Supplementary Figure S27, for the 'PY and PEAR', 'DPM and PEAR' and 'mean PSM' methods a good approximation is achieved with only 4 randomly chosen chains.

The PSMs allow the computation of (potentially overlapping) groups of genes with high pairwise co-clustering probabilities. We use a threshold of 80% for the identification of groups of genes with high pairwise co-clustering probability. The choice of 80% for the threshold is chosen to ensure that it is sufficiently stringent to allow meaningful groups to be identified, but low enough to allow reasonably sized groups to be identified. The word pairwise is used here to emphasise that this is not the probability of all the genes being in the same cluster, but that for any two genes in such a group the probability of these two genes being in the same cluster is above 80%. It should be noted that this approach is different from trying to find a single summary clustering, and that the groups will usually overlap.

GPseudoClust identifies four groups with pairwise co-clustering probabilities of more than 80%, three of which, however, have a large overlap. Therefore, we refer to the groups as 1, 2a, 2b, and 2c.

Group 1: *Bcl2l11, Flrt3, Nfkbid, Ralgds, Rasgef1b, Socs3.* All genes in this group belong to a 'peaked inflammatory module' identified in Shalek *et al.* (2014), which shows a 'rapid, yet transient induction' to LPS stimulation.

Group 2: *Ddx60, Dhx58, E030037k03rik, Iigp1, Irf7, Mpa2l, Ms4a4c, Nlrc5, Nos2, Phf11, Slco3a1*

Group 2a: Group 2 and *D14ertd668e, Il15.* Except for *Nos2*, all genes in this group belong to a 'core antiviral module, enriched for annotated antiviral and interferon response genes' (Shalek *et al.*, 2014). *Nos2* is part of the 'sustained inflammatory module; exhibiting continued rise in expression under LPS'.

Group 2b: Group 2 and *D14ertd668e, Procr.* This group consists of genes from the 'core antiviral module', except for *Nos2* and *Procr.*

Group 2c: Group 2 and *Il15, Procr.* This group consists of genes from the 'core antiviral module', except for *Nos2* and *Procr.*

We also applied the other clustering methods mentioned, see Section 3.4, to the Shalek data set. The importance of quantifying the uncertainty of inferred cluster structures as done by GPseudoClust is highlighted by Figure 4, where the various clustering methods resulting in a single clustering disagree quite significantly, with most ARIs between pairs of results obtained by different methods less than 0.6. In addition, Figure 4 also shows that when the two-stage method of combining GPclust with a pseudotime method is used, the clustering result depends on the choice of the pseudotime method.

4.6 Detecting branch-dependent clustering structures

The analysis of the Moignard data set shows very different clustering structures in the trunk, the endothelial and the erythroid branch, see Figure 5, which shows summary PSMs obtained using the 'PY

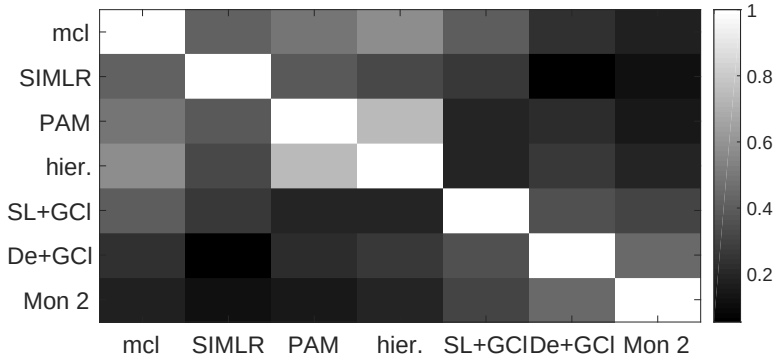


Figure 4: ARI between results obtained by different clustering methods: Shalek data set. A score of 1 shows that the two clusterings are identical, a score of 0 that they are no more related than expected by random chance. mcl = mclust, hier. = hierarchical clustering, SL+GCl = SLICER+GPclust, De+GCl = DeLorean+GPclust, Mon 2 = Monocle 2.

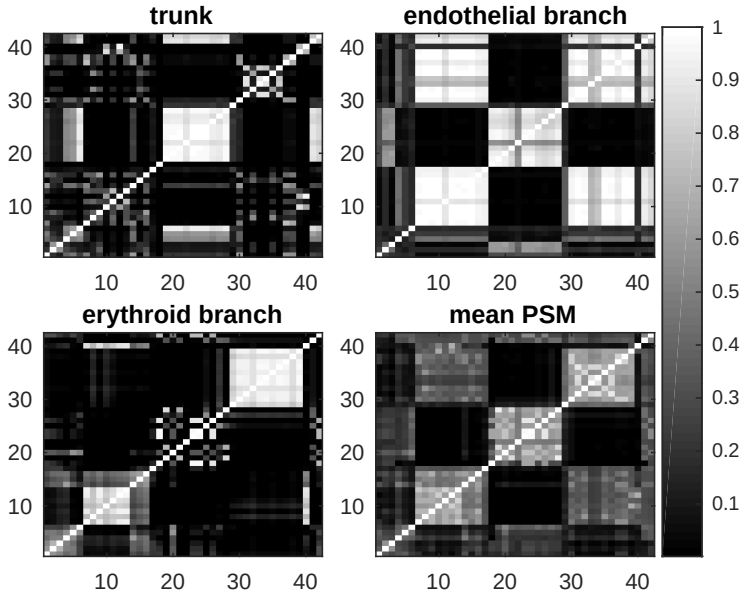


Figure 5: Moignard data: PSMs for branches. PSMs for each branch are obtained by the 'PY and PEAR' method. 'Mean PSM' refers to the unweighted mean of the summary PSMs of the three branches. A summary clustering was obtained from the mean PSM to order each of the matrices in the same way.

+ PEAR' method for the different branches. In this figure, the rows and columns of the four PSMs displayed are ordered in the same way to illustrate the differences in the clustering structures between the different branches.

For the trunk, *Fli1*, *Tal1*, *Etv2*, and *Kdr* have high posterior co-clustering probabilities, mirroring the fact that they are switched on early in the developmental process (see Supplementary Figures S14 and S15). Genes with very low expression levels in the trunk (*Gata1, Gfi1, Gfi1b, Hbbbh1, HoxB2, HoxD8, Ikaros, Itga2b, Mecom, Mitf, Myb, Nfe2, Sfpi1*) also have high co-clustering probabilities, see Supplementary Figure S16, similarly genes with relatively constant higher expression levels (*Ets2, FoxH1, FoxO4, Ldb1*, Supplementary Figure S17). For the endothelial branch, there is a group of genes with relatively constant higher expression level throughout the endothelial branch, which have high posterior co-clustering probabilities (*Cbfa2t3h, Cdh5, Egfl7, Erg, Ets1, Ets2, Etv6, Fli1, Hhex, Itga2b, Kdr, Kit,*

Ldb1, Lyl1, Mecom, Meis1, Notch1, Pecam1, Sox17, Sox7, Tal1, Supplementary Figure S20), and a group of genes which have very low expression levels or are not expressed (*Cdh1, Gata1, Gfi1, Gfi1b, HoxB2, HoxD8, Ikaros, Myb, Nfe2*, Supplementary Figure S21). For the erythroid branch GPseudoClust identifies again a group of genes with relatively constant higher expression levels (*Cbfa2t3h, Ets2, Etv6, FoxH1, FoxO4, Kit, Ldb1, Lyl1, Pecam1, Runx1, Tal1*, Supplementary Figure S22). *Gata1* and *Nfe2* are switched on at similar pseudotimes in the erythroid branch, whereas *Cdh5, Ets1, Etv2, Hhex, Kdr* and *Sox7* (Supplementary Figure S23) have a marked decrease in expression around a similar pseudotime (Supplementary Figure S24). For a detailed analysis and illustrations of the pseudotemporal dynamics of clusters of genes in different branches, see Section S7 of the supplementary materials and Supplementary Figures S13 to S26.

4.7 Combining multiple data sets

The subsampling methods proposed in Section 3.3 are also particularly useful in situations where we need to integrate data that were not obtained in exactly the same way, for instance because they were obtained from different cell lines or generally in slightly different experimental conditions. Instead of just blending the data sets, the subsampling method allows us to run chains for the different cell lines separately, then combine them in a principled way. The 'PY and PEAR' and 'DPM and PEAR' methods show particularly good agreement (see Supplementary Figure S28), but we also considered the "mean psm" and "lmkk" methods (see again Supplementary Figure S28).

Figure 6 illustrates the downweighting of those subsamples which are inconsistent with the integration of the two cell lines to a joint overall structure (weights close to 0 in Figure 6), and highlights again the high level of agreement between the 'PY + PEAR' and 'DPM + PEAR' methods.

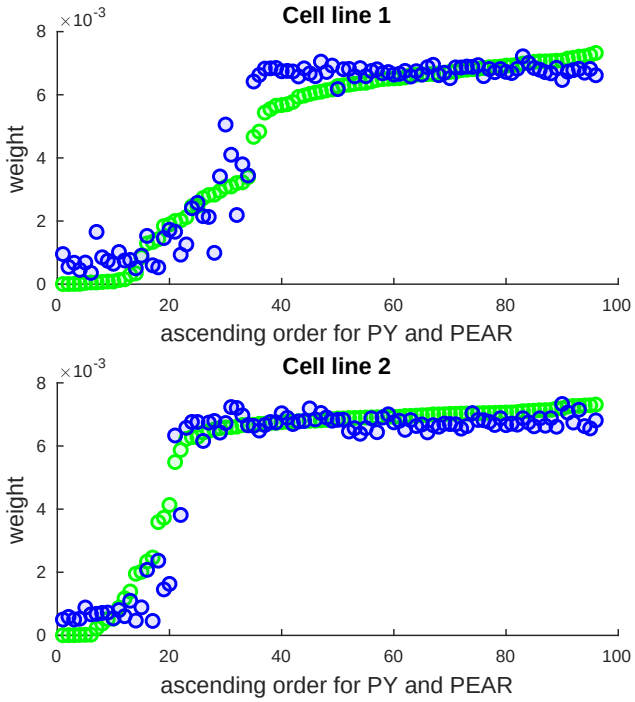


Figure 6: Stumpf data: comparison of weights for 'PY and PEAR' (green) and 'DPM and PEAR' (blue) methods. The weights are plotted along the y-axes and are sorted in the same way for both methods (ascending for 'PY and PEAR').

5 Discussion

GPseudoClust is a Bayesian nonparametric method for the clustering of genes for single-cell RNA-seq and RT-qPCR data in terms of latent shared pseudotime trajectories. Applying the method to simulated data shows that unless the clusters are very clearly separated from each other, clustering methods ignoring the pseudotemporal nature of the data may not be effective. While it is possible to combine pseudotime ordering and clustering methods in a two-step process, applications to both simulated and experimental data lead to clustering results with a dependence on the pseudotime method used, see Figures 2 and 4. GPseudoClust, a one-step ordering and clustering method, avoids this problem by sampling from a full posterior distribution of cluster allocations, fully exploring not only one clustering solution, but providing probabilities of genes being clustered together. GPseudoClust combines nonparametric Bayesian methods (Gaussian and Dirichlet processes) with efficient proposal distributions for MCMC, subsampling of cells (see Section 3.3), and novel methods for the combination of output from MCMC chains on subsampled data. This allows GPseudoClust to be applied to data sets with large numbers of cells.

In an application to dendritic cells GPseudoClust identifies clusters of genes closely associated with their biological function, and shows that there is considerable uncertainty in the clustering structures. GPseudoClust captures this uncertainty by providing a distribution of posterior co-clustering probabilities rather than just one single "point estimate" of a clustering. An application to branching data from early hematopoeitic cells demonstrates the ability of the method to identify strong differences between the clustering structures of the different branches. GPseudoClust identifies genes switched on or off at similar times in pseudotime as being co-clustered with a high probability. The uncertainty of clustering structures learned from the posterior distribution as represented by the PSM allows us to understand similarity of genes in terms of pairwise co-clustering probabilities.

An application to data obtained from different cell lines illustrates the ability of the method to analyse different data sets studying the same developmental process. GPseudoClust can be used to combine studies with different experimental protocols with different levels of measurement noise. The methods for finding weighted averages from multiple PSMs proposed here are designed to discard chains inconsistent with the overall clustering structure. We note that GPseudoClust could also be used to perform meta-analyses of previous studies, thanks to its ability to integrate data sets obtained under different experimental conditions. This may be of interest beyond the study of single-cell gene expression data.

While the computational efficiency of the subsampling methods makes it feasible to apply GPseudoClust to data sets with several thousand genes, the method is most suitable for the clustering of genes with high pseudotemporal variation. The stability of the methods used is demonstrated to be good. In particular, the final summary PSMs are shown to be robust to whether we use the 'DPM and PEAR' or 'PY and PEAR' method. Except for relative measurements like RT-qPCR, GPseudoClust is applied to log-transformed data. This is a frequent procedure for many pseudotime methods: see among many others Ahmed *et al.*, 2018; Haghverdi *et al.*, 2016; Ji and Ji, 2016; Reid and Wernisch, 2016; Welch *et al.*, 2016. Modelling count data directly in GPseudoClust could be achieved by a change in the likelihood function to a zero-inflated negative binomial distribution with GPs modelling the mean. However, this would further increase the computational complexity and make it much more difficult for the MCMC to achieve convergence.

Acknowledgements

We would like to thank Sascha Ott and William Astle for feedback and insightful comments.

References

Ahmed, S. *et al.* (2018). GrandPrix: Scaling up the Bayesian GPLVM for single-cell data. *Bioinformatics*, **35**(1), 47–54.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*, **11**(10), R106–R106.

Cabassi, A. and Kirk, P. D. W. (2019a). Kernel learning approaches for summarising and combining posterior similarity matrices. *In preparation*.

Cabassi, A. and Kirk, P. D. W. (2019b). Multiple kernel learning for integrative consensus clustering. *In preparation*.

Campbell, K. and Yau, C. (2016). Order under uncertainty: robust differential expression analysis using probabilistic models for pseudotime inference. *PLoS Comput Biol*, **12**(11), e1005212.

Coifman, R. *et al.* (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *P Natl Acad Sci USA*, **102**(21), 7426–7431.

Cooke, E. *et al.* (2011). Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements. *BMC Bioinformatics*, **12**(1), 399.

Eisen, M. B. *et al.* (1998). Cluster analysis and display of genome-wide expression patterns. *P Natl Acad Sci USA*, **95**(25), 14863–14868.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Ann Statist*, **1**(2), 209–230.

Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis and density estimation. *J Am Stat Assoc*, **97**, 611–631.

Fritsch, A. and Ickstadt, K. (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Anal*, **4**(2), 367–392.

Gilks, W. *et al.* (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Gönen, M. and Margolin, A. A. (2014). Localized data fusion for kernel k-means clustering with application to cancer biology. In *Advances in Neural Information Processing Systems 27*, pages 1305–1313.

Griffiths, J. A. *et al.* (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol Syst Biol*, **14**(4).

Haghverdi, L. *et al.* (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, **31**(18), 2989–2998.

Haghverdi, L. *et al.* (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nat Meth*, **13**(10), 845–848.

Hensman, J. *et al.* (2015). Fast nonparametric clustering of structured time-series. *IEEE T Pattern Anal*, **37**(2), 383–393.

Hensman, J. a. (2013). Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC Bioinformatics*, **14**(1), 252.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *J Classif*, **2**(1), 193–218.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J Am Stat Assoc*, **96**(453), 161–173.

Ji, Z. and Ji, H. (2016). TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res*, **44**(13), e117–e117.

Kaufman, L. and Rousseeuw, P. (2008). *Partitioning Around Medoids (Program PAM)*, chapter 2, pages 68–125. Wiley-Blackwell.

Kirk, P. *et al.* (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, **28**(24), 3290–3297.

Kiselev, V. *et al.* (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nat Meth*, **14**, 483–486.

Kunz, D. J. *et al.* (2018). Immune cell dynamics unfolded by single-cell technologies. *Frontiers in Immunology*, **9**, 1435.

Kvalseth, T. (1987). Entropy and correlation: Some comments. *IEEE T Syst Man Cy-S*, **17**(3), 517–519.

Lin, P. *et al.* (2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome Biol*, **18**(1), 59.

Liu, Z. *et al.* (2017). Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nat Commun*, **8**(1), 22.

Macaulay, I. C. *et al.* (2016). Single-cell RNA-sequencing reveals a continuous spectrum of differentiation in hematopoietic cells. *Cell Reports*, **14**(4), 966–977.

Maechler, M. *et al.* (2017). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.6.

McDowell, I. C. *et al.* (2018). Clustering gene expression time series data using an infinite Gaussian process mixture model. *PLoS Comput Biol*, **14**(1), 1–27.

Medvedovic, M. *et al.* (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, **20**(8), 1222–1232.

Moignard, V. *et al.* (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat Biotechnol*, **33**, 269 EP –.

Nestorowa, S. *et al.* (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*, **128**(8), e20–e31.

Pierson, E. and Yau, C. (2015). Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol*, **16**(1), 241.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann Probab*, **25**(2), 855–900.

Qiu, X. *et al.* (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nat Meth*, **14**, 979–982.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc*, **66**(336), 846–850.

Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.

Reid, J. and Wernisch, L. (2016). Pseudotime estimation: deconfounding single cell time series. *Bioinformatics*, **32**(19), 2973–2980.

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*, **20**(C), 53–65.

Sasagawa, Y. *et al.* (2013). Quartz-seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol*, **14**(4), 3097.

Scrucca, L. *et al.* (2017). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, **8**(1), 205–233.

Shalek, A. *et al.* (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, **510**, 363–369.

Stegle, O. *et al.* (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet*, **16**(3), 133–145.

Strauß, M. *et al.* (2018). GPseudoRank: a permutation sampler for single cell orderings. *Bioinformatics*, **35**(4), 611–618.

Stumpf, P. *et al.* (2017). Stem cell differentiation as a non-Markov stochastic process. *Cell Systems*, **5**(3), 268–282.e7.

Tenenbaum, J. *et al.* (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**(5500), 2319–2323.

Vallejos, C. *et al.* (2015). BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol*, **11**(6), 1–18.

Wang, B. *et al.* (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat Meth*, **14**, 414–416.

Welch, J. *et al.* (2016). SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biol*, **17**(1), 106.

Welch, J. *et al.* (2017). MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol*, **18**(1), 138.